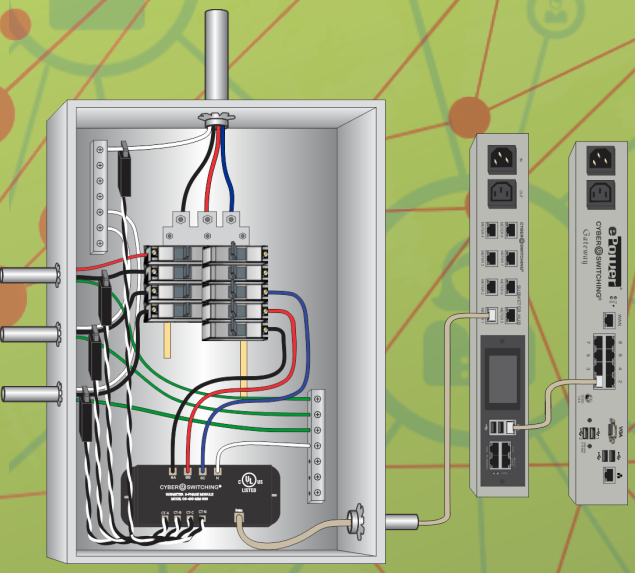
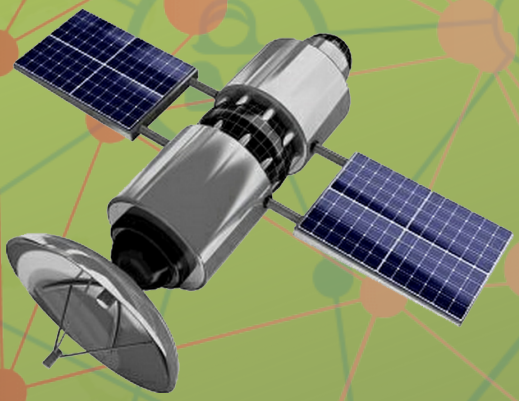
ASTES

Advances in Science, Technology & Engineering Systems Journal





VOLUME 2-ISSUE 3|MAY-JUNE 2017

www.astesj.com
ISSN: 2415-6698

EDITORIAL BOARD
Editor-in-Chief

Prof. Passerini Kazmerski
University of Chicago, USA

Editorial Board Members

Prof. Rehan Ullah Khan
Qassim University, Saudi Arabia

Dr. Nguyen Tung Linh
Electric Power University, Vietnam

Mohamed Mohamed Abdel-Daim
Suez Canal University, Egypt

Prof. María Jesús Espinosa
Universidad Tecnológica Metropolitana, Mexico

Tariq Kamal
University of Nottingham, UK

Sakarya University, Turkey

Dr. Omeje Maxwell
Covenant University, Nigeria

Dr. Hongbo Du
Prairie View A&M University, USA

Dr. Mohmaed Abdel Fattah Ashabrawy
Prince Sattam bin Abdulaziz University, Saudi Arabia

Prof. Majida Ali Abed Meshari
Tikrit University Campus, Iraq

Dr. Heba Afify
MTI university, Cairo, Egypt

Regional Editors

Dr. Hung-Wei Wu
Kun Shan University, Taiwan

Dr. Ahmet Kayabasi
Karamanoglu Mehmetbey University, Turkey

Mr. Aamir Nawaz
Gomal University, Pakistan

Dr. Shagufta Haneef
Aalborg University, Denmark

Aamir Nawaz
Gomal University, Pakistan

Dr. Maryam Asghari
Shahid Ashrafi Esfahani, Iran

Dr. Ebubekir Altuntas
Gaziosmanpasa University, Turkey

Dr. Gomathi Periasamy
Mekelle University, Ethiopia

Dr. Gomathi Periasamy
Mekelle University, Ethiopia

Abdullah El-Bayoumi
Cairo University, Egypt

Dr. Shakir Ali
Aligarh Muslim University, India

Dr. Sabry Ali Abdallah El-Naggar
Tanta University, Egypt

Dr. Walid Wafik Mohamed Badawy
National Organization for Drug Control and Research, Egypt

Dr. Walid Wafik Mohamed Badawy
National Organization for Drug Control and Research, Egypt

Ayham Hassan Abazid
Jordan university of science and technology, Jordan

Dr. Abhishek Shukla
R.D. Engineering College, India

Editorial

Advances in Science, Technology and Engineering Systems Journal (ASTESJ) is an online-only journal dedicated to publishing significant advances covering all aspects of technology relevant to the physical science and engineering communities. The journal regularly publishes articles covering specific topics of interest.

Current Issue features key papers related to multidisciplinary domains involving complex system stemming from numerous disciplines; this is exactly how this journal differs from other interdisciplinary and multidisciplinary engineering journals. This issue contains 220 accepted papers in Computer, Electrical and Electronics domains.

Editor-in-chief
Prof. Passerini Kazmersk

ADVANCES IN SCIENCE, TECHNOLOGY AND ENGINEERING SYSTEMS JOURNAL

Volume 2 Issue 3 **May-June 2017**

CONTENTS

Transmit / Received Beamforming for Frequency Diverse Array with Symmetrical frequency offsets 01
Shaddrack Yaw Nusenu

Detailed Analysis of Amplitude and Slope Diffraction Coefficients for knife-edge structure in S-UTD-CH Model 07
Eray Arik, Mehmet Baris Tabakcioglu

Applications of Case Based Organizational Memory Supported by the PAbMM Architecture 12
Martín, María de los Ángeles, Diván, Mario José

Low Probability of Interception Beampattern Using Frequency Diverse Array Antenna 24
Shaddrack Yaw Nusenu

Zero Trust Cloud Networks using Transport Access Control and High Availability Optical Bypass Switching 30
Casimer DeCusatis, Piradon Liengtiraphan, Anthony Sager

A Derived Metrics as a Measurement to Support Efficient Requirements Analysis and Release Management 36
Indranil Nath

Feedback device of temperature sensation for a myoelectric prosthetic hand 41
Yuki Ueda, Chiharu Ishii

Deep venous thrombus characterization: ultrasonography, elastography and scattering operator 48
Thibaud Berthomier, Ali Mansour, Luc Bressollette, Frédéric Le Roy, Dominique Mottier

Improving customs' border control by creating a reference database of cargo inspection X-ray images 60
Selina Kolokytha, Alexander Flisch, Thomas Lüthi, Mathieu Plamondon, Adrian Schwaninger, Wicher Vasser, Diana Hardmeier, Marius Costin, Caroline Vienne, Frank Sukowski, Ulf Hassler, Irène Dorion, Najib Gadi, Serge Maitrejean, Abraham Marciano, Andrea Canonica, Eric Rochat, Ger Koomen, Micha Slegt

Aviation Navigation with Use of Polarimetric Technologies 67
Arsen Klochan, Ali Al-Ammouri, Viktor Romanenko, Vladimir Tronko

Optimization of Multi-standard Transmitter Architecture Using Single-Double Conversion Technique Used for Rescue Operations 73
Riadh Essaadali, Said Aliouane, Chokri Jebali and Ammar Kouki

Singular Integral Equations in Electromagnetic Waves Reflection Modeling 82
A. S. Ilinskiy, T. N. Galishnikova

Methodology for Management of Information Security in Industrial Control Systems: A Proof of Concept aligned with Enterprise Objectives 88
Fabian Bustamante, Walter Fuertes, Paul Diaz, Theofilos Toulqueridis

Dependence-Based Segmentation Approach for Detecting Morpheme Boundaries 100
Ahmed Khorsi, Abeer Alsheddi

Paper Improving Rule Based Stemmers to Solve Some Special Cases of Arabic Language 111
Soufiane Farrah, Hanane El Manssouri, Ziyati Elhoussaine, Mohamed Ouzzif

Medical imbalanced data classification 116
Sara Belarouci, Mohammed Amine Chikh

ADOxx Modelling Method Conceptualization Environment 125
Nesat Efendioglu, Robert Woitsch, Wilfrid Utz, Damiano Falcioni

GPSR+Predict: An Enhancement for GPSR to Make Smart Routing Decision by Anticipating Movement of Vehicles in VANETs 137
Zineb Squalli Houssaini, Imane Zaimi, Mohammed Oumsis, Saïd El Alaoui Ouatik

Optimal Synthesis of Universal Space Vector Digital Algorithm for Matrix Converters 147
Adrian Popovici, Mircea Băbăiţă, Petru Papazian

Control design for axial flux permanent magnet synchronous motor which operates above the nominal speed 153
Xuan Minh Tran, Nhu Hien Nguyen, Quoc Tuan Duong

A synchronizing second order sliding mode control applied to decentralized time delayed multi-agent robotic systems: Stability Proof 160
Marwa Fathallah, Fatma Abdelhedi, Nabil Derbel

Fault Diagnosis and Tolerant Control Using Observer Banks Applied to Continuous Stirred Tank Reactor 171
Martin F. Pico, Eduardo J. Adam

Development and Validation of a Heat Pump System Model Using Artificial Neural Network 182
Nabil Nassif, Jordan Gooden

Assessment of the usefulness and appeal of stigma-stop by psychology students: a serious game designed to reduce the stigma of mental illness 186
Adolfo J. Cangas, Noelia Navarro, Juan J. Ojeda, Diego Cangas, Jose A. Piedra, José Gallego

Kinect-Based Moving Human Tracking System with Obstacle Avoidance 191
Abdel Mehsen Ahmad, Zouhair Bazzal, Hiba Al Youssef

A security approach based on honeypots: Protecting Online Social network from malicious profiles 198
Fatna Elmendili, Nisrine Maqran, Younes El Bouzekri El Idrissi, Habiba Chaoui

Pulse Generator for Ultrasonic Piezoelectric Transducer Arrays Based on a Programmable System-on-Chip (PSoC) 205
Pedro Acevedo, Martín Fuentes, Joel Durán, Mónica Vázquez, Carlos Díaz

Enabling Toy Vehicles Interaction With Visible Light Communication (VLC) 210
M. A. Ilyas, M. B. Othman, S. M. Shah, Mas Fawzi

Analysis of Fractional-Order 2xn RLC Networks by Transmission Matrices 217
Mahmut Ün, Manolya Ün

Fire extinguishing system in large underground garages 221
Ivan Antonov, Rositsa Velichkova, Svetlin Antonov, Kamen Grozdanov, Milka Uzunova, Ikram El Abbassi

Directional Antenna Modulation Technique using A Two-Element Frequency Diverse Array 227
Shaddrack Yaw Nusenu

Classifying region of interests from mammograms with breast cancer into BIRADS using Artificial Neural Networks 233
Estefanía D. Avalos-Rivera, Alberto de J. Pastrana-Palma

Magnetically Levitated and Guided Systems 241
Florian Puci, Miroslav Husak

Energy-Efficient Mobile Sensing in Distributed Multi-Agent Sensor Networks 245
Minh T. Nguyen

Validity and efficiency of conformal anomaly detection on big distributed data 254
Ilia Nouretdinov

S-Parameters Optimization in both Segmented and Unsegmented Insulated TSV upto 40GHz Frequency 268
Juma Mary Atieno, Xuliang Zhang, HE Song Bai

Synthesis of Important Design Criteria for Future Vehicle Electric System 277
Lisa Braun, Eric Sax

Gestural Interaction for Virtual Reality Environments through Data Gloves 284
G. Rodriguez, N. Jofre, Y. Alvarado, J. Fernández, R. Guerrero

Solving the Capacitated Network Design Problem in Two Steps 291
Meriem Khelifi, Mohand Yazid Saidi, Saadi Boudjit

A Computationally Intelligent Approach to the Detection of Wormhole Attacks in Wireless Sensor Networks 302
Mohammad Nurul Afsar Shaon, Ken Ferens

Real Time Advanced Clustering System 321
Giuseppe Spampinato, Arcangelo Ranieri Bruna, Salvatore Curti, Viviana D'Alto

Indoor Mobile Robot Navigation in Unknown Environment Using Fuzzy Logic Based Behaviors 327
Khalid Al-Mutib, Foudil Abdessemed

Validity of Mind Monitoring System as a Mental Health Indicator using Voice 338
Naoki Hagiwara, Yasuhiro Omiya, Shuji Shinohara, Mitsuteru Nakamura, Masakazu Higuchi, Shunji Mitsuyoshi, Hideo Yasunaga, Shinichi Tokuno

The Model of Adaptive Learning Objects for virtual environments instanced by the competencies 345
Carlos Guevara, Jose Aguilar, Alexandra González-Eras

An Overview of Traceability: Towards a general multi-domain model 356
Kamal Souali, Othmane Rahmaoui, Mohammed Ouzzif

L-Band SiGe HBT Active Differential Equalizers with Variable, Positive or Negative Gain Slopes Using Dual-Resonant RLC Circuits 362
Yasushi Itoh, Hiroaki Takagi

Moving Towards Reliability-Centred Management of Energy, Power and Transportation Assets 369
Kang Seng Seow, Loc K. Nguyen, Kelvin Tan, Kees-Jan Van Oeveren

Secure Path Selection under Random Fading 376
Furqan Jameel, Faisal, M Asif Ali Haider, Amir Aziz Butt

Security in SWIPT with Power Splitting Eavesdropper 384
Furqan Jameel, Faisal, M Asif Ali Haider, Amir Aziz Butt

Performance Analysis of Phased Array and Frequency Diverse Array Radar Ambiguity Functions 389
Shaddrack Yaw Nusenu

Adaptive Discrete-time Fuzzy Sliding Mode Control For a Class of Chaotic Systems 396
Hanene Medhaffar, Moez Feki, Nabil Derbel

Fault Tolerant Inverter Topology for the Sustainable Drive of an Electrical Helicopter 401
Igor Bolvashenkov, Jörg Kammermann, Taha Lahlou, Hans-Georg Herzog

Computational Intelligence Methods for Identifying Voltage Sag in Smart Grid 412
Turgay Yalcin, Muammer Ozdemir

A Highly-Secured Arithmetic Hiding cum Look-Up Table (AHLUT) based S-Box for AES-128 Implementation 420
Ali Akbar Pammu, Kwen-Siong Chong, Bah-Hwee Gwee

Service Productivity and Complexity in Medical Rescue Services 427
Markus Harlacher, Andreas Petz, Philipp Przybysz, Olivia Chaillié, Susanne Mütze-Niewöhner

Principal Component Analysis Application on Flavonoids Characterization 435
Che Hafizah Che Noh, Nor Fadhillah Mohamed Azmin, Azura Amid

A Reconfigurable Metal-Plasma Yagi-Uda Antenna for Microwave Applications 441
Giulia Mansutti, Davide Melazzi, Antonio-Daniele Capobianco

Verifying the Detection Results of Impersonation Attacks in Service Clouds 449
Sarra Alqahtani, Rose Gamble

Image Segmentation Using Fuzzy Inference System on YCbCr Color Model 460
Alvaro Anzueto-Rios, Jose Antonio Moreno-Cadenas, Felipe Gómez-Castañeda, Sergio Garduza-Gonzalez

Segmented and Detailed Visualization of Anatomical Structures based on Augmented Reality for Health Education and Knowledge Discovery 469
Isabel Cristina Siqueira da Silva, Gerson Klein, Denise Munchen Brandão

Intrusion detection in cloud computing based attack patterns and risk assessment 479
Ben Charhi Youssef, Mannane Nada, Bendriss Elmehdi, Regragui Boubker

Optimal Sizing and Control Strategy of renewable hybrid systems PV-Diesel Generator-Battery: application to the case of Djanet city of Algeria 485
Adel Yahiaoui, Khelifa Benmansour, Mohamed Tadjine

RFID Antenna Near-field Characterization Using a New 3D Magnetic Field Probe 492
Kassem Jomaa, Fabien Ndagijimana, Hussam Ayad, Majida Fadlallah, Jalal Jomaah

Design, Fabrication and Testing of a Dual-Range XY Micro-Motion Stage Driven by Voice Coil Actuators 498
Xavier Herpe, Matthew Dunnigan, Xianwen Kong

Self-Organizing Map based Feature Learning in Bio-Signal Processing 505
Marwa Farouk Ibrahim Ibrahim, Adel Ali Al-Jumaily

A delay-dependent distributed SMC for stabilization of a networked robotic system exposed to external disturbances 513
Fatma Abdelhedi, Nabil Derbel

Modelization of cognition, activity and motivation as indicators for Interactive Learning Environment 520
Asmaa Darouich, Faddoul Khoukhi, Khadija Douzi

Homemade array of surface coils implementation for small animal magnetic resonance imaging 532
Fernando Yepes-Calderon, Olivier Beuf

An Encryption Key for Secure Authentication: The Dynamic Solution 540
Zubayr Khalid, Pritam Paul, Khabbab Zakaria, Himadri Nath Saha

Multi-Domain Virtual Network Embedding with Coordinated Link Mapping 545
Shuopeng Li, Mohand Yazid Saidi, Ken Chen

Semantic-less Breach Detection of Polymorphic Malware in Federated Cloud 553
Yahav Biran, George Collins, Borky John M, Joel Dubow

Impact of PSS and SVC on the Power System Transient Stability 562
Mohammed Omar Benaissa, Samir Hadjeri, Sid Ahmed Zidi

The Search for a Convergent Option to Deploy Smart Grids on IoT Scenario 569
Hamilton da Gama Schroder Filho, José Pissolato Filho, Vinicius Luciano Moreli

Optimizing the Decoding Complexity of PEG-Based Methods with an Improved Hybrid Iterative/Gaussian Elimination Decoding Algorithm 578
Reem Alkanhel, Marcel Adrian Ambroze

FAG – a management support tool with BI techniques to assist teachers in the virtual learning environment Moodle 587
Luciana Silva Zapparolli, Itana Stiubiener

A comparative between CRISP-DM and SEMMA through the construction of a MODIS repository for studies of land use and cover change 598
Herman Jair Gómez Palacios, Robinson Andrés Jiménez Toledo, Giovanni Albeiro Hernández Pantoja, Álvaro Alexander Martínez Navarro

Power requirements and battery life measurement for wireless transmission between two nodes in different mediums 605
Radouane Karli, Abdellatif Bouchalkha, Khalid Alhammadi

A Novel Dual-Band Rectenna for Ambient RF Energy Harvesting at GSM 900 MHz and 1800 MHz 612
Dinh Khanh Ho, Van-Duc Ngo, Ines Kharrat, Tan Phu Vuong, Quoc Cuong Nguyen, Minh Thuy Le

Optimisation of Software-Defined Networks Performance Using a Hybrid Intelligent System 617
Ann Sabih, Yousif Al-Dunainawi, H. S. Al-Raweshidy, Maysam F. Abbod

Data Error Detection and Recovery in Embedded Systems: a Literature Review 623
Venu Babu Thati, Jens Vankeirsbilck, Jeroen Boydens, Davy Pissoort

Augmented Reality Prototype HUD for Passenger Infotainment in a Vehicular Environment 634
Shu Wang, Vassilis Charissis, David K. Harisson

Incorporating Spatial Information for Microaneurysm Detection in Retinal Images 642
Mohamed M. Habib, Roshan A. Welikala, Andreas Hoppe, Christopher G. Owen, Alicja R. Rudnicka, Adnan Tufail, Catherine Egan, Sarah A. Barman

Human Robot Interaction for Hybrid Collision Avoidance System for Indoor Mobile Robots 650
Mazen Ghandour, Hui Liu, Norbert Stoll, Kerstin Thurow

Multi Level Integrity Management in LTE/LTE-A Networks 658
Ouafae Kasmi, Amine Baina, Mostafa Bellafkih

Development of an automated vehicle stop system for cardiac emergencies 669
Tung T. Nguyen, Issey Takahashi, Takahiro Tanaka, Eijiro Takeuchi, Shinpei Kato, Tatsuya Suzuki, Hitoshi Kanamori, Yoshiki Ninomiya, Hirofumi Aoki

Computer Aided Classification using Support Vector Machines in Detecting Cysts of Jaws 674
Veena Divya K, Anand Jatti, Revan Joshi P, Sabah Meharaj

Spatiotemporal Traffic State Prediction Based on Discriminatively Pre-trained Deep Neural Networks 678
Mohammed Elhenawy, Hesham Rakha

Low-cost Hybrid Wheelchair Controller for Quadriplegias and Paralysis Patients 687
Mohammed Faeik Ruzaij Al-Okby, Sebastian Neubert, Norbert Stoll, Kerstin Thurow

Preparation of Ni-C Ultrafine Composite from Waste Material 695
Mahmoud A. Rabah, S.M. Abdelbasir

Schizophrenia Prediction Using Integrated Imaging Genomic Networks 702
Su-Ping Deng, Wenxing Hu, Vince D. Calhoun, Yu-Ping Wang

IGS: The Novel Fast IC Power Ground Network Optimization Flow Based on Improved Gauss-Seidel Method 711
Qinghao Ye, Xiaoxiao Wang, Jiansong Gong, Songhao Ru, Donglin Su

GNSS Positioning in Non-line-of-Sight Context—a Survey for Technological Innovation 722
Julia Breßler, Marcus Obst

Collision avoidance between a wheelchair front wheels and a step wall during step climbing using a care robot 732
Hidetoshi Ikeda, Kazuki Hashimoto, Daisuke Murayama, Rikuto Yamazaki, Eiji Nakano

Adapting Model Predictive Control for a PV Station and Evaluating two different MPPT Algorithms P&O and FLC 741
Imad Elzein, Yury Petrenko

Human Sit Down Position Detection Using Data Classification and Dimensionality Reduction 749
Paul Rosero-Montalvo, Daniel Jaramillo, Stefany Flores, Diego Peluffo, Vanessa Alvear, Milton Lopez

Designing a Smart Car Parking System (PoC) Prototype Utilizing CCTV Nodes: A vision of an IoT parking system via UCD process 755
Muftah Fraifer, Mikael Fernström

Multilingual evaluation of voice disability index using pitch rate 765
Shuji Shinohara, Yasuhiro Omiya, Mitsuteru Nakamura, Naoki Hagiwara, Masakazu Higuchi, Shunji Mitsuyoshi, Shinichi Tokuno

Efficacy of multiple anticancer therapies may depend on host immune response 773
Kritika Karri, Dhundy R Bastola

Interdisciplinarity to Integrate Knowledge in Engineering 788
Stella Abreu, Amélia Caldeira, Alexandra R. Costa, Tiago Gomes, Luís A. C. Roque

Treadmill and Vision System for Human Gait Acquisition and Analysis 796
Paulo A. Ferreira, João P. Ferreira, Manuel Crisóstomo, A. Paulo Coimbra

Assessment of the heat sinking effect of a human hand that holds a flexible phototherapy device for use in Kangaroo Mother Care 805
Luis Jimenez, Luis Vilcahuaman, Jorge Galdos

Network Intrusion Detection System using Apache Storm 812
Muhammad Asif Manzoor, Yasser Morgan

Towards Optimality in Online Learning – The OLeCenT Approach 819
Carl Beckford, Ezra K. Mugisa

An efficient model to improve the performance of platelet inventory of the blood banks 839
Annista Wijayanayake, Manisha Dandunna

Performance Evaluation of Associative Classifiers in Perspective of Discretization Methods 845
Zulfiqar Ali, Waseem Shahzad

A novel model for Time-Series Data Clustering Based on piecewise SVD and BIRCH for Stock Data Analysis on Hadoop Platform 855
Ibgtc Bowala, Mgnas Fernando

Analysis of Wireless Traffic Data through Machine Learning 865
Muhammad Ahsan Latif, Muhammad Adnan

Video-frame Quality Improvement before Shot boundary Detection by using Logarithm, Wavelet and Contoulet Transform 872
Songpon Nakharacruangsak

Electromagnetic Field and Lumped Circuit Analysis of Planar Transformer at Medium Frequency for Converter Application 878
Chengjian Lian, Daming Zhang

Indoor monitoring in Public Buildings: workplace wellbeing and energy consumptions. An example of IoT for smart cities application 884
Stefania Nanni, Elisa Benetti, Gianluca Mazzini

EEG Mind Controlled Smart Prosthetic Arm – A Comprehensive Study 891
Taha Beyrouthy, Samer Al Kork, Joe Akl Korbane, Mohamed Abouelela

Features based approach for indexation and representation of unstructured Arabic documents 900
Mohamed Salim El Bazzi, Driss Mammass, Abdelatif Ennaji, Taher Zaki

The Analysis and Comparison of Multiport Converter used for Renewable Energy Sources 906
Sarab Al-Chlaihawi, Ammar Al-Gizi, Aurelian Craciunescu

Structural Optimization of Wavy FinFET for Leakage Reduction and Performance Enhancement 913
Anju Chakkikavil, Nisha Kuruvilla, Ayoob Khan, Shahul Hameed

Program with Ixquic 918
Carmen Ramos, Tania Patiño

A Design of SDN Based IP Mobility Management Considering Inter-Domain Handovers and Its Evaluation 922
Misumi Hata, Mustafa Soylu, Satoru Izumi, Toru Abe, Takuo Suganuma

Computer Tomography from Micro-Electronics to Assembled Products 932
Keith Bryant, Ragnar Vaga

Mathematical Modeling and Digital Control of A Hybrid Switching Buck Converter 937
Muhammad Umar Abbasi

Effect of Ambient Temperature Variations on Particle Dimesions in Ultrasonic Nebulizers during Cold Vaporization 946
Ali Akpek

Graphics on demand: the automatic data visualization on the WEB 951
Ramzi Guetari, Maha Mallek

Spectral Reflectance Estimation and Color Reproduction Based on Sparse Neugebauer Model 958
Kohei Inoue, Kenji Hara, Kiichi Urahama

3D Senor-based Library Navigation System 967
Abdel-Mehsen Ahmad, Zouhair Bazzal, Roba Al Majzoub, Ola Charanek

Deterministic Approach to Detect Heart Sound Irregularities 974
Richard Mengko, Amy Hamidah, Ricky Saputra

Spatial Modulation Technique For Filtered-OFDM Based Wireless Transmission 981
Shaddrack Yaw Nusenu

Template-like Tensor Domain Operations to Enhancing Diffusion Datasets Quality 987
Fernando Yepes-Calderon, Flor M. Medina

Ensemble of Neural Network Conditional Random Fields for Self-Paced Brain Computer Interfaces 996
Hossein Bashashati, Rabab Kreidieh Ward

Optimal Plant Growth in Smart Farm Hydroponics System using the Integration of Wireless Sensor Networks into Internet of Things 1006
Nathaphon Boonnam, Jumras Pitakphongmetha, Siriwan Kajornkasirat, Teerayut Horanont, Deeprom Somkiadcharoen, Jiranuwat Prapakornpilai

Skeletonization in Natural Image using Delaunay Triangulation 1013
Vicky Sintunata, Terumasa Aoki

A High Efficiency 0.13μm CMOS Full Wave Active Rectifier with Comparators for Implanted Medical Devices 1019
João Ricardo de Castilho Louzada, Leonardo Breseghello Zoccal, Robson Luiz Moreno, Tales Cleber Pimenta

Multiclass Myoelectric Identification of Five Fingers Motion using Artificial Neural Network and Support Vector Machine 1026
Jawad Ahmad, Ammar Mohsin Butt, Muhammad Tanveer Riaz, Shoaib Bhutta, Muhammad Zeeshan Khan, Inam-Ul-Haq

Group of UAVs Moving on Smooth Control Law with Fixed Obstacles 1034
Dmytro Pavlovich Kucherov, Andrei Nikolayevich Kozub, Oleg Nikolayevich Kostyna

Predictive Technology Management for the Identification of Future Development Trends and the Maximum Achievable Potential Based on a Quantitative Analysis 1042
Michael Fries, Markus Lienkamp

Improving the Performance of Fair Scheduler in Hadoop 1050
Shou-Chih Lo, Ya-Wen Cheng

Ranking of Two Multi Criteria Decision Making Cases with Evidential Reasoning under Uncertainty 1059
Farzaneh Ahmadzadeh

Applied Horizontal and Vertical Geothermal Heat Exchanger with Heat Pump System to Provide Air Conditioning for an Academic Facility in Mexico 1064
Daniel Alcantar Martínez, Oskar Javier González Pedraza, Crisanto Mendoza Cavarrubias, J. Jesus Pacheco Ibarra, Jorge Alberto Rangel Arista

Space Time – Track Circuits with Trellis Code Modulation 1070
Marius Enulescu

Support Vector Machine based Vehicle Make and Model Recognition System 1080
Muhammad Asif Manzoor, Yasser Morgan

Estimation of voltage unbalance at an AC traction substation with different train operational scenarios 1086
Thanatchai Kulworawanichpong, Nattapong Mingpruk, Tosaphol Ratniyomchai

A Microwave Reflectometer Prototype for Detection of Body Through the Rubble of Collapsed Buildings 1092
Syed Daniyal Ali Shah

Rotary Cutting of Electrical Steel Laminations – A Contrast to Traditional Stamping 1107
Markus Hubert, Simon Kutter, Marco Ziegler, Michael Schneider, Joerg Franke

Design for Smaller, Lighter and Faster ICT Products: Technical Expertise, Infrastructures and Processes 1114
Omer Aydin, Orhan Uçar

A Novel Ultra High Speed and Configurable Discrete Wavelet Packet Transform Architecture 1129
Mouhamad Chehaitly, Mohamed Tabaa, Fabrice Monteiro, Abbas Dandache

Cognitive Artificial Intelligence Method for Interpreting Transformer Condition Based on Maintenance Data 1137
Karel Octavianus Bachri, Bambang Anggoro, Arwin Datumaya Wahyudi Sumari, Adang Suwandi Ahmad

Selective Detrending using Baseline Drift Detection Index for Task-dependant fNIRS Signal 1147
Jinung An, Gihyoun Lee, Seung Hyun Lee, Sang Hyeon Jin

Autonomous Micro-Air-Vehicle Control Based on Visual Sensing for Odor Source Localization 1152
Kenzo Kurotsuchi, Mitsuharu Tai, Hiromasa Takahashi

EMI Shielding Performance For Varies Frequency by Metal Plating on Mold Compound 1159
Min Fee Tai, Swee Leong Kok, Kenichiroh Mukai, Stephan Hotz, Patrick Brooks, Diego Reinoso Cocina

Comparative Study of Adaptive Consensus Control of Euler-Lagrange Systems on Directed Network Graph 1165
Yoshihiko Miyasato

Future Contract Selection by Term Structure Analysis 1172
Vasco Grossmann, Manfred Schimmler

Optimal Rotor Design of Line Start Permanent Magnet Synchronous Motor by Genetic Algorithm 1181
Bui Minh Dinh

Fault-Tolerant in Embedded Systems (MPSoC): Performance Estimation and Dynamic Migration Task 1188
Kamel Smiri, Habib Smei, Nourhen Fourati, Abderrazak Jemai

Incremental Control Techniques for Layout Modification of Integrated Circuits 1196
Patrik Vacula, Vlastimil Kotě, Adam Kubačák, Milan Lžíčař, Radek Zelený, Miroslav Husák, Jiří Jakovenko

Effect of Cold Drawing Pass Schedule on Mechanical Properties and Microstructure of ST 52 during Cold Drawing of Seamless Tubes and its Influence on Springback 1202
Dadabhau Baban Karanjule, Sunil Sukhdeo Bhamare, Thota Hanumant Rao

IoT: Their Conveniences, Security Challenges and Possible Solutions 1211
Davar Pishva

Recognition and Position Estimation for Multiple Labware Transportation Using Kinect V2 and Mobile Robots 1218
Mohammed Myasar Ali, Hui Liu, Norbert Stoll, Kerstin Thurow

Grasping and Placing Operation for Labware Transportation in Life Science Laboratories using Mobile Robots 1227
Mohammed Myasar Ali, Hui Liu, Norbert Stoll, Kerstin Thurow

Face Recognition and Tracking in Videos 1238
Swapnil Vitthal Tathe, Abhilasha Sandipan Narote, Sandipan Pralhad Narote

Efficiency of Photovoltaic Maximum Power Point Tracking Controller Based on a Fuzzy Logic 1245
Ammar Al-Gizi, Sarab Al-Chlaihawi, Aurelian Craciunescu

Model-based Development of Safety-critical Functions and ISO 26262 Work Products using modified EAST-ADL 1252
Bülent Sari, Hans-Christian Reuss

On-Chip Testing Schemes of Through-Silicon-Vias (TSVs) in 3D Stacked ICs 1260
Shadi MS. Harb, William R. Eisenstadt

Stability and nonlinear controller design of Fast-Lock Phase-Locked Loop in 0.18-μm CMOS 1266
Fatima Al-Mutairi, Reem Al-Mutairi

Towards Deployment Strategies for Deception Systems 1272
Daniel Fraunholz, Marc Zimmermann, Hans Dieter Schotten

Classification of patient by analyzing EEG signal using DWT and least square support vector machine 1280
Mohd Zuhair, Sonia Thomas

Proposal for a Visualization System of Purchase Relationship Using ID-POS Data 1290
Ko Hashimoto, Kohei Otake, Takashi Namatame

Public transportation network design: a geospatial data-driven method 1298
Alexey Golubev, Maxim Shcherbakov

A Secure Communication Framework for ECUs 1307
Ali Shuja Siddiqui, Yutian Gui, Jim Plusquellic, Fareena Saqib

Content Reordering Based on Mouse-tracking for Web Applications 1314
Md. Tanjim-Al-Akib, Lutfullahil Kabir Ashik, Krishanu Chowdhury, Md. Zahid Hasan, Hosne Al Walid

Multi-attribute Reverse Auction Design Based on Fuzzy Data Envelopment Analysis Approach 1323
Deyan Chen, Dingwei Wang

A Category Based Threat Evaluation Model Using Platform Kinematics Data 1330
Mustafa Çöçelli, Ethem Arkın

The effect of employing knowledge workers from technologically advanced countries: The knowledge spillover caused by the mobility of knowledge workers in electronic industries in Asia 1342
Ayano Fujiwara

Compact wireless control network protocol with fast path switching 1350
Yasutaka Kawamoto, Makoto Kubomi, Yuichi Kado

A Meta-analysis of Content and Language Integrated Learning in English 1358
Jong-Keol Kim, Jeong-ryeol Kim

Hybrid Ontology-based knowledge with multi-objective optimization model framework for Decision Support System in intercropping 1363
Kornkanok Phoksawat, Massudi Mahmuddin

Evaluating the Effect of Mozart Music and White Noise on Electroencephalography Pattern toward Visual Memory 1372
Syarifah Noor Syakiylla Sayed Daud, Rubita Sudirman

A Way for Measuring the Temperature Transients of Capacitors 1381
Zoltan Sarkany, Marta Rencz

Radiation Hybrid Mapping: A Resampling-based Method for Building High-Resolution Maps 1390
Raed Seetan, Jacob Bible, Michael Karavias, Wael Seitan, Sam Thangiah

Two-axis Gimbal with Passive Revolute Joints for Motion Isolation of a Stabilizing Mechanism 1401
Erick Moya, Keizo Araki, Masaki Okada, Akira Mizuno, Kenjiro Tadakuma, Riichiro Tadakuma

Modeling and Performance Evaluation of a Top Gated Graphene Withdrawn
MOSFETJith Sarker, Abu Shami Md. Zadid Shifat,Rezoan Ahmed Shuvro

CAM-Infrastructure: a novel solution for service advertisement in Cooperative Intelligent Transportation Systems 1422
Badis Hammi, Houda Labiod, Gérard Segarra, Alain Servel, Jean Philippe Monteuuis

Use of machine learning techniques in the prediction of credit recovery 1432
Rogerio Gomes Lopes, Marcelo Ladeira, Rommel Novaes Carvalho

Monocular Height Estimation Method with 3 Degree-Of-Freedom Compensation of Road Unevennesses 1443
Alex Masuo Kaneko, Kenjiro Yamamoto

Scalability Dilemma and Statistic Multiplexed Computing — A Theory and Experiment 1453
Justin Yuan Shi, Yasin Celik

Embedded C Programming Using FRDM to Enhance Engineering Students' Learning Skill 1461
Zulzilawati Jusoh, Mohamad Taib Miskon, Hasnorhafiza Husni, Nur Hafizah Rabi'ah Husin

A Modified Patch Antenna with Square-Open Loop Resonator Slot for Improved Bandwidth Performance in WiFi Applications 1467
Rida Gadhafi, Mihai Sanduleanu

A comparative study for using the LBC format for compressing static medical images 1472
Ildikó-Angelica Szöke, Vasile Stoicu-Tivadar, Diana Lungeanu

Image Navigation Using A Tracking-Based Approach 1478
Stephen Craig Stubberud, Kathleen Ann Kramer, Allen Roger Stubberud

Analytical and Algorithmic Approaches to Determine the Number of Sensor Nodes for Minimum Power Consumption in LWSNs 1487
Nukhet Sazak, Ali Soner Kilinc

System for the Automatic Estimation of the Tilt Angle of a Flat Solar Collector 1491
Jorge Fonseca-Campos, Paola N. Cortez-Herrera, Israel Reyes-Ramírez

Source Code Vulnerabilities in IoT Software Systems 1502
Saleh Mohamed Alnaeli, Melissa Sarnowski, Md Sayedul Aman, Ahmed Abdelgawad, Kumar Yelamarthi

Lookup Tables-based mean level detection of spatially distributed targets in non Gaussian clutter 1508
Nabila Nouar, Atef Farrouki

A Machine Learning based Framework for Parameter based Multi-Objective Optimisation of Video CODECs 1515
Maryam Al-Barwani, Eran A. Edirisinghe

2-D and 3-D Visualization of Many-to-Many Relationships 1527
SeungJin Lim

On the Performance of a Clustering-based Task Scheduling in a Heterogeneous System 1540
Hidehiro Kanemitsu, Masaki Hanada, Emilia Ndilokelwa Weyulu, Moo Wan Kim

Modeling and Control of Collaborative Robot System using Haptic Feedback 1549
Vivekananda Shanmuganatha, Lad Pranav Pratap, Pawar Mansi Shailendrasingh

Lifelong Learning Application: Mobile Application to Promote Lifelong Learning and Introduce Educational Institutions 1556
Wongpanya Nuankaew, Pratya Nuankaew, Kanakarn Phanniphong

Planning an Availability Demonstration Test with Consideration of Confidence Level 1565
Frank Müller, Bernd Bertsche

A Reliable, Non-Invasive Approach to Data Center Monitoring and Management 1577
Moises Levy, Jason O. Hallstrom

An Ontology-Based Temporal Scheduling and Complexity Reduction Model for Timetabling Knowledge Management 1585
Patience Usoro Usip, Enobong Moses Umoren, Udoinyang Godwin Inyang, Mfonobong Maurice Ntekop

COTS – Harsh Condition Effects Considerations from Technology to User Level 1592
Kirsten Weide-Zaage, Guillermo Payá-Vayá

A Secure Automated Elevator Management System and Pressure Sensor based Floor Estimation for Indoor Mobile Robot Transportation 1599
Ali Abduljalil Abdulla, Hui Liu, Norbert Stoll, Kerstin Thurow

Call Arrival Rate Prediction and Blocking Probability Estimation for Infrastructure based Mobile Cognitive Radio Personal Area Network 1609
Neeta Nathani, Gopal Chandra Manna

Soft Handoff Evaluation and Efficient Access Network Selection in Next Generation Cellular Systems 1616
Moses Ekpenyong, Daniel Asuquo, Samuel Robinson, Imeh Umoren, Etebong Isong

Proposal of an Embedded Methodology that uses Organizational Diagnosis and Reengineering: Case of bamboo panel company 1626
Eva Selene Hernández Gress, Antonio Oswaldo Ortega Reyes, Jaime Garnica González, José Ramón Corona Armenta

Three-Dimensional EEG Signal Tracking for Reproducible Monitoring of Self-Contemplating Imagination 1634
Qinglei Meng, Deepa Gupta, Abenezer Wudenhe, Xiaoming Du, L. Elliot Hong, Fow-Sen Choa

Optimizing the front-end power solution for automotive ADAS systems 1647
Beniamin Dragoi

Emotional state recognition in speech signal 1654
Krystian Kapala, Dawid Krawczyk, Stefan Brachmanski

Factors Affecting the Adsorption of Trivalent Chromium Ions by Activated Carbon Prepared from Waste Rubber Tyres 1660
Sylvia E. Benjamin, Muhammad Ashfaq Sajjid

A prospective design of a social assistive electronic system for the elderly 1665
Arsénio Reis, Hugo Paredes, Isabel Barroso, Maria João Monteiro, Vitor Rodrigues, Salik Ram Khanal, João Barroso

Theoretical developments for interpreting kernel spectral clustering from alternative viewpoints 1670
Diego Peluffo-Ordóñez, Paul Rosero-Montalvo, Ana Umaquinga-Criollo, Luis Suárez-Zambrano, Hernan Domínguez-Limaico, Omar Oña-Rocha, Stefany Flores-Armas, Edgar Maya-Olalla

Modeling Double Subjectivity for Gaining Programmable Insights: Framing the Case of Uber 1677
Loretta Henderson Cheeks, Ashraf Gaffar, Mable Johnson Moore

1-D Wavelet Signal Analysis of the Actuators Nonlinearities Impact on the Healthy Control Systems Performance 1693
Nicolae Tudoroiu, Sorin Mihai Radu, Elena-Roxana Tudoroiu, Wilhelm Kecs, Nicolae Ilias, Liana Elefterie

Pluggable Controllers and Nano-Patterns in Java with Lola 1711
Yossi Gil, Ori Marcovitch, Matteo Orru, Ori Roth

On Modeling Affect in Audio with Non-Linear Symbolic Dynamics 1727
Pauline Mouawad, Shlomo Dubnov

Retrieving Dialogue History in Deep Neural Networks for Spoken Language Understanding 1741
Myoung-Wan Koo, Guanghao Xu, Hyunjung Lee, Jungyun Seo

A Novel Approach to Sensor and Actuator Integrity Monitoring and Fault Magnitude Reconstruction 1748
Weihua Li, Chiman Kwan

A Comparison of Mean Models and Clustering Techniques for Vertebra Detection and Region Separation from C-Spine X-Rays 1758
Anum Mehmood, M. Usman Akram, Anum Tariq, Ayesha Fatima

Monitoring power breakers using vibro acoustic techniques 1771
Horia Balan, Traian Varodi, Mircea I. Buzdugan

A Cost Effective Security Technology Integrated with RFID Based Automated Toll Collection System 1777
Rafiya Hossain, Moonmoon Ahmed, Hasan Uz Zaman

Investigating the Expertise Indicators of Vulnerability Discovery Professionals 1784
Mortada Al-Banna, Boualem Benatallah, Moshe Chai Barukh

Optimization of Depth-Based Routing for Underwater Wireless Sensor Networks through Intelligent Assignment of Initial Energy 1799
Tanveer Ahmed, Muhammad Kaleem, Khurram Saleem Alimgeer, Mustafa Shakir, Sajid Nazir

Improved Hybrid Opponent System for Professional Military Training 1804
Michael Pelosi, Michael Brown, Kinza Ahmad

Smartphone Based Heart Attack Risk Prediction System with Statistical Analysis and Data Mining Approaches 1815
M. Raihan, Saikat Mondal, Arun More, Pritam Khan Boni, Md. Omar Faruqe Sagor

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1-6 (2017)

www.astesj.com

Special issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Transmit / Received Beamforming for Frequency Diverse Array with Symmetrical frequency offsets

Shaddrack Yaw Nusenu*

University of Electronic Science and Technology of China, School of Communication and Information Engineering, China.

Koforidua Technical University, Electrical / Electronic Department, Faculty of Engineering, Ghana.

ARTICLE INFO

Article history:
Received: 17 February, 2017
Accepted: 07 March, 2017
Online: 15 March, 2017

Keywords:
Frequency diverse array
Linear frequency increment
Symmetrical frequency offset
Transmit/received beamforming

ABSTRACT

Frequency diverse array (FDA) employs a small frequency increment across its array elements to provide range, angle and time dependent beampattern. In most radar systems, FDA with linear frequency offsets have been employed. But, its transmit beampattern is coupled in range and angle dimensions, which may degrade the output signal-to-interference-plus-noise ratio performance. Note that nonlinear offset is also beneficial in FDA radar. In this paper, symmetrical frequency offsets based FDA is investigated which removes the inherent periodicity of FDA beampattern to achieve a single maxima in the target area. In doing so, the signal information at the receiver is improved significantly. Transmit/received beampatterns, signal-to-interference-plus-noise ratio (SINR) and detection probability of the proposed method are compared with the conventional FDA. Computer simulations are used to verify the effectiveness of the proposed method.

1. Introduction

Frequency diverse array (FDA) utilizing a small frequency increment across each antenna elements was proposed by [1], to steer the beam electronically in angle and range dimensions. The FDA concept has range dependent beamforming capability. It was observed that frequency increment makes the array beampattern changes as a function of the range, angle and time [1]–[3]. In [4], FDA applications for various modes of operations in radar systems has been presented. The periodicity of beam pattern in time, range and angle was investigated in [5]. In [6], a linear FDA was proposed to address the range ambiguous clutters and improves the detection ability in a relatively slow moving targets. The radiation capabilities of FDA was presented in [7] to show its beam scanning features. Also it was proven that the scanning speed was related to frequency offset employed between two neighboring antenna elements. In [8], the range and angle coupled beamforming with frequency diverse chirp signals was presented. In addition, [9] investigated FDA range-angle dependent beamforming to suppress interferences at different ranges and directions.

FDA has a lot of promising application [10] which has sparked many interesting investigations in finding a suitable frequency offset to provide range-angle dependent. It is important to mention that frequency offset across the FDA elements plays essential role in improving the overall performance of an FDA radar, especially in controlling range-angle dependency and spatial distribution of generated beam pattern [11], [12]. Hence, researchers have shown great interest to investigate frequency offset between the adjacent elements of FDA to improve its performance. Since FDA offers a range-angle dependent beampattern, it is of great significance as this provides a potential for range-angle localization of targets, however, the beampattern of the conventional FDA is coupled in range-angle dimension.

In this regard, a nonuniform linear array was proposed in [13] to decouple the FDA transmit beampattern, however, the carrier frequency and/or frequency increments cannot be altered in real-time because it requires relocating the elements mechanically. In [14] and [15], logarithmically increasing frequency offsets and time-dependent frequency offsets, respectively are reported to decouple the range-angle beampattern. However, these methods resulted in poor beamforming performance in the range dimension. Note that the best decoupling way is to form a dot-shaped beampattern rather than S-shaped beampattern.

In this paper, transmit/received beamforming FDA radar with symmetrical frequency offset to decouple range and angle beampattern in order to produce a single maxima is investigated. The idea is to use the symmetrical frequency offset to provide extra

*Shaddrack Yaw Nusenu, UESTC-China and Koforidua Technical University-Ghana, Email: nusenu2012gh@yahoo.com

degree of freedom in terms of producing single maxima for a particular target of interest in a given region. Due to single maxima the received signal at the receiver array will be better than the signal for the conventional FDA. Furthermore, the proposed transmit/received beamforming are evaluated by the signal-to-interference-plus-noise ratio (SINR) and detection probability. Results shows the superiority of the proposed method.

The remaining sections are organized as follows: Section II presents the signal model of the conventional FDA, followed by the proposed symmetrical frequency offset FDA radar. Section III presents proposed method transmit / received beamforming. Section IV derives the SINR and detection probability. Section V presents results and discussions and concluding summaries are drawn in Section VI.

2. Background

2.1. Conventional FDA Range-Angle Dependent Transmit Beampattern

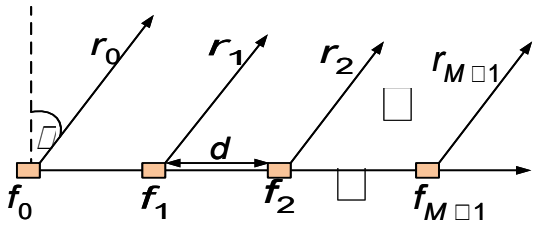


Figure 1: Illustration of linear frequency offset conventional FDA.

Figure 1 shows the conventional FDA antenna array structure. The narrow-band monochromatic signal radiated from each element is identical but with a frequency increment Δf. The radiated frequency of the mth element can be expressed as

$$f_m = f_0 + m\Delta f, \quad m = 0,1,\ldots,M-1 \tag{1}$$

where f_0 is the carrier frequency and M is the number of array elements. Then, the signal transmitted by the mth element can be represented by

$$s_m(t) = \exp\left(-j2\pi f_m t\right) \tag{2}$$

The signal arriving at a far-field point (r,θ) can be expressed as

$$s_m\left(t-\frac{r_m}{c_0}\right) = \exp\left[-j2\pi f_m\left(t-\frac{r_m}{c_0}\right)\right] \tag{3}$$

where c_0 is the speed of light and $r_m \approx r - md\sin\theta$, with d being the element spacing, is the target slant range for the mth element. If uniform weights (all ones) are applied, the array factor can be approximately derived as [16]

$$AF\left(t,r,\theta\right) = \sum_{m=0}^{M-1}\exp\left[-j2\pi f_m\left(t-\frac{r_m}{c_0}\right)\right]$$

$$\approx e^{-j\Phi_0}\frac{\sin\left[M\pi\left(\Delta f t-\frac{\Delta f r}{c_0}+\frac{df_0\sin\theta}{c_0}+\frac{d\Delta f\sin\theta}{c_0}\right)\right]}{\sin\left[\pi\left(\Delta f t-\frac{\Delta f r}{c_0}+\frac{df_0\sin\theta}{c_0}+\frac{d\Delta f\sin\theta}{c_0}\right)\right]} \tag{4}$$

where

$$\Phi_0 = 2\pi f_0\left(t-\frac{r}{c_0}\right)-p\left(M-1\right)\frac{\Delta f r}{c_0}+p\left(M-1\right)\frac{df_0\sin\theta}{c_0}+p\left(M-1\right)\frac{d\Delta f\sin\theta}{c_0}$$

Equation (4) implies that FDA has range, angle and time dependent transmit beampattern. However, the beampattern is coupled in the range and angle dimensions and consequently it will have multiple maximums. To avoid the dependence of range gain on the angle, the FDA parameters should be properly designed.

2.2. Proposed Symmetrical Frequency offset FDA for Transmit / Received Beamforming

In this paper, the symmetrical frequency offsets is proposed to decouple range and angle beampattern in order to localize a target in a given region. In designing symmetrical frequency offset, the central element is chosen as the symmetric point depicted in Figure 2.

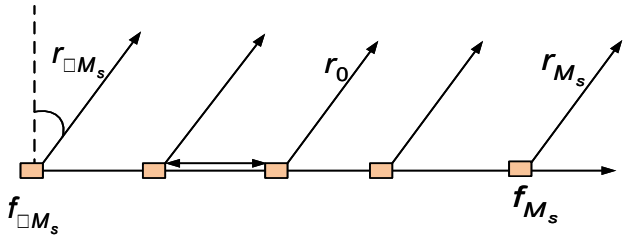


Figure 2: Illustration of proposed symmetrical frequency offset FDA.

In this paper, odd number of elements are assumed, namely, $M = 2Ms - 1$. The frequency offset for the kth element can be expressed as

$$\Delta f_k = |k|\Delta f, \quad k = -M_s+1,\ldots,0,1,\ldots,M_s+1 \tag{5}$$

Note that in order to show distinct representations, a new element index $k \in [-M_s+1,\cdots,0,\cdots,M_s-1]$ instead of m defined in (1) is adopted. Reformulate (4) as

$$AF\left(\theta,r,t\right) = \sum_{k=-M_s}^{M_s}\frac{1}{r_k}\exp\left[-j2\pi\left(f_0+|k|\Delta f\right)\left(t-\frac{r_k}{c_0}\right)\right]$$

$$\approx \sum_{k=-M_s}^{0}\frac{1}{r}\exp\left[-j2\pi\left(f_0-k\Delta f\right)\left(t-\frac{r_k}{c_0}\right)\right]$$

$$+\sum_{k=1}^{M_s}\frac{1}{r}\exp\left[-j2\pi\left(f_0+k\Delta f\right)\left(t-\frac{r_k}{c_0}\right)\right] \tag{6}$$

Similarly, $r_k = r - kd\sin\theta$ is used, where r denotes here as the reference range to the central element. Equation (6) can be rewritten in vector notation:

$$AF(\theta,r,t) \approx \left[\mathbf{w}^H \mathbf{a}(\theta,r,t)\right] \quad (7)$$

Where ^{H} is the conjugate transpose operator, w is the $M \times 1$ weighting vector, and transmit steering vector a (θ,r,t) is

$$\mathbf{a}(\theta,r,t) = \frac{1}{r}\left[\exp(-j\Phi_{-M_s}),...,1,...,\exp(-j\Phi_k),...,\exp(-j\Phi_{M_s})\right]^T \quad (8)$$

with ^{T} being the transpose operator and

$$\Phi_k = 2\pi\left(f_0 + k\Delta f\right)\left(t - \frac{r_k}{c_0}\right) \quad (9)$$

The $\boldsymbol{w}$ can be optimally designed to synthesize the desired transmit beampattern.

3. Transmit / Received Beamforming for Symmetrical frequency offset FDA

In this section transmit and received beamforming for proposed method is presented. The transmit beamforming can be expressed as

$$B_T(\theta,r) = \frac{\left|\mathbf{w}^H a(\theta,r)\right|^2}{\left|\mathbf{w}^H a(\theta_d,r_d)\right|^2} \quad (10)$$

The normalized received beampattern can be written as

$$B_r(\theta,r) = \frac{\left|\mathbf{w}_r{}^H \mathrm{v}(\theta,r)\right|^2}{\left|\mathbf{w}_r{}^H \mathrm{v}(\theta_d,r_d)\right|^2} \quad (11)$$

where $\mathrm{v}(\theta,r) = \left(\mathbf{w}^H a(\theta,r)\right)\mathrm{b}(\theta)$ denotes the virtual steering vector v (θ,r) and b(θ) being the received steering vector due to the propagation delays from a source to the receive elements.

4. Performance Analysis

4.1. SINR

The performance analysis of the proposed method are evaluated by signal-to-interference-plus-noise ratio (SINR) and probability of detection. First of all SINR can be expressed as

$$SINR = \frac{\sigma_d^2\left|\mathbf{w}_r{}^H \mathrm{v}(\theta_d,r_d)\right|^2}{\mathbf{w}_r{}^H Q_{i+n} \mathbf{w}_r} \quad (12)$$

where σ_d^2 is the variance of the target reflection coefficient and Q_{i+n} denotes the interference-plus-noise covariance matrix which is given as

$$Q_{i+n} \approx \sum_{i=1}^{F} \sigma_i^2 \mathrm{v}(\theta_i,r_i)\mathrm{v}^H(\theta_i,r_i) + \sigma_n^2 I \quad (13)$$

where σ_i^2 is the variance of the ith interference reflection coefficient, F is the number of interferences and $\sigma_n^2 I$ is the covariance matrix with I being an identity matrix. Plugging in (13) and $\mathbf{w}_r = \mathrm{v}(\theta_d,r_d)$ into (12) for any radar system, will yields SINR for that radar system.

4.2. Probability of Detection (Pd)

In this section, the probability of detection (Pd) is employed to evaluate the performance of a radar system. For the proposed method, the hypothesis problem can be written as

$$\begin{cases} H_0: \mathrm{x}(t) = n(t) \\ H_1: \mathrm{x}(t) = h(t) + n(t) \end{cases} \quad (14)$$

The noise process is assumed to be Gaussian and independent and identically distributed (i.i.d). The probability density function (PDF) can be expressed as

$$p(\mathrm{x}(t);H_0) = e^{\left(\frac{\|\mathrm{x}(t)\|^2}{\sigma_n^2}\right)} \quad (15)$$

$$p(\mathrm{x}(t);H_1) = e^{\left(\frac{\|\mathrm{x}(t)\|^2}{\sigma_n^2}\right)} \times e^{\left(\frac{\|h(t)+n(t)\|^2}{2\sigma_n^2}\right)} \quad (16)$$

The likelihood ratio test can be given as

$$g = \frac{p(\mathrm{x}(t);H_1)}{p(\mathrm{x}(t);H_0)} \underset{H_0}{\overset{H_1}{\gtrless}} x \quad (17)$$

The probability of detection p_d and probability of false alarm p_{fa}, respectively, is expressed as in [17]

$$p_d = p(g > x|H_1) = 1 - S_{y^2_{(2)}}\left(\frac{\sigma_n^2 S^{-1}_{y^2_{(2)}}(1-p_{fa})}{M^2N + \sigma_n^2}\right) \quad (18)$$

$$p_{fa} = p(g > x|H_0) = 1 - S_{y^2_{(2)}}\left(\frac{2x}{\sigma_n^2}\right) \quad (19)$$

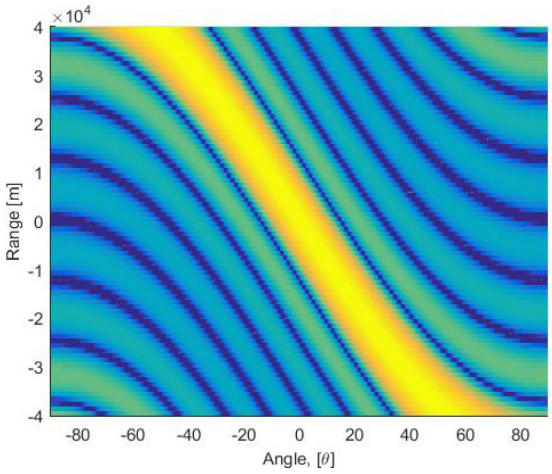
where S $(\cdot)$ denotes the cumulative distributive function, $y^2_{(2)}$ being the chi-square distribution, $P - N + k$ are the number of elements in kth array and R are the number of receiver array elements.

5. Results and Discussions

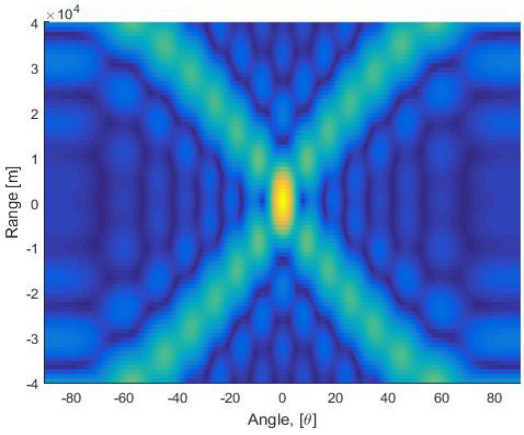
In the simulations, the carrier frequency f_0 = 10 GHz, frequency offset Δf = 3 KHz. Both transmit and receive arrays M and N, respectively, have 8 antenna elements spaced by half wavelength $d = \frac{c_0}{2f_0}$. The target is assumed to be located at (θ_d, r_d) = (50°,50 km). Gaussian noise is utilized as additive noise with zero mean spatially and temporally white random sequence having same variance on each array elements.

A. Transmit Beampattern

First, normalized transmit beampattern for FDA and the proposed symmetrical FDA is presented to show the effectiveness of symmetrical FDA. Note that the two arrays are steered to instantaneous angle θ = 0° and the beam direction ranges are also normalized to zero-range. Figure 3(a) illustrates the conventional FDA beampattern which is coupled in the range-angle dimension.



(a)



(b)

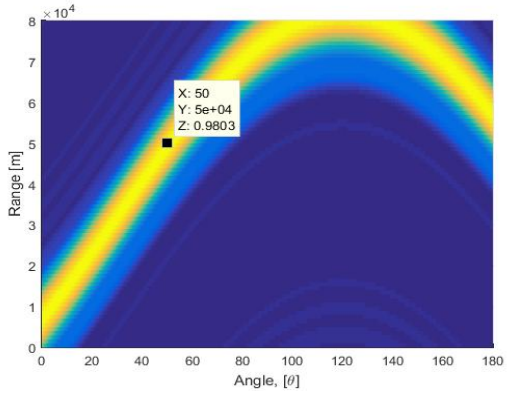
Figure 3: Comparisons of normalised transmit beampattern when Δf = 3 KHz is adopted: (a) Conventional FDA, (b) Proposed symmetrical FDA.

In contrast, Figure 3(b) shows the proposed symmetrical FDA which has decoupled range-angle beampattern. And this is beneficial for targets localization.

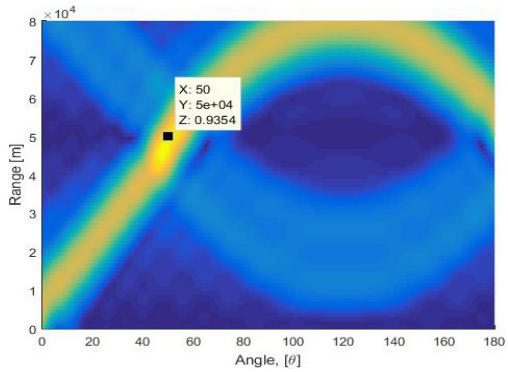
As shown in Figure 4(a), the conventional FDA beampattern shows the problem of periodic maxima. This maxima are undesirable for proper target localization. On the other hand, Figure 4(b) shows the proposed method beampattern which produce exactly one maxima at target location and no other maxima is seen in the entire region of observation.

B. Received Beampattern

In Figure 5(a) and 5(b) both radars produce maxima at the intended range-angle pair, but the conventional FDA has more serious ambiguity problem shown by the arrow in Figure 5(a). It is important to mention that because the conventional FDA has larger total frequency offset, it achieved higher range resolution than the proposed method. In the case of the proposed method, the target is better localized than the conventional FDA. It can be noticed that there is no spread of beampattern in both range and angle dimensions.

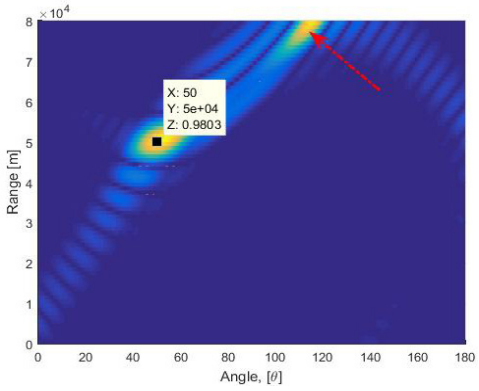


(a)

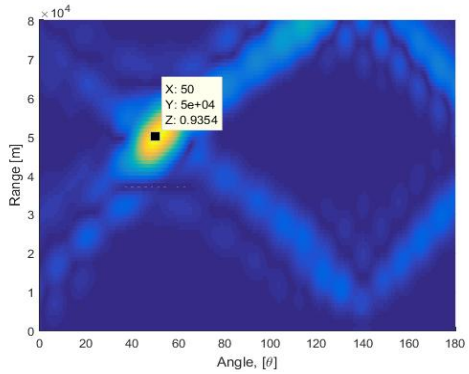


(b)

Figure 4: Comparisons of transmit beampattern when the target is present at (θ_d, r_d) = (50°,50 km): (a) Conventional FDA, (b) Proposed symmetrical FDA.



(a)



(b)

Figure 5: Comparisons of transmit beampattern when the target is present at (θ_d, r_d) = (50°,50 km): (a) Conventional FDA, (b) Proposed symmetrical FDA.

Finally, SINR and probability detection performance of the proposed method and the conventional FDA have been plotted. It can be observed that in Figure 6, the SINR of proposed symmetrical FDA is better than the conventional FDA. Thus proposed method has better robustness against the interferences.

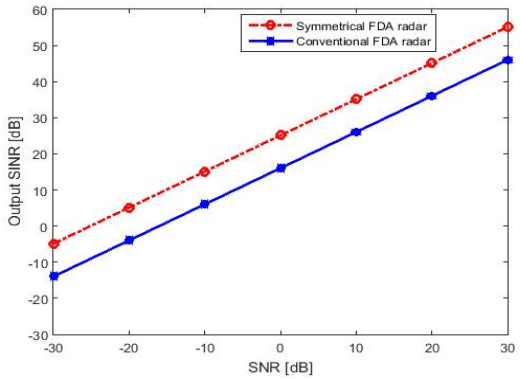


Figure 6: SINR versus SNR performance.

Figure 7 shows the detection probability versus SNR for the proposed method and the conventional FDA radar. Proposed method exhibits better detection performance compared to the conventional FDA radars.

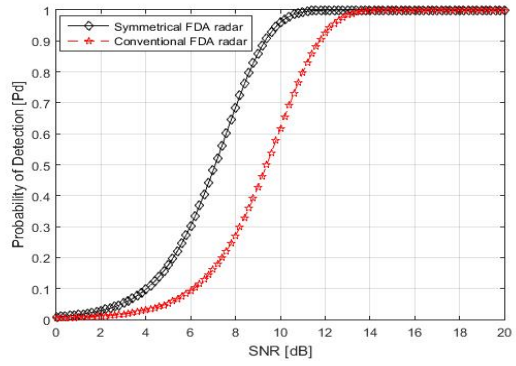


Figure 7: Probability of detection versus SNR performance.

Note that the improvement in performance in terms of SINR and detection probability can be attributed to single maxima due to symmetrical frequency offset employed across the transmit antenna array.

6. Conclusion

A symmetrical frequency offset based FDA has been presented. The proposed method avoids the periodic maxima presented in conventional FDA radar and focus the target with multiple beams to get single maxima for better target localization. Numerical results shows that the proposed method exhibits better robustness against interference as well as better detection performance than the conventional FDA radar due to the employment of symmetrical frequency offset across the transmit antenna array. Overall, the proposed method received beampattern show a reasonable improvement compared to the conventional FDA.

Conflict of Interest

The author declare no conflict of interest.

Acknowledgment

The author would like to thank Respected Professor Wen-Qin Wang (*Senior Member - IEEE*) of UESTC for his great contributions, support and encouragement.

References

[1] Antonik, P., Wicks, M.C., Griffiths, H.D., et al. "Frequency diverse array radars," Proc. of the IEEE Radar Conf., Verona, NY, April 2006, pp. 215-217

[2] Wicks, M.C., Antonik, P. "Frequency diverse array with independent modulation of frequency, amplitude, and phase," U.S.A Patent 7,319,427, 15 January 2008

[3] Wicks, M.C., Antonik, P. "Method and apparatus for a frequency diverse array," U.S.A Patent 7.511, 665B2, 31 March 2009

[4] Antonik, P., M. C. Wicks, H. D. Griffiths, and C. J. Baker. "Multi-mission multi-mode waveform diversity," Proc. IEEE Radar Conf. Dig., 580582, Verona, NY, USA, Apr. 2006, pp. 24-27.

[5] Secmen M., Demir S., Hizal A., Eker T. "Frequency diverse array antenna with periodic time modulated pattern in range and angle," IEEE Conference on Radar, 2007

[6] Wen-Qin Wang. "Range-Angle Dependent Transmit Beampattern Synthesis for Linear Frequency Diverse Arrays," IEEE Transactions Signal Processing. 61 (8), 2013, pp. 4073–4081.

[7] Huang S., Tong K. F., Baker C. J. "Frequency diverse array with beam scanning feature," Proc. IEEE Antennas Propag. Conf., USA, July 2008.

[8] Higgins T. and S. Blunt. "Analysis of range-angle coupled beamforming with frequency diverse chirps," Proceedings of the 4th International Waveform Diversity and Design Conference, Orlando, FL, Feb. 2009, pp. 140-144.

[9] Huaizong Shao, Jingchi Li, Hui Chen, and Wen-Qin Wang. "Adaptive Frequency Offset Selection in Frequency Diverse Array Radar," IEEE Antennas and Wireless Propagation Letters, vol. 13, 2014.

[10] W.-Q. Wang. "Overview of frequency diverse array in radar and navigation applications," IET Radar, Sonar and Navigation, DOI: 10.1049/ietrsn.2015.0464, vol. 10, 2016.

[11] Zhuang, L. and X. Z. Liu. "Precisely beam steering for frequency diverse arrays based on frequency offset selection," Proc. Int. Radar Conf., pp. 1-4, 2009.

[12] Chen, Y.-G., Y.-T. Li, Y.-H. Wu, and H. Chen. "Research on the linear frequency diverse array performance," Proc. IEEE 10th Int. Conf. on Signal Processing, 23242327, Beijing, Oct. pp. 24-28, 2010.

[13] W.-Q. Wang, H. C. So, and H. Z. Shao. "Nonuniform frequency diverse array for range-angle imaging of targets," IEEE Sensors Journal, vol. 14, no. 8, pp. 2469-2476, August 2014.

[14] W. Khan, I. M. Qureshi, and S. Saeed. "Frequency diverse array radar with logarithmically increasing frequency offset," IEEE Antennas and Wireless Propagation Letters, vol. 14, 2015.

[15] W. Khan and I. M. Qureshi. "Frequency diverse array radar with time dependent frequency offset," IEEE Antennas and Wireless Propagation Letters, vol. 13, 2014.

[16] Sammartino, P.F., Baker, C.J., Griffiths, H.D. "Frequency diverse MIMO techniques for radar," IEEE Trans. Aerosp. Electron. Syst., 2013, 49, (1), pp. 201-222.

[17] W. Q. Wang and H. Shao. "Range-angle localization of targets by a double-pulse frequency diverse array radar," Sel. Top. Signal Process IEEE J., vol. 8, no. 1, pp. 106-114, 2014.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 7-11 (2017)

www.astesj.com

Special issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Detailed Analysis of Amplitude and Slope Diffraction Coefficients for knife-edge structure in S-UTD-CH Model

Eray Arik[1,*], Mehmet Baris Tabakcioglu[2]

[1]Ermaksan, EMC Division, TURKEY

[2]Bursa Technical University, Electrical- Electronics Engineering, 16310, TURKEY

ARTICLE INFO

Article history:

Received :01 March, 2017

Accepted: 17 March, 2017

Online: 25 March, 2017

Keywords:

Diffraction coefficient

S-UTD-CH model

Wave Diffraction

ABSTRACT

In urban, rural and indoor applications, diffraction mechanism is very important to predict the field strength and calculate the coverage accurately. The diffraction mechanism takes place on NLOS (non-line-of-sight) cases like rooftop, vertex, corner, edge and sharp surfaces. S-UTD-CH model computes three type of electromagnetic wave incidence such as direct, reflected and diffracted waves, respectively. As obstacles in diffraction geometry are in the same or closer height, contribution of the diffraction mechanism is dominant. To predict the diffracted fields accurately, amplitude and slope diffraction coefficients and the derivative of these coefficients have to be taken correctly. In this paper, all the derivations about diffraction coefficients are made for knife edge type structures and extensive simulations are performed in order to analyze the amplitude and diffraction coefficients. In plane angle diffraction, contributions of amplitude and slope diffraction coefficient are maxima.

1. Introduction

Diffraction means bending of electromagnetic waves from a sharp edge, wedge, corner and narrow slits. The diffraction takes place in the case of relatively higher wavelength of electromagnetic wave from obstacle or slit aperture. This paper is an extension of work originally presented in Electromagnetic wave theory symposium [1]. In case that wavelength of electromagnetic wave is less than the height of obstacle; high frequency asymptotic techniques are used [2]. Geometrical optics model runs for refraction, reflection, direct propagation and have been used for a long time in wave propagation in lit region. Geometrical optics model cannot explain diffraction concept [3], and thus cannot give accurate field prediction particularly behind an obstacle. In order to solve the diffraction problem in the field prediction, geometrical theory of diffraction (GTD) model is introduced [4]. However, GTD model gives inaccurate outcomes in optical boundaries [5]. GTD gives inaccurate results in the situation that source, diffraction and observation points are in the same line [6]. GTD model is an extension to geometric optic model by adding diffracted fields terms and used in high frequency spectrum [7]. UTD model is introduced in order to overcome singularity problem of GTD model [8]. UTD model gives appropriate and exact outcomes in case of single diffraction [5]. UTD model is also an extension to geometric optic model and reflected waves are also considered in field prediction [6]. UTD model gives inaccurate outcomes in field prediction in the situation of closer building height. In order to predict the field strength more exactly, the second order diffracted field terms are added to total field at the receiver [9]. Slope diffraction model is introduced in [10, 11]. However, there are some singularities in field prediction at shadow boundary points [12]. Higher order diffraction coefficient and diffracted fields are developed to predict the relative path loss at the receiver for wedge and knife edge structures [13-15]. Amplitude and slope diffraction coefficients are developed via providing amplitude, slope and phase continuities in optical boundaries [16]. Slope UTD model applied multi-shaped cascaded obstacles [17] and adapted into time domain and all formulas are

*Corresponding Author: Mehmet Barış Tabakcioglu, Bursa Technical University, Mimar Sinan Campus, Yildirim, Bursa, TURKEY, mehmet.tabakcioglu@btu.edu.tr

www.astesj.com
https://dx.doi.org/10.25046/aj020302

derived in time domain [18]. A new GTD model is introduced for equal heights case [19]. After 10 diffractions slope UTD model loses the accuracy and requires more computation time. S-UTD-CH model is proposed to predict the field strength accurately after 10 diffractions for knife edge and wedge structures [20, 21]. Amplitude and slope diffraction coefficients are used in S-UTD-CH model developed in MATLAB environment. Firstly location and height data of transmitter, obstacle and receiver is entered to program. And then a ray tracing is made and all contribution to total path loss at the receiver is determined. Afterwards, by using amplitude and slope diffraction coefficient path loss is calculated. As an example a simple test scenario is used in order to show that effect of diffraction coefficient is maximum in case of plane angle diffraction. All the simulations are performed in MATLAB.

2. Amplitude Diffraction Coefficient

To calculate the total field behind an obstacle formula [8] is given by

$$E = [E_i D]A(s)e^{-jks} \quad (1)$$

where, E_i is incoming field, D is amplitude diffraction coefficient. $A(s)$ stands for spreading factor. s and k represent travelled distance and wave number, respectively. The amplitude diffraction coefficient for the knife edge type structure is given [12] by

$$D(\alpha) = -\frac{e^{-j\pi/4}}{2\sqrt{2\pi k\cos(\alpha/2)}}F[x] \quad (2)$$

where, α is diffraction angle ($\alpha \equiv \varphi - \varphi'$) and illustrated in Fig. 1. k is wave number and $F[x]$ refers to transition function.

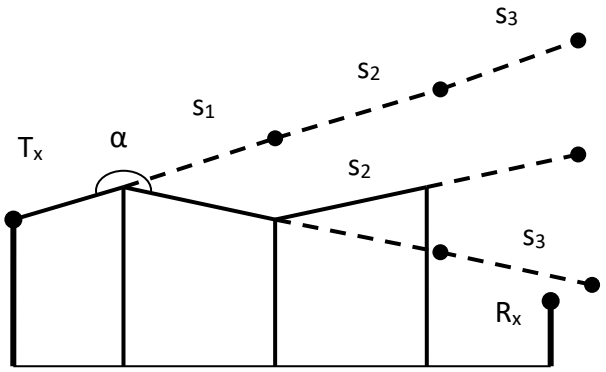


Fig.1. Multiple Knife edge Diffraction Scenario

Amplitude diffraction coefficient used in calculation of the diffracted fields behind the obstacle. The amplitude diffraction coefficient changes with respect to diffraction angle. The diffraction angle is measured between incoming and diffracted fields as depicted in Fig. 1. Changing of the amplitude diffraction coefficient with regard to diffraction angle is shown in Fig. 2.

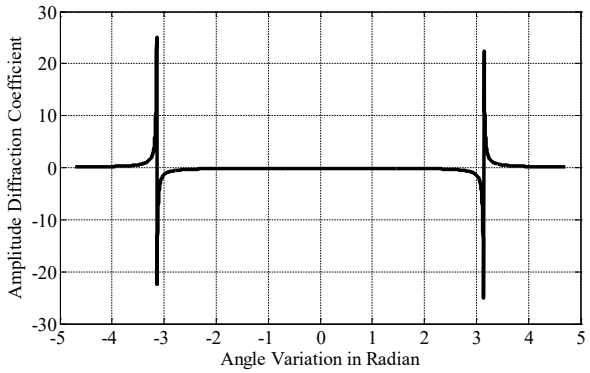


Fig. 2. Amplitude diffraction coefficient vs diffraction angle

As it is seen in Fig. 2, the diffraction angle variation is indicated on horizontal axis. Variation of the amplitude diffraction coefficient is indicated on vertical axis. Also it can be seen in Fig. 2, the magnitude of amplitude diffraction coefficient has the maximum value at π and - π radian (180° and -180°). Plane angle diffraction takes place in the case that source, diffracting and observation points are in the same line.

3. Slope Diffraction Coefficient

In order to calculate the diffracted field behind multiple obstacles, slope diffraction coefficient have to be used to predict the field strength accurately. The slope diffraction coefficient can be calculated by the formula [12] expressed by

$$d_s(\alpha) = \frac{1}{jk}\frac{\partial D(\alpha)}{\partial \alpha} \quad (3)$$

where, k is wave number, $D(\alpha)$ is amplitude diffraction coefficient and j is complex number. The slope diffraction coefficient obtained by partial derivation of amplitude diffraction coefficient with respect to diffraction angle [12] is given by,

$$d_s(\alpha) = -\frac{e^{-j\pi/4}}{\sqrt{2\pi k}}L_s \sin(\alpha/2)\big(1 - F(x)\big) \quad (4)$$

where, k is wave number, L_s is distance parameter for slope diffraction coefficient and $F[x]$ refers to transition function.

Slope diffraction coefficient used in calculation of the diffracted fields behind multiple obstacles. The slope diffraction coefficient changes according to the diffraction angle. Again the diffraction angle is measured between incoming and diffracted fields as depicted in Fig. 1. UTD model fails in predicting of the diffracted field in case of equal or closer obstacle height. In order to remove the discontinuities in the predicted field, derivatives of the amplitude diffracted fields have to be added on the receiver. Changing of the slope diffraction coefficient with respect to diffraction angle is shown in Fig. 3.

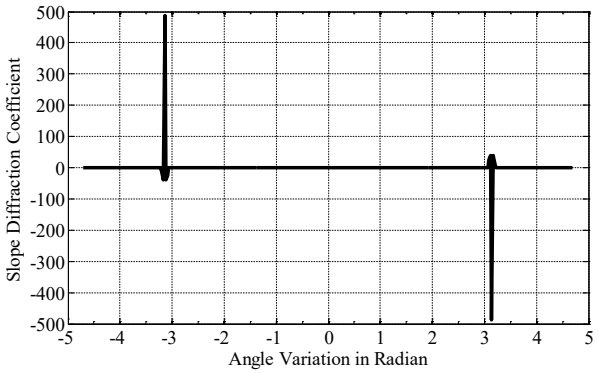


Fig. 3. Slope diffraction coefficient vs diffraction angle

As it is seen in Fig. 3, the diffraction angle variation is indicated on x-axis. Variation of the amplitude diffraction coefficient is indicated on y-axis. Also it can be seen in Fig. 3, the magnitude of the slope diffraction coefficient has the maximum value at π and - π radian (180° and -180°). By adding the slope diffracted fields, the deficiency of UTD model in equal height case is removed [22-23].

If there are 3 or more obstacles in the propagation scenario, normal derivative of the slope diffraction coefficient have to be used in prediction [12] is given by,

$$\frac{\partial d_s(\alpha)}{\partial n} = -\frac{1}{2s}\frac{e^{-j\pi/4}}{\sqrt{2\pi k}}\left\{L_s\cos(\alpha/2)[1-F(x)] + 4L_s^2 k sin^2(\alpha/2)\cos(\alpha/2)\,F'(x)\right\} \quad (5)$$

where, k is wave number, s is travelling distances, is L_s is distance parameter for slope diffraction coefficient and *F[x]* refers to transition function. Also F'[x] is derivative of transition function.

Derivative of the slope diffraction coefficient used in calculation of the diffracted fields behind multiple obstacles, when the obstacle number is greater than 2. The derivative of the slope diffraction coefficient changes with the diffraction angle. Again the diffraction angle is measured between incoming and diffracted fields as depicted in Fig. 1. Changing of the derivative of the slope diffraction coefficient with respect to diffraction angle is shown in Fig. 4.

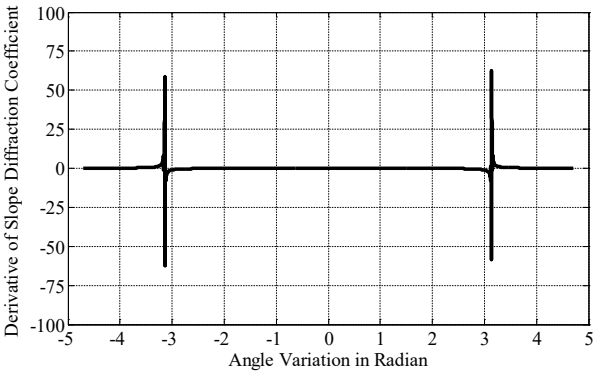


Fig. 4. Derivative of slope diffraction coefficient vs diffraction angle

As it is seen in Fig. 4, the diffraction angle variation is indicated on abscissa. Variation of the amplitude diffraction coefficient is indicated on ordinate. Also it can be seen in Fig. 4, the magnitude of the slope diffraction coefficient has the maximum value at π and - π radian (180° and -180°).

4. Results and Discussion

In order to verify the contribution of diffraction is maxima in the plane angle diffraction, test scenario in Fig. 5 is used.

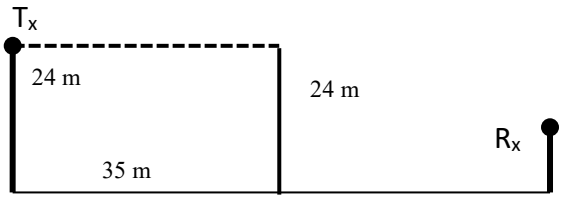


Fig. 5. Plane angle test scenario

As it is depicted in Fig. 5, the transmitter antenna height is 24 m, obstacle is 35 m apart from the transmitter and whose height is 24 m. Operating frequency is 300 MHz. Path loss variation for direct and diffracted fields with respect to the receiver height just behind the obstacle (36 m) is illustrated in Fig. 6

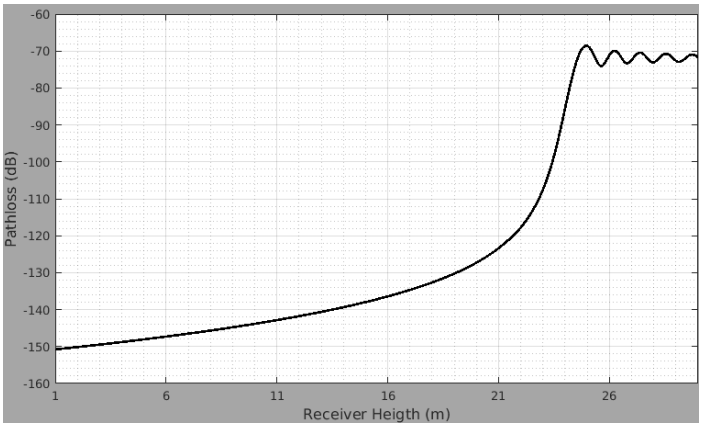


Fig. 6. Path loss of direct and diffracted fields together

As can be seen in Fig. 6, diffraction effect is maxima in the plane angle diffraction case. As the receiver height is 24 m, the angle between transmitter and receiver becomes 180°. After this height of receiver, direct field is dominated.

Path loss variation for only diffracted fields with respect to the receiver height just behind the obstacle (36 m) is illustrated in Fig. 7.

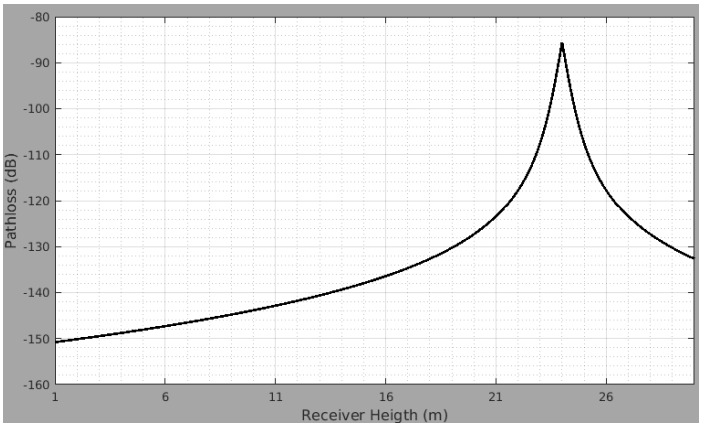


Fig. 7. Path loss of diffracted fields

As can be seen in Fig. 7, the diffraction effect is maxima in the plane angle diffraction case. As the receiver height is 24 m, the angle between transmitter and receiver becomes 180̊.

Path loss variation for reflected and diffracted fields with respect to the receiver height just behind the obstacle (36 m) is illustrated in Fig. 8.

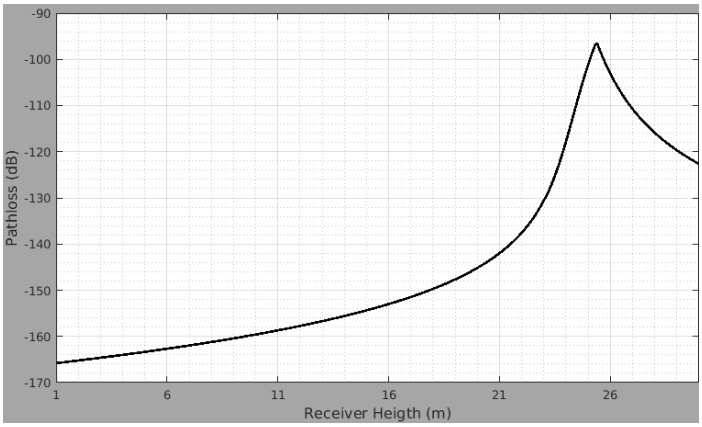


Fig. 8. Path loss of reflected and diffracted fields together

As can be seen in Fig. 8, diffraction effect is very much on the receiver position. In that position ground reflected fields make some contribution to total field. If the receiver height is increased the diffracted field contribution is decreased.

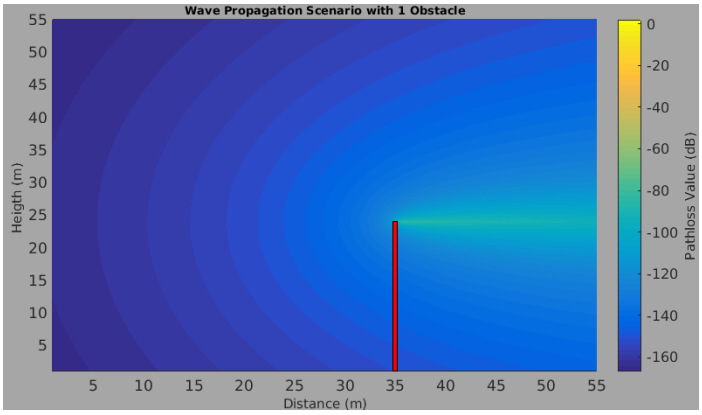


Fig. 9. Coverage Map

Furthermore, a coverage map is generated by using the simple test scenario in Fig. 5. By using only diffracted field, a coverage map is indicated in Fig. 9.

As it is seen in Fig. 9, the diffraction contribution is maxima in the case of plane angle diffraction. Plane angle diffraction takes place in the case that the receiver height is 24 m.

5. Conclusions

The diffraction mechanism is significant to predict the field strength and it is used for coverage prediction in urban, rural and indoor environment behind an obstacle like rooftop, edge, corner, and vertex. This field generated by direct, reflected and diffracted waves. The diffracted fields take place in the case of that wavelength of the electromagnetic wave is close or greater than the slit. In order to solve the diffraction problem in shadow region, a lot of electromagnetic wave propagation models are presented, such as GTD, UTD, S-UTD and S-UTD-CH. Amplitude and slope diffraction coefficient directly affect the total electromagnetic fields at the receiver. As the diffraction angle is close to plane angle, the contribution of slope diffracted field is so high. In the case of diffraction angle π radian; in other words observation, diffraction and field points are in the same line, contribution of amplitude and slope diffraction coefficient is maximum. In future, amplitude and slope diffraction coefficient will be developed for rectangular type structure.

Acknowledgment

This work is supported partially by TÜBİTAK (The Scientific and Technological Research Council of Turkey) under the grant no. 215E360.

References

[1] M.B. Tabakcioglu, "Amplitude and Slope Diffraction Coefficients for S-UTD-CH Mode", Electromagnetic Theory Symposium, Espoo, Finland, pp. 768-771, 2016.

[2] C.A., Balanis, C.A., Advanced Engineering Electromagnetics. John Wiley & Sons, 981 p, New York, USA, 1988.

[3] P.Y., Ufimtsev, Fundamentals of the Physical Theory of Diffraction, John Wiley & Sons, 329 p, New Jersey, USA, 2007.

[4] J.B., Keller, "Geometrical Theory of Diffraction". J. Opt. Soc. Amer., vol. 52, no 2, pp. 116-130, 1962.

[5] R.J. Luebbers, "A Heuristic UTD Slope Diffraction Coefficient for Rough Lossy Wedges", IEEE Trans. Antenna Propagation.,vol. 37, no 2, pp. 206-211, 1989.

[6] R.J., Luebbers, "Finite conductivity uniform GTD versus knife edge diffraction in prediction of propagation path loss". IEEE Trans. Antennas Prorpagat, vol. 32, pp. 70-76, 1984.

[7] C.A., Balanis, L. Sevgi, L. P.Y. Ufimtsev, "Fifty years of High Frequency Diffraction", International Journal of RF and Microwave Computer-Aided Engineering, vol. 23, no. 3, pp. 394-402, 2013.

[8] R.G.Kouyoumjian, P.H. Pathak, "A uniform geometrical theory of diffraction for an edge in a perfectly conducting surface", Proceedings of the IEEE, *vol.* 62, pp. 1448–1461, 1974.

[9] R.Tiberio, R. R.G. Kouyoumjian, "An analysis of diffraction at edges illuminated by transition region fields", *Radio Science*, vol. 17, no. 2, pp. 323-336, 1982.

[10] J.B. Andersen, "Transition zone diffraction by multiple edges", *IEEE Proceedings Microwave Antennas and Propagation*, vol. 141, no. 5, pp. 382-383,1994.

[11] H. Wang, "Modeling and Wideband characterization of Radio wave propagation", Ph.D. Thesis. *The University of Texas at Austin*, Austin, USA, 2005.

[12] C. Tzaras, S.R. Saunders, "An improved heuristic UTD solution for multiple-edge transition zone diffraction", *IEEE Transactions on Antennas and Propagation*, vol. 49, no: 12, pp. 1678–1682, 2001.

[13] P.D. Holm, "UTD-Diffraction Coefficients for Higher Order Wedge Diffracted Fields", *IEEE Transactions on Antennas and Propagation*, vol. 4, no: 6, pp. 879-888, 1996.

[14] P.D. Holm, "A new heuristic UTD Diffraction Coefficients for nonperfectly conducting wedges", *IEEE Transactions on Antennas and Propagation*, vol. 48, pp. 1211-1219, 2000.

[15] P.D. Holm, " Calculation of Higher Order Diffracted Fields for Multiple-Edge Transition Zone Diffraction", *IEEE Transactions on Antennas and Propagation*, vol. 52, no:5, pp.1350-1355, 2004.

[16] K. Rizk, R. Valenzuela, D. Chizhik, F. Gardiol, "Application of the slope diffraction method for urban microwave propagation prediction", *IEEE Vehicular Technology Conference*, pp. 1150-1155, Canada, 1998.

[17] G.Koutitas, C. Tzaras, "A Slope UTD Solution for a Cascade of Multishaped Canonical objects", *IEEE Transactions on Antennas and Propagation*, vol. 54, no: 10, pp. 2969-2976, 2006.

[18] A. Karousos, C. Tzaras, "Multi Time-Domain Diffraction for UWB Signals", *IEEE Transactions on Antennas and Propagation*, vol. 56, no: 5, pp. 1420-1427, 2008.

[19] A. Tajvidy, A. Ghorbani, " A New Uniform Theory-of-Diffraction-Based Model for the Multiple Building Diffraction of Spherical Waves in Microcell Environments", *Electromagnetics*, vol. 28, no: 5, pp. 375-387, 2008.

[20] M.B. Tabakcioglu, A. Kara, "Comparison of improved slope UTD method with UTD based methods and physical optic solution for multiple building diffractions", *Electromagnetics*, vol. 29, no:4, pp. 303–320, 2009.

[21] M.B.Tabakcioglu, A. Kara. "Improvements on Slope Diffraction for Multiple Wedges", *Electromagnetics*, vol. 30, no: 3, pp. 285-296, 2010.

[22] M.B.Tabakcioglu, D. Ayberkin, A. Cansiz, "Comparison and Analyzing of Propagation models with Respect to Material, Environmental and Wave Properties", *ACES Journal*, vol. 29, no:6, pp. 486-491, 2014.

[23] M.B. Tabakcioglu, "S-UTD-CH model in multiple diffractions", Internetional Journal of Electronics, vol. 103, no: 5, pp. 765-774, 2016.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 12-23 (2017)

www.astesj.com

Special issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Applications of Case Based Organizational Memory Supported by the PAbMM Architecture

Martín, María de los Ángeles[1], Diván, Mario José[2]

[1]Engineering School,Universidad Nacional de La Pampa, General Pico, 6360, Argentina

[2]Law and Economy School, Universidad Nacional de La Pampa, Santa Rosa, 6300, Argentina

ARTICLEINFO

Article history:
Received: 03 March, 2017
Accepted: 25 March, 2017
Online: 02 April, 2017

Keywords:
Organizational Memory
Data Stream Processing
Big Data
Remote Sensing
Measurement & Evaluation

ABSTRACT

In the aim to manage and retrieve the organizational knowledge, in the last years numerous proposals of models and tools for knowledge management and knowledge representation have arisen. However, most of them store knowledge in a non-structured or semi-structured way, hindering the semantic and automatic processing of this knowledge. In this paper we present a more detailed case-based organizational memory ontology, which aims at contributing to the design of an organizational memory based on cases, so that it can be used to learn, reasoning, solve problems, and as support to better decision making as well. The objective of this Organizational Memory is to serve as base for the organizational knowledge exchange in a processing architecture specialized in the measurement and evaluation. In this way, our processing architecture is based on the C-INCAMI framework (Context-Information Need, Concept model, Attribute, Metric and Indicator) for defining the measurement projects. Additionally, the proposal architecture uses a big data repository to make available the data for consumption and to manage the Organizational Memory, which allows a feedback mechanism in relation with online processing. In order to illustrate its utility, two practical cases are explained: A pasture predictor system, using the data of the weather radar (WR) of the Experimental Agricultural Station (EAS) INTA Anguil (La Pampa State, Argentina) and an outpatient monitoring scenario. Future trends and concluding remarks are extended.

1. Introduction

The organizational knowledge management represents a key asset to support decision-making processes by different organizational stakeholders. The main aim of knowledge management systems is to manage, store and retrieve the organizational knowledge, so that it can be used later to learn, share knowledge, solve problems, and ultimately to support better decision-making processes. To ensure an efficient management of organizational knowledge, it is necessary to have technological platforms that support it. In the previous work, we proposed architecture based on data flow processing, for this purpose. Specifically, the Processing Architecture based on Measurement Metadata (PAbMM) [1, 2, 3].

Nowadays, there are processing architectures which allows the real-time data processing through configurable topologies such as Apache Storm [4, 5] and Spark [6]. In this type of architectures, you can dynamically define the processing topology over the data streams and adjust it to different computation necessities, being possible delegating the data structural definition and its meaning inside of the application. In this context, we summarize Processing Architecture based on Measurement Metadata [1, 2], which supported by the framework for measuring and evaluating called C-INCAMI (Context-Information Need, Concept model, Attribute, Metric and Indicator) [7], incorporates metadata to the measurement process, promoting repeatability, comparability and consistency. From the point of view of the semantic and formal support for measurement and evaluation (M&E), the conceptual framework C-INCAMI establishes an ontology that includes the needed concepts and relationships for specifying the data and metadata for any M&E project. Moreover, unlike other strategies for the processing of data streams [8, 9], with the addition of metadata, PAbMM is reliable for guiding the processing of the Measures from heterogeneous data sources, analysing each one in context of origin, as well as its significance within the proposed M&E in which defined.

[1] María de los Ángeles Martín, Street 9 and 110, martinma@ing.unlpam.edu.ar
[2] Mario Diván, Coronel Gil 353, 1st floor, mjdivan@[eco|ing].unlpam.edu.ar

The data streams are structured under C-INCAMI/MIS (Measurement Interchange Schema) [10], which allows gather data and metadata jointly inside the same stream. In this metadata, we can describe the measurement context additionally to the entity under measurement, which permits avoid analysing the measure in isolation way.

The PAbMM evolves the original strategy [1] incorporating support to the big data repositories in contexts of distributed computation. This implies the necessity of gather Big Data and Data Stream Processing technologies; which ensures powerful large data volumes processing, allowing efficient management of knowledge in the Organizational Memory.

The Organizational Memory that integrates the Processing Architecture based on Measurement Metadata serve as base for the organizational knowledge exchange and to be used in recommender systems in decision making processes.

Therefore, by having a well-developed organizational memory that supports the structuring, reusing and processing of organizational knowledge is a primary decision (and likely a success factor) to achieve such an effective management.

Nonaka and Takeuchi have said that an organization cannot create knowledge itself. Conversely, the knowledge creation basis for an organization is the individual's tacit knowledge; and tacit knowledge is shared through interpersonal interactions [11].

Therefore, in order to reach and maintain the organizational effectiveness and competitiveness, an organization needs to learn from past and present experiences and lessons learnt and to formalize organizational memories for enabling to make explicit the individual's tacit knowledge -and why not community's tacit knowledge as well.

One of the main goals of an organizational knowledge management strategy is to make explicit the individuals' implicit knowledge, to try to formalize the informal knowledge in order to allow machine-processable semantic inferences. A way of alleviating this problem from the knowledge representation standpoint is to store the knowledge in a more structured and formal way. We have followed this approach by using the case-based organizational memory strategy. It combines organizational knowledge storage technology with case-based reasoning (CBR) to represent each item of informal knowledge. In general, the organizational memories are intended to store the partial formal and informal knowledge present in an organization with automatic processing capabilities. In particular, by structuring an organizational memory in cases can also facilitate the automatic capture, recovery, transfer and reuse of knowledge for problem solving.

The main goal of this research is the integration of the Organizational Memory and case-based reasoning into the Processing Architecture based on Measurement Metadata as conceptual foundation for any organizational knowledge management. Also, the discussion about the added value of organizational memories; Thus, data, information, and knowledge from heterogeneous and distributed sources can be automatically and semantically processable by web-based applications, for instance, an 'intelligent' recommendation system to support a more effective decision-making process.

This article is organized in six sections. The Section 2 summarizes the conceptual framework C-INCAMI. The section 3 outlines the Processing Architecture based on Measurement Metadata. The Section 4 outlines the case based organizational memory. The Section 5 illustrates the application of the organizational memory to a practical case: a pasture predictor system using the data of the weather radar of INTA EEA Anguil. The Section 6 shows the application of the PAbMM Architecture and organizational memory to an outpatient monitoring and diagnosis case. The Section 7 discusses related work and finally section 8 summarizes the conclusions.

2. C-INCAMI Overview

C-INCAMI is a conceptual framework [12], which defines the concepts and their related components for the M&E area in software organizations. It is an approach in which the requirements specification, M&E, and analysis of results are performed for satisfying a specific information need in a given context. C-INCAMI is structured in the following main components: i) M&E project management, ii) Non-functional Requirements specification, iii) Context specification, iv) Design and implementation of measurement and v) Design and implementation of evaluation. Most components are supported by the ontological terms defined in [13].

The M&E project definition component defines and relates a set of project terms needed for dealing with M&E activities, methods, roles and artefacts.

The Non-functional requirements component allows specifying the Information need of any M&E project. The information need identifies the purpose and the user viewpoint; in turn, it focuses on a Calculable Concept and specifies the Entity Category to evaluate. A calculable concept can be defined as an abstract relationship between attributes of an entity and a given information need. This can be represented by a Concept Model where the leaves of an instantiated model are Attributes. Attributes can be measured by metrics.

Regarding measurement design, a Metric provides a Measurement specification of how quantifying a particular attribute of an entity, using a particular Method (i.e. procedure), and how to represent its values, using a particular Scale. The properties of the measured values in the scale with regard to the allowed mathematical and statistical operations and analysis are given by the scale Type. Two types of metrics are distinguished. Direct Metric is that for which values are obtained directly from measuring the corresponding entity's attribute, by using a Measurement Method. On the other hand, the Indirect Metric value is calculated from other direct metrics' values following a formula specification and a particular Calculation Method.

For measurement implementation, a Measurement specifies the task by using a particular metric description in order to produce a Measure value. Other associated metadata is the data collector name and the timestamp in which the measurement was performed. The Evaluation component includes the concepts and relationships intended to specify the evaluation design and implementation. It is worthy to mention that the selected metrics are useful for a measurement tasks as long as the selected

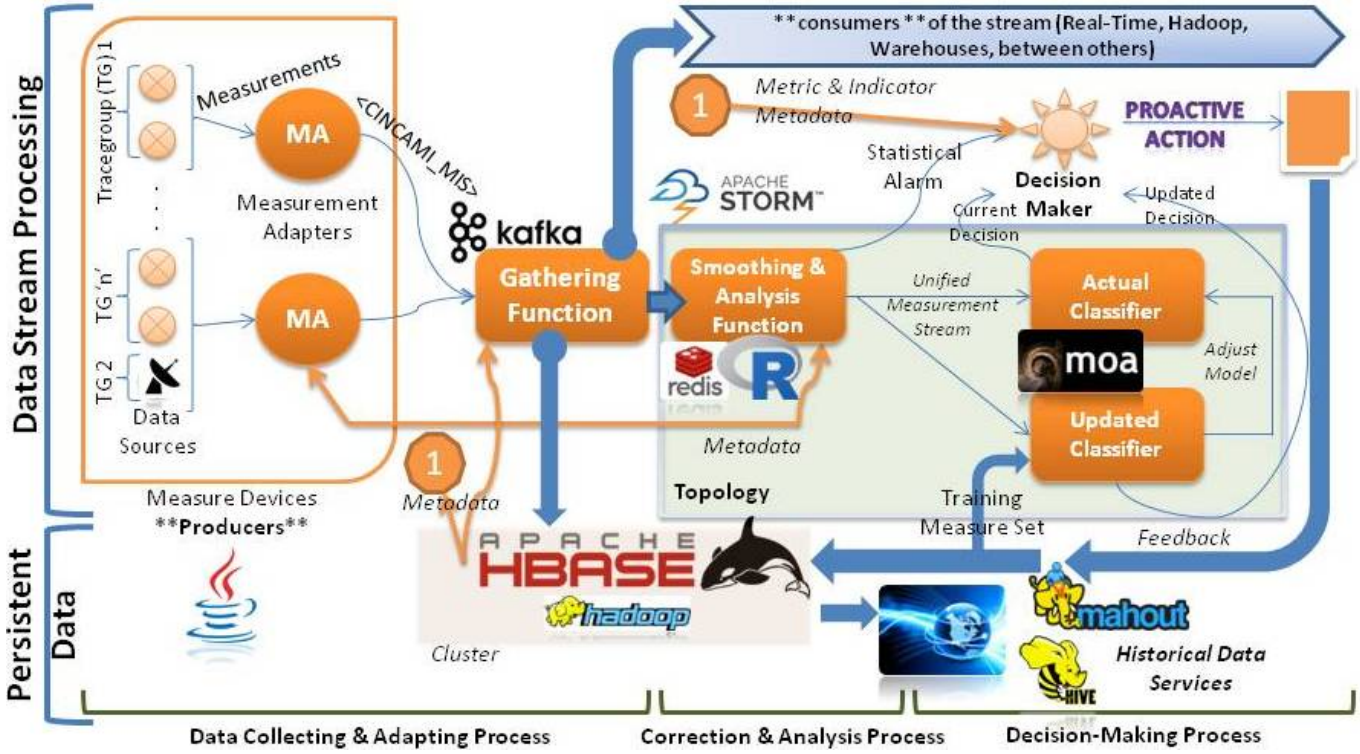


Figure 1. Extensions for the measurement component of C-INCAMI.

indicators are useful for an evaluation tasks in order to interpret the stated information need.

With the formalization of the project M&E based on C-INCAMI, the fact labelling process flow allows structuring the contents of a consistent and aligned with the project objective way. This structuring of the measurements within the PAbMM, maintains the concept that each measure is associated; for example, if a measure of contextual attribute or property.

Inside the stream and jointly with each measure associated with an attribute, the measures associated with the contextual properties are labeled. Thus, it enriches the statistical analysis because it is possible verify the formal and syntactic consistency of each measure against their formal definition, prior to moving forward with more complex statistical techniques.

3. The Processing Architecture Based on Measurement Metadata

The PAbMM is a manager of semi-structured measurements streams, enriched with metadata supported by C-INCAMI, specialized in M&E projects, which incorporates detective and predictive behaviour at online with the ability to manage and provide large volumes of data on demand.

As shown in Figure 1, the conceptual model in terms of stream processing it is as follows. The measurements are generated in the heterogeneous data sources (for example, The INTA weather radar [14]), which supply a module called Measurements Adapter (MA in Figure 1) generally embedded in measurement devices.

MA incorporates together with the measured values, the metadata of the M&E project and reports it to a central gathering function (Gathering Function -GF). GF incorporates the measurements streams in parallel in: a) The big data repository in persistent way, b) The C-INCAMI/MIS stream for the subscribers wishing to process information at the time when it is generated (for example, for INTA weather radar data, a consumer could be the National Meteorological Service), and c) Inside a buffer organized by monitoring groups -dynamic way of grouping data sources defined by the M&E project manager- in order to allow consistent statistical analysis at level of monitoring group or by geographic region where the data sources are located, without incurring in additional processing overhead. Additionally, GF incorporates load shedding techniques, which allows manage the queue services associated with the measurements, mitigating the risks of overflow regardless how they are grouped.

Thus, the C-INCAMI/MIS stream, is incorporated into the big data repository with measurements and metadata, and remains available to meet requests for services associated with data on historical measurement (Big Data Repository and The Historical Data Services in Figure 1). In addition to the measurement stream is sent to the subscribed consumers, a copy of this continues within the data stream processor and applies descriptive, correlation and principal components analysis (Analysis & Smoothing Function -ASF - in Figure 1) guided by their own

metadata, in order to detect inconsistent situations with respect to its formal definition, trends, correlations and / or identify system components that contribute most in terms of variability.

If a situation is detected in ASF, a statistical alert is triggered to the decision maker (DM) for evaluating whether or not to trigger external warning (via email, SMS, etc.) for reporting to the monitoring staff responsible of the situation (for example, in the case of the INTA WR, it could signal a possible device error or WR calibration). In parallel, the new measurements streams are reported to current classifier (Current Classifier -CC-), who must classify the new measurements whether or not correspond to a risk and report its decision to the DM. Simultaneously, the CC is reconstructed online, incorporating new measurements to the training set and producing a new model with them (Updated Classifier -UC).

The UC classifies the new measurements and produce a current decision which will also be communicated to the DM. The DM determines whether the decisions referred by classifiers (CC and UC) correspond to a risk and in which case, what probability of occurrence, act accordingly as defined in the minimum threshold of probability of occurrence defined by the project manager. Finally, regardless of the decisions taken, the UC turns into replacing the previous CC, just if an improvement in their ability for classifying is obtained according to the adjustment model based on ROC curves (Receiver Operating Characteristic) [15].

From a technological point of view, PAbMM evolves the solution proposed in the SDSPbMM, because now it is possible the management of the big data repositories in the distributed computation context, the data provision to thirds by subscription, and maintaining the real-time measurement processing in parallel (See Figure 1). Thus, we prioritize the use of the open source technology with the aim of promoting the extensibility, dynamism and broadcast of the architecture.

In this way and how you can see in the Figure 1, the data sources continue implementing the original interface of SDSPbMM called "*DataSource*", which establishes the responsibilities that a data source must satisfy to provide data in the architecture. Additionally, each MA in PAbMM (See Figure 1) becomes in a producer in terms of Apache Kafka [16, 17]. Thus, the data sent from the producers (for example, the INTA WR) will be processed by a subscription service inside of the processing cluster, under the concept called "Gathering Function" (GF). In this function, the data are collected from the data source, and the measures coming from the same entity under analysis are gathered and organized in a common buffer.

From the common buffer, we use Apache Kafka [17] to generate the unified measurement streams addressed to the consumers (See Figure 1), understanding by this to: a) The real-time subscribers (for example, Argentinian Air Force), b) The real-time processing topology which runs on Apache Storm [4], and c) The big data repository which store the streams in Apache HBase [18, 19] to support the historical data services.

Now, because the PAbMM runs on the Apache Storm and uses Apache Kafka, it is possible to consume and to process the measurement streams from the GF in a continuous way. The latter incorporates flexibility, scalability and dynamism in relation to the configuration of the processing topology, because both the Analysis and Smoothing Function (ASF) and the classifiers are *Bolts* [4], which may be reorganized in agile way and on demand. Moreover, inside the Storm topology, the PAbCMM continues using R [20] for the statistical computing (analysis and smoothing) and Redis [21] as NoSQL database for: *i) Cache Management, ii) Use of intermediate results from R, iii) Storing of snapshot with the last known state of each entity under monitoring.* Finally, the classifiers use the Hoeffding Tree from Massive Online Analysis (MOA) [22] to get a learning incremental strategy which allows to process big streams with limited computation resources.

At persistent level and by a side, PAbMM uses Apache Hadoop and HDFS (Hadoop Distributed File System) [23, 24] to promote the distributed computing, and Apache HBase [19] as columnar database which allows the monolithic scalability and random access to the measurements stored as C-INCAMI/MIS. From this repository, we use Apache Hive [25, 26] to support ad-hoc query on distributed contexts. Moreover, we use Apache Mahout [27, 28] to run classification and clustering analysis according to the needs, which allows one way of feedback the organizational memory.

The organizational memory is case-based and run over the Apache HBase. As the case-based reasoning engine uses MapReduce [24] programs in its recommendation subsystem, each case <fact,solution> is structured in the form of <key,value> to get the advantage from the distributed computing and the parallelism.

Finally and on the one hand, PAbMM is oriented to the distributed computing, scalability and extensibility over mature and open source technologies for supporting big data repositories; and on the other hand, it is possible broadcast the information by Apache Kafka while the stream engine runs over Apache Storm. Thus, PAbMM can be viewed as a simple topology, which allows us to manage the changes, make the adjustments/fixes that being necessaries, and make easier its interoperability.

In the case study that will be presented in section 5, where the architecture is applied for the processing of the data streams from a weather radar, our focus is in: a) Collect the measurements in structured and interoperable way by C-INCAMI/MIS, b) Store all the measures under a single repository, c) Use the stored measures to answer queries or to support a cluster/classification analysis, d) Provide data in real-time and historical on demand, e) Get and maintain an organization memory with the previous experience of the geographical region, and f) Prevent, or at least detect, the potential problematical situations from the documented experience.

4. Case Based Organizational Memory

Once the data (with the associated metadata), are incorporated from the data sources to persistent Big Data repository, it is desirable to structure them in an Organizational Memory, so that can later be exploited and used for recommendation during decision-making processes.

With the aim to manage and retrieve the organizational knowledge, in the last years numerous proposals of models and tools for knowledge management and knowledge representation have arisen. However, most of them store knowledge in a non-structured or semi-structured way, hindering the semantic and

automatic processing of this knowledge. In this section we specify a case-based organizational memory, so that it can be used to learn, reasoning, solve problems, and as support to better decision making as well.

A way of alleviating this problem from the knowledge representation standpoint is to store the knowledge in a more structured and formal way. We have followed this approach by using the case-based organizational memory strategy. It combines organizational knowledge storage technology with case-based reasoning (CBR) to represent each item of informal knowledge. In particular, by structuring an organizational memory in cases can also facilitate the automatic capture, recovery, transfer and reuse of knowledge for problem solving.

4.1. Case Based Organizational Memory Ontology Overview

Although the benefits of applying the knowledge management systems are well known, and the idea of applying case-based reasoning methods to lessons learned and best practices are not new in the knowledge representation area, there is almost no consensus yet on many of the concepts and terminology used in both knowledge management and case-based reasoning areas. In order to reach this aim we have constructed a common conceptualization for case-based organizational memory where concepts, attributes and their relationships should be explicitly specified; such an explicit specification of a conceptualization is one of the core steps for building ontology.

In the following sections, we will describe case-based organizational memory ontology and his application (using the architecture) to the construction of a pasture predictor system using the data of the wheather radar of INTA EEA Anguil.

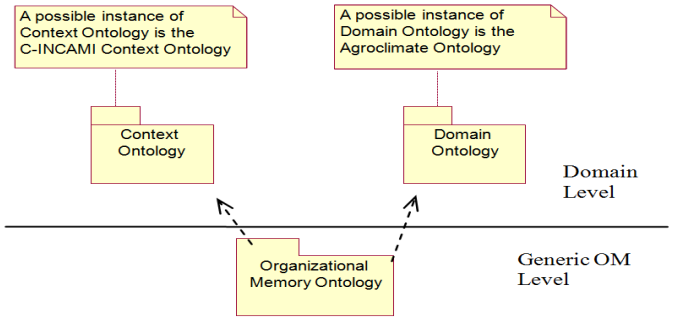


Figure 1. The relationship between ontologies at the specific domain level and at the generic organizational memory level components

The organizational memory ontology aims to be at a generic level from which other representations for specific domain applications can be formulated (see Figure 2). On the one hand, the case-based organizational memory ontology is defined at a generic organizational memory level, and on the other hand, for characterizing the cases according to the specific knowledge domain and its context [29], a domain and context ontologies should also be provided.

The objective of our ontology is to serve as a foundation for the organizational knowledge exchange with semantic power, which in turn facilitates the reuse, the interoperability and the automatic processing by agents [30].

The main concepts of the ontology, which are illustrated in the UML diagram of the Figure 3, are described in the following text, highlighted in italic.

An *organizational memory* is the way in which an organization stores and keeps track of what it knows, i.e., about the past, present and future knowledge. An organizational memory can have one or more *knowledge bases* which are intended to achieve different information needs of an organization –recalling that data, information and knowledge are useful assets for decision making. In addition, an organizational memory may be seen as a repository that stores and retrieves the whole specified, explicit, formal and informal knowledge present in an organization. Thus, a knowledge base is an organized body of related knowledge; taking into account that knowledge is a body of understanding and/or lessons learnt from skills and experiences that is constructed by people.

A type of knowledge base is a *case knowledge base* which stores the acquired knowledge in past experiences, good practices, learned lessons, heuristics, etc. to different domains; that is, it stores cases. A *case* is a contextualized *knowledge item* (i.e., an atomic piece of knowledge) representing an experience by means of a problem and its solution.

The representation of the knowledge through cases facilitates the reuse of the knowledge acquired in past problems to be applied to a new problem in similar situations [31].

A case can be seen as an ordered pair < P, S >, where P is the *problem* space, and S is the *solution* space.

There exists a general description of problems as $P(x_1, x_2, \ldots, x_n)$, where each individual problem is an instance $P(a_1, a_2, \ldots, a_n)$; also a general description of solutions as $S(y_1, y_2, \ldots, y_n)$, and every individual solution $S(b_1, b_2, \ldots, b_n)$ is an instance of that general description. The x_i are variables that characterize the problem (*problem feature*), and the y_i are variables that characterize the solution (*solution feature*), where both are features. A *feature* or attribute is a measurable physical or abstract property of an entity category. Since the stored cases refer to a specific knowledge domain, the features that characterize the problems and solutions are defined by a *domain concept* term; for example, the concepts coming from the meteorology domain ontology (i.e. *Precipitation Accumulation*, *Hail*, *Hail Damage, Environmental Temperature, Environmental Pressure* and so on, as we will illustrate in Section 5.2).

The case-based reasoning process consists in assigning values to problem variables and finding the adequate instances for solution variables. To find the appropriate values for the instances of a solution, the similarity assessment of cases should be performed, so that for each case knowledge base a *similarity assessment model* should be specified.

Greater detail of the functionality of the Organizational Memory can be found in [32], where processes have been formalized by the SPEM metamodel to promote its communicability and extensibility.

4.2. Similarity Assessment Model Representation

Most of the case-based reasoning applications have been focused on problems of specific domains. However, in order to be useful to an organization, a case-based reasoning system should be fitted in with the main knowledge sources that can stem from

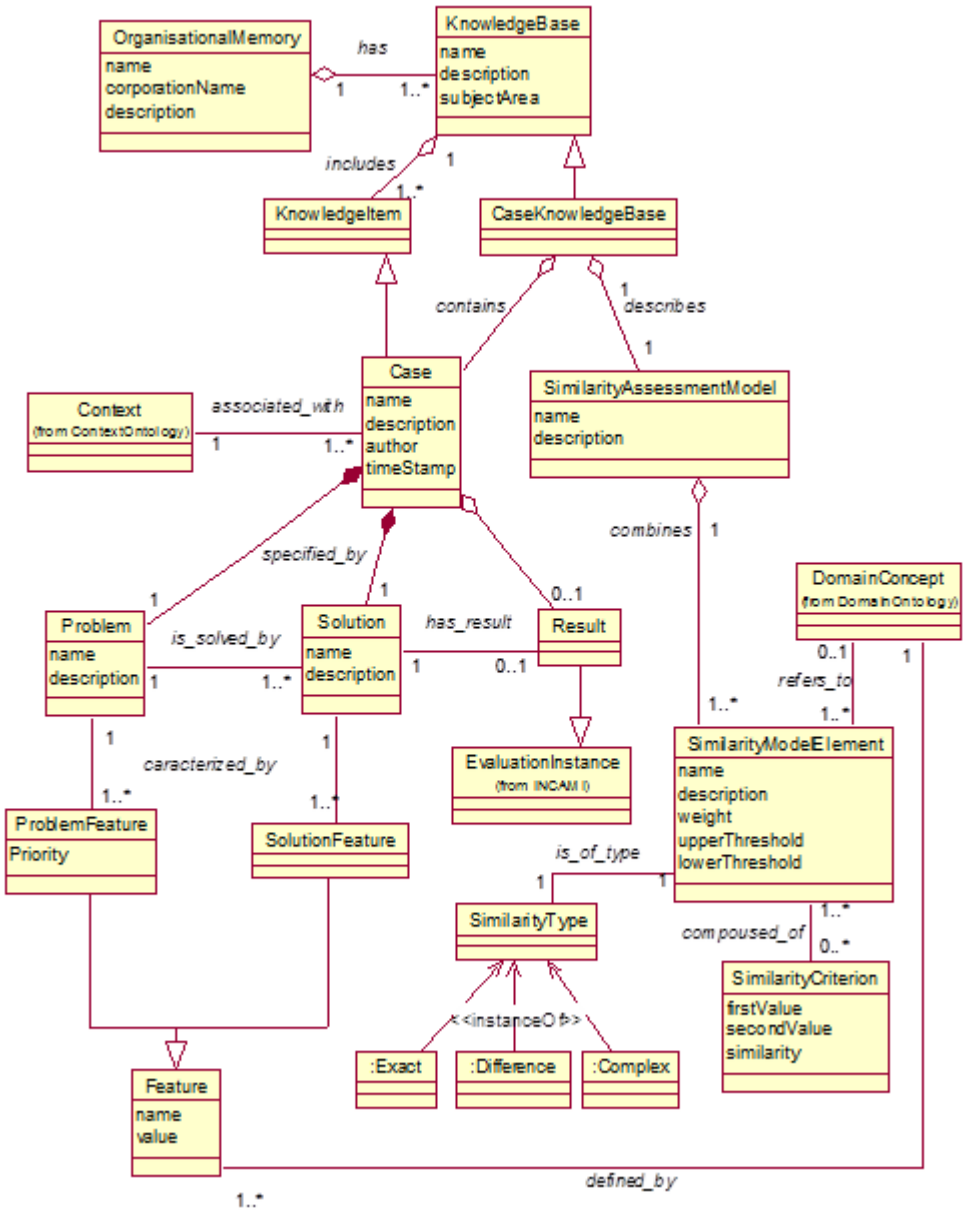


Figure 3. UML diagram that specifies the main terms, attributes and relationships to the Case-based Organizational Memory Ontology.

diverse domains, and so the similarity functions appropriate to each case knowledge base.

Usually, the similarity between a recovered case R and a new case C is defined as the sum of the similarities among its constituent feature values multiplied by their weights, i.e. the so-called Nearest Neighbor formula:

$$Similarity(R, C) = \sum_{f \in F} w_f * sim_f(f_R, f_C) \quad (1)$$

Where w_f is the weight of the feature f, and sim_f is the similarity measurements function to the feature. Therefore, in order to provide an appropriate representation of the similarity model, it is necessary to represent the weights that model the relative relevance, and the similarity function type for each specific feature. The weights are represented as an attribute inside each similarity element (see Figure 3), and the similarity type is restricted to be one out of three general types, namely: Exact, Difference and Complex.

The inclusion of these three types of functions is based on the analysis of numerous investigation works in the case-based

reasoning area, as well as taking into account they cover the similarity representation needs of most cases in the Software and Web Engineering area. Particularly:

- The Exact similarity function returns 1 if two feature values are the same and 0 otherwise.
- The Difference similarity function is inversely proportional to the difference between a feature values. It can only be applied when it is possible to define the value difference; for instance, between numerical values the difference similarity function returns 1 if both features are equals, and return $1/|f_c - f_n|$ in other case.
- The Complex similarity function solves the similarity for all those situations where the two previous functions are not applicable; for example, the semantic difference between two synonymous terms that is neither completely the same nor completely different. If the number of a feature values is finite, it is feasible to have beforehand the similarity measure values for all possible values' pairs. In our model, these parameters are represented in the Similarity Criterion class, which is defined as the assessment pattern used to determine the semantic similarities between two feature values.

Ultimately, an exhaustive glossary of terms, attributes and relationships are shown in [29], where the terminology for the case-based organizational memory ontology is explicitly described.

5. Practical Case 1: A Pasture Predictor

In order to illustrate the above main concepts, attributes and relationships, we will elaborate on an example of case-based knowledge base and its similarity assessment model for a specific domain: a pasture predictor system, using the data of the weather radar of INTA EEA Anguil. This case base stores a body of related knowledge about the growth of pasture based on a range of data, including current weather conditions and forecasts, rainfall events and past climate records, processed by the PAbMM architecture and taking the radar as a data source.

5.1. The Weather Radar of INTA EEA Anguil

Weather radars are active sensors of remote sensing that emit pulses of electromagnetic energy into the atmosphere in the range of microwave frequencies. These sensors are tools to monitor environmental variables, and specifically, the identification, analysis, monitoring, forecasting and evaluation of hydro meteorological phenomena, as well as physical processes that these involve, given the risk that can cause severe events. Its main applications are: a) Weather description, forecasting and nowcasting, b) Forecasting and monitoring of environmental contingencies (hail, torrential rain, severe storms, etc.), c) Security in navigation and air traffic control, d) Studies of atmospheric physics, e) studies of agro climatic risk, f) Provision of basic data for scientific and technological research, and g) Provision of input data for hydrological models (i.e. floods) [14]. The information recorded by the WR is collected through volumetric scans and today, the data are stored in separate files called volumes. The data contains the different variables: reflectivity factor (Z), differential reflectivity (ZDR), polarimetric correlation coefficient (RhoHV), differential phase (PhiDP), specific differential phase (KDP), radial velocity (V) and spectrum width (W).

Two types of data are distinguished: a) raw data and b) some level of data processing or "products". In both cases, the sampling units are 1 km^2 and 1º and each variable are stored in separate files.

Under normal operation, in a full day (00:00h to 23:50h), 8640 files are generated only for the range of 240 km just for one WR. In this sense and for each day, just the WR of the EAS Anguil produces daily a volume of 17GB of data, which represents about 6.2 Tb annually and just for one WR.

From raw data using the proprietary software Rainbow 5, different processing can be obtained, for example, some hydrological products estimating precipitation characteristics as SRI (Surface Rain Intensity), which generates values intensity or PAC (Precipitation Accumulation), which calculates a cumulative rain in a predefined time interval. These products can be formatted in XML or raster image. Also, INTA developed software that can generate more products from raw radar data, for example applications of models to estimate occurrence of hail and hail damage to crops [14].

The users of radar data and radar products, with free and open access are: i) National System of Weather Radar (SiNaRaMe), ii) SMN, iii) Sub secretary of Nation Water Resources (SSRH), iv) National Water Institute (INA), v) Civil Defence vi) Argentinean Air Force, vii) Commercial and General Aviation, viii) Directorate of Agriculture and Climate Contingencies (DACC, Mendoza) ix) Agro climatic Risk Office (ORA), xi) Universities, xii) Research Groups and related product development, xiii) Insurance Companies, xiv) Media, xv) INTA.

In this case, PAbMM define a M&E project which is useful between others things, for detecting decalibration in the WR in each data stream processed. So, and in terms of PAbMM, each WR is a heterogeneous data source that uses a MA. The architecture store the C-INCAMI streams from each WR, provides by subscription the data and allows the monitoring of the stream to make decisions in real time. In this way and from the processed data, you can build knowledge management systems for supporting the decision making in the agricultural production domain, based in the experience stored through the Organizational Memory.

5.2. Knowledge Base for Pasture Production

This case base stores a body of related knowledge about pasture growth in relation to current, past and future weather conditions, so that it serves as the basis to a recommendation system that support the new cycle of pasture production regarding similar past ones.

This knowledge base, collect heterogeneous data from different sources (One source of data is the Weather radar of INTA EEA), as well as manual measurements made by agricultural producers, for example the estimated Daily production of pasture.

The aims of the knowledge base are to support productivity, efficiency and continuing growth in this important industry.

It saves past knowledge about weather conditions (as problems) and their pasture production (as solutions).

By providing 60-day forecasts for the growth of pasture, the aim to help farmers make better decisions in managing their herds, production and costs.

To illustrate the knowledge base and for easier understanding, a simplified model of the Case structure is shown. This case knowledge base characterizes the problem situation through various characteristics including current weather conditions 7-day forecasts, rainfall and hail events, and past climate records. Also the context data of location and time is taken into account. In this example, the case is characterized through four features, namely: Precipitation Accumulation, Hail, Hail Damage, and Rain Forecast that are defined in the meteorology domain ontology.

Table I Example of similarity assessment model for a case base

Feature	Description	Type	Wht
Precipitation Accumulation	Accumulative rain in Last 10 days	Difference	0.40
Hail	Indicates the occurrence of hail in Last 10 days (Possible values are *yes* or *no*)	Exact	0.15
HailDamage	Indicates the occurrence of hail damage in Last 10 days (Possible values are *yes* or *no*)	Exact	0.25
RainForecast	Indicates the next 7 days rain forecast	Difference	0.20

Analogously, the solution will be characterized by the PastureProduction feature; which indicates the daily kg of dry matter produced in one hectare.

For each feature that characterizes a case, we should establish its weight and its similarity function type (see Table 1). These design decisions could be made by an expert taking into account which features are considered more relevant from the similarity point of view to evaluate in the end the global similarity of two cases.

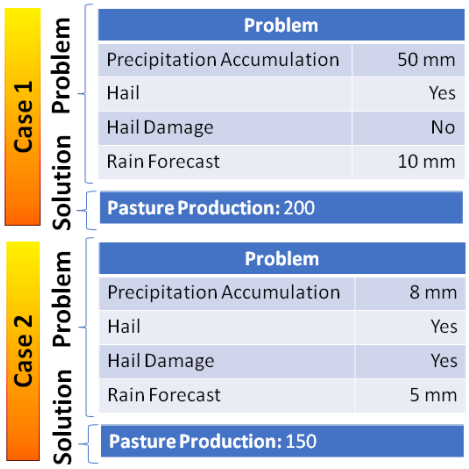


Figure 2. Example of representation of two stored past cases

Once defined the case structure and its similarity assessment model, each case is stored with all the feature values that characterize it and its solution accordingly. Two case examples are shown in Figure 4.

A new decision in herds managing can benefit from the case-based organizational memory by recovering the pasture prediction information of the most similar environmental conditions. Let us suppose that we want to buy animals to fatten and need to know if we have enough grass, and that the case base stores the two cases shown in Figure 4, among others.

In order to minimize risks, we can take advantage of the recorded knowledge by retrieving and reusing the most similar past experience. Therefore, table 2 shows the similarity calculation of each feature of the new case compared to the previous past ones, i.e., "Case 1" and "Case 2". Hence the global similarity calculations give as outcomes:

$$Similarity(Case\ 1, New) = 0.4 * 0.03 + 0.15 * 1 + 0.25 * 1 + 0.2 * 0.2 = \mathbf{0.453}$$

$$Similarity(Case\ 2, New) = 0.4 * 0.083 + 0.15 * 1 + 0.25 * 0 + 0.2 * 0.1 = \mathbf{0.203}$$

Table II. Example of Similarity Assessment between the previous two past cases and the new one

Feature	Case 1	Case 2	New Case	Similarity Case1/New	Similarity Case2/New
Precipitation Accumulation	50	8	20	$\frac{1}{\lvert 50-20\rvert} = 0.03$	$\frac{1}{\lvert 8-20\rvert} = 0.083$
Hail	Yes	Yes	Yes	1.00	1.00
HailDamage	No	Yes	No	1.00	0.00
RainForecast	10	5	15	$\frac{1}{\lvert 10-15\rvert} = 0.2$	$\frac{1}{\lvert 5-15\rvert} = 0.1$

Resulting "Case 1" as the most similar to the new case, and therefore the pasture prediction is 200 daily kg of dry matter produced.

6. Practical Case 2: Outpatient Monitoring and Diagnosis

In this section we show another use of PAbMM and Case Based Organizational Memory oriented to an outpatients monitoring scenario.

Considering a healthcare centre, the doctors need monitor the outpatients for avoiding adverse reactions or major damage in general. In this sense, PAbMM bring to doctors of a mechanism by which they can be informed about unexpected variations and/or inconsistencies in health indicators defined by them (as experts). So, the core idea is: there exists some proactive mechanism based on health metrics and indicators that produces an alarm for each risk situation associated to the outpatient under monitoring.

Considering C-INCAMI, the information need is "to monitor the principal vital signs of an outpatient when he/she is given the

medical discharge from the healthcare centre". The entity under analysis is the outpatient.

According to medical experts, the corporal temperature, the systolic arterial pressure (maximum), the diastolic arterial pressure (minimum) and the cardiac frequency represent the relevant attributes of the outpatient vital signs to monitor. They also consider as necessary monitoring the environmental temperature, environmental pressure, humidity, and patient position (i.e. latitude and longitude) contextual properties.

Indicator Code	ETEMP	
Indicator Name	Level of corporal temperature	
Domain		
(hyperpyrexia, very high fever, fever, normal, hypothermia risk, hypothermia)		
Scale		
Type of Scale	Categorical	
The Value Domain	Ordinal	
Unit	-	
Calculation Method (Elementary Model)		
Name	Temperature Analysis	
Specification	Axillary Temperature (Direct Metric)	Indicator Value
	>= 41.50	Hyperpyrexia
	[38.30 , 41.50)	Very high fever
	[37.50 , 38.30)	Fever
	[36.50 , 37.50)	Normal
	[35.00 , 36.50)	Hypothermia risk
	< 35.00	Hypothermia (While the temperature is descending there is risk of plain encephalogram).
References	1. Loscalzo, J., Fauci, A., Braunwald, E., Dennis L., Hauser, S., Longo, D. (2008). "Harrison's principles of internal medicine". McGraw-Hill Medical. pp. Chapter 17, Fever versus hyperthermia. ISBN 0-07-146633-9.	
	2. Marx, John (2006). "Rosen's emergency medicine: concepts and clinical practice". Mosby/Elsevier. p. 2239. ISBN 9780323028455.	
Decision Criteria		
Indicator Value	Level of Acceptability	
Hyperpyrexia	Not Acceptable	
Very high fever	Not Acceptable	
Fever	Poorly Acceptable	
Normal	Acceptable	
Hypothermia risk	Poorly Acceptable	
Hypothermia	Not Acceptable	

Figure 3. Details of the Level of Corporal Temperature elementary indicator specification

The definition of the information need, the entity, its associated attributes and the context are part of the "*Non-functional requirements specification*" and "*Context specification*" components as discussed in section 2. For monitoring purposes, the metrics that quantify the cited attributes, were selected from the C-INCAMI DB repository (Implemented by Apache HBase in figure 1); likewise, the metrics that quantify the cited contextual properties.

After the set of metrics and contextual properties for outpatient monitoring has been selected, the corresponding elementary indicators for interpretation purposes have also to be selected by experts.

In this way, they have included the following elementary indicators: the level of corporal temperature, the level of pressure, the level of cardiac frequency and the level of difference between the corporal and the environmental temperature. The concepts related to indicators are part of the Evaluation Design and Implementation component. Figure 5 shows the specification of the level of corporal temperature elementary indicator; for example, the different acceptability levels with their interpretations are shown, among other metadata.

Besides, considering that ranges of the acceptability levels (shown in Figure 5) are in an ordinal scale type, then the target variable for the stream mining function (classification) is feasible. So, both classifiers, CC and UC, act relying on the values of the given indicators and their acceptability levels.

6.1. Implementation Issues

Once all the above project information was established, it is necessary for implementation issues to choose a concrete architecture to deploy it. Fig. 6 depicts an abridged deployment view for the outpatient monitoring system considering the PAbMM approach.

Let's suppose we install and set up the MA in a mobile device –the outpatient device-, which will work in conjunction with sensors as shown in Fig. 6. Therefore, while the data collecting and adapting processes are implemented in a mobile device by the MA, the gathering function and other processes can reside in the medical centre computer. The MA component informs the measures (streams) to the gathering function (GF) in an asynchronous and continuous way. MA takes the measures from sensors (in fact, the data sources) and incorporates the associated metric metadata for attributes and contextual properties accordingly. For instance, it incorporates the contextual property ID for the environmental temperature joint to the value to transmit; and so, for every attribute and contextual property. Note that data (values) and metadata are transmitted through the C-INCAMI/MIS schema to the gathering function (GF inside of PAbMM in figure 6), as discussed in section 3.

Although all the values of metrics and contextual properties from monitored outpatients are simultaneously received and analyzed, let's consider for a while, for illustration purpose, that the system only receives data for the axillary temperature attribute and the environmental temperature contextual property from one outpatient, and that also the system visualizes them.

The measures and, ultimately, the acceptability level achieved by the level of corporal temperature elementary indicator, indicate a normal situation for the patient.

Nevertheless, the online decision-making process (Inside of PAbMM, See Section 3), apart from analyzing the level of acceptability met for attributes; also it analyzes the interaction with contextual properties and their values.

At first glance, what seemed to be normal and evident, it was probably not because the processing model can detect in a proactive form a correlation between axillary temperature and environmental temperature. This could cause the triggering of a preventive alarm from the medical centre to doctors, because the increment on the environmental temperature can drag in turn the increment in the corporal temperature, and therefore this situation can be associated to a gradual raise in the risk likelihood for the outpatient.

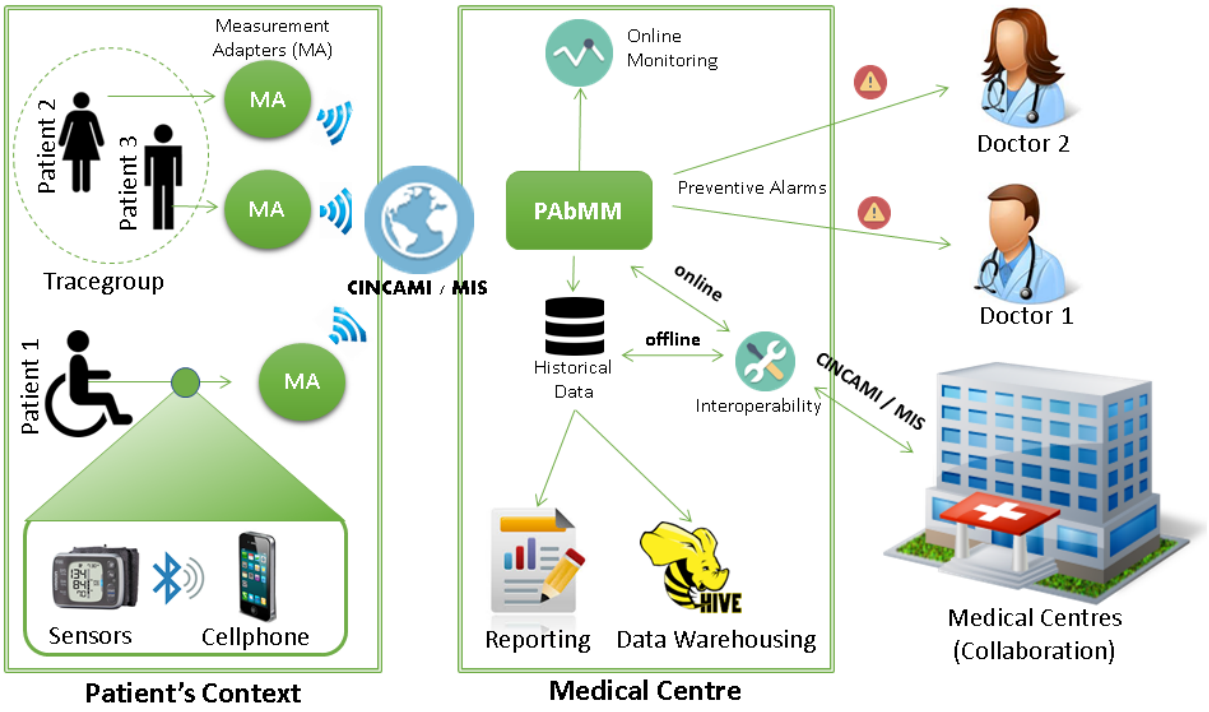


Figure 6. A deployment view for the Outpatient Monitoring System.

6.2. Diagnosis System

Despite of outpatient monitoring task, using the knowledge base of the organizational memory, and the outpatient monitoring data streams, is possible recommend disease diagnosis for a better decision-making. In a similar way to pasture prediction presented in section 5.2, for designing the diagnostic system based on the MOBC, the Organizational Memory must be customized, defining the domain ontology according to the type of knowledge to be stored (See Figure 2).

Case 1

Problem

Problem	
Corporal temperature	38.9°
Systolic arterial pressure	120 mmHg
Diastolic arterial pressure	80 mmHg
Cardiac frequency	95 bpm

Solution

Diagnosis: Influenza **Treatment:**	
Drug	Amoxicillin
Concentration	500 mg
Dose	1 capsule
Daily Frequency	4

Figure 7. Example of a diagnostic case stored in the MOBC.

A problem (possible illness) can be characterized by its symptoms. In order to simplify our example, and for this study, the complete ontology was not developed. We will consider that corporal temperature, the systolic arterial pressure (maximum), the diastolic arterial pressure (minimum) and the cardiac frequency represent the relevant attributes that characterize the problem in the patient, but this list must be completed and adapted to each health institution, according to your needs. Each symptomatic table may have a diagnosis associated with it, and each diagnosis may have different treatments.

Figure 7 shows an example of a case that represents the diagnosis "Influenza", and its respective solution (or corrective action) based on the conceptual model of Figure 2.

For illustrating how the PAbMM classifier can be trained in the arrival of new measures from the monitoring, an example is shown in which the MO stores among its data, two cases: the Influenza case (See Figure 7) and the hypertension case, (See Figure 8). In the case of new data, the classifier will generate a new case (data + metadata using C-INCAMI / MIS) and try to establish an alarm (if applicable) interacting with the MO through the recommendation function.

Measures reported from PAbMM to MO suggest a temperature of 36.9°, a diastolic arterial pressure of 110 mmHg, a systolic arterial pressure of 200 mmHg and a heart rate of 98 bpm. The PAbMM indicates an alarm because it has detected a progressive increase in the pressure, but he does not know how to proceed in terms of medical treatment, and it resorts to organizational memory. The CBR engine of the MO will look for a more similar case. For our example we apply the similarity function Difference, which values the similarity equal to 1 if both characteristics are equal and $1/|f_c - f_n|$ in other case.

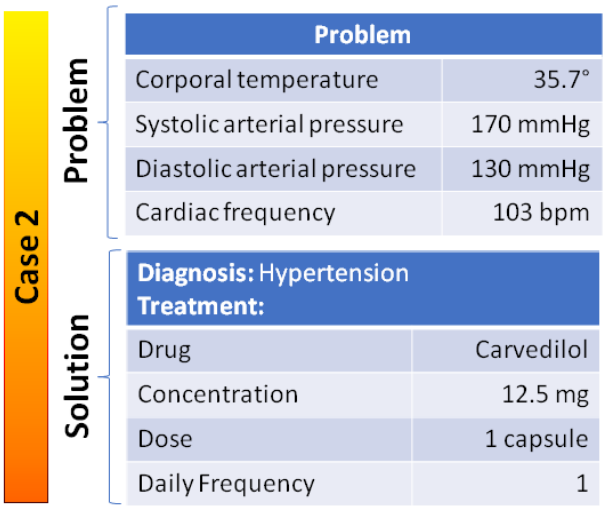


Figure 8. Example 2 of a diagnostic case stored in the MOBC.

Table 3. Medical Recommendation. Example of Similarity Assessment between the previous two past cases and the new one

Feature	Case 1	Case 2	New Case	Similarity Case1/New	Similarity Case2/New
Corporal temperature	38.9	35.7	36.9	0.5	0.83
Systolic corporal pressure	120	170	200	0.0125	0.03
Diastolic arterial pressure	80	130	110	0.033	0.05
Cardiac frequency	95	103	98	0.33	0.20

We also apply equal weight (0.25) to each of the four characteristics. In order to calculate the overall similarity of each pre-existing case with respect to the new case (NC), the formula (1) in section 4.2 is applied, replacing the weights and similarity values for each characteristic (See Figures 7 and 8). As set out below:

$$Similarity(Case\ 1, New) = 0.25 * 0.50 + 0.25 * 0.0125 + 0.25 * 0.033 + 0.25 * 0.33 = \mathbf{0.219}$$

$$Similarity(Case\ 2, New) = 0.25 * 0.83 + 0.25 * 0.033 + 0.25 * 0.05 + 0.25 * 0.2 = \mathbf{0.279}$$

In this way, the most similar case to the new one is the hypertension, which will allow triggering together with the alarm, the recommended treatment (Carvedilol concentration 12.5, 1 capsule per day) for the detected situation.

7. Related Work

In our specific work we have illustrated the use of the case-based reasoning approach to develop a case-based organizational memory; using the advantage of having the PAbMM architecture, which manage large volumes of structured data together with their metadata. The main goal is to exploit the knowledge that con be extracted from organizational memory under a key-value structure (i.e. case-solution structure) stored in the Big Data repository, allowing to incorporate more experience for recommending the courses of actions to the decision-making process.

Firstly, from the organizational memory point of view, there are numerous proposals in the knowledge management area, for example the ones documented in [33, 34]. Most of them capture and store the knowledge in repositories of documents like manuals, memos, and text files systems, etc. where structured or semi-structured storage strategies are seldom used. These approaches usually do not employ powerful mechanisms of semantic and automatic knowledge processing based on ontologies therefore causing very often loss of time and high investment in human resources.

Secondly, from the stream processing architecture point of view, there are recent works which make focus on the data stream processing from a syntactic point of view [8, 9]. In this context, the data model of the stream is based in a key-value structure and incorporates techniques for the adaptive management of high-rate streams [35]. Our architecture incorporates metadata based on a M&E framework, which allows to guide the organization of measures (for example, through snapshots in memory and the last known state of the entities under analysis), facilitating consistent and comparable analysis from the statistical point of view, being able to fire alarms based in the interpretation of the decision criteria of the indicators whose value was got from the data.

8. Conclusions

The organizational knowledge management represents a key asset to support a more effective decision-making process by different stakeholders. In this direction, by having an IT-based organizational memory that supports the structuring, reusing and processing of organizational knowledge is a primary decision to achieve that effective management.

In the previous work [3], we had specified a case-based organizational memory for the stream processing architecture based on measurement metadata. In this paper we deepen the former research and considerably expand the development of the proposed technologies. As a result, we obtained a robust and mature platform for the processing of heterogeneous sources data streams and the management of organizational knowledge.

The knowledge representation through cases facilitates the reuse of knowledge acquired in past problems to be applied to a new problem in similar situations, in addition facilitates the automatic knowledge processing as well.

In this way, the proposed case-based organizational memory can benefit from processing power of the PAbMM architecture, which is supported by distributed Big Data technologies. Also, the PAbMM and the importance about how the measures can be informed jointly with the metadata by C-INCAMI/MIS has been shown. The metadata help guide the processing strategy through the definition of the M&E project and it's possible thanks to the

C-INCAMI framework, which establishes the concepts and relations necessaries to support any M&E project.

Finally, we have illustrated these models and approach with two practical cases. On the one hand a pasture predictor system, using the data of the weather radar of INTA EEA Anguil. This case base stores a body of related knowledge about the growth of pasture based on a range of data, including current weather conditions and forecasts, rainfall events and past climate records, processed by the PAbMM architecture and taking the radar as a data source. On the other side the outpatient monitoring and diagnosis case, shows the processing power of the proposed PAbMM Architecture.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This research is supported by the project 09/F068 of the Engineering School and partially through the technical cooperation contract between the Engineering School and the EAS INTA Anguil (Argentina).

References

[1] M Martín and M Diván, "Case Based Organizational Memory for Processing Architecture based on Measurement Metadata," in 5th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO), Noida, India, 2016.

[2] M Diván and L Olsina, "Process View for a Data Stream Processing Strategy based on Measurement Metadata," Electronic Journal of Informatics and Operations Research, vol. 13, no. 1, pp. 16-34, June 2014.

[3] M Diván et al., "Towards a Data Processing Architecture for the Weather Radar of the INTA Angui," in International Workshop on Data Mining with Industrial Applications, Asunción, Paraguay, 2015.

[4] A. Jain and A. Nalya, Learning Storm. Create real-time stream processing applications with Apache Storm. Birmingham, United Kingdom: Packt Publishing Ltd., 2014.

[5] Apache Software Foundation. Apache Storm. [Online]. http://storm.apache.org/index.html. Last acces: march 28 of 2017.

[6] M. Frampton, Mastering Apache Spark. Birmingham, United Kingdom: Packt Publishing Ltd, 2015.

[7] F. Papa, H. Molina and L. Olsina, "How to Measure and Evaluate Web Applications in a Consistent Way" in Web Engineering: Springer, 2007, ch. 13, pp. 385–420.

[8] M. Ali, W. Aref, R. Bose, A. Elmagarmid, A. Helal, I. Kamel & M. Mokbel, "NILE-PDT: A Phenomenon Detection and Tracking Framework for Data Stream Management Systems," in VLDB, Trondheim, Norway, 2005, pp. 1295-1298.

[9] D. Abadi, Y. Ahmad, M. Balazinska, U. Cetintemel, M. Cherniack, J. Hwang, W. Lindner, A. Maskey, A. Rasin, E. Ryvkina, N. Tatbul, Y. Xing & S. Zdonik, "The Design of the Borealis Stream Processing Engine," in Conference on Innovative Data Systems Research (CIDR), Asilomar, CA, 2005, pp. 277-289.

[10] M. Diván, "Strategy for Data Stream Processing based on Measurement Metadata", PhD Thesis, Computer Science School. National University of de La Plata (Argentina), 2011.

[11] I. Takeuchi and H. Nonaka, The Knowledge-Creating Company. New York: Oxford University Press Inc., 1995.

[12] H. Molina and L. Olsina, "Towards the Support of Contextual Information to a Measurement and Evaluation Framework," in QUATIC, Lisboa, Portugal, 2007.

[13] L. Olsina and M. Martín, "Ontology for Software Metrics and Indicators," Journal of Web Engineering (JWE), vol. 3, no. 4, pp. 262-281, 2004.

[14] Y. Bellini Saibene, M. Volpaccio, S. Banchero, and R. Mezher, "Development and use of free tools for data exploitation of the INTA Weather Radars (in spanish)," in 43º Jornadas Argentinas de Informática - 6º Congreso Argentino de Agroinformática, Buenos Aires, 2014, pp. 74-86.

[15] C Marrocco, R Duin, and F. Tortorella, "Maximizing the area under the ROC curve by pairwise feature combination," ACM Pattern Recognition, pp. 1961-1974, 2008.

[16] Apache Software Foundation. Apache Kafka. [Online]. http://kafka.apache.org/. Last access: march 28 of 2017.

[17] N. Garg, Apache Kafka. Set up Apache Kafka clusters and develop custom message producers and consumers using practical, hands-on examples. Birmingham, United Kingdom: Packt Publishing Ltd., 2013.

[18] Apache Software Foundation. Apache HBase. [Online]. http://hbase.apache.org/. Last access: march 28 of 2017.

[19] J Spaggiari and K O'Dell, Architecting HBase Applications (Early Release). New York, USA: O'Reilly Media, 2015.

[20] R Core Team, R: A Language and Environment for Statistical Computing. Vienna, Austria: The R Foundation for Statistical Computing, 2016.

[21] M Da Silva and H Tavares, Redis Essentials. Birmingham, United Kingdom: Packt Publishing, 2015.

[22] A. Bifet, G. Holmes, R. Kirkby, and B. Pfahringer, "MOA: Massive Online Analysis," Journal of Machine Learning Research, vol. XI, pp. 1601-1604, 2010.

[23] Apache Software Foundation. Apache Hadoop. [Online]. http://hadoop.apache.org/. Last access: march 28 of 2017.

[24] A Holmes, Hadoop in Practice, 2nd ed. New York, USA: Manning, 2015.

[25] Apache Software Foundation. Apache Hive TM. [Online]. https://hive.apache.org. Last access: march 28 of 2017.

[26] J Rutherglen, D Wampler, and E Capriolo, Programming Hive. California: O'Reilly Media Inc., 2012.

[27] Apache Software Foundation. Apache Mahout: Scalable machine learning and data mining. [Online]. http://mahout.apache.org/. Last access: march 28 of 2017.

[28] A Gupta, Learning Apache Mahout Classification. Birmingham, United Kingdom: Packt Publishing, 2015.

[29] M Martín, "Organizational Memory Based on Ontology and Cases for Recommendation System", PhD Thesis, Computer Science School, National University of de La Plata (Argentina)., 2010.

[30] M. Martín and L. Olsina, "Added Value of Ontologies for Modeling an Organizational Memory," in Building Organizational Memories: Will You Know What You Knew ?: IGI Global, pp. 127-147.

[31] H Chen and Z Wu, "On Case-Based Knowledge Sharing in Semantic Web," in XV International Conference on Tools with Artificial Intelligence, California, 2003, pp. 200-207.

[32] M Diván and M Martín, "Strategy of Data Stream Processing sustained by Big Data and Organizational Memory (In Spanish)," in In proc. of National Conference of Informatics Engineering/ Information Systems Engineering, Buenos Aires, Argentina, 2015.

[33] W Karl-Heinz, "A Case Based Reasoning Approach for Answer Reranking in Question Answering," CoRR, vol. abs/1503.02917, 2015. [Online]. http://arxiv.org/abs/1503.02917

[34] M Lee, M Lee, S Hur, and I Kim, "Load Adaptive and Fault Tolerant Distributed Stream Processing System for Explosive Stream Data", Transactions on Advanced Communications Technology, vol. 5, no. 1, pp. 745-751, 2016.

[35] J Samosir, M Indrawan-Santiago, and P Haghighi, "An evaluation of data stream processing systems for data driven applications," Procedia Computer Science, vol. 80, pp. 439-449, June 2016.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 24-29 (2017)

www.astesj.com

ASTES Journal
ISSN: 2415-6698

Special Issue on Recent Advances in Engineering Systems

Low Probability of Interception Beampattern Using Frequency Diverse Array Antenna

Shaddrack Yaw Nusenu*

[1] *Koforidua Technical University, Electrical / Electronic Department, Faculty of Engineering, Ghana.*

ARTICLEINFO

Article history:
Received: 12 February, 2017
Accepted: 31 March, 2017
Online: 02 April, 2017

Keywords:
Frequency diverse array
costa sequence
Low probability of intercept
Transmit beamforming

ABSTRACT

Frequency diverse array (FDA) antenna employs a small frequency increment, as compared with the carrier frequency, across antenna array elements. The use of a frequency increment generates an array factor that is dependent on angle, time, and the range and this yields a range and angle dependent beampattern. In this paper, a transmit array beamforming FDA employing costa sequence frequency increments is proposed that offers low probability of interception (LPI) for radar systems. Because active radars are highly visible to intercept receivers, FDA beam with uniform linear array and non-linearly increasing frequency offsets to reduce the system visibility is replaced by the conventional high gain phased-array antenna beam. The proposed method produce random-like peak distribution without obvious peaks making it difficult to be detected by unfriendly detector or intercept receivers without known of the specified coding sequence for the frequency increments. Numerical simulation results verify the proposed method.

1 Introduction

Phased-array antenna has been extensively employed in many applications such as radar, electronic warfare, radio astronomy and airport safety etc. This is because it can steer the high-gain beam electronically with high effectiveness towards any desired direction [1, 2, 3]. The directional gain obtainable by employing phased-array antenna is beneficial for detecting, tracking weak targets and suppressing sidelobe interferences from other directions [4]. Conversely, the high gain beam is visible to intercept receivers which will degrade the capability of the surveillance system [5]. Furthermore, active surveillance radars are highly susceptible to detection and interception by opposing forces. The electronic support measures (ESM) can be used to intercept the radar transmission and possibly facilitate harmful electronic counter-measures (ECM) techniques [6]. Therefore it is essential to develop low probability of intercept (LPI) radar systems.

LPI radar system employs a distinct emitted waveform intended to prevent a non-cooperative intercept receiver from intercepting and detecting its emission. On the other hand, low probability of identification (LPID) radar system uses a special radiated waveform intended to prevent a non-cooperative intercept receiver from intercepting and detecting its emission but if intercepted, identification of the radiated waveform modulation and its parameters becomes difficult. It is worth mentioning that LPID radar is an LPI radar [7].

Several techniques have been introduced to reduce radar visibility and improve its LPI ability [8]. Mostly, three main areas are considered in order to reduce radar visibility [6]: (i) spreading the energy in time with high duty cycle waveforms (ii) spreading the energy in frequency with wide bandwidth waveforms (iii) spreading the energy in space through broad transmitter antenna beams. In [9], frequency hopping, orthogonal frequency division multiplexing (OFDM), and random waveforms have been put forward for LPI radars. In [10], switched antenna technique called antenna hopping method which uses irregular scan patterns to reduce the susceptibility to receivers was introduced to enhance the LPI performance. Although with specific beamforming technique, the probability of being detected in sidelobe region can be reduced. In [6], LPI transmit array beamforming method was proposed which employed phased array antenna. Even though phased-array antenna has been widely employed for numerous applications, it produce fixed steering beam at an angle for all the ranges [11].

A new flexible antenna array called frequency diverse array (FDA) has been introduced in [12, 13]. The authors reported that the range-angle-time dependent far-field ar-

*Shaddrack Yaw Nusenu, Koforidua Technical University, Ghana, Email: nusenu2012gh@yahoo.com

https://dx.doi.org/10.25046/aj020304

ray pattern formation are the core attributes of the concept. FDA was defined in [15] as an antenna aperture where the frequency of each antenna element is controlled independently to steer the main beam, depending on the time domain of each frequency component in the far-field. The most significant feature of FDA, as opposed to phased-array, is that the former employs a small frequency increment across its antenna array elements. This frequency increment makes the array beampattern to change as a function of the range, angle and time [11, 12, 13]. In [14], FDA radar using costas sequence modulated frequency increments was proposed to decoupled the range and angle beampattern, in which only FDA applications are considered.

In this paper, LPI FDA transmit beamforming with costa sequence frequency increments is proposed. The proposed method transmitted beampattern will have a random-like peak distribution without obvious peaks. This implies that it will be difficult for the intercept receiver from intercepting and detecting without known of the specified coding sequence for the frequency increments.

The rest of the paper is organized as follows: Section 2 proposes the LPI beamforming using costas sequence FDA antenna, which is verified in Section 3 with numerical results. Finally Section 4 concludes the paper.

2 LPI FDA BEAMFORMING

Consider an FDA with M elements which can be excited by either the same waveform or different waveforms. In this paper, we assume costas sequence frequency increments used for the mth element is Δf_m as shown in Figure 1. In the costas coding technique frequencies for the subpulses are chosen in a random pattern, according to some predetermined rule. Consider the frequency-time matrix shown in Figure 2(a), (b) and (c). In this case, the rows are indexed from $i = 1, 2, ..., M$ and the columns are indexed from $j = 0, 1, 2, ..., (M-1)$. The rows and columns are used to denote subpulses and distinct frequencies, respectively. As shown in Figure 2(a), (b) and (c) a dot indicates the distinct frequency value assigned to the associated subpulse. In this fashion, the frequency assignments are chosen randomly. The element of the difference matrix in row i and column j is given as

$$D_{i,j} = a_{i+j} - a_j,\ i + j \leq M \tag{1}$$

where a_i is the ith element of the coding sequence. The remaining locations (where $i + j > M$) are left blank. Using (1) the first row is formed by taking differences between adjacent elements in the coding sequence, then second row by taking differences between next-adjacent elements, and so forth. We employ this technique for FDA antenna, thus, only one carrier frequency is transmitted by any one of the M FDA elements and each carrier frequency is permitted only once. In [16], the details for the construction algorithms for costas signals can be found.

The radiated frequency fed to the mth of the LPI FDA using costas frequency increments can be given as

$$f_m = f_0 + \Delta f_m,\ m = 0, 1, 2, \ldots, M-1 \tag{2}$$

where f_0 and Δf_m are the carrier frequency and frequency increment, respectively, and M is the number of array elements. Specifically, standard FDA is achieved when $\Delta f_m = \mathrm{m}\Delta f$. Suppose a far-field position with range and angle (r, θ), the phase of the transmitted signal of mth element is represented by

$$\alpha_m(r, \theta) = (f_0 + \Delta f_m)\, r_m \beta \tag{3}$$

$$r_m = r + md\sin\theta,\ m = 0, 1, 2, \ldots, M-1 \tag{4}$$

where d is the element spacing. Hence, the phase difference between the signals transmitted by the mth and first elements can be approximately given as

$$\psi_m(r, \theta) = -\beta f_0 md\sin\theta + \Delta f_m r\beta - \beta m\Delta f_m d\sin\theta \tag{5}$$

Next, we apply a set of costas frequency coding, Δf_m and phases, $\delta_m(r)$ to each element to achieve a "spoiled' low-gain beamforming. Hence the transmit array factor (AF), which can be seen as the fundamental basis pattern [17, 18] is expressed as

$$AF_0(r, \theta) = \sum_{m=0}^{M-1} e^{-j\beta f_0 md\sin\theta} . e^{j\Delta f_m \beta r} . e^{-j\beta\Delta f_m md\sin\theta} . e^{-j\delta_m(r)} \tag{6}$$

where Δf_m is designed using Welch-constructed costas coding sequence and $\delta_m(r)$ can be designed using optimization algorithms. By applying a linear phase progression to (6) in increments of $\kappa = \frac{2\pi}{N}$ for first pattern, 2κ for the second pattern and so forth, additional patterns can be created. In such a fashion, $(N-1)$ patterns can be generated. Therefore (6) can be formulated as (7) [18] at the top of the next page, where $n = 1, 2, ..., N$.

It is worth mentioning that, these N steered versions of the fundamental guarantee linear independence. Moreover, they all exhibit low gain and broad beamwidth. Therefore their peak power can be significantly reduced in any directions to ensure LPI property.

Furthermore, to retain the array radar performance and simultaneously achieving LPI property, we synthesize the high gain patterns along the desired direction by linear combinations of the basis patterns. Suppose the nth basis pattern is weighted by the complex coefficient $w_{0,n}$, then N transmit beampatterns can be expressed in matrix formulation as given in (8), [6, 18].

This implies that using (8), high gain beampatterns can be formed by linearly combinations of the N basis patterns.

3 NUMERICAL RESULTS

In the simulations, we assume the LPI FDA operating at a carrier frequency of $f_0 = 10\ GHz$. An 32 elements uniform linear array is used for the standard FDA and phased-array beamforming (see Figure 4). The transmit array element spacing is designed to be $d = \frac{c_0}{2f_0}$. We consider the Welch constructed costas coding sequence for $M = 10$, with the chosen positions as 1,2,4,8,5,10,9,7,3,6. Similarly, for $M = 12$, we have 1,2,4,8,3,6,12,11,9,5,10,7 and $M = 15$, as 2,8,9,12,4,14,10,15,13,7,6,3,11,1,5. These costas sequence coding with $M = 10$, $M = 12$, $M = 15$, respectively, was used for LPI FDA transmit beamforming for the arrays and

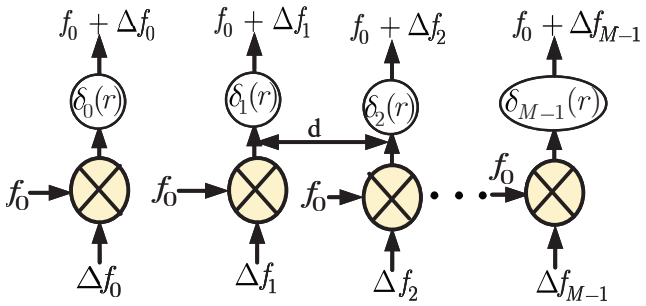
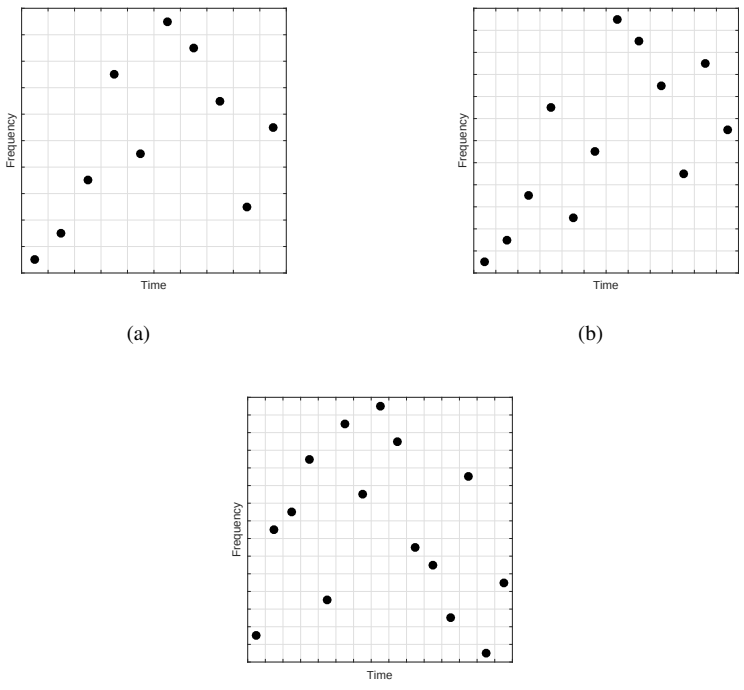


Figure 1: Illustration of LPI FDA. δ_m and Δf_m denotes phase shifters and costas coding frequency increments for the mth element, repectively.

$$AF_{N-1}(r,\theta) = \sum_{m=0}^{M-1} e^{-(N-1)m\kappa} e^{-j\beta f_0 md\sin\theta} . e^{-j\Delta f_m \beta r} . e^{-j\beta\Delta f_m md\sin\theta} . e^{-j\delta_m(r)} \tag{7}$$

$$\begin{bmatrix} A_0(r,\theta) \\ \vdots \\ A_{N-1}(r,\theta) \end{bmatrix} = \begin{bmatrix} w_{0,0} & \cdots & w_{0,N-1} \\ \vdots & \ddots & \vdots \\ w_{N-1,0} & \cdots & w_{N-1,N-1} \end{bmatrix} \begin{bmatrix} AF_0(r,\theta) \\ \vdots \\ AF_{N-1}(r,\theta) \end{bmatrix} \tag{8}$$



(a) (b)

(c)

Figure 2: Frequency-time matrix representations of costas sequence coding: (a) M = 10, (b) M = 12, (c) M = 15.

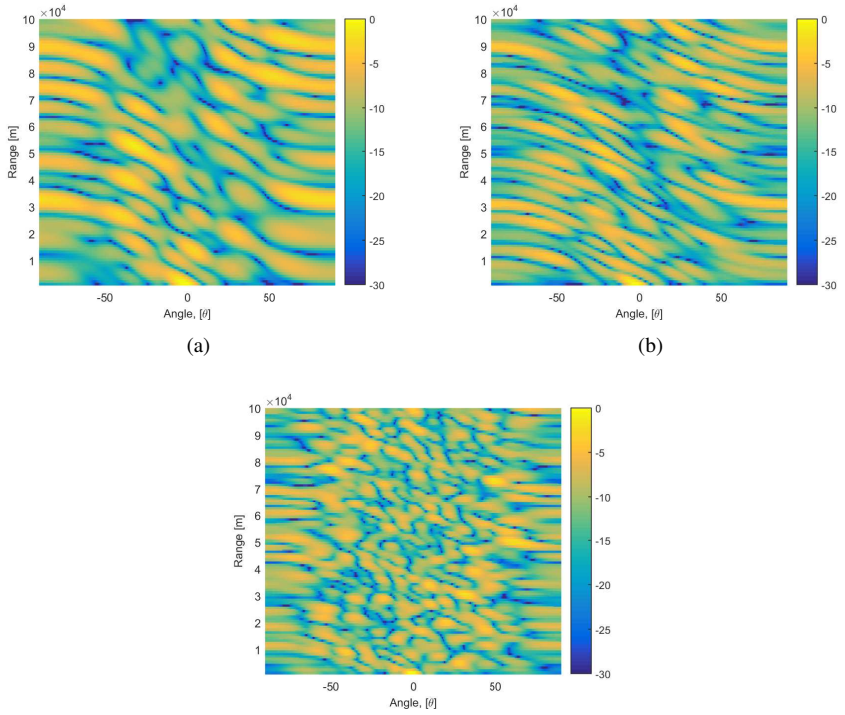


Figure 3: Illustration of LPI FDA beampattern using costas sequence coding: (a) $M = 10$, (b) $M = 12$, (c) $M = 15$.

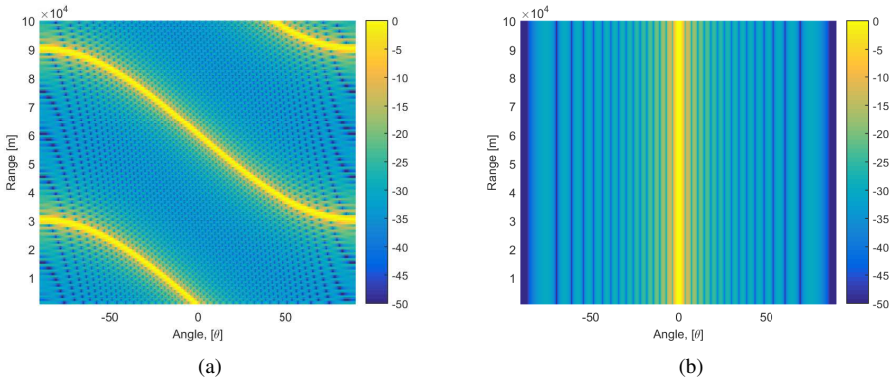


Figure 4: Illustration of transmit beampattern: a) Standard FDA, b) Phased-array.

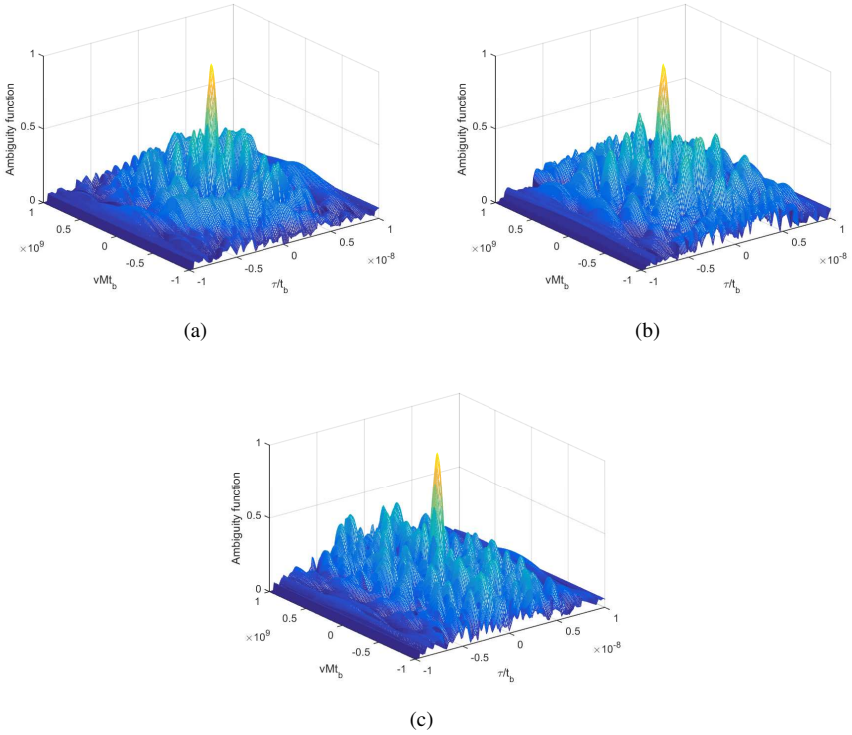


Figure 5: Illustration of three-dimensional ambiguity function LPI FDA using costas sequence coding: (a) M = 10, (b) M = 12, (c) M = 15.

their frequency-time matrix representation is shown in Figure 2.

Next, we show the LPI FDA transmit beamforming employing costas sequence with M = 10, M = 12, M = 15, respectively, in Figure 3. It can be seen that for the respective costas sequence coding employed, the LPI FDA transmit beamforming has random-like peak distributions without noticeable peaks. This characteristics implies that it is very difficult to be intercepted and detected by intercept receivers or unfriendly detectors without the knowledge of the specified costas coding sequence used for the frequency increments.

Suppose $\Delta f = 5\ kHz$, Figure 4(a) shows the standard FDA when specifically $\Delta f_m = \mathrm{m}\Delta f$. It can be observed that it produces range and angle dependent and range-angle coupling transmit beamforming. This feature can be exploited for range and angle localization of radar targets. Differently, the phased-array produce angle-dependent only transmit beamforming, as shown in Figure 4(b). However, the phased-array provides no range information because of its range-independent beamforming.

Finally, the ambiguity function of LPI FDA costas coding sequence is shown in Figure 5 for distinct M. It can be seen that the ambiguity functions accumulates into a spike, thus clearly resembles thumb-tack response.

4 CONCLUSION

This paper proposed a transmit beamforming technique using FDA which offers LPI for surveillance systems. The high-gain phased-array antenna beam is replaced with the low gain FDA beams with nonlinear frequency increments to reduce the system visibility. The proposed technique achieves the same performance as the traditional high-gain beam. The obvious advantage of the proposed technique presented here is the random-like peak energy distributions without noticeable peaks. The results show that the proposed technique has the advantages of both FDA radar and guarantee LPI property. The frequency increments are chosen from Welch constructed costas coding sequence across the array element to ensure the LPI property for radar systems. Overall, the proposed technique can possibly be combined with other LPI waveform design techniques to further reduce the susceptibility and visibility of LPI radar systems.

Conflict of Interest The author declare no conflict of interest.

Acknowledgment The author would like to thank Respected Professor Wen-Qin Wang (Senior Member - IEEE) of UESTC for his great contributions, support and encouragement.

References

[1] T. A. Lam, D. C. Vier, J. A. Nieslsen, C. G. Parazzoli, and M. H. Tanielian. “Steering phased array antenna beams to the horizon using a buckyball NIM lens,”Proc. IEEE, vol. 99, no. 10, pp. 1755-1767, Oct. 2011.

[2] B. Bauman, A. Christianson, A. Wegener, and W. J. Chappell. “Dynamic visualization of antenna pat-

terns and phased-array beam steering,"IEEE Antennas Propag. Mag., vol. 54, no. 3, pp. 184-193, Jun. 2012.

[3] C. J. Reddy. "Phased array-based systems and applications,"IEEE Antennas Propag. Mag., vol. 39, no. 5, pp. 102-103, Oct. 1997.

[4] Li, J., Stoica, P. "The phased array is the maximum SNR active array,"IEEE Signal Process. Mag., 27, (2), pp. 143-144, 2010.

[5] O. G. Vendik and D. S. Kozlov. "Phased antenna array with a sidelobe cancellation for suppression of jamming,"IEEE Antennas and Wireless Propagation Letters, vol. 11, 2012.

[6] D. E. Lawrence. "Low probability of intercept antenna array beamforming,"IEEE Transactions on Antennas and Propagation, vol. 58, no. 9, pp. 2858-2865, September 2010.

[7] Phillip E. Pace. "Detecting and Classifying Low Probability of Intercept Radar,"Second Edition 2009.

[8] V. Krisnamurthy. "Emission management for low probability intercept sensors in network centric warfare,"IEEE Transactions on Aerospace and Electronic Systems, vol. 41, no. 1, pp. 133-151, January 2005.

[9] W.-Q. Wang. "MIMO SAR OFDM chirp waveform diversity design with random matrix modulation,"IEEE Transactions on Geoscience and Remote Sensing, vol. 53, no. 3, pp. 1615-1625, March 2015.

[10] E. J. Baghdady. "Directional signal modulation by means of switched spaced antennas,"IEEE Transactions on Communications, vol. 38, no. 4, pp. 399-403, April 1990.

[11] W.-Q. Wang. "Frequency diverse array antenna: New opportunites,"IEEE Antennas and Propagation Magazine, vol. 57, no. 2, pp. 145-152, April 2015.

[12] P. Antonik, M. C. Wicks, H. D. Griffiths, and C. J. Baker. "Range dependent beamforming using element level waveform diversity," in Proceedings of the International Waveform Diversity and Design Conference, Las Vegas, USA, pp. 1-4, January 2006.

[13] P. Antonik, M. C. Wicks, H. D. Griffiths, and C. J. Baker. "Frequency diverse array radars," in Proceedings of the IEEE Radar Conference, Verona, NY, April pp. 215-217, 2006.

[14] Shaddrack Yaw Nusenu, Zhe Wang, Wen-Qin Wang. "FDA radar using Costas sequence modulated frequency increments," in Proceedings of CIE International Radar Conference, pp. 1058–1061. 2016.

[15] Wen-Qin Wang, H. C. So, and Alfonso Farina. "An Overview on Time/Frequency Modulated Array Processing ," DOI 10.1109/JSTSP.2016.2627182, IEEE Journal of Selected Topics in Signal Processing.

[16] S. W. Golomb and H. Taylor. "Constructions and properties of costas arrays,"Proceedings of the IEEE, vol. 72, no. 9, pp. 1143-1163, September 1984.

[17] Wen-Qin Wang. "Adaptive Rf Stealth Beamforming For Frequency Diverse Array Radar,"23rd European Signal Processing Conference (EUSIPCO) pp. 1158-1161, 2015.

[18] Wen-Qin Wang. "Moving-Target Tracking by Cognitive RF Stealth Radar Using Frequency Diverse Array Antenna,"IEEE Transactionson Geoscience And Remotesensing, vol. 54, no. 7, pp. 3764–3773, July 2016.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 30-35 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems 2017

ASTESJ

ISSN: 2415-6698

Zero Trust Cloud Networks using Transport Access Control and High Availability Optical Bypass Switching

Casimer DeCusatis [*,1], Piradon Liengtiraphan [1], Anthony Sager [2]

[1]*Marist College, School of Computer Science and Mathematics, Poughkeepsie, NY 12603 USA*

[2]*BlackRidge Technologies, Research & Development Division, Reno NV 89433 USA*

ARTICLE INFO

Article history:
Received: 02 March, 2017
Accepted: 25 March, 2017
Online: 04 April, 2017

Keywords:
Cybersecurity
Authentication
Bypass

ABSTRACT

Cyberinfrastructure is undergoing a radical transformation as traditional enterprise and cloud computing environments hosting dynamic, mobile workloads replace telecommunication data centers. Traditional data center security best practices involving network segmentation are not well suited to these new environments. We discuss a novel network architecture, which enables an explicit zero trust approach, based on a steganographic overlay, which embeds authentication tokens in the TCP packet request, and first-packet authentication. Experimental demonstration of this approach is provided in both an enterprise-class server and cloud computing data center environment.

1. Introduction

Network-based cybersecurity attacks against cloud data centers have been increasing in both frequency and severity, far outstripping traditional defense methods. In an effort to combat this problem, this paper is an extension of work originally presented at the recent IEEE International Conference on Smart Cloud [1] which introduced the use of first packet authentication in a zero-trust cloud network. We further develop the use of transport access control (TCP) for authentication in cloud environments, such as Amazon Web Services (AWS). We then introduce a high-speed optical bypass switch into the system architecture, and characterize its performance under realistic load conditions.

Security remains a significant concern for cloud computing environments. According to recent industry survey data [2], 27% of connected third-party cloud applications introduced into enterprise environments in 2016 posed a high security risk. Authentication is a particular problem in cloud environments, since the traditional security perimeter is effectively virtualized. Many cloud applications cannot be effectively segregated from the corporate infrastructure, and communicate freely with software-as-a-service (SaaS) platforms [2]. Cloud data centers are under constant attack from a variety of bad actors; a moderately-sized commercial data center network can experience over 100,000 security events per day [2,3]. These attackers range from individual hacktivists and cyber-gangs motivated by creating social disruption to large, well organized groups with political or financial motivations who are backed by nation-states. Increasingly, these attacks have multiple goals, including compromising critical network resources such as the network controller.

In response to the growing number and sophistication of cybersecurity threats, a United States Presidential Executive Order tasked the National Institute of Standards and Technology (NIST) with creating a set of voluntary policies and guidelines to help develop the U.S. cybersecurity framework. This eventually led to the so-called "zero trust model" for information security [4]. Traditional security models are based on a perimeter security model (also known as an implicit trust model or "trust but verify" approach), in which all communication is trusted between devices within a specified security group. This relatively static approach is based on segmenting the network and creating a demilitarized zone (DMZ) between trusted and untrusted portions of the network. Perimeter security breaks down in modern cloud computing environments, leading to a new approach that explicitly verifies all network connections. These zero trust networks assume that all traffic is a threat until it is authenticated and inspected. Zero trust is intended to provide a scalable security infrastructure that redefines the approach to resource

*Corresponding Author C. DeCusatis, Marist College, Poughkeepsie, NY, Casimer.decusatis@marist.edu

segmentation, a fundamental principle in which resources to be protected are grouped together and securely isolated or partitioned to limit unauthorized access.

While a full zero trust network has not yet been commercially realized, significant progress has been made towards the development of enabling technologies. It is generally recognized that long-standing segmentation techniques such as VLANs no longer provide sufficient cloud network security [2-6]. Many organizations attempt to segment their networks in a coarse granularity fashion to reduce risk, subject to limitations imposed by legacy hardware, complex virtualization software, and a lack of programmable resources [6]. By contrast, a zero trust network security architecture incorporates an explicit trust model and dynamic, automated security policy that extends across conventional security boundaries but still provides fine granularity segmentation and isolation of critical resources. All traffic needs to be validated, even between virtual machines (VMs) sharing a common physical host. Explicit security is part of a layered, defense-in-depth approach, which avoids kill chains and thus prevents single points of failure from compromising the entire security defense system. Fine grain segmentation improves management visibility and makes it feasible to disrupt network attacks as early as possible in the attack process, preferably to prevent data reconnaissance techniques from even identifying the resources which are being protected, much less being able to fingerprint these resources in preparation for an attack.

Micro-segmentation of a zero-trust network should preferably include authentication of not just users or applications, but also extend to the level of authenticating individual packets. Conventional networks assert the identity of a user or application based on a series of attributes such as network addresses, which may be forged [5]. Such networks may decide to trust a user or application based on some criteria, but the concept of trust does not apply to conventional network packets, which are the fundamental building blocks of any network. Our research implements a form of authentication with packet level granularity, which offers several advantages. First, a finely grained approach (at the packet level) improves visibility, particularly when combined with analysis of management plane data logs. Other potential benefits of our approach include simpler, vendor agnostic architectures, better scalability, and improved application portability.

Our approach helps avoid unauthorized awareness (a request for access to the network should not only be denied, it should avoid providing the requestor with any information about the nature of resources that are connected to the network). For example, modern data center networks are subjected to a constant stream of access requests, since even a denied TCP connection request will return some information about the nature of the network, thereby assisting attackers in fingerprinting the target system [5, 6]. This means that potential attackers can gather useful information about a potential target by repeatedly trying to complete a connection request, even if the request attempt itself does not succeed. The information collected in this manner may be used to plan future attacks or identify weaknesses in the perimeter defenses. The approach demonstrated in this paper prevents error message reconnaissance information from reaching a potential attacker, without compromising performance of the remaining system.

While the theoretical approach to zero trust data center architectures encompasses a wide range of components, for the remainder of this paper we will concentrate on disruptive network technologies. For example, zero trust networks benefit from the centralized management plane and dynamic configurability offered by software-defined networks (SDN). While the definition and basic operating principles of SDN networks are well known [7], we will briefly mention several of their useful features. Programmable SDN controllers are able to implement dynamic network segmentation based on data collected from sources outside the network itself, such as honeypots, security analytic engines, and other sources. The application of security analytics to monitoring or management data sets enables the creation of actionable threat intelligence, allowing an SDN network to proactively discourage security threats and respond in near real time when new threats become apparent. This approach is particularly effective when combined with virtualized network functions (VNFs) such as virtual routers, firewalls, or other appliances [8, 9].

In this paper, we describe an approach that enables zero trust networks by providing first-packet based authentication, combined with transport layer access control, and experimentally demonstrate the use of this approach in defending an SDN controller from cyberattacks. We describe a steganographic overlay that embeds network authentication tokens in a TCP connection request, and blocks unauthorized traffic from completing a request. Resources protected in this manner are effectively concealed from reconnaissance attempts by attackers. We also demonstrate an approach to transport layer identity management and authentication. We show that this approach prevents fingerprinting of key network resources (such as the SDN controller) by blocking any response to unauthorized packets at the transport layer and below. We experimentally demonstrate this approach in both large, enterprise-class servers and a cloud computing test bed.

This paper is organized as follows. After the introduction, we describe the operation of a transport layer identity management scheme in section 2. We then present experimental results for the enterprise server and cloud test bed use cases in section 3. Finally, section 4 presents our conclusions and recommendations for future work.

2. Transport Access Control Architecture

Our approach independently authenticates each network session at the transport layer, prior to granting any access to the network. This is implemented through a combination of two technologies, namely transport access control (TAC) and first packet authentication. To our knowledge, these approaches have not previously been combined in this manner. Unauthorized session requests are completely rejected, and there is no feedback to a potential attacker attempting to fingerprint the network. Explicit trust is established by generating a network identity token during session setup. The network token is a 32 bit, cryptographically secure, single use object which expires after four seconds. Tokens are associated with identities from existing Identity Access Management (IAM) systems and credentials, such as Microsoft Active Directory or the IAM system used by Amazon Web Services [10]. Explicit trust is established by authenticating these identity tokens as early as possible, namely on the first packet of a TCP connection request (see Figure 1).

Tokens are generated for each unique entity (user or device) requesting access to network resources. An in-line virtual appliance (the TAC gateway) is installed between the equipment

being protected (the protected resource) and the rest of the network; a second gateway is installed between the authorized or trusted user and the rest of the network. When the trusted user attempts to access the protected resource, the first gateway inserts an identity token into the first packet of the TCP connection request. When the second gateway receives a connection request, it extracts and authenticates the inserted identity token and then applies a security policy (such as forward, redirect, or discard) to the connection request based on the received identity. This gateway acts as a policy enforcement point transparent to the rest of the system architecture and backwards compatible with existing network technologies. If the network access token for a TCP request fails to resolve to an identity or resolves to an identity that lacks the authority to access the requested resource, then the connection request is rejected without providing any further response to the requestor, effectively cloaking the presence of the protected resource. Failed access attempts are logged in an external Syslog server, which allocates enough memory to avoid wrapping and over-writing log entries. Existing tools such as SIAM can be used to analyze the logs or generate alerts of suspicious activity. We note that continuous logging of all access attempts is consistent with the approach of a zero-trust network (i.e. not allowing any access attempts to go unmonitored). Both the identity insertion gateway and identity authentication gateway appliances can be hosted as VNFs hosted on a virtual server, router, or other compatible piece of networking equipment.

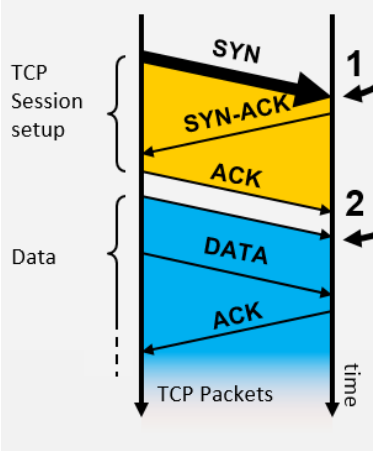


Figure 1: Transport Layer Authentication handshake; first packet authentication is performed at position 1 (as early as possible); traditional session authentication is performed at position 2 (after the session is already established).

This approach offers several advantages, including separation of security policy from the network design (addresses and topologies). This approach works for any network topology or addressing scheme, including IPv4, IPv6, and networks that use the Network Address Translation (NAT) protocol, and is compatible with dynamic addressing often used with mobile devices. This approach extracts, authenticates, and applies policy to the connection requests, not only protecting against unauthorized external reconnaissance of the network devices but also stopping any malware within the protected devices from calling home (exfiltration). Security policies can be easily applied at the earliest possible time to conceal network-attached devices from unauthorized awareness. By preventing unauthorized awareness and access, transport access control blocks both known and unknown attack vectors. This approach is low latency and high bandwidth since packet content is not inspected. Since the network tokens are embedded in the TCP session request, they do not consume otherwise useful data bandwidth. The combination of transport access control and a segmented, multi-tenant network implements a layered defense against cybersecurity threats, and contributes to non-repudiation of archival data. These techniques are also well suited to protecting public and hybrid cloud resources, or valuable, high performance cloud resources such as enterprise-class mainframe computers. Further, this approach can be applied to protecting the centralized SDN network controller from unauthorized access, and enable only authorized SDN controllers to manage and configure the underlying network. TAC uses an innovative identity token cache to provide high scalability and low, deterministic latency. The token cache is tolerant of packet loss and enables TAC deployments in low bandwidth and high packet loss environments.

3. Cloud test bed experimental results

The experimental cybersecurity cloud test bed is illustrated in Figure 2. A protected resource (in this case, the SDN controller) is intended to be accessible only from either of two trusted clients. A BlackRidge hardware appliance gateway, which implements TAC with first packet authentication [11], is placed in-line with the trusted clients, where it inserts tokens in the transport frame headers. Tokenized packets flow through the untrusted network, which eventually routes them to the SDN controller. A virtual gateway appliance is placed between the network and the SDN controller, which will authenticate the tokenized packets and only allow authenticated and authorized packets to pass through to the SDN controller. Any packets without Tokens or identified traffic without the authority to access the SDN controller will be dropped. Our test configuration uses a hardware appliance to insert Tokens and a Virtual Appliance running on VMWare ESXi to authenticate Tokens. The hardware and software appliances are only addressable through their management ports and use the management ports to access the required network time protocol (NTP) Servers. A list of trusted devices to be allowed access is provisioned in the TAC gateways, and the list of trusted devices can be edited using the gateway management ports.

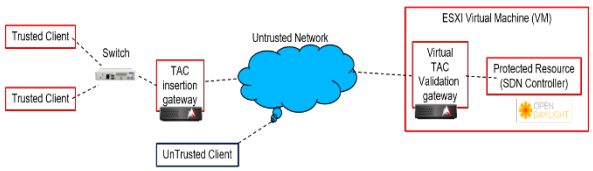


Figure 2 – Experimental cybersecurity cloud test bed

The specifications for the gateway used in this testbed are given in Table 1. This gateway has three modes of operation, known as Bridge, Enforce, and Monitor. In Bridge mode the gateway does not perform authentication or insert tokens into the data packets; rather, it simply functions as a two port, Layer 2 bridge device. Enforce mode will perform authentication and insert tokens into the 32 bit sequence and acknowledgement number fields of a TCP frame, according to the established address list policy. Monitor mode has the same functionality as Enforce mode, with the exception that it does not enforce the security policy. Monitor mode is useful to validate configurations during installation and setup for a new gateway. By toggling a configured gateway between Bridge and Enforce modes, it is possible to observe the effects of turning token-based authentication off and on. The gateway architecture is a “bump

in the wire" approach, and the gateway device is only addressable through its own dedicated device management access port.

Table 1 – TAC gateway capabilities

Performance	Branch Gateway	1G VM, z/VM, and AWS Gateway[1]	1G 1RU Gateway	10G VM Gateway[1]	10G 1RU Gateway	z Systems LPAR Gateway[1]
Throughput	1 Gb/sec	1 Gb/sec	1 Gb/sec	10 Gb/sec	10Gb/sec	1Gb/sec
Latency	<200 µsec	<100 µsec	<80 µsec	<100 µsec	<12 µsec	TBD
Max Capacity	**Branch Gateway**	**1G VM/KVM Gateway**	**1G 1RU Gateway**	**10G VM Gateway**	**10G 1RU Gateway**	**z Systems LPAR Gateway**
Identities[2]	1K	10K	10K	30K	40K	10K
Concurrent Sessions	100K	1M	1M	4M	4M	1M

1. Virtual appliance performance is server platform dependent. Required resources:
   - 1GbE VM, z/VM, AWS: 4 core CPU, 4-GB memory, 8GB storage
   - Branch Gateway 2 cores 8gb memory, 256GB SSD storage
   - 10GbE VM: 8 core CPU, 16 gb memory, 8GB storage
   - z Systems LPAR Gateway: 1 IFL, 16 GB memory, 32GB storage
2. Total Identities is limited by the Enterprise IDMS

We configured the test bed and toggled the gateway between Bridge and Enforce modes. This allowed us to verify that Enforce mode would only permit tokenized packets from one of the trusted clients to reach the SDN controller. We then attempted a reconnaissance scan of the SDN controller from an untrusted client. These scans were conducted using several industry standard tools, including Metasploit, HTTPrint, Firewalk, and PuTTY [5, 6]. When the TAC gateways were in Bridge mode, we were able to successfully fingerprint the SDN controller, and determine that it was an instance of Open DayLight (Helium release) running OpenFlow 3.2 in this example. We then repeated the scans with the gateway in Enforce mode, and as expected we were unable to identify even the presence of an SDN controller since The TAC authentication gateway blocks all potential responses at and below the transport layer. We cannot even determine if there is a TAC gateway present, as TAC was also used to protect the management port of the gateway. This implements both packet level authentication and unauthorized awareness, both desirable properties in a zero trust architecture.

To evaluate the effectiveness of the TAC gateway in defending against denial of service (DoS) attacks, we launched a DoS attack at the network protocol layer from the untrusted client in Figure 2. Common DoS simulation tools require knowledge of the target IP address, but as previously demonstrated the TAC gateway effectively cloaks the IP address for our SDN controller. For test purposes, we assume that an attacker has somehow obtained the SDN controller IP address through outside channels (perhaps a spear phishing attack on the network administrator) and we proceed to launch a DoS attack against the controller. Using a standard tool such as Low Orbit Ion Cannon (LOIC), we launched an attack against the IP address of the gateway data port, management port, and SDN controller. All packets were blocked by the TAC gateway without providing any additional intelligence to the attacker.

The gateway can also be used in a Layer 3 operating mode, which performs NAT for selected ports on the gateway. This is useful in cloud computing environments, allowing the gateway to present a public IP address on its client facing, untrusted port and a private IP address on its trusted port. In this case, the insertion and authentication of tokens is performed before NAT. We have demonstrated this approach in a public cloud deployment, similar to figure 2 except that the protected resource was located within an Amazon Web Services cloud. The cloud service provider infrastructure is located on the right-hand side of figure 2 (replacing the ESXi network), with the public Internet acting as the untrusted network and the cloud users on the left side of figure 2. Public IP addresses are used on ingress ports facing the untrusted network and the cloud service provider's protected resource connect to a trusted egress port.

The virtual TAC gateways were also tested in a enterprise-class environment using the IBM zCloud (a cloud based on a highly virtualized IBM Z Systems mainframe). These servers are commonly used in public, private, and hybrid cloud environments for Fortune 500 applications (particularly in the financial markets) as well as within cloud service providers such as SoftLayer. As shown in figure 3, an IBM model z13 enterprise server was provisioned into four logical partitions, with two partitions running the z/OS operating system, and two partitions running zLinux. For each operating system, one partition served as the protected resource while the other served as the trusted host. All four partitions share common physical network interfaces, provided by an Open System Adapter (OSA) card running 1 Gbit/s Ethernet. The virtual appliances were hosted in two additional logical partitions (LPARs), interconnected with the protected resources, trusted hosts, and OSAs as shown in figure 3. Additional OSA cards were provisioned to serve as interfaces for the network management ports on the virtual appliances.

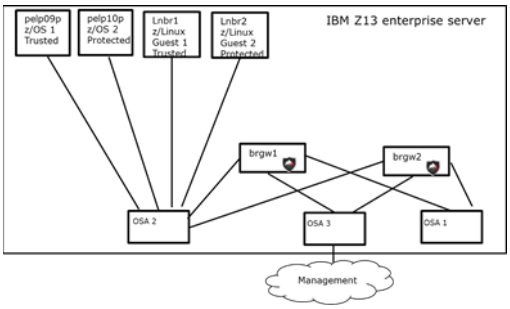


Figure 3 – Enterprise server use case test configuration

Three use cases were tested using this configuration. First, the gateways were configured to allow connectivity only between Linux partitions 1 and 2. We confirmed normal operation of resource connectivity including SSH, SCP, iperf, sftp, and wget functions, and verified that untrusted hosts such as zOS-1 could not access the Linux partitions. Second, the gateways were configured to allow connectivity between Linux-1 and zOS-2 partitions (note that the gateway authentication is independent of the operating system running in either the supplicant or the trusted resource). As before, we verified basic functionality (including multiple sftp file transfers between the trusted host and protected resource) and confirmed that other partitions, such as zOS-1, could not access the protected \resource in this configuration. Third, the gateway was configured to allow connectivity between the two zOS partitions. As in the previous use cases, we verified basic functionality (including multiple sftp file transfers between the protected resource and trusted hosts) and confirmed that the Linux-1 partition was unable to access protected resources in this configuration. These three test cases established that the gateways could be configured to enable or disable applications running between any two partitions on the same physical server, even if the gateway itself is hosted in a partition on the same physical server. This approach directly supports the zero-trust architecture we intended to implement.

The gateway functionality was further demonstrated in a pre-production test environment at Marist College, part of the New York State Cloud Computing and Analytics Center (CCAC).

In this case, the network spans multiple buildings on the campus MAN (about 1 km apart) as shown in figure 4. In the first building, ten sysadmin terminals were interconnected to a Layer 2 switch, and one switch port was connected to the insertion gateway. In this configuration, all terminals connected to the trusted switch port receive authentication tokens and will be allowed to access the protected resource. In a second building, the second gateway was configured in Layer 3 mode, which was then attached to the protected resource (a sysadmin application (Syswiki) running in a SUSE Linux guest VM on an IBM z144 enterprise server (mainframe).

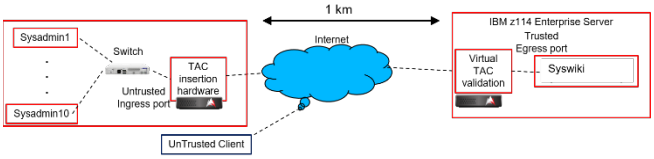


Figure 4 – syswiki BlackRidge

Using this configuration, we verified that the gateway allowed the Sysadmin trusted host terminals to perform operations including SSH, scp, and http post/get functions to the protected resource. We also verified that other terminals (untrusted hosts) on the same network were unable to access the protected resource when the gateways were set to Enforce mode, but could access the protected resource when the gateway was put into Bridge mode. The Sysadmins showed no measurable degradation in response time or performance of the protected resource application with and without the gateway operating in enforce mode.

Further, there was no measurable performance impact when accessing other resources on the campus network or accessing Internet resources when the gateway was in Enforce or Bridge mode. We also conducted a port scan of the gateway using Nmap/Zenmap tools from an untrusted terminal, and were unable to identify any open ports or fingerprint the Syswiki when the gateways were in Enforce mode.

4. Optical Bypass Switch Testing

For the next phase of testing, we incorporate a high-speed optical bypass switch into the system architecture, so that in the event of a hardware or software failure in the gateway, the TAC appliance will not obstruct network traffic. One advantage of an optical bypass switch is that it continues to pass traffic even if a power failure or firmware problem disables the functionality of the TAC gateway. Similar approaches have been suggested in high performance computing applications, where network traffic bypasses firewalls and other security features in order to maintain throughput for the application [12]. For these experiments, we used a 10 Gbit/s traffic stream while evaluating the impact of protection switching on the gateway.

In our environment, the TAC gateway incorporated a 10 Gbit/s multimode Silicom brand optical switch (model PE210G2BP) with optical bypass switch driver ixgbe ver 4.3.15. The optical bypass testbed is shown in figure 5. We configure two ESXi server instances, each driving 10 Gbit/s of traffic (loadgen 1 and loadgen 2, respectively). These servers run the Iperf benchmark test, enabling us to test switch times and throughput in both directions. As shown, the virtual switch on server loadgen1 connects to an untrusted port on the TAC gateway, while the virtual switch on loadgen2 connects to the corresponding trusted port.

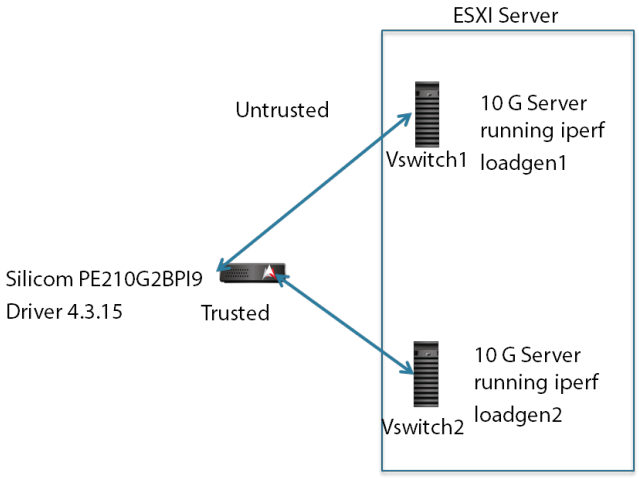


Figure 5 – Optical Bypass Switch testbed

Each test was repeated five times to obtain a meaningful average result, as shown in table 2. We can see that the optical bypass can sustain a data rate of about 8.3 Gbit/s while running Iperf, and about 8.9 Gbit/s in bypass mode (as expected, the bandwidth in bypass mode is somewhat higher).

Table 2 – optical bypass switch sustainable bandwidth measurements

Bypass mode, Gigabits per second	Iperf, Gigabits per second
8.88	8.39
8.95	8.24
8.92	8.33
8.91	8.05
8.85	8.31

In order to determine how long it would take for the optical bypass to tansfer traffic under different conditions, another server was inserted as a tap on the untrusted link, as shown in figure 6. By running WireShark on a 10 Gbit/s interface, this server is able to measure the effective switching times. A representative WireShark trace, shown in figure 7, confirms that the gateway bypass switch exhibits 110 ms recovery time upon a power failure to the gateway. We found that recovery time was faster when simulating a firmware outage (53 ms) and slightly longer when the TAC cloaking function was suspended due to a firmware problem (257 ms).

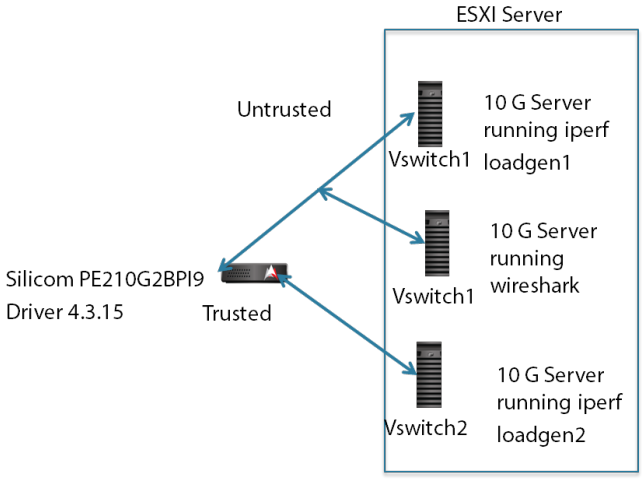


Figure 6 – Optical Bypass swithing time test bed

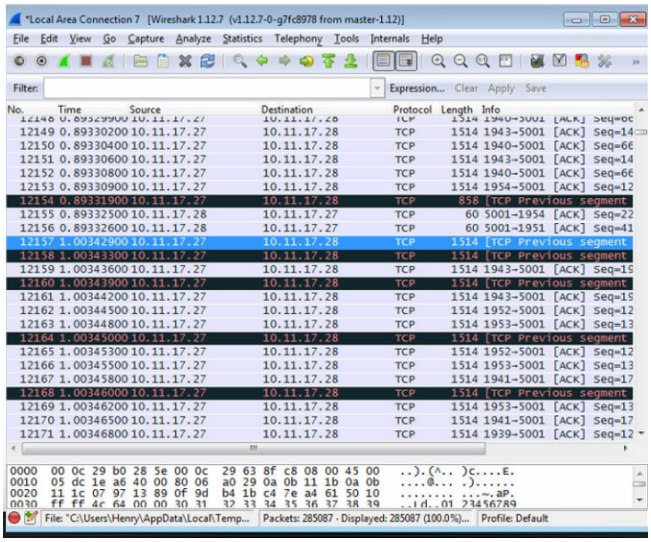


Figure 7 – Representitive WireShark trace from Optical Bypass switching time test bed

Finally, we reconfigured the gateway to conduct syslog benchmark testing, as shown in figure 8. The WireShark server is now monitoring traffic through the TAC gateway with optical bypass. We can run an nmap port scan simultaneously on servers loadgen1 and loadgen 2 using this configuration. For this testing, we configured the gateway to automatically blacklist any IP address which makes more than 100 access attempts in 60 seconds (an effective defense against some types of brute force dictionary attacks). We can keep an address blacklisted for different periods, ranging from 30 seconds to hours or even 24 hours. A blacklisted address generates a specific message type in the syslog, so we can easily determine the number of blacklist events in a given period.

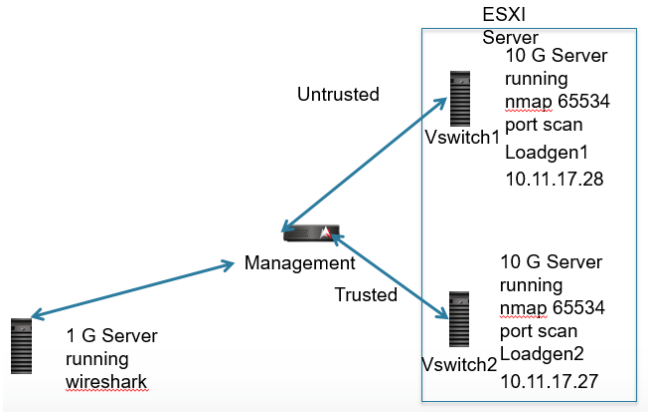


Figure 8 – Optical Bypass syslog test bed

For our tests, the initial default setting was a rate limit of 10 messages within 5 seconds, or a peak rate of 5 Mbit/second. Based on existing data about typical syslog messaging patterns [2, 3, 12], we expect the performance of our system to be well within normal commercial operating parametes. The peak data throughput measured during this experiment was 14 Mbit/second.

5. Conclusions

The growing cybersecurity treat requires an architectural redesign of the data center network, based on the principles of an explicit zero trust network. We have demonstrated several principles of zero trust using a transport access control system, based on a steganographic overlay, which embeds authentication tokens in the TCP packet request and first-packet authentication. The system was tested on both x86 and Z Systems platforms in private cloud environments, and using AWS in a public cloud environment. This system can provide enhanced security in both enterprise computing and cloud environments as part of a defense-in-depth strategy and prevents unwanted fingerprinting of protected resources. An optical bypass switch was also characterized as part of a high availability architecture.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The authors gratefully acknowledge the support of the National Science Foundation (NSF) grant 1541384, Campus Cyberinfrastructure – Data, Networking and Innovation Program (CC-DNI), per NSF solicitation 15-534, for the project entitled CC-DNI (Integration (Area 4): Application Aware Software-Defined Networks for Secure Cloud Services (SecureCloud). The authors also gratefully acknowledge the support of Marist College and the New York State Cloud Computing and Analytic Center (CCAC).

References

[1] C. DeCusatis, P. Liengtiraphan, A. Sager, and M. Pinelli, “Implementing zero trust networks with transport access control and first packet authentication”, Proc. IEEE International Conference on Smart Cloud 2016, New York, NY (November 18-20, 2016)

[2] “Cisco 2017 annual security report”, published by Cisco System Inc., https://www.cisco.com/web/offer/gist_ty2_asset/Cisco_2017_ASR.pdf . (Last accessed February 9, 2015)

[3] “IBM X-Force trend and risk report”, published by IBM Corporation, October 2013http://www-03.ibm.com/security/xforce/downloads.html (last accessed December 18, 2015)

[4] NIST report, “Developing a framework to improve critical infrastructure cybersecurity”, submitted by Forrester Group, 18 pp. http://csrc.nist.gov/cyberframework/rfi_comments/040813_forrester_research.pdf (April 2013; last accessed January 5, 2015).

[5] R. Smith, Elementary Information Security, 2nd edition, Jones and Bartlett, Burlington, MA (2016)

[6] S. Oriyano, Hacker Techniques, Tools, and Incident Handling, 2nd edition, Jones and Bartlett, Burlington, MA (2014)

[7] S. Cherukuri, “NFV architecture and orchestration for cloud based virtual managed services”, Cisco Live 2015 paper BRKSDN-2065 https://www.ciscolive.com/online/connect/sessionDetail.ww?SESSION_ID=83663&tclass=popup (last accessed May 20, 2016)

[8] M. Casado, N. Foster, A. Guha. “Abstractions for Software-Defined Networks”. Vol. 57, pp 86-95. Communications of the ACM. 2014

[9] J. Kindervag and R. Harrison, “Orchestrate a zero trust network”, https://www.brighttalk.com/webcast/9591/186577?utm_campaign=communication_missed_you&utm_medium=email&utm_source=brighttalk-transact&utm_content=webcast (last accessed April 28, 2016)

[10] Amazon Web Services Identity and Access Management, April 2016 https://aws.amazon.com/iam/ (last accessed May 20, 2016)

[11] BlackRidge white paper, “Dynamic network segmentation”, August 2012http://www.blackridge.us/images/site/page-content/BlackRidge_Dynamic_Network_Segmentation.pdf (last accessed April 27, 2016)

[12] T. Hwang, “Secure Science DMZ using event driven SDN”, Proc. Internet 2 symposium, https://meetings.internet2.edu/media/medialibrary/2015/10/05/20151005-hwang-Cisco-Secure-Science-DMZ.pdf (October 5, 2015, Denver, CO; last accessed March 3, 2016).

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 36-40 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ
ISSN: 2415-6698

A Derived Metrics as a Measurement to Support Efficient Requirements Analysis and Release Management

Indranil Nath*

Chartered Fellow, BCS, The Chartered Institute for IT, United Kingdom

ARTICLEINFO

Article history:
Received: 11 March, 2017
Accepted: 27 March, 2017
Online: 04 April, 2017

Keywords:
Software Requirements Management
Software Engineering
Software Development Process
Requirements Specification
Lifecycle Model
Release Planning, Analysis Metrics

ABSTRACT

This paper presents a Release Management model to support requirements management. Requirements development and management can be integrated with a release-planning approach to achieve lesser Requirements spillover problems which is an innovative way to capture, control and evolve the user requirements based on integer linear programming.

1. Introduction

The Capability Maturity Model (CMM) and its successor Capability Maturity Model Integration (CMMI) have been widely accepted as a process model to steer software process improvement (SPI) activities [1]. In some parts of the software industry, like the defense industry, software suppliers are obliged to have their software processes at least at CMM [2] level 4 or 5 in order to get any contracts. In many organizations, software has become such a critical part of the product portfolio, that senior management at the strategic level has defined objectives and allocated budgets for improvements. Despite these successes to realize sustainable improvements, the percentage of failures is however probably still higher. It is true of any system that there are several Critical Success Factors [3]. They include both performance requirements (such as serviceability, reliability, portability, and usability) and limited resource requirements (such as people, time and money) In my experience, they either failed because of poor estimation or poor change management or even poor requirements including release plan.

Although modeling activities [4] like Data Flow Diagram, Entity Relationship Diagram, State Charts, Object Oriented Method and Unified Modeling Language helped to better understand the Requirements and to represent them in clear and comprehensive manner, it still failed to adapt to an evolutionary approach for incremental software release and an opportunity for prioritizing requirements. This paper presents a Release Management model to support requirements management. Requirements development and management can be integrated with a release-planning approach to achieve lesser Requirements spillover problems that are an innovative way to capture, control and evolve the user requirements. A metric coupled with the incremental software release that offers organizations an opportunity for prioritizing requirements, an optimization technique, based on integer linear programming, to support software vendors in determining the next release of a software product developed and demonstrated in Section 3. The technique is based on the assumption that a release's best set of requirements

*Corresponding Author: Indranil Nath, Chartered Fellow, BCS, The Chartered Institute for IT, United Kingdom
Email: indranil.nath@bcs.org.uk

is the set that results in maximum projected revenue against available resources in a given time period.

2. General Process Metrics

Today measuring processes and products, objective, quantified information can be gathered to support decision-making at all levels. The objective measurements provided allow management and software product engineers to:

1. Identify inefficient processes that need improvement.
2. Reengineer ineffective processes that do not add quality to products.
3. Evaluate process changes to measure process improvement.
4. Utilize historic measurements of productivity, complexity, size, effort, and cost during project planning.
5. Monitor a current project plan's risk by measuring the schedule, cost, size, and effort of the products under development and comparing them to the assumptions and estimates used as the basis for the plan.
6. Monitor product development by measuring current cost, effort, and schedule data to determine if corrective action, such as re-planning, is needed.

A typical Process Flow of the Software Process is shown in Fig1 that is based on Humphrey [4] with only minor changes. In the first phase Analysis the Requirements are studied, where after estimates are made for the size of the resulting program, the productivity, the effort distribution over the subsequent phases, and other product and process metrics. These estimates are recorded on the Project Plan Summary.

During the subsequent phases, the actual defect and time data are logged on the Defect Log Form and the Time Log Form. The Design Review Checklist and the Code Review Checklist are used to structure the review process during the Design Review phase and the Code Review phase. When the program has been tested and is ready to be released as a Finished Product, all recorded data is summarized in the Project Plan Summary, enabling the software engineer to make an analysis in the Post Mortem phase prior to the program release.

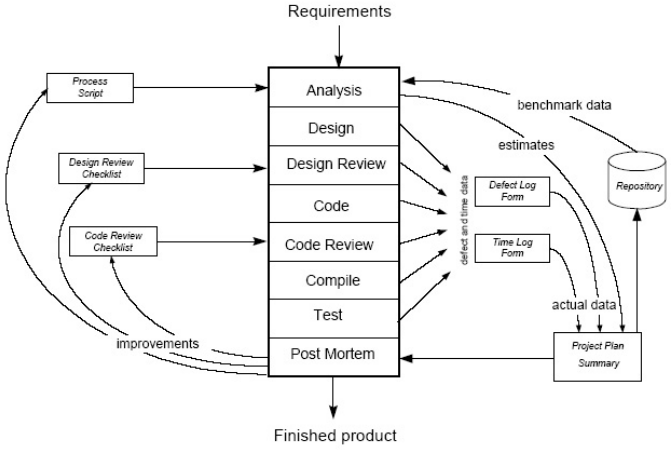


Fig 1: Process flow of the Software Process (based on [HUM 1997]).

The purpose of this analysis is to investigate how the process used can be further improved. The result may be that the Process Script, the Design Review Checklist and/or the Code Review Checklist are adjusted. Further, the data from the Project Plan Summary can be used to calculate accumulated values for productivity, time or effort distribution, and so on. This information can be stored in a Repository to support the estimation process for future programming assignments.

By following the Process Flow of the Software Process [5] Humphrey proved that valuable product and process data could be obtained and the analysis of the data would enable a software engineer to calculate for instance the following metrics for better decision making in software engineering:

2.1. Estimation Accuracy

Estimates are collected for the size of the program to be developed and the process to be followed (time distribution, productivity, etcetera). These estimates are compared with the actual figures when the program has been implemented and tested. The result is the estimation inaccuracy, which can be calculated with the following

Equation:

$$\text{Estimation Inaccuracy} = \frac{\text{Actual Value} - \text{Estiamted Value}}{\text{Estimated Value}} \times 100$$

2.2. Relating Size and Effort: Productivity

By measuring the size and effort of programs, the productivity can be calculated in retrospect. When the productivity has been measured, however, it can be used to proactively calculate the expected effort or time needed to develop a program when the size of the program has been estimated, the equation being:

$$\text{Calculated Effort} = \frac{\text{Estimated Size}}{\text{Measured Productivity}}$$

2.3. Yield

When developing a program subsequent development phases are passed through. In each development phase defects will be injected and removed. Removed defects might have been injected in the current development phase; however, they might also have been injected in previous development phases. The yield of a development phase can be defined as a number, expressing the number of defects removed in a particular phase divided by the sum of the number of defects inherited from previous phases and the defects injected in the phase itself:

$$\text{Yield (X)} \frac{\#\text{removed in } P_X}{\left(\sum_{P_0}^{P_{X-1}} \#\text{inherited in } P_X\right) + (\#\text{injected in } P_X)} \times 100$$

Where #= Number of Defects and P_x is ”Phase x” The yield of a phase can vary from 0% to 100% (if the denominator equals 0, the yield of the phase is equal to 100%). When the yield of a phase equals 0 it means that in that particular phase no defects have been removed. They are all carried forward to the next phase. When the yield of a phase equals 100%, it means that in that particular phase

all existing defects are removed. These are both the defects inherited from previous phases and the defects injected in that particular phase itself. In the ideal case, the yield of each development phase is equal to 100%. This means that in each phase all injected defects are also removed and that the final product is free of defects.

But in practice, not all defects will always be removed. Defects will be carried forward to the subsequent phases. If they are not found in, they will be shipped being part of the final product. Note that the yield of the different development phases can only be calculated when all defects have been found. Only then it is known how many defects were injected and removed in the different phases.

If an injected defect is carried forward to subsequent phases, it is very important to detect and remove the defect as early as possible. The rework effort is proportional to the difference between the injection phase and the detection and removal phase. If an early injected defect is found in one the latest phases of development, the rework effort will extend over more phases. Very expensive defects are the ones that are injected during the Analysis and Design phase and detected during the Test phase. It is assumed here, that there are no defects in the requirements. Another approach is now to calculate the Yield of the phases Analysis and Design (A&D) combined:

$$Yield_{AD} = \frac{\#\text{ removed in A\&D Phase}}{\#\text{ injected in A\&D Phase}} * 100$$

This metric denotes the quality of the early phases before handing over of the product for further implementation and testing.

2.4. Appraisal/Failure Ratio (A/FR)

Striving for a high combined Yield means investing effort in appraisal to detect defects as early as possible. An example of an appraisal action is, for instance, a peer review. The objective of appraisal actions is to pass forward a reliable product to the next phase. Advantages are twofold. In the first place, this will probably lead to a more reliable final product. It is a utopia to think that all transported defects will be found at a later stage. In the second place, the total failure cost in subsequent phases due to fixing detected defects will probably decrease as the number of expensive defects decreases. A new metric is introduced, called the Appraisal/Failure Ratio or A/FR. This is the ratio between the appraisal cost in the early development phases and the failure cost in subsequent development phases, its equation:

$$\text{Appraisal / FAilure Ratio} = \frac{\text{appraisal effort in A\&D Phase}}{\text{failure effort in subsequent phase}}$$

The A/FR can vary from 0 to infinite (if the denominator equals 0, the A/FR is equal to 0 if defects are found but not removed and the A/FR is undefined if no defects are found). An A/FR equal to 0 means that no appraisal effort is spent in detecting and removing defects in the early phases of development. The failure cost to be made in subsequent phases later will presumably be very high, assuming that the reliability of the final product is important.

2.5. Defect Injection and Defect Removal Rate

Collecting defect data enables one to calculate the defect injection and defect removal rate in each development phase. It can, for instance, be related to the effort spent in that particular phase:

$$\text{Injection Rate (X)} = \frac{\#\text{ injected in } P_X}{\text{effort in } P_X}$$

$$\text{Removal Rate (X)} = \frac{\#\text{ removed in } P_X}{\text{effort in } P_X}$$

2.6. Defect Density

The defect density can be calculated by relating the total number of defects found during development to the resulting program size:

$$\text{Defect Density} = \frac{\#\text{ found}}{\text{program size in KLOC}}$$

A defect density equal to 0 means that no defects are introduced during development; in other words, the Yield of each development phase equals 100%. These metrics are powerful to analyze the software process and adjust the process in order to implement improvements. Most metrics are self-explaining with regard to their best values. The best values for Estimation Inaccuracy and Defect Density are for instance 0%, for Yield the best value would be 100%. This is however not evident for a metric like A/FR. But there is more. The best value does not necessarily mean that it is the optimal value. There might be practical limitations to obtain a Yield of 100% for instance. Human work is not perfect, causing the injection of defects, and a project budget is normally limited, meaning that time for appraisal and failure expenditures is limited. It is worth considering the question whether there are optimal combinations of a set of metrics, that there is no metric that allows helping management to decide on prioritizing requirements in the 3-Degree-Of-Freedom [6] of Software Engineering (Software Development Process, System Components and View onto the Systems).

3. The Derived Metrics

This section will have all the derived metric from above and how the derived metrics can be used to support the re-engineering. The set of derived metrics will offer the possibility to analyze and re-engineer processes and their respective work products. Finally, this re-engineering will allow for efficient Requirement Analysis and bi-directional requirement traceability [7-9].

One of the most common life-cycle approaches adopted by most software development organizations is the evolutionary approach. This approach, coupled with incremental software release offers the opportunity for prioritizing requirements [10]. Customers receive part of the full working system early on as a beta release and it is easier to schedule and estimate for each delivery. This approach also enables user feedback early on (during development stage) and changes and additions to requirements are easily handled. Release management is mathematically modeled as follows.

Let us consider a set of requirements $\Re$

$$\Re = \{r_1, r_2, ..., r_n\}$$

Assign $\Re$ into an evolutionary life cycle release categories C as:

I. Next Release (R_{n+1})
II. Next but one release (R_{n+2})
III. Not yet assigned (R_{n+m})

Requirement Dependencies D and requirement precedence within the domain $\Re$ are specified as:

- Requirement Coupling (RC): r_i must be in the same release as r_k
- Requirement precedence (RP): r_i must be implemented before r_k

Effort constraints Ω are formulated as:

- The total effort for implementing $\Re$ is Ω
- For $r_i \in \Re$, an estimated effort is assigned ω_i
- $\sum_{I=1}^{n} \omega_i \leq \Omega$

Stakeholders Prioritization S is formulated as:

- Set of stakeholder: $S = \{s_1, s_2, ..., s_m\}$
- Weight of each stakeholder $s_k \in S : w(s_k)$
- Each stakeholder s assigns a priority to the situation that r_i is assigned to the option $k : P(s, r_i, k)$

Thus, in the evolutionary life-cycle, a software release is defined as:

$$\text{Release } (\Re) = \int (\Re, C, D, \Omega, S)$$

4. Discussion

Release Re-engineering using Requirement Analysis [11]. For software vendors, the process to determine the requirements [12] for the next release of a software product is often difficult. The mathematical formalization of release composition with a corresponding optimization tool that aims to support product managers and development project managers during release planning is discussed as above. The tool is based on integer linear programming and assumes that an optimal set of requirements is the set with maximum projected revenue against available resources in a given time period. The input for the optimization is twofold. Input data like the list of candidate requirements estimated revenue and required team resources per requirement, whether or not a requirement is mandatory, comprise the first type of input. Secondly, several managerial steering mechanisms provide flexibility in the optimization environment.

Companies developing software products face complex challenges when determining requirements for upcoming releases. Often the wish list of requirements extends the capacity of available resources, requirements may not be unambiguous, they may be difficult to prioritize, etc. As a matter of fact, many aspects influence the definition of an optimal set of requirements for a next release. Several scholars have presented lists of such aspects including importance or business value, personal preference of certain customers and other stakeholders, the penalty if not developed, the cost of development in man days, development lead-time, requirement volatility, requirement dependencies, the ability to reuse, and requirements quality.

In order to deal with this multi-aspect optimization problem, different techniques and procedures have been applied. The analytical hierarchy process [13] assesses requirements according to certain criteria by taking all possible requirement pairs, relatively valuing each pair, and subsequently using matrix calculations to determine a weighted list of requirements. Jung extended the work of Karlsson and Ryan by reducing the complexity of applying the analytical hierarchy process to large amounts of requirements using linear programming techniques [14]. Through cumulative voting (Leffingwell and Widrig (2000)) different stakeholders are asked to distribute a fixed amount of units (e.g. euros) between all requirements, from where an average weighted requirement list is constructed. With discrete event simulation, the effect of the development of requirements is modeled, allowing what-if analysis.

The set of derived metrics shown in Section 3 offers the possibility to analyze projects, Software Requirement, and their respective work products in detail [15]. The following can be concluded:

- **Predictability:** There is a considerable budget overrun (+50%) and schedule overrun (+19%), but the realized product size is less than predicted (-24%). The realized overall productivity is nearly half the predicted value (-49%). These data need to be further analyzed in order to find out the reasons for these deviations.
- **Effectiveness:** The cumulative Yield at the Detailed Design phase and the Implementation phase is very low, 10% and 19% respectively. This implies that many defects are transported to subsequent phases and that the failure cost in these phases will probably be high. This is confirmed by the fact that the total amount of rework equals 38% of the overall lifecycle effort.
- **Efficiency:** The efficiency is very low because many defects have been found in the later phases. As a consequence, the overall failure cost is high (38% of overall life cycle effort). This is confirmed by the fact that the cumulative A/FR before the Detailed Design phase and Implementation phase is very low (0.02 and 0.03 respectively).

This analysis gives an organization the possibility to select improvements for future projects, assuming that this project has

not been very exceptional but comparable with similar projects. Improvements might be:

A) Improve the predictability by using a lower productivity figure when estimating projects. In this way, more realistic estimates will be obtained.
B) Increase the quality of the resulting product by increasing the Yield in the earlier development phases. This can be done by increasing the review effort (A/FR higher) or by adjusting the review method.
C) Improve the efficiency by increasing the A/FR in the various development phases before the Integration Test phase. Special interest might be given to the Implementation phase and Unit Test phase as in these phases the Detection Rate and Removal Rate are very high.

In all cases, the available data can be used to make predictions about the expected effects in terms of return on investment. This is not always easy, as the various parameters might be correlated (for instance, a higher A/FR will probably contribute to a higher Yield).

However, by building a record of experiences, the capabilities in this area can be stepwise improved. In this example, the following approach might be applied in an iterative way:

1) Increase the cumulative A/FR by increasing the appraisal cost in the earlier development phases.
2) Estimate what the effect will be on the cumulative Yield at the Detailed Design phase: what will be the reduction of defects injected in the Requirements phase and Architecture phase?
3) Take measures to improve the inspection and review process during the Implementation phase and Unit Test phase and estimate the effect on the Defect Detection Rate and the Defect Removal Rate in these phases.
4) Use the cumulative Yield at the Detailed Design phase to predict the number of defects that will be passed forward to subsequent phases and use the Defect Detection Rate and Defect Removal Rate to calculate the failure cost.
5) Calculate the effect on effort, time, and expected corrective maintenance cost.
6) Repeat the steps 1 till 5 until satisfactory (and realistic) values have been obtained.

5. Conclusion

The derived metrics can be considered to be useful for organizations still having to build a measurement foundation for their software projects and products as well as for organizations looking for further improvements of their software processes. The mature organizations will analyze their available data, select the most promising improvements and re-engineer and re-measure their performance over time. The less mature organizations will not be able to implement re-engineering from day one but need to define a strategy how to implement the model in subsequent steps. The order of the steps can be derived from the most urgent problems the organization is faced with:

I. Predictability. In this case, the first step might be to collect estimates and actual for size, effort and time.
II. Effectiveness. In this case, the focus is on the output of the project, being the quality of the product. Important measurements to be considered first are defects (Yield values per phase and cumulative).
III. Efficiency. In this case, attention must be paid to appraisal cost related to failure cost (A/FR) and defect rates (Defect Detection Rate and Defect Removal Rate).

In all cases, supporting methods and tools might be selected to facilitate the measurements, as long as it is understood that "a fool with a tool still remains a fool". A method or tool will not solve the problem; they can only support a defined process.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] J.W.E. Greene, "Software Process Improvement: Management Commitment, Measure, and motivation", QSM Ltd. SPI98ADR
[2] CMMI-SVC, V1.3, Improving processes for providing better services, Software Engineering Process Management Program, Software Engineering Institute, November 2010
[3] Tom Gilb, "Competitive Engineering, A Handbook for Systems Engineering, Requirements Engineering and Software Engineering Using Planguage", Page 4. Elsevier Butterworth-Heinemann, 2006.
[4] R.S. Pressman, "Software Engineering – A Practitioner's Approach", Page 300, 301, 309-310, 305, 793, 438, 646. McGraw-Hill, 2001.
[5] Humphrey, W. Managing the Software Process, Addison-Wesley Publishing Company, 1990.
[6] I. Nath, S Kundu, "Is Software Requirements a Corporate Asset?" Fig.2. Degrees of Freedom in the Software Engineering Components of SE-CUBE, IEEE International Engineering Maangement Conference Proceedigs, Innovation and Entrepreneurship for Sustainable Development, Volune 1, 2004
[7] Davis, A. M., "Software Requirements: Analysis and Specification", Prentice Hall, 1990.
[8] Kovitz, B. L., "Practical Software Requirements: A Manual of Contents & Style". Manning, 1999.
[9] Potts, C., "Requirements Models in Context", 3rd International Symposium on Requirements Engineering (RE'97), Annapolis, USA, 6-10 January 1997, pp. 102-104, 1997.
[10] Nancy Mead., "Requirements Prioritization Introduction, Software Engineering Institute, Carnegie Mellon University, September 2006.
[11] A.G. Sutcliffe & M. Ryan, "Experience with SCRAM, a Scenario Requirements Analysis Method", 3rd Int'l. Conf. on Requirements Eng., pp. 164-171, 1998.
[12] Nguyen, L. and Swatman, P.A., "Managing the Requirements Engineering Process", Working Paper 2000/15, School of Management Information Systems, Deakin University, Australia, 2001.
[13] Melvin Alexander, "Decision-Making using the Analytic hierarchy Process (AHP) and SAS/IML", Social Security Administration, Baltimore, MD, 2012.
[14] John W. Chinneck, "Practical Optimisation " a Gentle Introduction", 2001
[15] Mack W. Alford, "Software Requirements Engineering Methodology (SREM) at the age of two", In 4th Int. Computer Software & Applications Conference, New York, pages 866–874. IEEE, 1980.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 41-47 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Feedback device of temperature sensation for a myoelectric prosthetic hand

Yuki Ueda, Chiharu Ishii*

Department of Mechanical Engineering, Faculty of Science and Engineering, Hosei University, Tokyo 180-0023, Japan

ARTICLE INFO

Article history:
Received: 21 February, 2017
Accepted: 28 March, 2017
Online: 05 April, 2017

Keywords:
Myoelectric prosthetic hand
Sensory feedback
Temperature sensation
Peltier element

ABSTRACT

In this study, a control system of a feedback device of temperature sensation for myoelectric prosthetic hand users was improved. The device is mounted on a user's upper arm in contact with the skin, and transfers the temperature sensation in the upper arm corresponding to the temperature detected by the temperature sensor at the fingertip to the user. However, since the feedback device developed in the previous study was controlled in an open loop system, it was difficult to maintain the temperature for a long period. To solve this problem, a closed loop control system was constructed for the feedback device by using a PID controller. In addition, in order to verify the performance of the feedback device controlled in the closed loop control system, performance evaluation experiments were performed. As a result, the sufficient capability of the feedback device controlled in the closed loop control system was indicated. Finally, temperature identification experiments were performed based on psychophysical method to verify an effectiveness of the feedback device. As a result, the difference threshold of the feedback device of temperature sensation, 0.62℃ was obtained.

1. Introduction

A myoelectric prosthetic hand is an electrically driven artificial hand in which the intended motion is identified and the movement is controlled based on biosignals generated when muscles are moved. However, myoelectric prosthetic hand users are unable to feel sensations when they touch an object using a prosthetic hand. Therefore, investigation of the functions that provide sensation feedback to the user is required. It is shown that sensory feedback is important for improvement in proficiency level of a myoelectric prosthetic hand [1].

Recent researches focused on developing a feedback device of sensations for myoelectric prosthetic hand users. These included our previous study [2] in which a feedback device of force sensation was developed. In this mechanism, when the myoelectric prosthetic hand held an object, the belt of the device was wound by a motor and tightened the operator's upper arm. This enabled the user to obtain a force sensation. Additionally, the user could feel a difference in the object hardness by changing the winding speed of the belt.

A study [3] also reported a feedback device using vibrations. The device changed the vibration intensity based on the size of the detected force. Another study [4] reported a device that perceives the temperature when touched by a myoelectric prosthetic hand using a hot and cold pad. This device obtains the temperature difference between an object and the user's skin temperature by the changes in the temperature of the hot and cold pads. A study [5] examined transfer of temperature sensation using the Peltier element.

In our previous study [6], a feedback device of temperature sensation was developed by using a Peltier element to perceive the object temperature when the user touches objects by a prosthetic hand. The object temperature is measured by a temperature sensor attached at fingertip of the prosthetic hand. However, the temperature measurement required several tens of seconds. Therefore, a temperature prediction algorithm was proposed to shorten the measurement time. The study [4] examined a device that transmits a temperature difference between skin temperature of cut section and object temperature. However, the temperature sensation differs at each body site. Therefore, the feedback device developed in our study transfers temperature sensation in the upper arm corresponding to the temperature detected at the fingertip of the prosthetic hand. To this end, an experiment to examine the difference in the temperature sensation felt by the fingertip and that felt by the upper arm was performed, and the feedback device reproduces the temperature sensation felt by the fingertip to the

*Corresponding Author: Chiharu Ishii, Email: c-ishii@hosei.ac.jp

upper arm based on the result of the experiment. Furthermore, an identification experiment for temperature was performed to verify the effectiveness of the feedback device. However, the feedback device developed in previous study was controlled in an open loop system. Therefore, it was difficult to maintain the temperature using this control system.

This paper is an extension of work originally presented in 2016 International Conference on Advanced Mechatronic Systems [6]. The key changes from the conference paper are summarized as follows.

In this paper, a closed loop control system was constructed to enhance the performance of the feedback device. In addition, to improve the operability of the myoelectric prosthetic hand, a new myoelectric prosthetic hand was built. Then, performance evaluation experiments were performed to verify the performance of the feedback device controlled in the closed loop control system using the new myoelectric prosthetic hand. Furthermore, psychophysics tests were conducted to measure the difference threshold of the feedback device in terms of the temperature difference.

2. Feedback Device of Temperature Sensation

2.1. Peltier Effect and Seebeck Effect

Exothermic and endothermic reactions occur when an applied current flows through a junction between two conductors. This phenomenon is called the Peltier effect. The Peltier element is an electronic device that changes surface temperature on the basis of the Peltier effect. It enables both cooling and heating by applying a voltage, and temperature can be adjusted by regulating the voltage. The feedback device has a Peltier element, and transfers the temperature sensation to the user. The Peltier element used in this study is "TEC1-12708" Peltier element made by HB Electronic Components.

In contrast to the Peltier effect, the Seebeck effect is a phenomenon that generates a voltage by providing a temperature difference to a conductor. When the temperature difference is provided to the conductor, voltage that is proportional to the temperature difference is generated due to the Seebeck effect. This principle is used to measure temperature. A thermocouple is a temperature sensor that uses the Seebeck effect. In this study, a K type "AD-1214" thermocouple manufactured by T & D Corporation was used as the temperature sensor.

2.2. Outline of Feedback Device

Figure 1 shows the feedback device of temperature sensation (Hereafter, FB device) developed in our study.

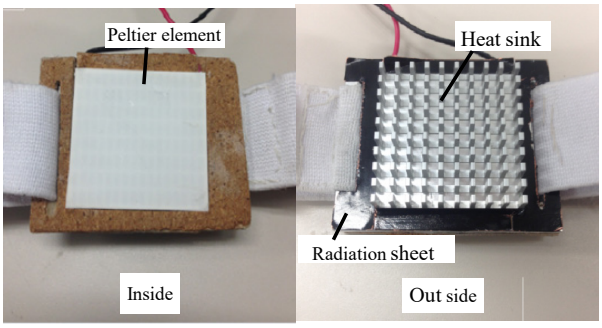


Figure 1: Feedback device of temperature sensation (FB device)

When the FB device is used for a long period, there is a possibility of high-temperature due to the accumulated heat. Hence, the continuous operating time of the FB device controlled in an open loop control system was limited to 5 s. Figure 2 shows relationship between the applied voltage and the temperature produced by the FB device.

The FB device enables refrigeration of surface temperature to a minimum of 15°C for 5 s. In addition, when the FB device increases the temperature, it enables heating exceeding 50°C. However, the temperature of the FB device was limited to 40°C for safety considerations.

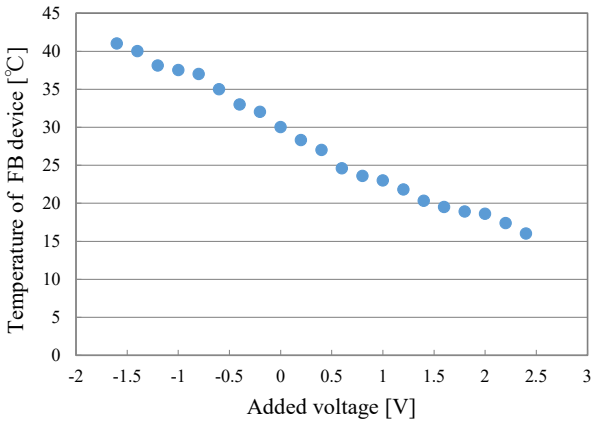


Figure 2: Added voltage and temperature of the FB device

3. Temperature Prediction and Temperature Sensation Investigation

After contact with the object, it takes several tens of seconds until the measured value settles. Thus, the temperature sensor needs a long time for measuring the temperature. Consequently, a temperature prediction was performed to shorten the measurement time.

3.1. Temperature Prediction Algorithm

Let $y(t)$ denote a sensor output and $u(t)$ denote an amount of temperature variation, and let relationship between $y(t)$ and $u(t)$ be given by the following equation in transfer function expression:

$$Y(s) = G(s)U(s) \tag{1}$$

In order to determine the transfer function $G(s)$, the parameters identification was performed by using the gradient method. As a result, $G(s)$ was given by the following equation:

$$G(s) = \frac{0.89s^3+0.66s^2+0.49s+0.06}{s^4+2.54s^3+1.7s^2+0.8s+0.06} \tag{2}$$

Furthermore, let ΔT denote the predicted temperature variation and $\hat{y}(t)$ denote the estimated output, then the estimated output is calculated as follows using the identified transfer function.

$$\hat{y}(t) = \frac{0.89s^3+0.66s^2+0.49s+0.06}{s^4+2.54s^3+1.7s^2+0.8s+0.06}\Delta T \tag{3}$$

Let $\varepsilon(t)$ denote the error between the estimated output $\hat{y}(t)$ and sensor output $y(t)$, which is represented by the following equation.

$$\varepsilon(t) = \hat{y}(t) - y(t) \quad (4)$$

The predicted temperature variation ΔT is updated by using $\varepsilon(t)$ as follows in each sampling period.

$$\Delta T_{new} = \Delta T_{old} + K \cdot \varepsilon(t) \quad (5)$$

where ΔT_{old} is the predicted temperature variation one step before and ΔT_{new} denotes an updated predicted temperature variation. In addition, K is an arbitrary constant, which was chosen as K=70 in this study, since noise was comparatively low and the reaction time was fast when the temperature prediction was performed.

Let T denote the predicted temperature and T_0 denote the room temperature, then the predicted temperature is given by the following equation.

$$T = \Delta T_{new} + T_0 \quad (6)$$

The predicted temperature is updated to reduce the error when the sensor detected a change of the temperature, thus the temperature when reached to the equilibrium state is given by Equation (6). As a result, the sensor can detect the temperature of the object in a short time.

3.2. Verification of Temperature Prediction Algorithm

In order to verify the effectiveness of the proposed temperature prediction algorithm, an experiment to investigate the behavior of the temperature prediction when the touched object was changed to an object with a different temperature during the temperature prediction was performed. When the sensor touched an object with a temperature of 20°C after touching an object with a temperature of 30°C under the room temperature of 25°C, the temperature was predicted by using the temperature prediction method. Figure 3 shows the result of the temperature prediction.

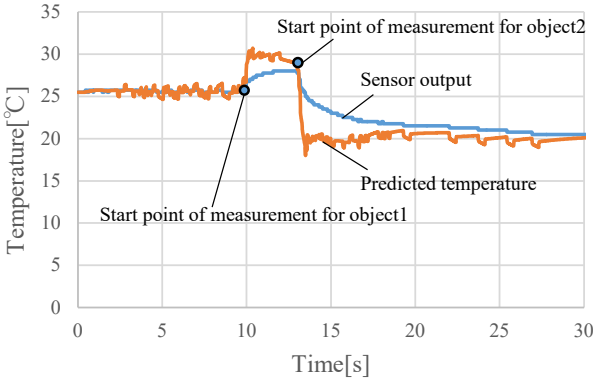


Figure 3: Temperature prediction for objects with different temperature

The results showed that it is possible to perform the temperature prediction even when an object with a different temperature was touched during the temperature prediction.

3.3. Purpose of Temperature Sensation Investigation

It is difficult for the FB device to reproduce the low temperature. As shown in Figure 2, the FB device cannot reproduce a temperature of lower than 15°C. However, the temperature sensations of an individual differ at each body site. For example, there is difference between the temperature sensations when an object with the same temperature is touched with a finger or with the upper arm. Therefore, an investigation of temperature sensation was performed, and a temperature that corresponds to the temperature at which an object was touched with a finger was determined as feedback temperature to the upper arm.

3.4. Method and Result

In order to investigate the difference in the temperature sensation between the fingertip and the upper arm, an experiment was performed in the following procedure using water from temperatures ranging from 5°C to 40°C for each 5°C. The details of the procedure of this experiment are addressed in the literature [6].

Table 1 shows the result of the temperature sensation investigation performed for 7 healthy subjects in their 20s.

Table 1: Result of temperature sensation investigation

Temperature (upper arm) [℃] Subject	Temperature (Fingertip) [℃]							
	5	10	15	20	25	30	35	40
A	15	18	22	27	29	30	34	37.5
B	15	18	22	28	29	30	33	37.5
C	15	18	22	28	29	30	33	37.5
D	15	18	22	27	29	30	33	37.5
E	15	18	22	27	29	30	34	37.5
F	15	18	22	27	29	30	33	37.5
G	15	18	22	27	29	30	33	37.5
Ave.	15	18	22	27.3	29	30	33.3	37.5

Based on these results, the FB device is controlled to reproduce the temperature experienced by the upper arm which is equivalent to the temperature experienced by the fingertip. Thus, the temperature sensation which should be reproduced to the FB device is determined by the temperature prediction at the fingertip. However, the temperature of FB device is not changed when the prosthetic hand touched an object at 30°C. This is because the temperature of FB device is approximately 30°C at normal temperature.

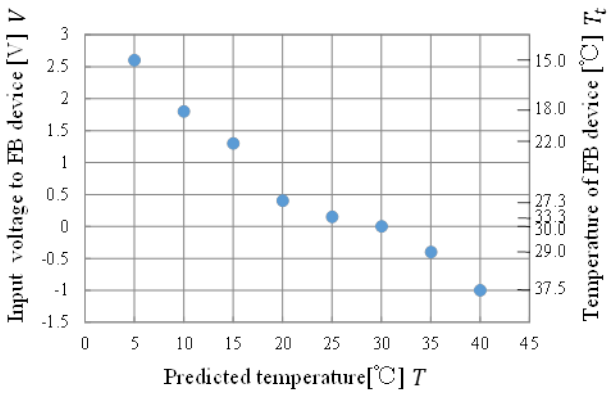


Figure 4: Predicted temperature and input voltage to the FB device

The temperature which should be presented at the upper arm and the voltage which should be applied to the FB device were determined as follows. The relationship between the temperatures experienced by the fingertip and by the upper arm was interpolated as shown in Figure 4. Then, the relationship between the predicted

temperature at the fingertip T and the target temperature in the upper arm T_t, and the relationship between the predicted temperature at the fingertip T and the input voltage to the FB device V are approximated as shown in Equations (7) and (8) respectively, by using the least-squares method.

$$T_t = 0.00008T^4 - 0.006T^3 + 0.1841T^2 - 1.1439T + 16.902 \quad (7)$$

$$V = 1.578 \times 10^{-7}T^6 - 2.135 \times 10^{-5}T^5 + 1.122 \times 10^{-3}T^4 - 2.894 \times 10^{-2}T^3 + 0.3819 \times T^2 - 2.543T + 8.743 \quad (8)$$

4. Temperature Identification Experiment

4.1. Purpose and Equipment

The FB device transmits the temperature sensation corresponding to the predicted temperature of the object which was touched by the fingertip of the myoelectric prosthetic hand to the user's upper arm. In order to verify the effectiveness of the developed FB device, the temperature identification experiment was performed using a myoelectric prosthetic hand developed in the literature [7].

The myoelectric prosthetic hand needs three types of operations in the experiment, namely bending the finger to touch the object, maintaining the posture, and releasing the finger. To distinguish these operations, electrodes were stuck on the flexor digitorum superficialis muscle and the extensor digiti minimi muscle to measure surface electromyogram (SEMG). The flexor digitorum superficialis muscle is used to flex the fingers, and the extensor digiti minimi muscle is used to extend the fingers. Figure 4 shows the positions where the electrodes were stuck on.

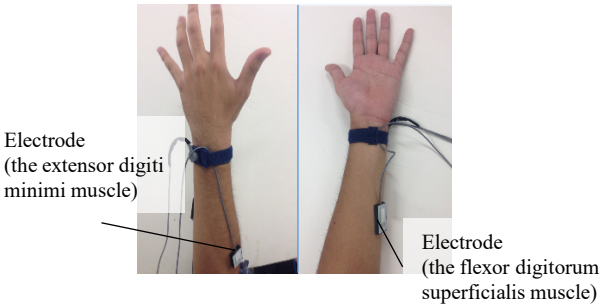


Figure 4: Positions of electrodes

4.2. Method of Experiment

The subjects executed the identification of the temperature transmitted by the FB device. In the experiment, another Peltier element which is different from that of the FB device was used as the object touched by the fingertip of the prosthetic hand. The following five variations in temperature were employed: hot (approximately 40°C), lukewarm (approximately 35°C), not much (25°C–30°C), a little cold (approximately 20°C), and cold (approximately 15°C).

The temperature identification experiment was performed in the following way. The temperature of the object (Peltier element) is randomly selected from the five variations, and the temperature of the object is adjusted to become the selected temperature. The subject operates the myoelectric prosthetic hand to touch the object. Then, the temperature prediction is performed and the voltage calculated from Equation (8) on the basis of the predicted temperature is applied to the FB device. Based on the corresponding temperature sensation reproduced by the FB device, the subject identifies the temperature of the object from the five variations. The experiment was executed for 10 subjects. The experimental environment is shown in Figure 5.

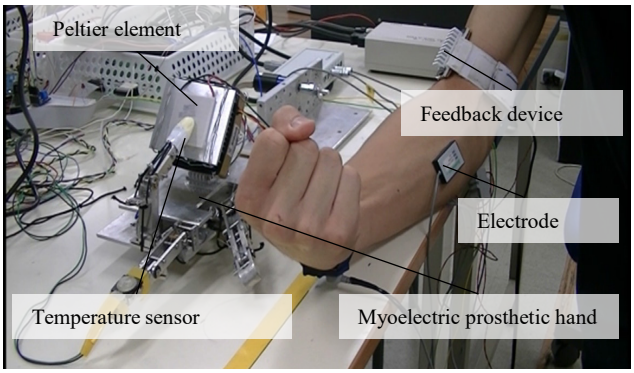


Figure 5: Experimental environment

Table 2 shows the success rates of temperature identification for each subject, and Table 3 shows the success rate of temperature identification for each presented temperature.

Table 2: Result of the temperature identification

Subject	Success rate of identification [%]
A	90
B	80
C	100
D	90
E	70
F	100
G	80
H	100
I	90
J	80
Average	88

Table 3: Success rates for each temperature

Presented temperature [°C]	Success	Failure	Success rate of identification [%]
40	17	0	100
35	23	5	85.2
25–30	9	1	90
20	22	4	84.6
15	17	2	89.5

The results show that the average success rate for 10 subjects is 88%. In addition, the success rate of each presented temperature exceeded 84%. The above results revealed that the difference in temperature can be recognized with high accuracy. The causes for the failure identification include the failure of the temperature prediction and the individual differences of the sensation of each subject.

5. Closed Loop Control System

5.1. Construction of the Control System

The FB device developed in the previous study is controlled in an open loop system, in which the constant voltage through Equation (8) computed from the predicted temperature is applied to the FB device. Hence, it is difficult to maintain the temperature using this control system, because if the voltage is continuously provided to the FB device the temperature keeps increasing or decreasing. Therefore, the continuous operating time of the FB device was limited to 5 s.

To solve this problem, a closed loop control system was constructed. To this end, a temperature sensor which is the same type as one attached at the fingertip of the prosthetic hand was additionally attached on the surface of the Peltier element of the FB device. The target temperature in the upper arm T_t corresponding to the predicted temperature at the fingertip of the myoelectric prosthetic hand T is determined by Equation (7). Then, the input voltage to the FB device is determined by a PID controller. The PID gains were determined by trial and error, and chosen as $K_P = 1$, $K_I = 0.05$ and $K_D = 0.01$.

In order to verify the validity of the closed loop control system with PID controller, sudden change in temperature when the target temperature is decreased from 40°C to 15°C was examined. The transition of the temperature of the FB device and the input voltage to the FB device are shown in Figure 6.

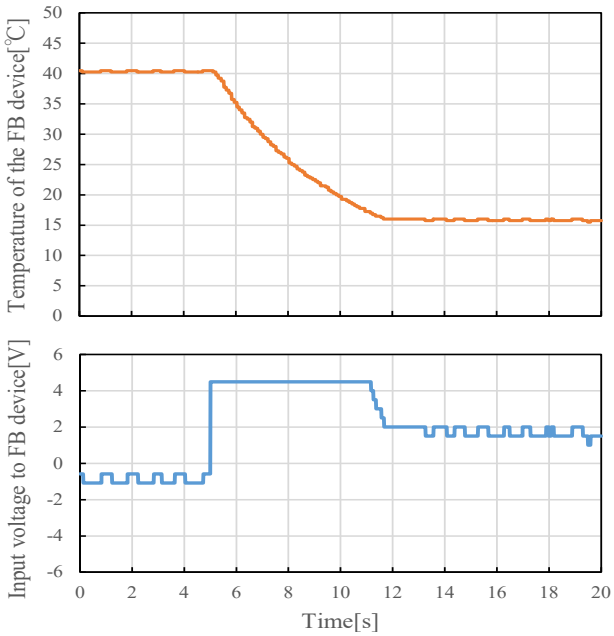


Figure 6: Temperature transition and input voltage to the FB device

As shown in Figure 6, the constructed closed loop system enabled the adjustment of the presenting temperature of the FB device according to the temperature change and also enabled a long time continuous operation of the FB device.

6. Performance Evaluation Experiments

6.1. Purpose and Equipment

In order to verify the performance of the FB device controlled in the closed loop control system, an evaluation experiment was performed.

To improve the operability of the myoelectric prosthetic hand, a new myoelectric prosthetic hand was designed and built by imitating the commercial prosthetic hand, which is shown in Figure 7. The temperature sensor was attached to the silicone finger installed on the fingertip of the index finger of the myoelectric prosthetic hand, and the pressure sensor was attached to the fingertip of the thumb of the prosthetic hand. Thus, when the prosthetic hand holds an object, temperature value of the object and grasping force of the finger are obtained. The new myoelectric prosthetic hand was used in this experiment.

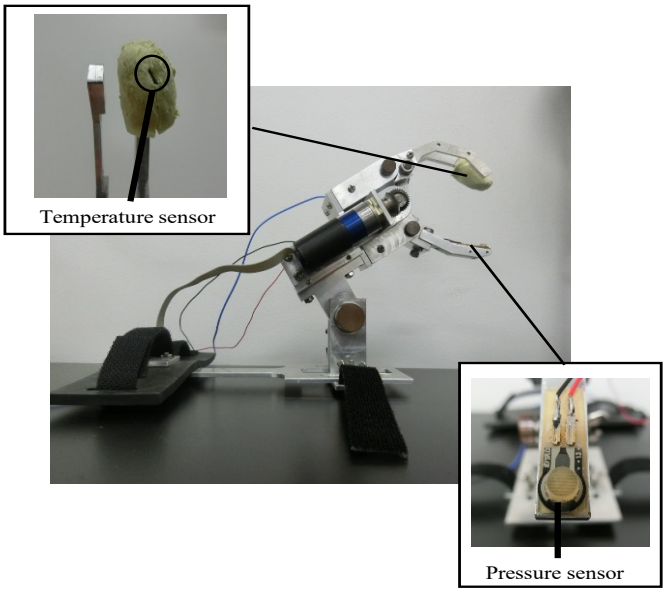


Figure 7: New myoelectric prosthetic hand

6.2. Method of Experiment

The experiment was performed in the following procedure. In this experiment, the following four variations in temperature were employed: approximately 40°C, approximately 35°C, approximately 15°C, and approximately 5°C.

1. The temperature of the object (Peltier element) is adjusted to become the selected temperature.

2. The temperature of the object T_{real} is recorded and the real value of the target temperature in the upper arm T_{t_real} is calculated through Equation (7) from T_{real}.

3. The object is touched by the fingertip of the myoelectric prosthetic hand.

4. The temperature prediction is performed and the target temperature corresponding to the predicted temperature T_t is calculated through Equation (7).

5. The temperature of the FB device is controlled by using the closed loop control system.

6. The temperature of the FB device T_{FB} and the target temperature corresponding to the predicted temperature T_t are recoded.

7. The error between the target temperature T_t and the real value T_{t_real} (prediction error $T_{t\_real} - T_t$), and the error between the target temperature T_t and the temperature of the FB device T_{FB} (control error of the FB device $T_t - T_{FB}$) are recorded for 10s.

The above procedure was repeated 5 times for each temperature.

6.3. Result

Table 4 shows the result of the performance evaluation experiment.

From the experimental result, a maximum error of 3.43°C was observed between the target temperature T_t and the real value T_{t_real}, and a maximum error of 2.86°C was observed between the target temperature T_t and the temperature of the FB device T_{FB}. In addition, the overall average of the prediction error and control error of the FB device for 10s were 1.09°C and 0.61°C, respectively. From this result, it is considered that the FB device controlled in the closed loop control system showed sufficient capability to present the temperature sensation.

6.4. Purpose of Difference Threshold Measurement Experiment

The usefulness of the FB device controlled by the closed loop system is objectively verified with a psychophysics experiment method. Namely, a temperature difference threshold of the FB Device is measured by this experiment. The difference threshold is the minimum difference between two stimulations that can be distinguished by the user. Therefore, in this study, the difference threshold means a change in the temperature that the user can recognize using the FB device.

6.5. Method of Experiment

In this experiment, the myoelectric prosthetic hand is not used, but each temperature is input directly to a computer, and then the difference threshold measurement of a single body of the FB device is conducted.

The temperature of 30°C is used as standard stimulation, and five kinds of temperature, 28°C, 29°C, 30°C, 31°C and 32°C, are used as comparative stimulation. The following describes the experimental procedure:

Table 4 Result of performance evaluation experiment

Temperature of the object T_{real} [°C]	Prediction error $T_{t\_real} - T_t$ [°C]		Control error of the FB device $T_t - T_{FB}$ [°C]	
	Ave.	Max	Ave.	Max
5.4	0.58	0.98	0.31	0.95
4.7	1.03	1.51	0.49	1.26
5.2	1.03	1.51	0.55	1.26
5.4	1.19	1.78	0.44	1.22
5.4	1.37	1.87	0.55	1.12
14.6	1.27	2.26	0.46	1.62
15	0.27	0.86	0.31	1.13
14.5	1.00	1.70	0.41	1.65
15	1.49	2.28	0.36	1.59
15.4	1.19	1.98	0.36	1.59
35	1.31	2.42	0.40	1.58
35	0.82	1.74	0.47	1.42
35.2	0.97	1.66	0.45	0.85
35.4	1.17	1.70	0.58	1.10
35.4	0.89	1.57	0.40	1.05
41.3	2.23	3.43	1.60	2.49
40.3	0.50	0.93	0.69	1.39
41.5	1.65	2.28	1.75	2.86
40.3	0.93	1.20	1.32	2.14
40.3	0.92	2.00	0.36	0.93

1. The FB Device is attached on the upper arm of the subject.

2. An experimenter inputs the standard stimulation (30°C) to the computer and standard stimulation is presented to the subject by the FB device.

3. An experimenter inputs the comparative stimulation that is randomly selected from the five kinds of temperatures to the computer and comparative stimulation is presented to the subject by the FB device.

4. The subject then answers which temperature is higher or whether the two are the same.

5. 25 sets of the work 2 ~ 4 are performed.

6. The works 2 and 3 was replaced and 25 sets of the work 2 ~ 4 are performed.

In the operations, all the comparative stimulations were used ten times in random orders. The experiment was executed for 5 healthy subjects in their 20s.

6.6. Result

Table 5 shows the result of difference threshold measurement experiment.

Table 5: Results of difference threshold measurement experiment

Temperature of comparison stimulus [°C]	Rate of the subject's answer[%]		
	Hot	Equal	Cold
28	0	14	86
29	4	36	60
30	6	82	12
31	60	34	6
32	86	10	4

When the method of three categories (large, almost equal, small) is used and when the stimulation value at which the occurrence ratio of "large" and the occurrence ratio of the others are both 50% is written as S_U, the stimulation value at which the occurrence ratio of "small" is 50% is written as S_L, the difference threshold ΔS can be obtained by the following equation:

$$\Delta S = \frac{S_U - S_L}{2} \quad (9)$$

The stimulation value S_U and the stimulation value S_L were calculated through a linear interpolation method. Table 6 shows the results.

Table 6: Calculation of difference threshold

S_U	30.81
S_L	29.58
Difference threshold ΔS	0.62

From the above calculation procedures, the difference threshold of the FB device was found to be ΔS = 0.62°C. Therefore, the experimental results suggest that the temperature change can be distinguished if the temperature is changed 0.62°C or more under the standard stimulation, namely temperature of 30°C.

7. Conclusion

In this paper, a control system of the FB device of the temperature sensation developed for myoelectric prosthetic hand users was improved. In the previous FB device controlled in the open loop control system, it was difficult to operate the FB device continuously. Also, the FB device could not adjust the temperature for the temperature change. To solve these problems, a closed loop control system was constructed for the FB device and was tested for sudden change of the temperature.

In addition, a new myoelectric prosthetic hand was built to improve the operability of the myoelectric prosthetic hand. Then, a performance evaluation experiment was carried out using the new myoelectric prosthetic hand to verify the performance of the FB device controlled in the closed loop control system. The overall average of the prediction error and control error of the FB device were 1.09°C and 0.61°C, respectively. Thus, the sufficient capability of the FB device controlled in the closed loop control system was shown.

Furthermore, psychophysics tests were conducted to measure the difference threshold of the feedback device. As a result, the difference threshold of the feedback device of temperature sensation, 0.62°C was obtained.

Conflict of Interest

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] H. Schmidt, "The importance of information feedback in prostheses for the upper limbs", Prosthetics and Orthotics International, I, 21-24, 1977.

[2] T. Morita，T. Kikuchi, and C. Ishii, "Development of Sensory Feedback Device for Myoelectric Prosthetic Hand to Provide Hardness of Objects to Users", Journal of Robotics and Mechatronics, **28**(3), 361-370, 2016.

[3] R. Okuno, M. Yoshid and K. Akazawa, "Basic study of infrasonic vibrotactile stimulus for prosthetic hand", in Medical & Biological engineering & Computing, Vol.35 Suppl., 661, World Congress on Medical Physics and Biomedical Engineering, 1997.

[4] A. Otsuka, T. Tsuji, S. Oki and M. Sakawa, "Development of a feed-back-system for thermal sensation", Bulletin of the Japanese Society of Prosthetic and Orthotic Education, Research and Development, **17**(2), 135-138, 2001. (in Japanese)

[5] H. Morimitsu and S. Katsura, "Heat inflow control of Peltier device based on heat inflow observer", Proceedings of SICE Annual Conference, 996–1001, 2010.

[6] Y. Ueda，and C. Ishii, "Development of a Feedback device of Temperature Sensation for a Myoelectric Prosthetic Hand by Using Peltier element", Proceedings of the 2016 International Conference on Advanced Mechatronic Systems, 488-493, 2016.

[7] A. Harada，T. Nakakuni，M. Hikita and C. Ishii, "Robot Finger Design for Myoelectric Prosthetic Hand and Recognition of Finger Motions via Surface EMG", Proceedings of the 2010 IEEE International Conference on Automation and Logistics, 273-278, 2010.

[8] T. Kikuchi，and C. Ishii, "Identification of Finger Operation using Support Vector Machine and Control of Myoelectric Prosthetic Hand based on Integrated Electromyogram", Proceedings of the 2014 IEEE International Conference on Robotics and Biomimetics, 1272-1277, 2014.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 48-59 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Deep venous thrombus characterization: ultrasonography, elastography and scattering operator

Thibaud Berthomier[1,2,3], Ali Mansour[1], Luc Bressollette[2], Frédéric Le Roy[1], Dominique Mottier[2,3]

[1]*LABSTICC (UMR 6285), ENSTA Bretagne, 2 rue François Verny, 29806 Brest Cedex, France*

[2] *EA 3878 (GETBO), CIC Inserm 1412, CHU de la Cavale Blanche, Bld Tanguy Prigent, 29609 Brest Cedex, France*

[3]*F-CRIN INNOVTE, CIC Inserm 1408, CHU Saint-Etienne, 42055 Cedex 2, France*

ARTICLE INFO

Article history:
Received: 05 March, 2017
Accepted: 20 March, 2017
Online: 05 April, 2017

Keywords:
Deep Venous Thrombosis
Thrombus
Ultrasonography
Elastography
Image preprocessing
Scattering Operator
Wavelets
Feature Extraction

ABSTRACT

A thrombus or a blood clot is the result of blood coagulation which is a natural process to prevent bleeding. An inappropriate formation of a thrombus in a deep vein is known as Deep Venous Thrombosis (DVT). The main complication of a DVT is a Pulmonary Embolism (PE) which occurs when a thrombus breaks loose and travels to the lungs. DVT, PE, or both are also known as Venous thromboembolism (VTE). It affects an estimated 300,000–600,000 Individuals just in the United States per year and can cause considerable morbidity and mortality. This multifactorial disease related to advanced age, immobility, surgery or obesity is an important public health issue. Our project is looking to link the VTE epidemiology (risk factors, patient history, PE) to the thrombus structure. To reach our goals, we are collecting ultrasonography (echogenicity) and elastography (stiffness) of human thrombus. This manuscript describes our approach to create and preprocess a database using Toshiba Aplio 500. Our approaches to characterize the thrombus structure with ultrasound images are also described. The feature extraction is made with the scattering operator. Obtained features are then reduced using Principal Component Analysis and are analyzed to evaluate our approach.

1. Introduction

This manuscript is an extension of two conference papers presented at 2016 2nd International Conference on Frontiers of Signal Processing [1, 2]. Deep Venous Thrombosis (DVT) is an outcome of an incongruous blood clot in a deep venous network. Due to the clot, the blood flow is slowed or stopped in the affected vessel. Therefore, the remaining blood finds its way back to the heart through the superficial network. This situation can oftenly occur in leg vessels and it may lead to warm, swelling, pain or edema of the affected leg. DVT is a serious disease and has to be treated as soon as possible to avoid the blood clot extent and irreparable damage to the veins. Moreover, it may also have deadly consequences when a clot fragment breaks off, travels to the lung and causes a Pulmonary Embolism (PE). Without an urgent care, a DVT has about a 50% risk of leading to a PE [3]. DVT has multifactorial causes: immobilization, surgery, age, cancers, genetic variations; and it is mostly diagnosed via ultrasound. More precisions on blood clot formation and risk factors are provided in section 2.

Our project is aiming to relate the blood clot structure to its age, its formation cause and PE risk. Ultimately that can lead us to the detection of a cancer in an early stage. Section 2 presents the imaging techniques – ultrasonography and elastography – and our acquisition system. Section 3 describes our databases get with Toshiba Aplio 500 and the preprocessing steps. Section 4 proposes a wavelet-based feature extraction algorithm, the scattering operator [4]. Finally, major energetic relevant features have been selected using Principal Component Analysis.

2. Background and Key issues

Here in after, our project on Deep Venous Thrombosis is described. Subsection 2.1 describes a blood clot and its major risk factors. Subsection 2.2 summarizes the state of art and describes

*Corresponding Author: Thibaud Berthomier, LABSTICC (UMR 6285), ENSTA Bretagne, 2 rue François Verny, 29806 Brest Cedex, France
Email: thibaud.berthomier@ensta-bretagne.org

the main imaging techniques. Subsection 2.3 outlines technical issues related to our acquisition process.

2.1. Medical Background

2.1.1. Blood circulation and coagulation

Blood flows one-way through the body [5]: the arteries (in red on Figure 1) distribute oxygenated blood from the heart to various organs and muscles. Veins (in blue on Figure 1) have thinner and elastic walls and take over from arteries to return back oxygen impoverished blood to the heart. The blood cycle is performed using: (a) Heartbeat to maintain a continuous flow, (b) Diaphragmatic or deep breathing, (c) Muscle pump system (calf muscular contraction) and (d) Venous pump of the foot which is a major factor in blood return activated by walking. Moreover, veins include a valve system to avoid blood reflux. Prolonged immobilization due to plaster cast, bed rest or long distance flights slow down blood circulation and raise the risk of a venous thrombosis.

The physiological process to keep blood within veins and stop bleeding in case of injuries is called hemostasis [6]. Hemostasis involves vascular spasm (reduction of the damaged vessel diameter) and coagulation which change the liquid blood to a gel. The mechanism of coagulation includes: activation, adhesion, and aggregation of platelets along with deposition and maturation of fibrin. When a broken vessel is cared, a process named fibrinolysis starts dissolving the clot and prevents it from growing and causing thrombosis. Hemostasis or fibrinolysis disorder is indeed one of the risk factors for venous thrombosis.

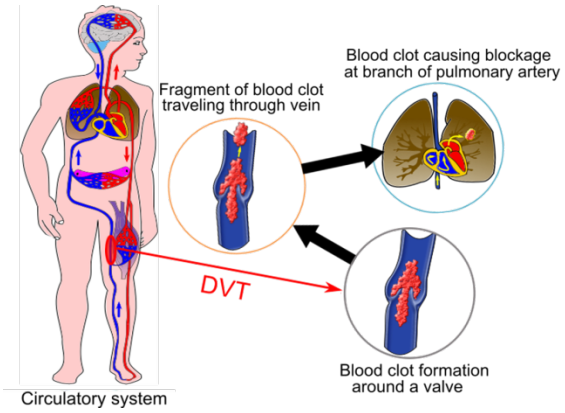


Figure 1: Deep Venous Thrombosis and its major complication, Pulmonary Embolism.

2.1.2. Thrombosis, symptoms and treatment

Thrombosis occurs when an abnormal blood clot is formed and/or when the fibrinolysis process becomes deficient. Thrombosis symptoms are an inflammation of a stamped or obstructed vessel [5], most commonly in the lower extremities. Superficial thromboses occur in the veins near the skin and usually don't cause fever, infection or swelling. On the contrary, DVT are very dangerous because a piece of the clot can break off, be carried by the blood to the lung and cause a PE, as seen in Figure 1.

The clot formation often begins in the calf veins, where the blood flow may slow down (valves, collateral veins) and extends to the knee or thigh veins [6]. The risk of a PE increases when the clot gets closer to the heart. When a deep vein is obstructed by a clot, the blood cannot flow anymore back to the heart through this vein; this creates a hypertension upstream part of the clot, therefore the blood finds another way through the superficial venous network.

Many cases are asymptomatic and resolve spontaneously. Otherwise, DVT can be revealed by three main signs: swelling of the calf or the ankle, leg warm to the touch and leg pain or tenderness. The patient could also have numbing, cramps, a sensation of heavy legs, pain on palpation and/or bluish skin discoloration. At a confirmed stage, DVT may go along with fever, edema, ulcer, tachycardia or even a complete functional impotence [6]. Asymptomatic DVT is about 5 to 20 times more common than symptomatic events [3]. Moreover, relapse is frequent and the patient can have post-thrombotic symptoms, *i.e.* long-term complications.

DVT treatment aims to prevent the clot from getting any bigger or breaking loose and causing a pulmonary embolism. In general, the anticoagulant drugs decrease the blood's ability to form a clot but they don't break up existing blood clots (to avoid a PE). The blood clot will be naturally destroyed by the body (fibrinolysis). Later on, physicians should treat the patient to reduce the risk of new thrombosis.

2.1.3. Epidemiology

Normally in our blood circulation system, there is a trade-off between coagulant and anticoagulant molecules. Many physiopathological mechanisms can unbalance the fragile trade-off and create a venous thrombosis. Virchow's triad [7] for thrombus formation is:

(a) Venous stasis,

(b) Endothelial injury,

(c) Hypercoagulable state.

Venous stasis gathers all factors responsible for the blood slow down: extended immobilization (bed rest, labor, and plaster cast), heart failure, varicose vein and venous constriction (prolonged sitting, long distance flight, cancer). For instance, statistics on venous thrombosis showed an incidence of two to four per ten thousand passengers of a flight over five hours [8]. Smoking also increases the risk as it stimulates the production of fibrinogen in the blood (and hence platelet aggregation and coagulation) and generates more viscous blood.

Endothelial injury brings factors degrading the venous walls. Therefore, an inflammation caused by an infection (related to surgery, catheter, traumatism and age) can produce a thrombosis. According to [8], statistics show that the risks, without preventive thrombosis in medicine, increase in the range of ten to twenty percent in the four to five weeks following a hospital discharge. In the same way, risks become twice larger every ten years after the age of 40.

Hypercoagulability incorporates all mechanisms reinforcing coagulation or disturbing anticoagulation. It can be acquired factors (such as pregnancy, hormone therapy, and cancer), inflammatory disorders or inherited factors (Factor V Leiden mutation, protein C or S deficiency, antithrombin deficiency). Pregnant women or people with hormonal treatment have two to five times more likely to develop a DVT [8].

2.1.4. Diagnosis and imaging techniques

An untreated DVT in the calf has about a 25% risk of extending and reaching the proximal veins (popliteal, femoral, or iliac) [3]. As well, half of untreated proximal DVT will lead to a PE. To reduce the risk of a long term complication (*e.g.* damage of venous valves and wall) and PE, it is vital to diagnose as soon as possible the DVT. However, this task seems to be difficult because the signs and symptoms associated with the DTV are not specific to this disease [6]. A blood test can exclude thromboembolic disease where the DVT probability is low. This test is based on the concentration of D-Dimers (*i.e.* fibrin degradation products) which are present in the blood when the coagulation system is activated. Nevertheless, a positive test does not always indicate thrombosis.

Imaging studies of the venous blood network are needed for diagnosis. Duplex ultrasonography is the standard imaging process. Physicians view the blood network and flow, and check on the veins' compressibility. A vein with a blood clot is relatively incompressible and is more echogenic than a free vein. Contrast venography is a gold standard but it is rarely used because its invasiveness [9]. Magnetic Resonance Angiography (MRA) and Computed Tomography Angiography (CTA) are rarely used: they are costlier and less than ultrasonography which is mostly sufficient to diagnose a DVT. In the event of a PE suspicion, CTA is the standard procedure.

2.2. Medical issues

2.2.1. Venous thrombus characterization

DVT is a multifactorial disease which can cause serious and irremediable health problems. Its early detection would limit the risk. Since this disease is often asymptomatic, prevention measures and surveillance might be used before and after any procedure or event that can increase the risk of a DVT. In case of doubt, the detection using ultrasound is relatively simple. Yet it is a lot more difficult to identify the thrombosis origins, age and to estimate the risk of PE. Thus, the objective of our project is to analyze the blood clot structure with the help of ultrasonography and elastography techniques in order to estimate the age, the origins and the risk of a PE. Elastography is an imaging technique, based on ultrasound, used with success to analyze human tissue stiffness (e.g. liver fibrosis detection).

In the literature, there are studies linking thrombosis maturity to the clot elasticity ([10, 11]), determining the impact of genetic variations on the onset of a DVT ([12, 13]) or estimating treatment efficiency ([14, 15]). But, currently, there is a little or no convincing evidence that there is a correlation between the blood clot structure and its triggering factors, hence making our research even more challenging. To characterize thrombi, *in vivo* data are collected thanks to Brest University Hospital. The following two subsections will respectively describe ultrasonography and elastography. In biomedical Engineering, these two techniques are widely studied in various research projects. In particular, the authors of [16] publish an instructive survey titled "Automated breast cancer detection and classification using ultrasound images". The recent and promising imaging technique, elastography, is mainly considered for detection and characterization of several tissues, such as liver [17], breast [18, 19] and thyroid [20, 21]. The tissue shiftiness is indeed a valuable source of information to diagnose diseases.

2.2.2. Ultrasonography

Ultrasonography (or Brightness-mode with 2-dimension imaging – B-mode 2D) is the default and the most intuitive mode: it is a 2-dimensional cross sectional view of the portions of the organs or the tissues that are intersected by the scanning plane. The shape of the field (sector, rectangular or trapezoid) depends on the used probes. The chosen probe to perform venous examination is a 5-10 MHz linear probe. On a gray scale, high reflectivity (as tissue) is colored white and for no reflection (as blood) is colored black. Two ultrasound imaging findings can point out the presence of a blood clot in a vein: the relative incompressibility of the vein and an echogenic intraluminal material [22].

In most case, the blood clot is echogenic (see Figure 2). At the beginning of a DVT episode (in the first hours or days), the clot is hypoechoic, homogeneous, elastic and it dilates the vein. On the contrary, an older clot (several weeks) becomes slowly more heterogeneous, harder and smaller. The clot echogenicity actually depends on its blood cell composition, distribution and on its fibrin mesh [23]. The incompressibility of the vein during the ultrasound test is considered as the main detection criterion of a DVT. Conversely, using the adjacent artery as a reference point, the full compression of the vein with a minimal deformation of the artery indicates an absence of thrombosis. Moreover, the vascular paresis (fixity of the venous valves) can indicate the presence of a blood clot as well.

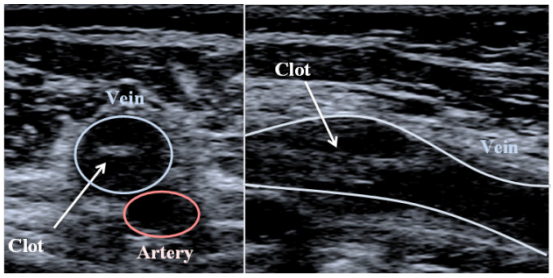


Figure 2: Transverse (left) and longitudinal (right) ultrasonographies of a blood clot in the femoral vein (Images obtained with Toshiba Aplio 500).

2.2.3. A new approach: elastography

Elastography consists in mapping the hardness, or the elasticity, of human tissues, e.g. their resistance when a mechanical force is applied on it: the harder a tissue is, the more elastic it is. A static external stress σ (in pascal Pa), applied to the surface of a solid, is linearly proportional to its fractional extension ε (non-dimensional) by the modulus of elasticity E (in Pa). This principle is named Hooke's law [24]:

$$\sigma = \varepsilon E \quad (1)$$

Human soft tissues can distort under the influence of two types of mechanical waves: compressional and shear waves. Compressional waves are also called longitudinal waves because the particle displacement is parallel to the direction of wave propagation. As well, shear waves are named transverse waves because, this time, the particle displacement is perpendicular to the direction of the wave propagation.

The velocity of these waves is directly connected to the elastic modulus E (or Young's modulus). In soft biological tissues [24], the compressional velocity is much higher (≈ 1500 m/s) than the

shear wave velocity (≈ 10 m/s), so the Young's modulus can be approximated using the following equation:

$$E \approx 3\rho c_s^2 \quad (2)$$

where ρ is the volume density (kg/m^3) and c_s the shear wave velocity (m/s). The volume density is assumed to be constant (1000 kg/m^3 which is the water density) even if it is actually different from one tissue to another (fat ≈ 950 kg/m^3, blood ≈ 1025 kg/m^3, liver ≈ 1060 kg/m^3, muscle ≈ 1070 kg/m^3 and bone between 1380 kg/m^3 and 1810 kg/m^3 [25]).

Elastography systems do not measure directly the hardness of the human tissue but estimate the velocity of the shear waves. These systems send either a mechanical or an acoustic impulse to generate the shear waves and follow their propagation using ultrasound, *i.e.* compressional waves. Ultrasonic echoes are analyzed in order to determine the velocity of the shear waves, and hence the elasticity, using equation (2).

2.2.4. Elastography and venous blood clot

In the medical context, the main application is the diagnosis of hepatic fibrosis because the liver gets harder when the fibrosis gets more severe. In old days, palpation was used to estimate the hardness. Recently, several systems can precisely measure and create a map of the organ hardness (elastography). In theory, the blood clot hardness (and hence elasticity) should increase with age thank to the loaded fibrin. The exact age of the thrombus can help determining the right treatment (anticoagulant, surgery). Several studies ([11, 26]) create an animal model to estimate the exact age of the venous thrombus *in vivo* using ultrasound elastography. On human patients, the treatment efficiency makes the experiment more difficult. However, for some patients still presenting a DVT after three months, the results show a correlation between the shear wave velocity measures and the age of the thrombus [27]. Furthermore, our project aims to characterize the thrombus structure by analyze his stiffness using elastography.

2.3. Technical issues

2.3.1. Acquisition system: Toshiba Aplio 500

Our research project is currently focused on ultrasonography and elastography to characterize the blood clot structures and emphasize the existing links to the patients' physiopathologies. Elastography is a recent imaging technique (about 10 to 20 years) but it becomes inescapable for biomedical manufacturers as it has been introduced in several systems: Toshiba (Aplio 500), Echosens (FIbroscan), Siemens (Acuson S3000), Supersonic Imagine (Aixplorer), Philips (Epiq), Hitachi Aloka (Arietta), General Electric (Ultrasonix, Logic SonixTouch), Esaote (MyLabEight).

At Brest Hospital, Toshiba Aplio 500 machines are used to diagnose DVT with ultrasonography and to practice elastography. This section describes our acquisition systems, the used procedure and relative data. Toshiba Aplio 500 delivers high resolution clinical images to help a physician making quick and reliable diagnostic [28]. As said before, the system provides both modes ultrasonography and elastography.

Three types of 2D-ultrasonography are available:

- Classic Brightness-mode (*cf.* paragraph 2.2.2 or fundamental ultrasound imaging: the probe transmits an impulsion at a certain frequency and measures echoes reflected by the different interaction with the medium. The system then interprets the echoes at the selected frequency to build the image.
- Tissue Harmonic Imaging (THI): the image is produced using higher harmonics. These harmonics are generated by the nonlinear propagation of the ultrasound beam through tissues. THI has a better lateral resolution and a better signal-to-noise ratio.
- Differential Tissue Harmonic Imaging (DTHI): Toshiba develops a new nonlinear sonographic imaging technique where a dual-frequency f_1 and f_2 is transmitted, then the system analyzes the echoes of the difference $f_2 - f_1$ and the second harmonic of f_1. DTHI can combine the advantages of fundamental and THI modes.

In the Shear Wave Elastography (SWE) mode, three display modes are available on Toshiba's system (see Figure 3):

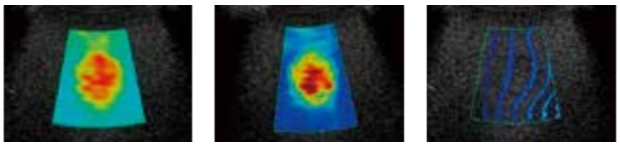
(a) Shear velocity named speed mode (m/s),

(b) Modulus of elasticity named elasticity mode (kPa),

(c) Propagation mode.

The relationship between shear velocity c_s (m/s) and modulus of elasticity E is given by equation (2).

Figure 4. If the shear waves, lines, are parallel, the reliability in this area becomes high. In contrast, distorted and non-paralleled shear waves correspond to a low reliability.



(a) Speed mode (m/s) (b) Elasticity mode (kPa) (c) Propagation mode

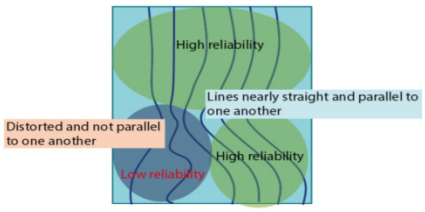


Figure 4: Reliability of the propagation display [29].

Previously, the used probe is primarily intended to observe the organs and their structures within the abdomen. That probe allowed us to use elastography which greatly helps to diagnose a liver or a pancreas fibrosis. Our ultrasound echo signals are acquired using a convex array ultrasound transducer (Toshiba PVT-375BT). With a variable carrier frequency in 1.5 to 6 MHz, the venous network can be observed in a depth of 0 up to 10 cm.

Since June 2016, the Toshiba Aplio 500 has been updated by a new probe: PWT-1005BT. This Linear Array Probe has a frequency range [5, 14] MHz and is intended, in particular, to

small-part imaging. The update and the new probe have improved the imaging quality of our acquisition. However, we ended up by having two distinct databases for feature extraction and classification. Next two subsections describe the acquisition procedure and the data

2.3.2. Data acquisition process with Toshiba Aplio500

In case of DVT suspicion, the medical expert locates the blood clot head using compression ultrasonography. During the diagnosis, the patient lies on his back. Then, the physician switches to the SWE mode and selects a Region Of Interest (ROI). A part of the tissue is deformed by the shear wave. For an existing obstacle (such as a tumor in the tissue or a clot in a vein), the shear wave velocity in that area should be different from the velocity of surrounded areas. A softer obstacle slows down the shear waves as opposed to a stiffer obstacle which makes them faster.

The system displays the elastography on the left side of the screen and the ultrasonography on the right side, see Figure 5. Later on, physicians select a shape (an ellipse or another shape) around the blood clot on the ultrasound image. Then, the system returns the shear wave velocities in this area and displays the mean and the standard deviation values at the bottom of the screen, see Figure 5. After checking the data reliability, several screenshots are made and added to our database.

Previously, and in order to reduce human error (*e.g.* probe pressure on the skin, localization), two experts make 10 measures, each patient. For each patient, a second check-up should be made three months later. In most cases, the treatment dissolves the blood clot but new measurements are performed again. The new system limits operator errors and makes elastographies more stable. Moreover, previously, the blood clot was only analyzed in a transverse view. However, a longitudinal observation is also considered, as shown in Figure 2 or Figure 5.

2.3.3. Data description

The data are exported from Aplio 500 in a format named Digital Imaging and COmmunication in Medicine (DICOM). This is a standard for handling, storing, and transmitting medical imaging information and related data. A DICOM file contains a certain number of attributes (metadata such as name, age, etc.) and also one special attribute containing the image pixel data (ultrasonography, X-rays, MRA, CTA, etc.).

In our case, the image data is a color screenshot as shown in Figure 5 which contains different information: the acquisition parameters (system, probes, and date), the patient's name (erased here for medical secrecy), the ultrasound image, the elastography, the elliptical ROI selected by an expert and the shear wave velocity in this area. Raw images can't be provided with Toshiba's system. For this reason, the ultrasonography and the elastography have to be extracted from the screenshot and then preprocessed.

The metadata provides us information about the patient (name, age, and gender), some acquisition parameters (e.g. probes, emitted frequency) and the images (size, depth). It is worth mentioning that in our data, the DICOM files do not contain any information about the shear wave velocity. Indeed, this information should be recovered from the image data. Thus, a Human Computer Interaction (HCI) is created and described in the next section.

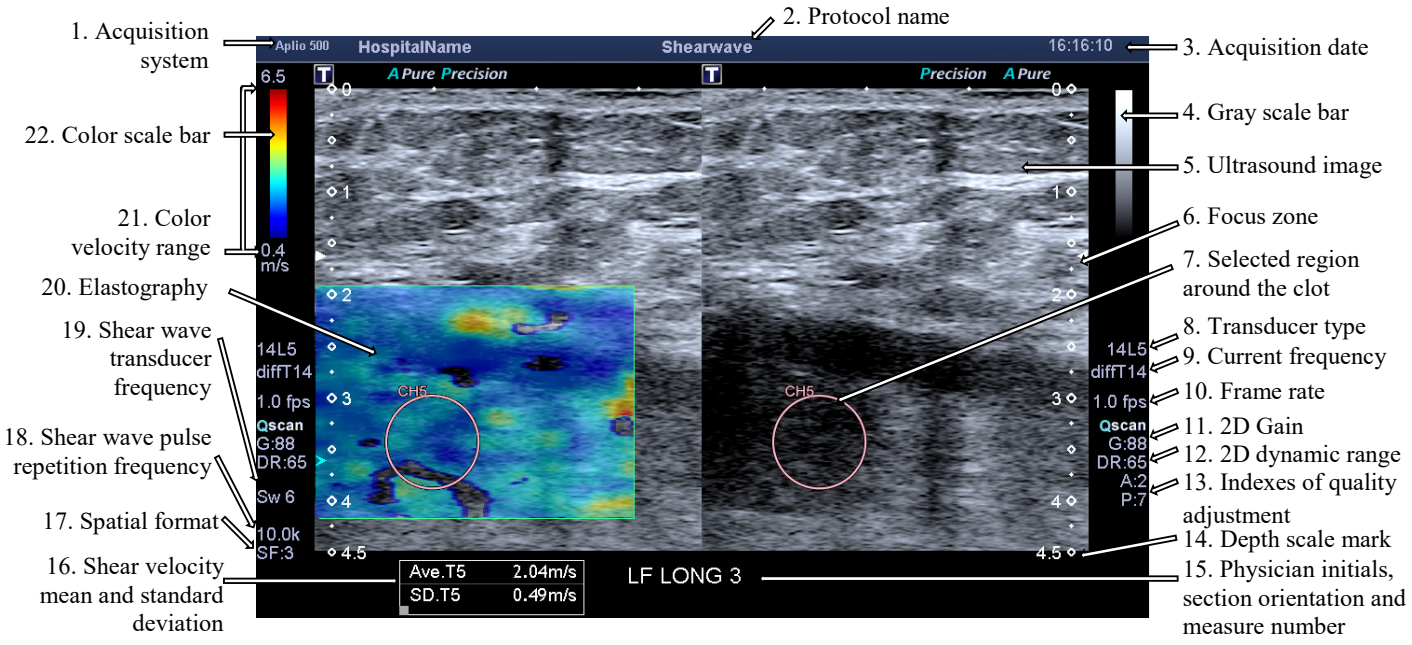


Figure 5: Image exported from Toshiba Aplio 500 and composed of the elastography and ultrasound images; the measurement parameters, the mean and the standard deviation of the shear wave velocity are displayed on the border of the image.

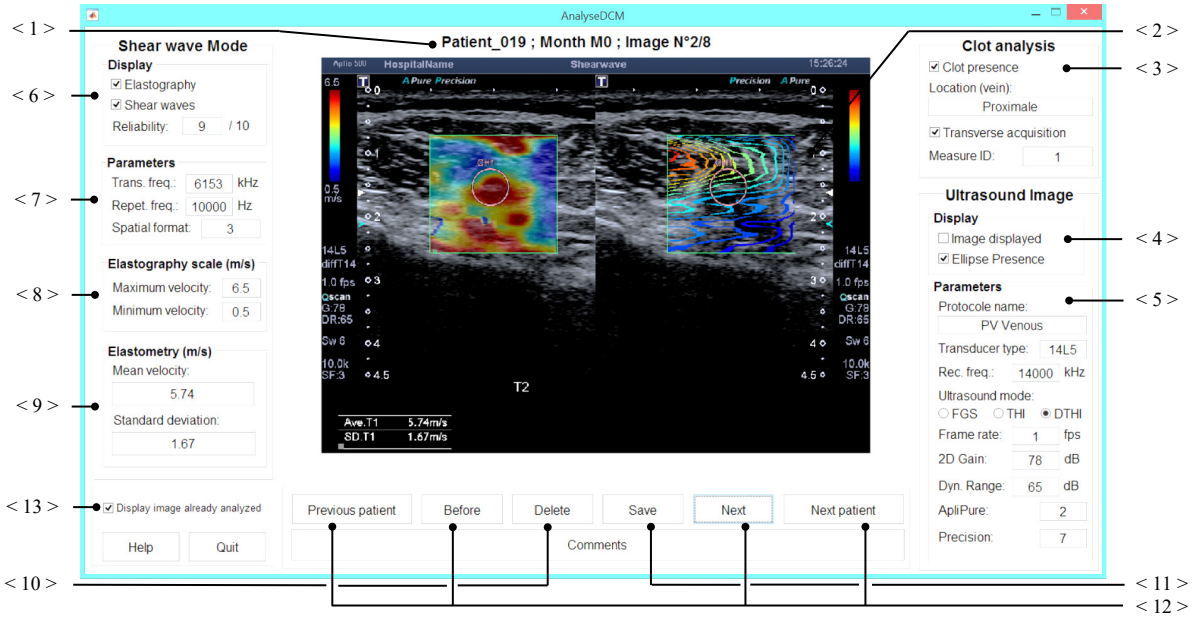


Figure 6: HCI of our application to manually extract information from the DICOM files exported from Toshiba Aplio 500.

3. Database and preprocessing

Previous sections explain the major goal of our project in the characterization of the blood clot structure in order to identify the DVT main triggering factor. To do this, two databases (before and after the updating of the system update) composed of ultrasonography and elastography in DICOM format are built. This current section describes the used process to prepare our data for further image processing (feature extraction and classification).

3.1. Data analysis using a Human Computer Interaction

The first step of the process is to read the DICOM file which is composed of metadata and an image. This image contains mostly two parts (mainly ultrasonography and elastography), the acquisition parameters and the elastometry values. Moreover, the patient physiopathologies and the presence of PE are collected in a spreadsheet enclosed with the database. To gather all those data in a single file, to identify the type of each part of the image and to recover the information unavailable in the metadata, a HCI is built. Our application browses through all DICOM files and is illustrated in Figure 6.

The DICOM image is displayed on the middle of the HCI (the < 2 > pointer of Figure 6). The patient's number, the month (approximation of clot age) and the image number are indicated just above the image (the < 1 > pointer of Figure 6). The right fields of the HCI are related to the global acquisition information and the ultrasonography. The user notices the acquisition orientation (transverse or longitudinal) and the thrombus presence in the two checkboxes (the < 3 > pointer of Figure 6). Since there can be several screenshots for one acquisition (with or without ellipse, shear wave propagation), the acquisition ID (or measure ID) should be specified too. The location of the clot is extracted from the spreadsheet. The two checkboxes indicated by the < 4 > pointer of Figure 6 have to be checked if the image contains an ultrasonography part and an ellipse around the clot. The different fields, indicated by the < 5 > pointer of Figure 6, correspond to the ultrasonography parameters (8. to 13. pointers of Figure 5) and they will be useful to evaluate the impacts of these parameters on our future simulations.

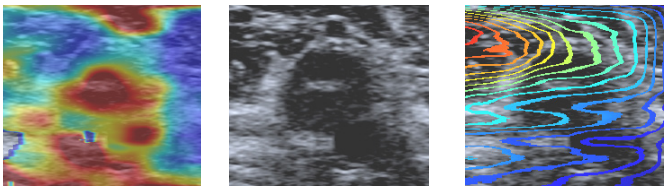
The left part of the HCI is dedicated to the shear wave elastography mode. In the box indicated by the < 6 > pointer of Figure 6, the user can indicate if the image displays the elastography or the shear wave propagation (see Figure 3). He can also estimate the reliability of the acquisition. The shear wave parameters (18. to 20. pointers of Figure 5) are entered fields indicated by the < 7 >pointer of Figure 6. The four next fields (< 8 > and < 9 > pointers of Figure 6) allow the user to enter the color range (the 19. pointer of Figure 5), the means and the standard deviation of the shear wave velocity (the 16. pointer of Figure 5).

Furthermore, the user can delete or save changes using the dedicated buttons (the < 10 > or < 11 > pointer of Figure 6). The image is then converted in the bitmap format ('.bmp'). The recovered information and the metadata are stored in a file. The navigation between the patient's DICOM files is made using the buttons 'Previous patient', 'Next patient', 'Before' and 'Next' (the < 12 > pointer of Figure 6). The already exported files can be skipped by checking the checkbox indicated by the <13> pointer of Figure 6.

3.2. Image extraction and sorting by type

DICOM image contains two parts among elastography, ultrasonography and shear waves propagation. This paragraph explains how these two parts are isolated and how the depth resolution is estimated. The boundaries of each part can be estimated using the metadata. The depth resolution is not contained in the metadata so the depth scale marks are segmented to estimate it (see tag 14. in Figure 5). These marks have a special pixel color whose value is used for the segmentation. Some marks are not filled. After filling them, an erosion removes the isolated pixels (pixel of the ultrasound image with the same color as the marks). Then, the middle mark indexes are calculated, the number of marks is counted and the depth resolution is estimated. The both extracted

images are finally cropped so that neither scale mark nor a black border is present on these images.



(a) Elastography (b) Ultrasonography (c) Propagation mode

Figure 7: Images extracted from a transverse acquisition: a DICOM file of this acquisition is display in Figure 6 (images get with Toshiba Aplio 500).

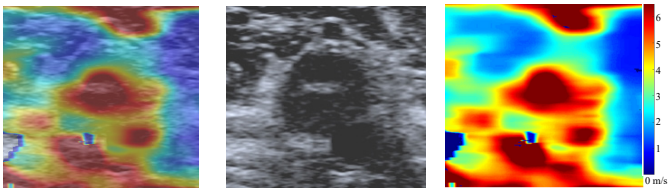
Moreover, for a single acquisition, there could be several DICOM files so our program gathers all files with the same measure ID. Each extracted image is store in the main file and its fieldname is defined by the image type and the ellipse presence. Until all file with the same ID are browsed, the images, if wanted by the user, are cropped to the elastography size. An example is shown in Figure 7.

3.3. Image preprocessing

3.3.1. Elastography

As you can see in Figure 8 (a), the elastography is a RGB map superimposed on the ultrasonography. For further analyses, the ultrasonography has to be removed and RGB colors have to be converted into velocities using the color scale displayed on main image (see Figure 6). Depending on the acquisition parameters, the transparency coefficient γ is either 0 or 0.5. The RGB map I_{RGB} can be computed from the elastography I_{Elasto} and the ultrasonography I_{Echo} with the following equation:

$$I_{RGB} = \frac{I_{Elasto} - \gamma\, I_{Echo}}{1-\gamma} \quad (3)$$



(a) I_{Elasto} (b) I_{Echo} (c) Velocity map

Figure 8: Estimation of the velocity map from the ultrasonography and the elastography (images get with Toshiba Aplio 500).

Then, each pixel of the RGB map is compared with color scale to obtain its velocity value. The final velocity map (see Figure 8 (c)) is stored in the main file.

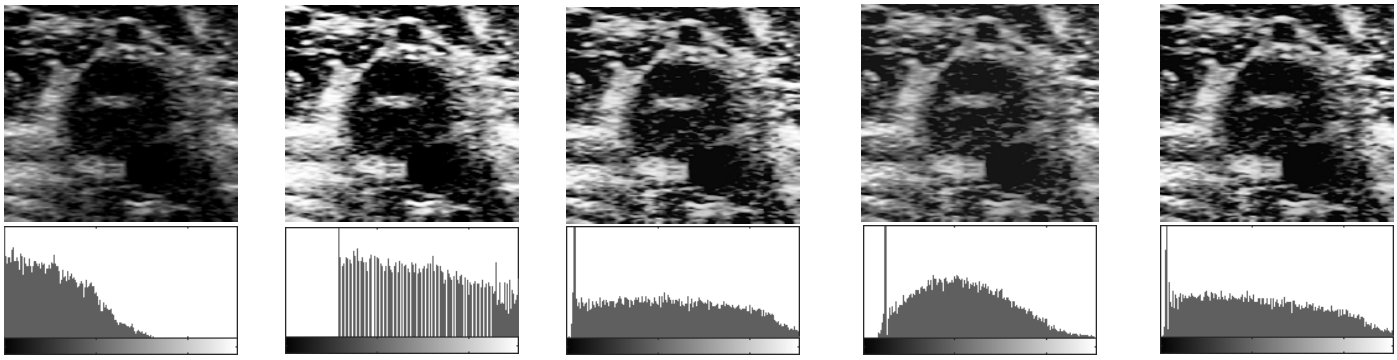
3.3.2. Ultrasonography

During the medical examination, the physician often needs to adjust some parameters to observe the blood clot. Before the system update, most ultrasonographies were acquired with the THI mode at the frequency 5MHz (setup A). After the system update, the Differential THI is preferred and there are two main setups: one with a 10MHz frequency (setup B) and another with 14MHz (setup C). Whatever the setup, the most changing parameter is still the gain which essentially affects the contrast and the brightness. To standardize our acquisition for each setup, this manuscript proposes the equalization of obtained histograms with the two following methods: ordinary Histogram Equalization (HE) and Contrast-Limited Adaptive Histogram Equalization (CLAHE) [30].

HE improves the contrast globally by adjusting intensity values. This process uses the same transformation derived from the image histogram to transform all pixels. The transformation function is proportional to the cumulative distribution function of pixel values. This works well when the distribution of intensity values is similar throughout the image. However, when the image contains regions that are significantly lighter or darker than most of the image, the contrast in those regions will not be sufficiently enhanced. In our images, the blood clot is mostly darker than the surrounding tissues so the contrast inside the blood clot may not be improved.

For improving the local contrast and enhancing the definitions of edges in each region of an image, HE can be applied over each region of the image with a transformation function derived from their neighbor pixels. This method is called Adaptive Histogram Equalization (AHE) and a major drawback concerning the overamplifying small amounts of noise in largely homogeneous regions of the image. This noise problem can be reduced by limiting contrast enhancement specifically in homogenous areas characterized by a high peak in the associated histogram because many pixels will be inside the same gray range.

With CLAHE, the slope associated with the gray level assignment function is limited by a threshold. The main disadvantage of CLAHE is that the ratio between the gray values of the original images and those of the resulted image is not



(a) Ultrasonography (b) HE (c) CLAHE - Uniform (d) CLAHE - Rayleigh (e) CLAHE - Exponential

Figure 9: Ultrasonography contrast enhancement of an image (get with Toshiba Aplio 500) (a) using respectively ordinary Histogram Equalization (HE (b) and Contrast Limited Adaptive HE (CLAHE with uniform- (c), Rayleigh- (d) and Exponential- (e) distribution functions.

constant, so CLAHE images are not suited for quantitative measurements. Figure 13 compares HE and CLAHE for a blood clot image. In this example, CLAHE methods give better results but it is hard to choose the best distribution function (uniform, Rayleigh or exponential) for CLAHE.

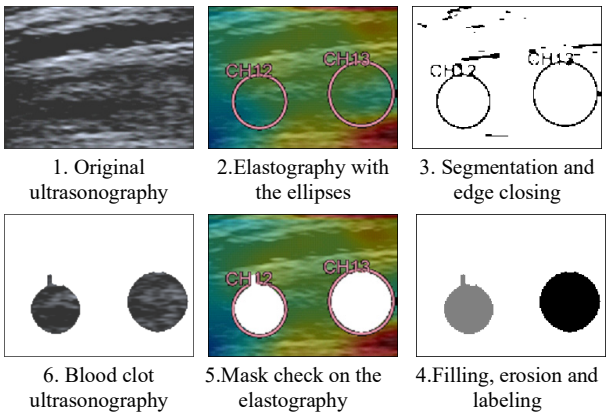
3.4. Blood clot for image extraction

To characterize the blood clot structure and identify the thrombosis triggering factors, hereinafter the scattering operator[1] is applied on the ultrasonographies containing a blood clot. To simplify and reduce the computing effort of the scattering operator, these extracted images should be squares and have a size of 2. The next paragraphs explain our strategy to create our database composed of blood clot resized square images

3.4.1. Blood clot localization

During the acquisition, the physician draws an ellipse to emphasize a blood clot. In the longitudinal observation, several ellipses can be drawn on different areas of the blood clot (see Figure 10 2.). For the structure characterization, the inside of the ellipse is analyzed using image processing. An image mask is created by exploiting the special color of the ellipses. Segmentation at this specific pixel color is made either on the ultrasonography (if available) or on the elastography. Therefore, the segmentation followed by a filing operation give the ellipse mask.

Sometimes, the ellipses are truncated (because of the previous cropping) or their edges are not closed. Thus, a validation step, either manual (display of the result and dialogbox) or automatic (based on the ellipse area), checks if the ellipses are well-segmented. Otherwise, their edges are closed and filled. This procedure can be repeated if necessary. Then, with an image processing toolbox, the ellipses are labelled (see Figure 10) and their properties estimated: area, centroid O, orientation α, major a and minor b axes (see Figure 12). Finally, this information is stored in the main acquisition file.



1. Original ultrasonography | 2.Elastography with the ellipses | 3. Segmentation and edge closing

6. Blood clot ultrasonography | 5.Mask check on the elastography | 4.Filling, erosion and labeling

Figure 10: Segmentation of the ellipses, creation of their mask and extraction of their position (Ultrasonography and elastography get with Toshiba Aplio 500).

3.4.2. Image extraction

At first, the blood clot image was identified as the biggest square inside an ellipse, as shown in Figure 11. This approach includes a rotation if the ellipse and the image have not the same orientation.

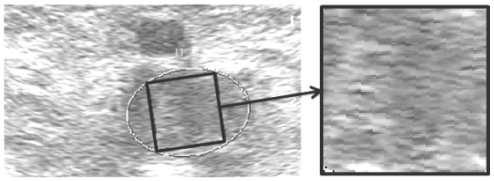


Figure 11: Extraction of the largest square inside an ellipse on an ultrasound image: the rotation distorts the image.

To avoid the artefacts generated by this rotation, this manuscript suggests to not rotate the global image and to find the biggest square in the image orthonormal system. Figure 12 and Figure 13 illustrate this idea, the different variables and the results without artefacts. The width c is computed using the following system:

$$\frac{X^2}{a^2} + \frac{Y^2}{b^2} = 1 \tag{4}$$

where $X = c(cos\alpha - sin\alpha)$ and $Y = c(cos\alpha + sin\alpha)$.

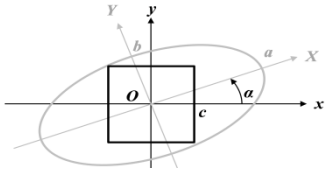


Figure 12: Extraction of the largest square inside an oriented ellipse.

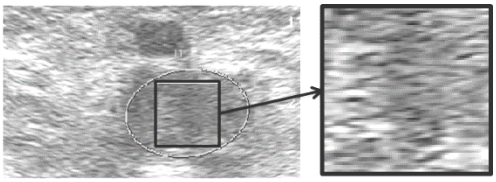


Figure 13: The extracted image is smaller with this second method but there is no artefact.

3.4.3. Image resizing

The size of the square differs according to the patient (its anatomy), the clot age, its size and its localization (vein). To compare their scattering transform, the images need to have the same size. The simulations were made with two different sizes (64x64 and 128x128 pixels) and four resizing methods: nearest neighbor interpolation [31], bicubic interpolation [31] and zero padding in the Fourier or in the cosine representation. If the goal size is relatively close of the original size, these four methods give very similar results. In the next section, the scattering operator is applied on 64x64 images resized with zero padding in cosine transform.

4. Feature extraction and classification

Previous sections describe the background of our project and the collection mechanisms of our data. This section presents our way to characterize thrombi structures with ultrasound images. The exposed simulations are made with our first database (about 200 images for 12 patients). The blood clot will be characterized

[1] The scattering operator is an algorithm based on wavelet transforms implemented in a deep convolution network [4].

with features extracted by applying an algorithm named scattering operator [4]. This technique is based on wavelet transforms and detailed in the subsection 4.1.

This project intends to correlate the images features and the Venous ThromboEmbolism (VTE) epidemiology (Risk factors, history of DVT, and presence of EP). Principal Component Analysis (PCA) is used to reduce the important size of the scattering features. Simulation results are presented in the subsection 4.2.

4.1. Feature extraction using the Scattering Operator

The challenge of automatically classify images resides in the high variability within a same class. This variability is often uninformative in the sense that it does not characterize a class change. The scattering operator aims at reducing this variability by creating a "translation invariant image representation, which is stable to deformations and preserves high frequency information for classification" [4].

4.1.1. Scattering wavelets

The scattering operator combines three three types of operations:

(a) A local average ϕ_J to recovered the low frequencies:

$$\phi_J(u) = 2^{-2J}\phi(2^{-J}u) \tag{5}$$

where 2^J is the maximum scale and u stands for the spatial position vector and ϕ is a scaling function named father wavelet.

(b) Localized band pass filters ψ_λ defined, for each scale $0 < j < J$ and an orientation θ, by the following equation:

$$\psi_\lambda(u) = 2^{-2j}\psi(2^{-j}r_\theta^{-1}u) \tag{6}$$

where, to simplify notation, $\lambda = 2^{-j}r_\theta$ and r_θ is the rotation matrix: $r_\theta = \begin{bmatrix}\cos(\theta) & -\sin(\theta)\\ \sin(\theta) & \cos(\theta)\end{bmatrix}$. The wavelet transform is stable to small deformation and invertible if the rotated and scaled wavelet filters cover the whole frequency plane (see Figure 15) so this operation recovers the information lost by the local average. In the simulation, the Morlet wavelet family is used.

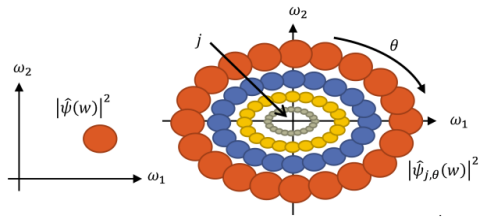


Figure 14: Frequency support of the mother wavelet on the left side and of the child wavelets on the right side: $\hat{\psi}_\lambda$ is the Fourier transform of ψ_λ.

(c) A non-linear operator to contract the space and move the energy to the low frequencies which is the modulus (see Figure 15).

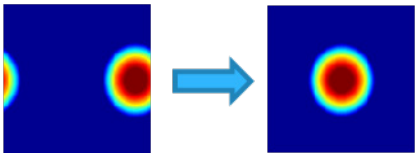


Figure 15: Frequency support (threshold at -3dB) of a localize pass band filter before (left) end after (right) the application of the modulus.

4.1.2. Scattering transform

The first step to compute the scattering transform S of an image x consists in applying the average filter ϕ_J on the image x:

$$S[\emptyset]x = x * \phi_J \tag{7}$$

To recover the high-frequency information lost by this operation, localized band pass filters ψ_λ are applied. Then, that information is extracted with a modulus and a low pass filter again:

$$S[\lambda_1]x = |x * \psi_{\lambda_1}| * \phi_J \tag{8}$$

$S[\lambda_1]x$ defines the first layer of the scattering transform S. To recover the high frequency information lost during the average operations is applied localized band pass filters $\psi_\lambda$to construct the second layer of S:

$$S[\lambda_1, \lambda_2]x = \big||x * \psi_{\lambda_1}| * \psi_{\lambda_2}\big| * \phi_J \tag{9}$$

If $p = (\lambda_1, \lambda_2, \ldots, \lambda_m)$ is a path of length m, then $S[p]x$ is an image called scattering coefficient of order m computed at the layer m of the scattering network (see Figure 16). For a large invariant representation, several layers are necessary to avoid losing crucial information. The number of rotations, scales and layers are key parameters of the scattering transform and need to be optimized.

Finally, each ultrasound image will be represented by the $L_1(\mathbb{R}^2)$ norm of each image $S[p]x$ of the scattering transform. The $L_1(\mathbb{R}^2)$ norm forms a crude signal representation which measures the scarcity of the wavelet coefficients and is defined by:

$$\|Sx\|_1 = \int_{\mathbb{R}} |Sx(u)| du \tag{10}$$

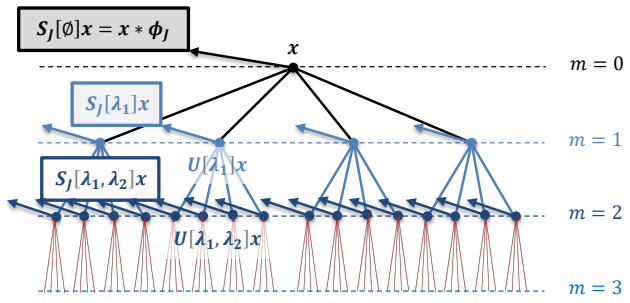


Figure 16: The scattering operator computed the first layer of wavelet coefficients modulus and its local average $S_J[\emptyset]x$ (black arrows). Then, the process is applied again on each wavelet coefficients modulus to forms the second layer of coefficients $S[\lambda_1]x$ (blue arrows). This process is repeated until the desired layer m is reached.

4.1.3. Feature reduction using Principal Component Analysis

Let P be the number of the scattering feature. Firstly, the data are centered around the origin by subtracting the mean of each variable $E(Sx_k)$ from the dataset. PCA algorithm approximates the difference $Sx_k - E(Sx_k)$ by projecting them in a linear space of lower dimension $d \ll P$. Let V_k stands for the linear space generated by the d PCA eigenvectors of Sx_k covariance matrix having the largest eigenvalues. Among all linear spaces of dimension d, V_k is the approximation space $Sx_k - E(Sx_k)$ with the smallest expected quadratic error.

Finally, PCA is equivalent to approximating Sx_k by its projection on an affine approximation space:

$$A_k = E(Sx_k) + V_k \quad (9)$$

This affine space selection is effective if $Sx_k - E(Sx_k)$ is well approximated by a projection in a low-dimensional space. The scattering parameters (number of scales, orientations and layers) and the dimension d must be adjusted so that Sx_k has a better approximation in the affine space A_k.

4.2. Simulations

4.2.1. Experimental process

Our databases are formed by three different categories: the old acquisition system, the new one with 10 MHz and 14 MHz. In this paper, the exposed simulations are made with the biggest possible database including ultrasound images get with exactly the same parameters. Indeed, the use of different presets can affect the scattering coefficients and biased our conclusions. Consequently, the test database contains about 200 images acquired with the older version of the system. Our main goal is to identify the main causes of the DVT and to evaluate the risk of a PE. Thus, the expected classes would include:

- Cancer
- Idiopathic: DVT with unknown pathogenesis of apparently spontaneous origin
- Idiopathic and PE
- Immobilization
- Pregnancy
- Surgery and PE

The scattering operator parameters (number of scales, orientations, layers and the size of the images) should be optimized. The optimization step is necessary; even though, there is no guarantee that it will lead to a satisfactory classification. We are not totally sure that the cause of the thrombosis and the presence of a PE are linked to the clot structure. In our simulations, our images are resized to 64 x 64 pixels because a power-of-2 square image reduces the calculation factor of the scattering operator. All coefficients are computed with the following parameters:

- Number of scales J: 2, 3, 4, 5, 6 and 7
- Number of orientations L: 1, 2, 4, 6 and 8.

The scattering energy has an exponential decay with respect to the order m [4]. Moreover, its energy converges to 0 as m increases and is below 1% when $m \geq 3$. Therefore, all obtained coefficients (for all J and L) are neglected when $m \geq 3$. To analyze the scattering feature for each order, each image will be represented by one vector per order (*i.e.* four vectors: orders m = 0, 1, 2 or m = 3). After that, the dimension of these four vectors is reduced using PCA. The affine approximation space will be projected in two dimensions: the two 2 best discriminant dimension among the three first principal components.

4.2.2. Results

Figure 17 shows the projection of the scattering features in two-dimension spaces. Figure 17 (a) considers coefficients at the order 0: the abscissa axis represents the first Principal Component (PC) and the ordinate axis the third PC (it is more discriminant than the second PC). The point cloud can be separate into four clusters:

- Cluster 1: some images of patient No .2 (immobilization.)
- Cluster 2: some images of patient No. 4 (idiopathic), patient No. 8 (idiopathic) and patient No. 19 (pregnancy);
- Cluster 3: patient No. 9 (surgery & EP)
- Cluster 4 (rest of the point cloud): No. 1, No. 10, No. 13, No. 15, No. 16, No. 17 and No. 18.7

However, it is difficult to link this observation with the main cause of the DVT. There is a pregnant woman in two clusters (No. 17 and No. 19). Idiopathic thromboses are also presented in the two clusters (cluster 1 and 4).

The points of patient No.2 show the impact of the acquisition conditions. Indeed, two levels can be easily distinguished. In fact, these images were taken by two different medical experts. To reduce the effect of the human factor, an articulated arm was used. We can notice as well that there are two levels for patient 19: the beginning of the DVT and three months later. The age of the blood clot seems to impact the scattering coefficients. Naturally, an old blood clot is stiffer, and thus more echogenic than a recent blood clot.

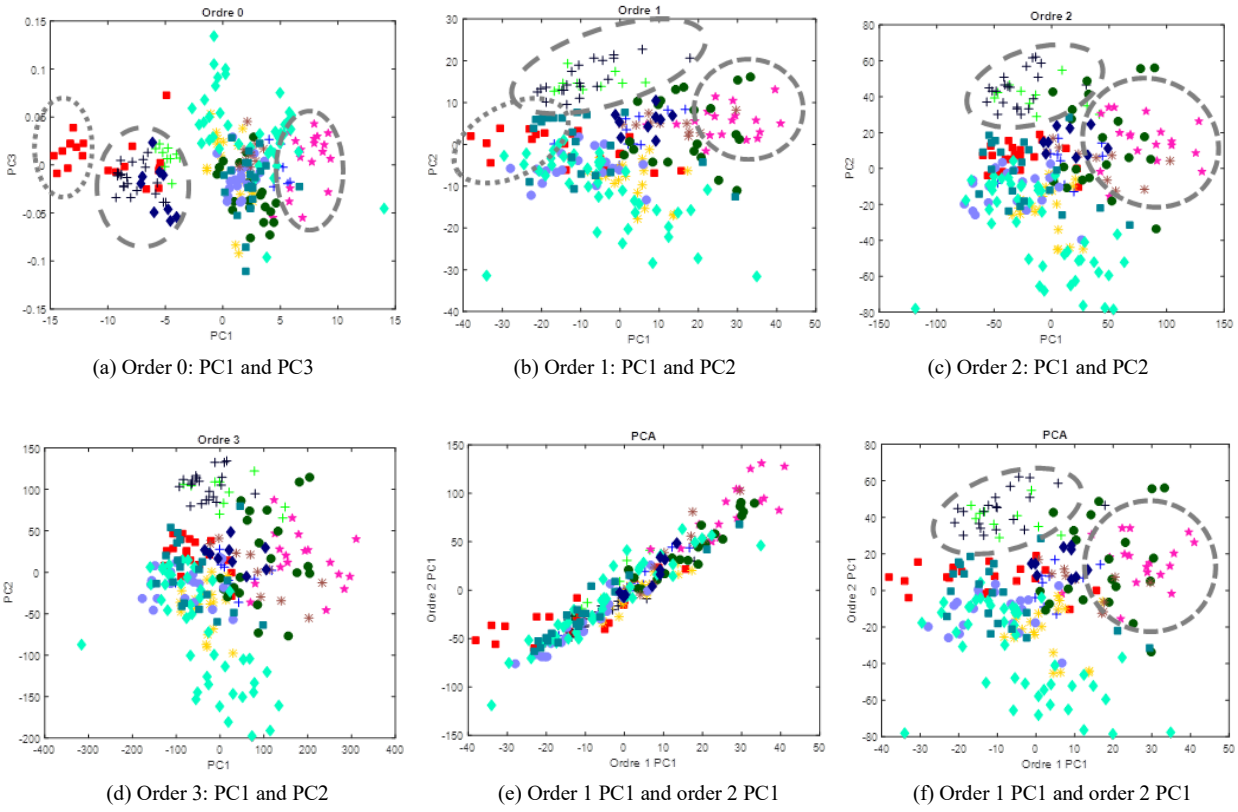
Moreover, the order 0 corresponds to low frequency information which greatly depends on the overall energy of the image. In order to characterize the clot structure, the higher scattering orders appear to be more appropriate. Nevertheless, looking at Figure 17 (b) to (d), the clusters observed at order 0 are less explicit. The scattering coefficients seem to give less information at the deeper order: order 2 (c) and order 3 (d) point clouds are very similar, so it is sufficient to explore until m=2. Finally, if orders 1 and 2 are represented on the same figure, then Figure 17 (e) and (f) show that order 2 do not add a lot of information comparing the order 1: Figure 17 (e) shows that the first PC of these orders are mostly correlated. As well, Figure 17 (c) and (f) are quite similar to Figure 17 (b).

These simulations did not reveal much about a correlation among the blood clot structure, the scattering coefficients and the VTE epidemiology. In our experimentations, we tried other metrics (*e.g.* Minkowski at different exponents), increased the size of the images (128x128 pixels) and applied the described contrast enhancement methods (HE and CLAHE). But, unfortunately, the new parameters do not improve our results.

5. Conclusion

Venous thromboembolism occurs with inappropriate blood clots, causing considerable morbidity and mortality in the world. It affects an estimated 300,000–600,000 individuals just in the

United States per year [32]. The term venous thromboembolism includes both deep vein thrombosis, clots in the deep veins and



(a) Order 0: PC1 and PC3 (b) Order 1: PC1 and PC2 (c) Order 2: PC1 and PC2

(d) Order 3: PC1 and PC2 (e) Order 1 PC1 and order 2 PC1 (f) Order 1 PC1 and order 2 PC1

Idiopathic	Idiopathic & PE	Cancer	Immobilization	Pregnancy	Surgery & PE
Patient No.1 Patient No.4 Patient No.8	Patient No.10 Patient No.16	Patient No.13 Patient No.15	Patient No.2 Patient No.18	Patient No.17 Patient No.19	Patient No.09

Figure 17: Projection of the scattering features in two-dimension spaces: (a) to (d) display the two best discriminants Principal Component (PC) at each order; (e) displays the first PC of the orders 1 and 2; (f) displays the first PC of the order 1 and the second PC of the order 2.

pulmonary embolism. Multiple factors can cause a venous thromboembolism: stasis, endothelial injury or/and hypercoagulability. Deep vein thrombosis is a common, serious and multifactorial disease. Our project aims at characterizing the blood clot structure in order to date it, explain its formation and estimate the risk of pulmonary embolism. Several techniques exist to analyze their structure: phlebography, magnetic resonance angiography, computed tomography angiography and ultrasonography. Ultrasonography and elastography have the advantage to be easy, repeatable, cheaper and non-invasive. This paper presents our approaches to generate databases using Toshiba Aplio 500. The outputs of Toshiba's system are preprocessed images exported in the DICOM format. Therefore, a human computer interaction is created to read the output files, store the acquisition parameters and extract both the ultrasonography and the elastography.

After some preprocessing steps, the scattering operator is applied to the ultrasonography in order to characterize the blood clot structure. Our simulations show that the scattering operator seems to be not suitable for the clot characterization. In recent work, we compared the obtained results with respect to different image sizes, resizing methods and with histogram equalization (ordinary HE or CLAHE). Even doing the later procedures, similar results are obtained. Currently, we are working on spectral clustering because it gave positive results in prostate and brain MRI images [27, 28]. Moreover, we are looking to increase the size of our database and to standardize the acquisition procedure. In future work, we want to compare the results obtained with the scattering operator, spectral clustering and deep learning. The later approach is recently widely used and shows promising results in breast tumor elastography classification [19].

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This research project is supported by the Brittany region (Région Bretagne, France) and French Clinical Research Infrastructure Network on Venous Thrombo-Embolism (F-CRIN INNOVTE). We would like to acknowledge Brest University Hospital, ENSTA Bretagne and Toshiba for the material support. Finally, our thanks go out to all the people involved in our project and, particularly, to L.Fréchier and B.Hermenault.

References

[1] T. Berthomier, A. Mansour, L. Bressollette, F. Le Roy, and D. Mottier, "Venous blood clot structure characterization using scattering operator," in *2016 2nd International Conference on Frontiers of Signal Processing (ICFSP)*, Warsaw, Poland, 2016, pp. 73–80.

[2] T. Berthomier, A. Mansour, L. Bressollette, F. Le Roy, and D. Mottier, "Deep Venous Thrombosis: Database creation and image preprocessing," in *2016 2nd International Conference on Frontiers of Signal Processing (ICFSP)*, Warsaw, Poland, 2016, pp. 87–92.

[3] N. L. Reyes, S. D. Grosse, and A. M. Grant, "Deep vein thombosis & Pulmonary Embolism," in *CDC Health Information for International Travel 2016*, Oxford University Press, Incorporated, 2015, pp. 132–135.

[4] J. Bruna and S. Mallat, "Invariant Scattering Convolution Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1872–1886, Aug. 2013.

[5] A. M. N. Gardner and R. H. Fox, *The Venous System in Health and Disease*. IOS Press, 2001.

[6] R. W. Colman, *Hemostasis and Thrombosis: Basic Principles and Clinical Practice*. Lippincott Williams & Wilkins, 2006.

[7] C. N. Bagot and R. Arya, "Virchow and his triad: a question of attribution," *British Journal of Haematology*, vol. 143, no. 2, pp. 180–190, Oct. 2008.

[8] P. Léger, D. Barcat, C. Boccalon, J. Guilloux, and H. Boccalon, "Thromboses veineuses des membres inférieurs et de la veine cave inférieure (article in French)," *EMC - Cardiologie-Angéiologie*, vol. 1, no. 1, pp. 80–96, Feb. 2004.

[9] J. P. Carpenter, G. A. Holland, R. A. Baum, R. S. Owen, J. T. Carpenter, and C. Cope, "Magnetic resonance venography for the detection of deep venous thrombosis: Comparison with contrast venography and duplex Doppler ultrasonography," *Journal of Vascular Surgery*, vol. 18, no. 5, pp. 734–741, Nov. 1993.

[10] J.-L. Gennisson, S. Lerouge, and G. Cloutier, "Assessment by transient elastography of the viscoelastic properties of blood during clotting," *Ultrasound in Medicine & Biology*, vol. 32, no. 10, pp. 1529–1537, Oct. 2006.

[11] B. Geier *et al.*, "Ultrasound elastography for the age determination of venous thrombi. Evaluation in an animal model of venous thrombosis," *Journal of Thrombosis and Haemostasis*, Jan. 2005.

[12] Smith NL *et al.*, "Asociation of genetic variations with nonfatal venous thrombosis in postmenopausal women," *Journal of the American Medical Association*, vol. 297, no. 5, pp. 489–498, Feb. 2007.

[13] F. Couturaud *et al.*, "Factors that predict thrombosis in relatives of patients with venous thromboembolism," *Blood Journal*, vol. 124, no. 13, pp. 2124–2130, Sep. 2014.

[14] F. Couturaud *et al.*, "Six Months vs Extended Oral Anticoagulation After a First Episode of Pulmonary Embolism: The PADIS-PE Randomized Clinical Trial," *The Journal of the American Medical Association*, vol. 314, no. 1, pp. 31–40, 2015.

[15] H. R. Büller *et al.*, "Edoxaban versus warfarin for the treatment of symptomatic venous thromboembolism," *The New England Journal of Medicine*, vol. 369, no. 15, pp. 1406–1415, Oct. 2013.

[16] H. D. Cheng, J. Shan, W. Ju, Y. Guo, and L. Zhang, "Automated breast cancer detection and classification using ultrasound images: A survey," *Pattern Recognition*, vol. 43, no. 1, pp. 299–317, Jan. 2010.

[17] I. Gatos *et al.*, "A new computer aided diagnosis system for evaluation of chronic liver disease with ultrasound shear wave elastography imaging," *Med. Phys.*, vol. 43, no. 3, pp. 1428–1436, Mar. 2016.

[18] K. Skerl, S. Vinnicombe, S. McKenna, K. Thomson, and A. Evans, "First step for computer assisted evaluation of qualitative supersonic shear wave elastography characteristics in breast tissue," in *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, 2016, pp. 481–484.

[19] Q. Zhang *et al.*, "Deep learning based classification of breast tumors with shear-wave elastography," *Ultrasonics*, vol. 72, pp. 150–157, Dec. 2016.

[20] J. Ding, H. D. Cheng, J. Huang, Y. Zhang, and C. Ning, "A novel quantitative measurement for thyroid cancer detection based on elastography," in *2011 4th International Congress on Image and Signal Processing*, 2011, vol. 4, pp. 1801–1804.

[21] F. Paparo *et al.*, "Real-Time Elastography in the Assessment of Liver Fibrosis: A Review of Qualitative and Semi-quantitative Methods for Elastogram Analysis," *Ultrasound in Medicine & Biology*, vol. 40, no. 9, pp. 1923–1933, Sep. 2014.

[22] A. Sarti and F. L. Lorini, *Echocardiography for Intensivists*. Springer Science & Business Media, 2012.

[23] S. Beuil, "Recherche d'une corrélation entre l'échographie 2D et l'Elastométrie sur les thromboses veineuses proximales profondes pour datation du thrombus," Diploma of Complementary Specialized Studies (thesis in French), UBO, Brest, 2013.

[24] M. E. Hachemi, S. Callé, and J.-P. Remeniéras, "Application of shear wave propagation to elasticity imaging of biological tissues," *Traitement du signal et cancérologie*, vol. 23, 2006.

[25] L. M. Roch, "Introduction à l'imagerie médicale (University course in French)," University of Montreal Hospital, Mar-2006.

[26] E. Mfoumou, J. Tripette, M. Blostein, and G. Cloutier, "Time-dependent hardening of blood clots quantitatively measured in vivo with shear-wave ultrasound imaging in a rabbit model of venous thrombosis," *Journal of Thrombosis and Haemostasis*, vol. 133, no. 2, pp. 265–271, Feb. 2014.

[27] L. Bressollette, P. Quehe, S. Gestin, and S. Beuil, "Imagerie du thrombus (article in French)," *Journal des Maladies Vasculaires*, vol. 40, no. 5, p. 286, Sep. 2015.

[28] Toshiba Medical Systems Corporation, Ed., "System Brochure : Aplio 500, picture perfect ultrasound." 2012.

[29] H. Lijima, "Approaches to the diagnosis of liver fibrosis," Japan, 2014.

[30] K. Zuiderveld, "Contrast limited adaptive histogram equalization," in *Graphics Gems IV*, P. S. Heckbert, Ed. San Diego, CA, USA: Academic Press Professional, Inc., 1994, pp. 474–485.

[31] H. S. Prasantha, H. L. Shashidhara, and K. N. Balasubramanya Murthy, "Image scaling comparison using universal image quality index," in *2009 International Conference on Advances in Computing, Control, and Telecommunication Technologies*, 2009, pp. 859–863.

[32] M. G. Beckman, W. C. Hooper, S. E. Critchley, and T. L. Ortel, "Venous Thromboembolism: A Public Health Concern," *American Journal of Preventive Medicine*, vol. 38, no. 4, Supplement, pp. S495–S501, Apr. 2010.

[33] G. Tartare, "Contribution à l'analyse de l'IRM dynamique pour l'aide au diagnostic du cancer de la prostate," Ph.D Thesis (in French with an abstract in English), Université du Littoral Cote d'Opale, 2014.

[34] P. Chuzel *et al.*, "Automatic clustering for MRI images, application on perfusion MRI of brain," in *2016 2nd International Conference on Frontiers of Signal Processing (ICFSP)*, 2016, pp. 63–66.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 60-66 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Improving customs' border control by creating a reference database of cargo inspection X-ray images

Selina Kolokytha*[1], Alexander Flisch[1], Thomas Lüthi[1], Mathieu Plamondon[1], Adrian Schwaninger[2, 3], Wicher Vasser[3], Diana Hardmeier[3], Marius Costin[4], Caroline Vienne[4], Frank Sukowski[5], Ulf Hassler[5], Irène Dorion[6], Najib Gadi[6], Serge Maitrejean[6], Abraham Marciano[6], Andrea Canonica[7], Eric Rochat[7], Ger Koomen[8], Micha Slegt[8]

[1]*Empa - Swiss Federal Laboratories for Materials Science and Technology, Center for X-ray Analytics, 8600 Dübendorf, Switzerland*

[2]*University of Applied Sciences and Arts Northwestern Switzerland, School of Applied Psychology,4600 Olten, Switzerland*

[3]*Center for Adaptive Security Research and Applications (CASRA), 8050 Zürich, Switzerland*

[4]*CEA, LIST, Department of Imaging & Simulation for Non-Destructive Testing, 91191 Gif-sur-Yvette, France*

[5]*Fraunhofer Institute for Integrated Circuits IIS, Development Center for X-ray Technology (EZRT), 90768 Fürth, Germany*

[6] *Smiths Heimann S.A.S. (SH), 94405 Vitry sur Seine, France*

[7] *Swiss Federal Customs Administration (FCA), 3003 Bern, Switzerland*

[8] *Dutch Customs Laboratory (DTCA), 6401 DN Heerlen, The Netherlands*

ARTICLE INFO

Article history:
Received: 16 February, 2017
Accepted: 22 March, 2017
Online: 06 April, 2017

Keywords :
X-Ray
Inspection
Cargo
Database
Imaging
Security
Training

ABSTRACT

Countries' effective and efficient border security is becoming increasingly important in today's global world of economy and trade. To ensure these, customs organisations are responsible for the prevention of illicit goods' transportation across borders, particularly upon entry. The predicament of the customs administrations will soon be aggravated by both the increase of global trade and the trend towards 100% screening. It is therefore a goal to advance inspection methods to enable successful cargo handling, a scope towards which this research was aimed at. This work was carried out as part of the project ACXIS "Automated Comparison of X-ray Images for cargo Scanning" a European research project within the seventh framework programme answering the call SEC-2012.3.4-1: « Research on Automated Comparison of X-ray Images for cargo Scanning », to improve the process with the largest impact to trade flow: the procedures of freight X-ray scanning. As such, this project was focused on to implementing a manufacturer independent reference database for X-ray images of illicit and non-illicit cargo, developing procedures and algorithms in order to uniform X-ray images of different cargo scanners, and developing a training simulator for inspection officers and a toolbox enclosing several assisted and automated identification techniques for potentially illicit cargo.

*Corresponding Author Selina Kolokytha, +41 (0) 587654184, selina.kolokytha@empa.ch

1. Introduction

Within the grand scheme of security, border control, and specifically the screening of border traffic cargo, has seen increasing focus in recent years. Given today's world of trade and the threats that exist to our societies, effective and efficient border controls are increasingly essential. The fiscal integrity and security of the movements of goods across these borders are the responsibility of customs administrations. Accordingly, they are also expected to have a positive impact on the economic competitiveness of their respective nation.

Indeed, customs administrations aim at identifying cargo that is inconsistent with its declaration, contain illicit goods, or that pose threats to society. Smuggling of counterfeited and pirated goods cause harm to a nation's wellbeing due to missed taxes that could otherwise be put to service for society. The Organization for Economic Co-operation and Development (OECD) estimates that the global international trade in counterfeits and pirated goods in 2007 may amount to USD 250 billion, which would equate to 1.95% of global trade [1]. Cigarettes lead the chart in the European Union with around 35% among all smuggled goods [2]. The increasing number of cargo transported through borders requires innovative inspection procedures. In this direction, automatic detection in cargo screening constitutes an important role.

Within the project ACXIS, cargo screening developments are made in supporting the inspection officers' work of illicit cargo detection. More specifically, ACXIS develops automated target recognition (ATR) functions to analyse X-ray images of cargo screened at border crossing points (land, air or sea). These functions can continuously self-improve through a central reference data base. This database contains X-ray images of illicit and non-illicit cargo, and it is designed to be manufacturer independent. All these database images, particularly the ones containing threats, constitute a significant tool for the training of screening officers and also for machine learning methods in automated detection algorithms [3,4].

2. Hardware Set-Up and Optimisation

In order to image illicit cargo mock-ups, laboratory X-ray scanners were used at Empa - the Swiss Federal Laboratories for Materials Science and Technology, at CEA - the French Alternative Energies and Atomic Energy Commission, and at EZRT - the German Development Center for X-ray Technology [5]. The Dutch and Swiss customs administrations used their own commercial cargo scanners, such as Smiths Heimann, Rapiscan and Nuctech systems.

The X-ray source used for the data acquisition at Empa was a portable dual energy 4 and 6 MeV linear accelerator; with a focal spot size of 2 mm (Fig. 1). Certain modifications were made on the existing high energy CT at Empa, in order to optimise the system for this project [6-8]. First, rails were placed on the linear accelerator and the detector towers to allow changing the source-to-detector distance up to 5.4 m. Then, the CT detector were replaced by a newly designed line detector consisting of 20 modules with 32 channels each, made of 30 mm deep CdWO4 scintillators with a pitch of 4.6 mm (X-Scan LCS 4.6, from Detection Technology). A 3D CAD of the new detector is shown in Fig. 2. Careful design and work was carried out in the detector assembly in reducing X-ray scattering. Additionally, new lead and tungsten source collimators were installed to shape the beam according to the detector geometry. An X-ray filter, specially designed, was fitted between the primary and secondary source collimators to flatten the intensity of the X-ray beam over the full detector width (approximately 3 m).

The laboratory scanner built at CEA consists of a linear accelerator, a line detector and a translation table (Fig. 3). The scanning geometry was chosen to match the commercial scanners, namely a horizontal translation of the object, using a vertical detector line. The linear accelerator is a Varian Linatron M9A model, delivering 5, 6 and 9 MeV with a maximal dose rate up to 30 Gy/min. The focal spot size is specified to less than 2 mm in diameter. The detector is made of 10 X-Scan LCS 4.6 modules, assembled on a straight line. A lead collimator is used in front of the active area and the backside is shielded by 5 mm of lead.

Figure 1 ACXIS laboratory set up of cargo container imaged by linear accelerator (left) and new detector (right; hidden behind container).

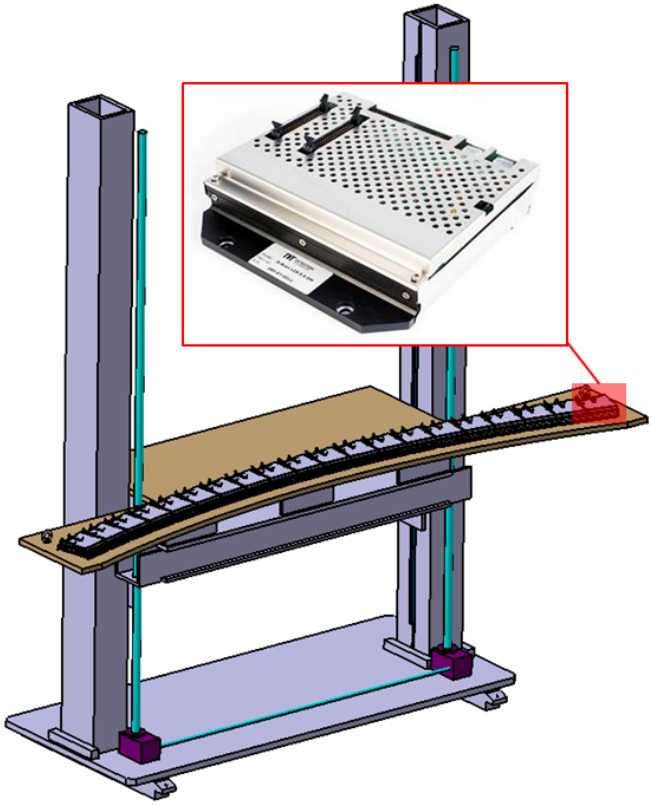


Figure 2 3D CAD of the new detector, shown without the casing, to allow visualisation of the individual detector modules, one of whose close up photo is shown. The CAD shows the detector fitted on adjustable towers.

Figure 3 Laboratory setup at CEA.

The translation table has a second plate that can be tilt with ±15°, as shown in Fig. 4. This option was used for mock-up objects for which different incident views were needed. Tilting the objects is equivalent to placing them at different heights in a full container, hence several combinations of various positions were operated more easily.

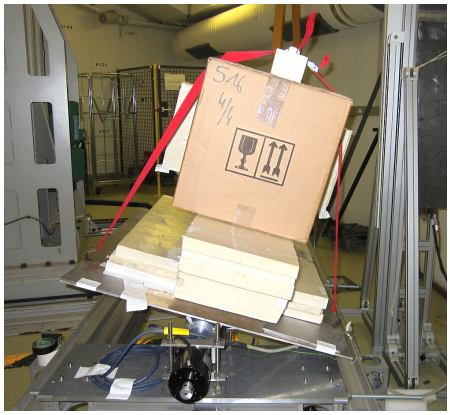


Figure 4 Translation table with tilt in the setup of CEA.

3. Reference Database: An Integration

One of ACXIS key goal is the creation of a reference database of realistic X-ray images of both non-illicit and illicit cargo. Such a database is essential for the development of training systems for customs officers and for providing assisted detection techniques of threats with the use of dedicated algorithms. To construct this database, we rely on historic images provided by customs. However, the number of threat images acquired through real detections remains limited and needs to be increased by additional threat images obtained through simulation. A very important aspect for this project is that all the pre-mentioned images must have a standard format so that images originating from different systems can be incorporated.

3.1. Creating Artificial Images of Illicit Cargo

To efficiently generate X-ray images for the database we can rely on two strategies. The first one consists in using simulation tools to generate artificial X-ray projections of illicit cargo. CEA develops the software CIVA, which integrates a module for modelling X-ray generation, interaction and detection, thus allowing the creation of realistic radiographic projections of any object that is put in the simulation scene. In this case, the shape of the object has to be described by a 3D CAD model and its material has to be specified. For multi-material objects however, each material component has to be segmented, which will result in highly complex objects in the case of a complete cargo. In practice, due to the complexity of container loadings, this approach has not been implemented. The second strategy consists in using real images of legal freights, which are widely available, and to combine these with one or several threat images. For this purpose a so-called "merging tool" has been developed and threat images have been either acquired or created by simulation.

Both approaches have advantages and disadvantages: for instance, cigarettes packs have a very specific texture that is difficult to model in the simulation software and real images have to be used in this case. On the other hand, some realistic 3D models of weapons have been collected and were used to easily generate multiple threat images with various orientations.

3.2. Threat Images

3.2.1 Acquisition from Mock-Up Scans

Based on risk analysis conducted by the customs administrations, threat objects and their locations in the container were identified. Following their recommendations, a set of mock-ups was created with real threat items and simulants, combined with common goods (Figure 5). These were scanned using laboratory systems, and some with a cargo scanner, chosen as reference.

For some illicit items the incidence of the X-ray paths through the object largely influences the produced image. Therefore mock-up X-ray scans were performed with different object tilts so that the best projection could be selected depending on the vertical position of the object in the container.

Figure 5 Photograph of a container of apples containing hidden Plastex (left) and hand grenade (right), above, and X-ray image of this threat container, below.

3.2.2 Generation from Simulation Software

The main objective of the mock-up scans was to enlarge the variety of threat images. For the same purpose, simulation tools

were also used in order to generate X-ray images from 3D models of various types of threat items (e.g. weapons).

The acquisition geometry, source spectrum and detector parameters of the high energy scanner were defined in the simulation software. 3D models of different weapons were then loaded into the software environment and multiple projections were created (Figure 6).

Figure 6 3D CAD model (left) and simulated X-ray projection (right) of a weapon.

The strength of the simulation approach is to create easily a large set of images of an object from different viewpoints. This variety of images of a same object is essential for training learning algorithms for detecting specific threats. Figure 7 displays simulated images of a small weapon, a larger weapon and a bottle that could represent a chemical product used for explosive devices for instance. Each threat has been rotated from side view to top view by steps of 30°.

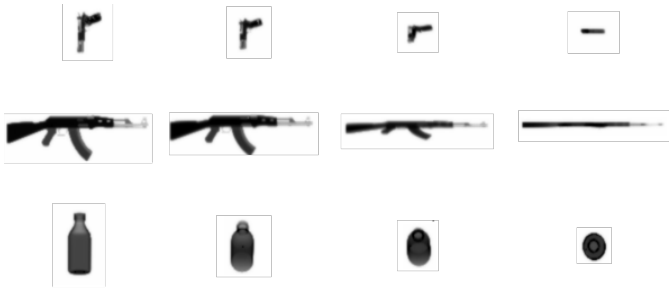


Figure 7 Simulated images of three different threats from multiple viewpoints.

3.3 Merging Threat Images with Historic Ones

The basic principle of X-ray imaging is the measurement of the X-ray photons which are attenuated when traversing matter. This process is modelled by the Beer-Lambert law, which can be approximated as follows:

$$I_{det} = I_0 \, e^{-\sum_{i=1}^{N} \mu_i \cdot x_i} \quad (1)$$

Where I_{det} is the beam intensity at the detection point after traversing N different materials, I_0 is the beam intensity at the initial point, μ_i and x_i are the linear attenuation coefficient and the thickness of each material, respectively. According to (1), the beam intensity at the detector point corresponds to the combination of the attenuation due to each material successively penetrated by the X-ray beam. It can be decomposed as:

$$I_{det} = \underbrace{I_0 \cdot e^{-\mu_1 \cdot x_1} \dots e^{-\mu_j \cdot x_j}}_{I_{bg}} \cdot \underbrace{e^{-\mu_{j+1} \cdot x_{j+1}} \dots e^{-\mu_N \cdot x_N}}_{e^{-\mu_{th} \cdot x_{th}}} \quad (2)$$

Where I_{bg} is the beam intensity attenuated by the background (non-illicit cargo without threat constituted of j materials) and the parameters μ_{th} and x_{th} represent the characteristics of the threat element (constituted of N-j materials).

Then, by injecting in (2) the additional measure of the incident beam intensity attenuated by the threat I_{th} it comes:

$$I_{det} = I_{bg} \cdot \frac{I_{th}}{I_0} \quad (3)$$

with:

$$I_{th} = I_0 \cdot e^{-\mu_{th} \cdot x_{th}} \quad (4)$$

According to (3), we can combine pixel to pixel the attenuation of the threat (threat image acquired as described in earlier subsection) to the grey value of the non-illicit cargo image to create an image of illicit cargo (Figure 8).

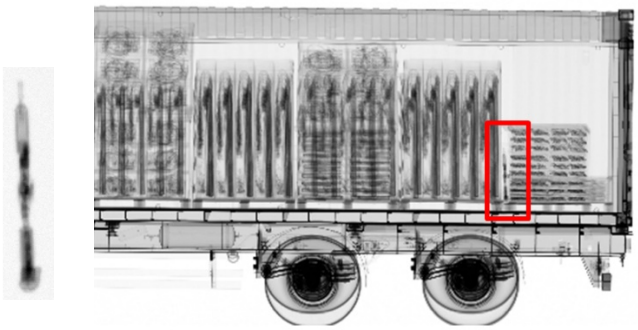


Figure 8 Simulated image of a weapon (left) inserted into a non-illicit cargo to create an artificial illicit cargo image (right).

The implemented tool allows the insertion of a same threat image multiple times at different positions in the background image and with different orientations. The major difficulty when performing the merging is in the selection of suitable and realistic spots for placing the considered threat.

A first set of illicit cargo images created artificially has been submitted to customs officers, who have assessed and validated the quality of the merging process.

3.4 Creating a Unified X-ray Image Standard

Upon visual inspection of X-ray images, the human brain can easily adapt to different systems, regardless the different parameter variations such as geometric deformations, textures, and levels of contrast or noise. This is not equally the case for automatic algorithms.

Given the large variety of X-ray scanners installed at border checkpoints, it is important that the acquired images are similar in terms of the permissioned parameters. As such, standardising these images enables an increase in the automatic detection performances. Therefore, all X-ray data can be converted into a standardised format and stored in a database, along with the results of the ATRs and useful annotations (Figure 9 centre). This standardisation enables the comparison between scans originating from different X-ray imaging systems. The database consists of set of approximately 30.000 scans as reference material, and new scans can be integrated continuously.

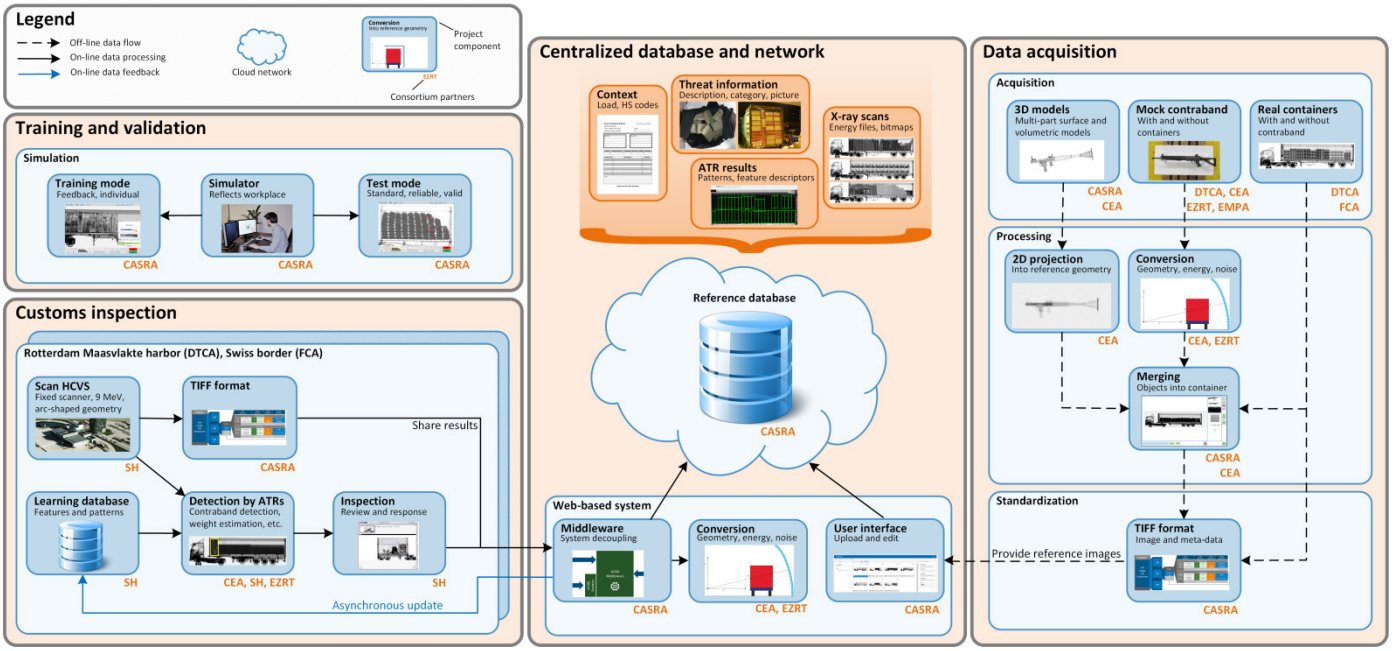


Figure 9 X-ray inspection procedure for border controls, with automated target recognition functions based on a centralised reference database.

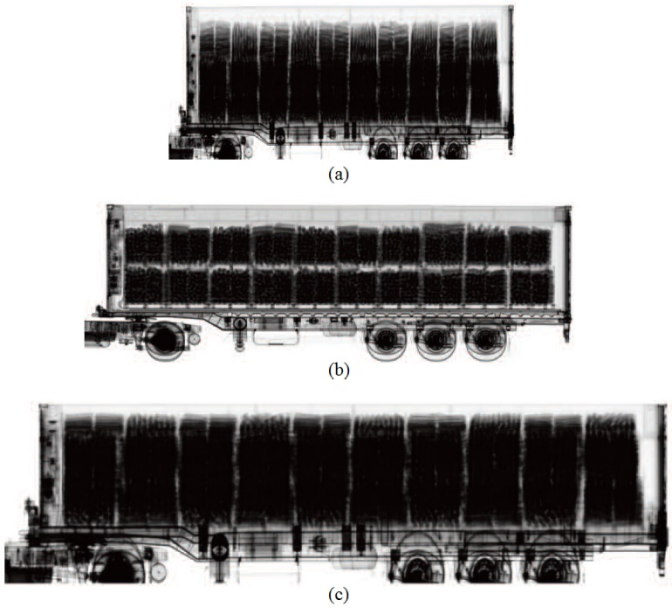


Figure 10 Real container scan with Euromax (a) and SH-HCVS (b) scanners. The Euromax scan was converted to the HCVS format (c).

Similarly, scanners with very different energy spectra such as mobile systems equipped with 2.5 MeV sources and stationary ones using 9 MeV accelerators produce very different images in terms of contrast and image quality [9]. Based on this analysis, it is not relevant to convert all scanner images to a unique reference configuration but for creating a homogeneous database, we propose to standardise images acquired with systems of energy in the range 4 MeV to 9 MeV. The strategy for the adaptation of contrast is based on the use of X-ray images of step wedges of different materials, which give a unique response curve for a given system. Using these images, we express, for a given thickness of material, the attenuation value of one system as a function of another. The equation of this curve corresponds then to the contrast conversion function, which is applied after the geometric conversion.

Finally, several noise reduction algorithms were evaluated in order to select the best one, *BM3D* [10] to reduce the noise to the level of the reference scanner, without affecting the image quality (e.g. the texture or the sharpness).

4. Customs Officers Training

information into consideration, or to ignore it when they believe that the ATR function has generated a false alarm. The waybill information is also provided to the officer, regarding the container contents.

The CBT software incorporates both training and testing modes. The training mode can be used for initial and recurrent X-ray image interpretation training using the large reference database. For each container image, the CBT asks the screening officer whether and where suspect goods are present. Following each image inspection, the officer is given feedback. When a screener correctly reports that an X-ray image contains an item that does not match the waybill or is illicit, the response is counted as a hit. If an incorrect illicit item present answer has been given, the response is categorised as a false alarm. Feedback on missing illicit items and correctly reporting that a container image is harmless is also provided. When an illicit item is present in the image, the feedback includes an additional window displaying the illicit item as a photograph and as an X-ray image. The training time and further details of the simulated scenario can also be examined. Upon completion of a training session, a summary feedback on detection performance in that session is provided. The testing mode is similar to the training mode but feedback is not provided. It can be used to measure X-ray image interpretation competency of customs officers and to conduct research for example on training progress and the value of ATRs.

5. ATR Algorithms Trained on Merged Threats

In comparison with the amount of data required to train and test ATR algorithms, a small sample of images containing illicit goods is available. Building a database derived from merging techniques is therefore crucial, especially for machine learning based algorithms where hundreds of thousands of images are often needed. Thus, a single sample can easily be projected according to different orientations and scales, increasing the variability of the dataset.

Our experience has shown that some differences may appear between the results yielded on artificially generated and real testing sets. Though, while tested on real data, ATR algorithms indicate highly satisfactory performances.

In the framework of the ACXIS project, two prominent ATR tools have been developed and trained on projected images: the Automatic Firearm Detection (Figure 12 (a)) and Reefer Unit Analysis (Figure 12 (b)) algorithms.

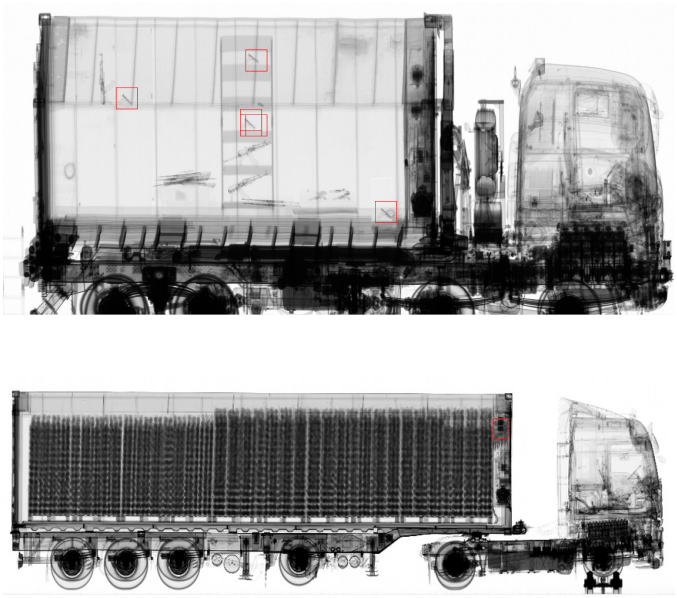


Figure 12 (a) Firearms automatically identified in the container scanned by an HCVM system (b) Narcotics automatically detected in the reefer unit, Rotterdam HCVS, 2016.

6. Conclusions

This work, which was part of the ACXIS project, developed a manufacturer independent reference database for X-ray images of illicit and non-illicit cargo, procedures and algorithms to uniform X-ray images of different cargo scanners and measurement parameters, and an automated identification method of potentially illicit cargo. Historic images of real detection scans, images of illicit cargo mock-ups as well as images of non-illicit cargo were integrated into the reference database. The resulting database was then used for the improved training and evaluation of customs' officers. The developed software incorporates ATR functions, which together with the impact of the systematic computer-based training is expected to confirm the benefits through a validation study. The foreseen results of the study will confirm that this software increases the success of border controls and reduces the required time necessary for cargo inspection.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

Project ACXIS is supported by the European Commission through the Seventh Framework Programme (FP7).

References

[1] S. Michel, M. Mendes, J. C. de Ruiter, G. C. Koomen, and A. Schwaninger, "Increasing X-ray image interpretation competency of cargo security screeners," Int. J. Ind. Ergon., vol. 44, no. 4, pp. 551–560, 2014.

[2] A. Bolfing and A. Schwaninger, "Measurement formulae for image-based factors in x-ray imagery," VICOREG Technical Report, 2007.

[3] A. Bolfing, T. Halbherr, and A. Schwaninger, How image based factors and human factors contribute to threat detection performance in X-ray aviation security screening. Springer, 2008.

[4] S. Michel, S. M. Koller, J. C. De Ruiter, R. Moerland, M. Hogervorst, and A. Schwaninger, "Computer-based training increases efficiency in X-ray image interpretation by aviation security screeners," in Security Technology, 2007 41st Annual IEEE International Carnahan Conference on, 2007, pp. 201–206.

[5] F. Sukowski, T. Fuchs, and P.-M. Keßling, "High energy X-ray CT of freight containers," in Fraunhofer Symposium Future Security. 4th Security Research Conference 2009, 2009, pp. 277–281.

[6] G. Chen, P. Bjorkholm, T. R. Fox, Z. Wilson, and X. Bonsergent, "X-Ray Cargo Inspection: Status and Trends," in Applications of Accelerators in Research and Industry: Twentieth International Conference, 2009, vol. 1099, pp. 570–573.

[7] E. A. Miller, J. A. Caggiano, R. C. Runkle, T. A. White, and A. M. Bevill, "Scatter in cargo radiography," Appl. Radiat. Isot., vol. 69, no. 3, pp. 594–603, 2011.

[8] N. Reims, T. Schoen, M. Boehnel, F. Sukowski, and M. Firsching, "Strategies for efficient scanning and reconstruction methods on very large objects with high-energy x-ray computed tomography," in SPIE Optical Engineering & Applications, 2014, pp. 921209–921209.

[9] A. Mackenzie, D. Dance, A. Workman, M. Yip, K. Wells and K. Young, "Conversion of mammographic images to appear with the noise and sharpness characteristics of a different detector and x-ray system," *Medical Physics,* vol. 39, no. 2721, 2012.

[10] K. Dabov, A. Foi, V. Katkovnik and K. Egiazarian, "Image denoising by sparse 3D transform-domain collaborative ltering," *IEEE Transactions on Image Processing,* vol. 16, no. 8, 2007.

[11] S. M. Koller, D. Hardmeier, S. Michel and A. Schwaninger, "Investigating training, transfer and viewpoint effects resulting from recurrent CBT of X-Ray image interpretation," Journal of Transportation Security, vol. 1, no. 2, pp. 81-106, 2007.

[12] T. Halbherr, A. Schwaninger, G. R. Budgell and A. Wales, "Airport Security Screener Competency: A Cross-Sectional and Longitudinal Analysis," International Journal of Aviation Psychology, vol. 23, no. 2, pp. 113-129, 2013.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 67-72 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Aviation Navigation with Use of Polarimetric Technologies

Arsen Klochan[1*], Ali Al-Ammouri[1], Viktor Romanenko[2], Vladimir Tronko[2]

[1]*Department of Electronics and Computing Techniques, Faculty of Transport and Information Technology, National Transport University, 01010, Ukraine*

[2]*Department of Avionic, Educational and Scientific Institute of Air Navigation, National Aviation University, 03058, Ukraine*

A R T I C L E I N F O

Article history:
Received: 06March, 2017
Accepted: 03 April 2017
Online: 09 April, 2017

Keywords:
Aircraft
Azimuth of polarization plane
Dielectric plate
Flight trajectory
Fresnel formulas
Mathematical modeling
Measurement method
Navigation
Navigation parameters
Polarimeter
Polarimetric navigation system

A B S T R A C T

The article deals with the problem of increasing the accuracy of determination of aircraft's navigation flight parameters. Increasing the accuracy and sensitivity of navigation parameters determination will improve the efficiency of air conveyance by optimizing the flight trajectories and reducing the flight time and simultaneously will ensure the sufficient level of air conveyance safety. The main purpose of this article is to develop a polarimetric navigation method for determining the aircraft's coordinates during flight. This method can potentially increase the accuracy and sensitivity of navigation parameters determination. The article gives a detailed analysis of the existing methods and systems for determination of aircraft's navigation parameters, their advantages and the sphere of application. Also the polarimetric navigation method and the system for implementing this method are proposed. The system envisages the use of two measurement channels and consists of two main blocks: the radiation unit and the measurement unit. The practical implementation of the proposed navigation method and system includes the development, creation and use of the polarimetric navigation system. The results of system's mathematical modeling are given and the optimum parameters of the optical channel elements of the proposed system are determined based on these results. Further research in this area envisages the development and creation of the polarimetric navigation system as well as experimental verifications of the concept. Using a high-precision measuring method, such as polarimetric measurement method, to determine the aircraft's navigation flight parameters will significantly improve the safety and efficiency of the airlines industry.

1. Introduction

Modern transportation market is characterized by the dynamic development of the airline industry. This is due to increasing mobility of population, development of national and international tourism, increasing enterprise activity and globalization process. According to ICAO statistics annual increase in the volume of regular freight and passenger traffic is observed. Thus, for example, during 2015 [3] there was the increase in the volume of passenger traffic by 7.1% and increase in the volume of freight traffic by 1.7%.The main advantages of aviation operations that distinguish it among other types of transportation are a high speed, high reliability, better safety of cargo, the ability to deliver cargo to remote areas, the shortest traffic routes and others. The vast majority of transcontinental cargo and passenger transportation over long distances is carried out by air. There are main disadvantages of aviation operations such as the high cost, high consumption of materials and high power consumption, dependence on weather conditions.

One of the priorities for increasing the efficiency and safety of air traffic is to ensure effective and overall air navigation services [4]. Improving the quality of air navigation services will optimize routes and reduce the duration of flights. This will reduce some of the above mentioned disadvantages of air transportation. Air navigation services include the air traffic service, the radio-

*Corresponding Author: Arsen Klocan, PhD student, Department of Electronics and Computing Techniques, Faculty of Transport and Information Technology, National Transport University, 46/2, Kyiv Way Street, Boryspil city, 08300, Ukraine
Phone: +3 8(068)-352-8126 | Email: varsenchuk@gmail.com

technical support, the aeronautical information service and the meteorological service. This article will focus on radio-technical support aspects of air navigation services.

Radio-technical support should be understood to mean a set of measures aimed at the timely formation of information about air situation and its issuance to aviation's control point and aircraft's crews to provide navigation, takeoff and landing of aircraft [5]. Radio-technical support includes providing communication, navigation and surveillance. This article will focus on providing navigation of the aircraft's flight.

Air navigation (Aeronavigation) is the realm of science about the methods and means of driving the aircraft and determination of its movement parameters during the flight within the atmosphere. The aircraft's flight includes several main stages, such as, takeoff, climb, route flight, landing approach, landing and others. During the flight, the following navigational tasks are performed: programming of flight trajectory; determining the coordinates of the aircraft in space; determining the vectors of air speed, ground speed and wind speed; determining the flight time to point of call, turning points; determining maneuver in the horizontal and vertical planes during flight, approach, landing and other tasks. Accurate and timely determination of aircraft navigation parameters at each stage is a prerequisite for safe and efficient air transportation.

Aircraft's navigation parameters include parameters and their derivatives, which describe orientation, location and movement of aircraft in space and around center of mass. There are the main navigational parameters defined during the flight such as coordinates (geographical, polar), height (absolute altitude, height above ground, true height), heading (magnetic, true, great-circle), speed (indicated, airspeed, ground), attitude (roll, pitch and yaw angles) and others.

One of the promising directions to improve the efficiency of air transportation is the concept of free flight. This concept involves transferring operating choice of the route trajectory, flight speed and profile to aircraft crews [6].This will allow using airspace more rationally and flexibly. Applying this concept can increase efficiency and provide annual savings of several billion dollars by reducing fuel consumption and flight time. Applying of this concept involves the compacting of the flow of air transport in airspace. This, in turn, requires improving the accuracy of determination of the coordinates and position of aircraft in air space, as well as providing early warning about possible conflicts to ensure the necessary level of safety.

2. Problem Statement

The dynamic development of the airline industry, the air traffic volume growth and implementation of air traffic management concepts «Free Flight» lead to the necessity for compacting the flow of air transport in airspace and lead to potential reduction in air transportation safety because of aircraft collision at the critical density of aircraft in the air space. Improving safety with air transport flow compacting is possible by increasing the accuracy of determination of the aircraft's coordinates and by predicting aircraft motion trajectory for the near future. The analysis of the motion trajectories of aircraft which are in locally restricted space will allow providing air traffic regulation and safe deconfliction of aircraft in the airspace. Thus, there is a need to develop new or improve existing methods for determining the coordinates of the aircraft to ensure improvement of measurement accuracy. It is critical to improve measurement accuracy of aircraft coordinates in locally restricted space: in the zone of the airport, around turning point, near the intersection zone of air-route and in other zones of airspace with a high density of aircraft.

3. Review

To determine navigation parameters, different technical means which can be divided into autonomous and non-autonomous systems are used. The non-autonomous navigation systems define navigation parameters by using information which coming on board from outside. The autonomous navigation systems are characterized by the use of onboard equipment for navigation parameters determination.

All navigation technique depending on their physical principles of action can be divided into 5 main groups: geotechnical, inertial, radio, astronomical, lighting. Each of these groups has certain operating peculiarities and is used to define aircraft's leg.

Geotechnical navigation means are based on the different physical properties of the Earth and its atmosphere, which include the gravity force, Earth's magnetism and the earth magnetic field, terrain, pressure change with flight altitude changing, air velocity, thermal properties of the Earth and others. This group of navigation technical means refers to autonomous navigation systems and is characterized by low accuracy and sensitivity. This group includes magnetic and gyromagnetic compasses, magnetometers, gravitometers, a barometric altimeter, an air speed meter, correlation-extreme navigation systems and others. This group of navigation technical means was widely used at the beginning of aviation development, and now they are used as emergency devices.

Radio navigation means are based on the physical properties of artificial electromagnetic waves that propagate in space. This group of means refers to non-autonomous navigation systems: the source of radio waves (radio beacon) is needed for their work. This group of means is characterized by high precision work in any weather conditions, at any time of the day, and at different distances. This group includes a short range radio navigation system, a long range radio navigation system, the VOR system, the DME system, a radio altimeter and others. Radio navigation means are widely used in modern aircraft in different stages of flight, to perform various tasks.

Astronomical navigation means are based on the use of objective laws of relative position and the Earth's motion relative to the Sun, the Moon and other celestial bodies. This group of navigation technical means refers to non-autonomous navigation systems: visibility of celestial bodies is needed for their work. This group of means is characterized by high accuracy and is used exclusively for navigation tasks. Equipment of astronomical navigation means is used as an additional information source of aircraft position during flight over the ocean and in the North and South Poles. This group includes astronomical compasses, sextants and others.

Inertial navigation means are based on measuring of angle rate and linear accelerations caused by the action of non-gravitational force (engine thrust, atmospheric effect, lift component, control force and others). This group of means refers to autonomous navigation systems. Modern inertial means are characterized by relatively high accuracy and sensitivity, the accumulation of errors over time, the relatively small weight and size. This group includes fiber optic gyros, ring laser gyros, MEMS-gyros, attitude and

heading reference systems, inertial navigation systems and others. This type of navigation means is widely used in modern aircraft in different stages of flight, to perform various tasks.

Lighting and visual navigation means are intended to marking of visual cues and provide ground-based aircraft navigation. This group of means refers to non-autonomous navigation systems: for their work is needed earth-based system for marking of visual cues. Means of this group is mainly used at leg of takeoff and landing. This group includes light beacons, searchlights, laser instrument landing systems of direct vision, laser instrument landing systems on the dispersion effect and others.

The current stage of aerotechnics development is characterized by integrated utilization of navigation technical means with different physical principles. It permits to reduce the measurement error of navigation parameters significantly and reduce the adjustment time of navigation system.

Inertial and radio navigation systems and methods are widely used to determine navigation parameters during end route flight. Inertial navigation systems allow obtaining information about linear and angle accelerations on aircraft's board. Onboard system determines the angle of turn and covered distance in some system of coordinates, as well as calculates aircraft's navigation parameters, such as coordinates, heading, attitude, altitude, speed and others. Radio navigation systems provide determination of aircraft's coordinates by measuring distance and azimuth to radio beacon. The holistic use of inertial and radio navigation systems permit to increase the accuracy of determining the navigation parameters.

Consider radio navigation systems in more detail. . According to the measured navigation parameter radio navigation systems are divided into the azimuth-measurement navigation system, range-measurement navigation system, rho-theta navigation system, differential-range measurement navigation system and others.

The systems which provide the navigational information about the distance to the radio beacon on board of the aircraft refer to the range measurement navigation systems. Using this type of navigation systems allows determining aircraft coordinates by the distance to three DME-beacons or the distance to two DME-beacon and altitude.

Azimuth-measurement radio navigation systems are the systems which provide the navigational information about relative angular position of aircraft and radio beacon on board of the aircraft. Using this type of navigation system allows determining aircraft coordinates by the azimuths of two VOR beacons and altitude.

The systems that provide the navigational information about relative angular position of the aircraft and radio sources and about the distance to the radio beacon on board of the aircraft are defined as rho-theta navigation system. Using this type of navigation system allows determining aircraft coordinates by the azimuth and distance to radio beacon and altitude.

The differential-range measurement navigation systems are the systems that provide on board of the aircraft the navigation information about the location of aircraft by determining the difference of the distances to several radio navigation points. To determine the coordinates of the aircraft, it is necessary to have three (four) independent differences of the distances to two (three) beacons.

So to perform navigation tasks during the flight route the most widely used radio and inertial navigation means. The radio navigation means characterized by high precision work in any weather conditions, at any time of the day, and at different distances, but needs availability of radiowaves source - radio beacon. The inertial navigation means are autonomous and characterized by relatively high accuracy and sensitivity, the relatively small weight and size, but they accumulate the errors over time. Integrated utilization of inertial and radio navigation systems permit to increase the accuracy of aircraft's navigation parameters determining.

4. Problem Solution

Optical measurement methods are widely used in various fields of science and technology. Application scope of optical measurement methods has expanded significantly in recent years. This is due to their advantages. Optical measurement methods have the following advantages: noncontact measurement, non-destructive testing, high speed of measurements, high measurement accuracy and sensitivity, and others. One of the application directions of optical measurement methods for determining the aircraft's coordinates is the use of the optical gyroscopes: fiber optic gyroscopes and ring laser gyroscopes. Their operation principle is based on the Sagnac effect. This effect consists in the change of the phase difference of two oppositely directed light waves, which propagate in a closed loop at its turn around a normal axis. Optical gyroscopes are autonomous means and refer to inertial navigation means. They are characterized by high accuracy, sensitivity and stability, relatively small gyro drift - 0,001 degrees per hour. The disadvantages of ring laser gyroscopes include high power, the complexity of their production technology, relatively large mass and dimensions. The disadvantages of fiber optic gyros include inverse scattering, effect of temperature, vibration, magnetic and electric fields on the quality of their work.

One of the most sensitive optical measurement methods is polarimetric measurement method. It allows measuring the azimuth plane polarization with accuracy 0,0005° [7]. Polarimetric measurement methods involve measuring the polarization properties of radiation after interaction of polarized radiation with the objects: the degree of polarization, azimuth plane of polarization, ellipticity and others. The use of polarimetric measurement methods to solve navigation problems is possible through the use of an isotropic dielectric plane-parallel plate in the optical channel of measurement. The dielectric plate rotates the plane of polarization of the incident radiation depending on his angle of incidence.

This paper is an extension of work originally presented in IEEE 4th International conference « Methods and Systems of Navigation and Motion Control (MSNMC) » [8]. This article continues studying the issue of using polarimetry in aircraft navigation systems and is aimed at a wider and more detailed examination of the polarimetric technology possibility to solve aviation navigation tasks. Polarimetric techniques are understood to mean the totality processes of collection, accumulation, processing, transmission, storage and display of information. They use polarimetric methods and devices to obtain basic information about the object. This article proposes the polarimetric navigation method and system, which implements this method. The operation principle of the

method is based on Fresnel's formulas. The proposed method and system are based on the use of physical properties of artificial electromagnetic waves in the optical band, which propagate in space. The polarimetric navigation method and system are classified as non-autonomous and in accordance with the classification above they can be attributed to radio navigation means.

Rotating the polarization plane of polarized radiation after falling on isotropic planar dielectric plate with weakly absorbing material is explained by various transmittance and reflectance for p- and s-polarized beam components, which are described by Fresnel's formulas. Studies, presented in [1] and [2], confirm this statement and show that at falling of linearly polarized beam on the dielectric plate the reflected and refracted beams will also be linearly polarized, and the angle of polarization plane will depend on the angle of incidence.

The values of azimuth of polarization plane for the reflected and refracted beam that has passed through the two faces of the plate can be determined by the following formulas:

$$\varphi_r = arctg\left(-\frac{\cos(i-r)}{\cos(i+r)}\cdot tg\varphi_e\right),$$
$$\varphi_d = arctg\left(\cos^2(i-r)\cdot tg\varphi_e\right).$$

where φ_r – azimuth of polarization plane of the reflected beam,
φ_e – azimuth of polarization plane of the incident beam,
φ_d – azimuth of polarization plane of the refracted beam,
i – angle of incidence,
r – angle of refraction,

Thus, when polarized radiation falls on the isotropic dielectric plate, the polarization plane of the refracted beam, reflected beam and beam passing through the two faces of the plate rotates. The azimuth rotation of polarization plate depends on the incidence angle of radiation and the azimuth of polarization plane of the incident beam.

The use of polarimetric technologies to solve the aeronavigation tasks is as follows. The emitter, which is located on navigation beacon, emits linearly polarized radiation with a certain dispersion aperture in the vertical plane and the minimum dispersion aperture in the horizontal plane. The polarization plane of radiation rotates simultaneously with the rotation of radiation diagram. The meter located on board of the aircraft receives radiation from the navigation beacon, measures the polarization plane azimuth of the beam that has passed through the two faces of the plate and determine the angle of incidence and polarization plane azimuth of the incident beam. This polarization plane azimuth of the incident beam is directly associated/ connected with the navigation beacon bearing and angle of incidence depends on the relative navigation beacon bearing and distance to it, as well as pitch angle, yaw angle and altitude of the aircraft.

Consider the block diagram of the optical measurement channel shown in Fig. 1.

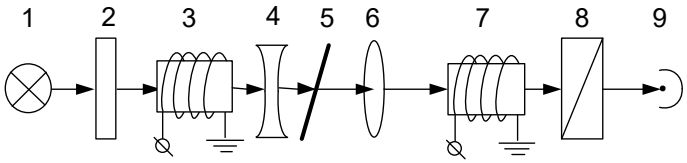


Figure 1. Block diagram of the optical channel

Radiation source 1 is intended to emit unpolarized light. The optical filter 2 is designed to filter radiation in order to highlight a certain wavelength. Faraday cell 3 is designed to polarize radiation and to rotate its plane of polarization according to the law of diagram rotation of beacon radiation. Scattering lens 4 is designed to dissipate polarized radiation with defined aperture dispersion. The dielectric plate 5 is designed to rotate the plane of polarization depending on the angle of incidence and polarization plane azimuth of the incident beam. It provides primary information about angle of incidence and polarization plane azimuth. Focusing lens 6 is designed to focus the polarized beam after dielectric plate on modulator 7. Analyzer 8 is designed to determine the plane azimuth of polarization radiation, which has passed through the dielectric plate. Photodetector 9 is designed to convert the intensity of radiation incident on it into an electrical signal. The resulted electrical signal is sent to the measurement unit for determining of polarization plane azimuth and incidence angle of radiation.

Accuracy measurement is defined by the relation of a signal to noise. The relation of a signal to noise for photodetector has the following appearance [7]:

$$\frac{S}{N} = \frac{U_s^2}{U_{TH}^2 + U_{SH}^2} = A_1(k_1+k_2)^2\cdot\Delta^2\cdot\frac{4\cdot P^2\cdot\sin^2 2\alpha}{\frac{U_{TH}^2}{A_2(k_1+k_2)^2}+1-P\cdot\cos 2\alpha}.$$

where S is signal; N is noise; U_S, U_{TH}, U_{SH} are voltages generated by signal, thermal and shot noises, respectively; A1, A2 are constants depending on the properties of the photodetector; k1, k2 are principal transmittances of polarizing prisms; Δ is system sensitivity; P is polarization degree of the light in the optical channel; α is the angular vibration amplitude of polarization plane, changing according to the periodic law: $\alpha = \alpha_0\Phi(t)$.

Using the polarimetric techniques to solve aeronavigation problems anticipates the use of polarimetric navigation system, which consists of two main units: radiation unit and measurement unit. Radiation unit is set on the beacon and designed to emit polarized beam with fixed value of polarization plane azimuth. At that the polarization plane azimuth depends on the direction of radiation. Measurement unit is set on the onboard and designed to measure the polarization plane azimuth of the passed beam and calculate navigation beacon bearing. Radiation unit has two channels and provides radiation of two linearly polarized beams with different wavelengths and polarization plane azimuths. Dependence of the polarization plane azimuth in the first and second channels (ϕ_{e1}, ϕ_{e2}) on radiation azimuth (θ) of the navigational beacon is shown in Fig. 2. Radiation unit consists of two identical channels, each of which is configured to corresponding channel of radiation unit.

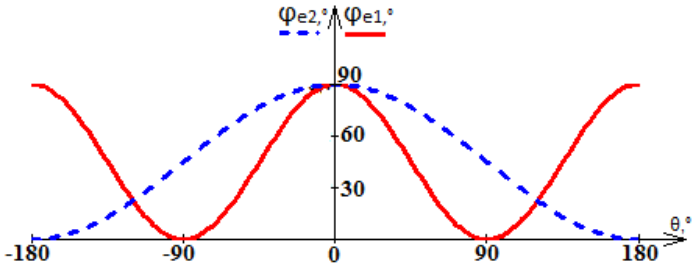


Figure 2. Dependence polarization plane azimuth on radiation station azimuth in the first and second channel

The block-diagram of one channel of polarimetric navigation system is shown in Figure 3. One channel of polarimetric navigation system includes radiation unit 1, measurement unit 2 and propagation medium 11. Unit 1 emits the polarized beam with certain polarization plane azimuth that passes through the propagation medium 11 and falls to the measurement unit 2 where the calculation of navigation beacon bearing takes place.

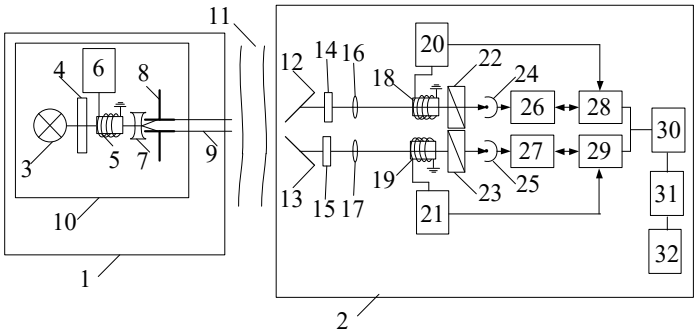


Figure 3. Block diagram of one channel of polarimetric navigation system

One channel of the radiation unit includes radiation source 3; monochromator 4; Faraday cell 5; sound generator 6; scattering lens 7, slit diaphragm 8. Both radiation channels are placed on the platform 10, which rotates. Monochromator 4 is designed to filter the incidence beam and miss out the wavelength on which the appropriate channel of measurement unit is tuned. Faraday cell 5 is designed for polarizing the beam and rotating its plane of polarization. Sound Generator 6 is intended to form control signals to the Faraday cell 5. Scattering lens 7 and a slit diaphragm 8 are designed for scattering plane-polarization radiation in the vertical plane with minimal divergence radiation angle in the horizontal plane. The platform 10 rotates around a vertical axis at constant frequency and provides changing the radiation direction of the beam 9. The beam polarization plane and platform rotate rotates simultaneously. This ensures communication of polarization plane azimuth and the radiation station azimuth. Two measurement channels differ by rotation frequency of the polarization plane: in the first channel it is equal to the speed of platform rotation, while in the second channel it is equal to the half of the platform rotation frequency.

The channel of measurement unit consists of planar dielectric plates 12 and 13, which are designed to rotate the polarization plane depending on the angle of incidence and polarization plane azimuth of the incident beam, it provides getting primary information about incidence angle and polarization plane azimuth; optical filters 14 and 15, which are designed to filter the incidence beam and miss out the wavelength on which the appropriate channel of radiation unit is tuned; focusing lens 16 and 17, which are designed to focus the radiation on the Faraday cell 18 and 19; Faraday cell 18 and 19, which are designed to modulate polarized radiation in an alternating magnetic field; sound generators 20 and 21, which are designed to form the control signals which are transmitted for Faraday cells 18 and 19; analyzers 22 and 23, which are designed to determine the polarization plane azimuth; photodetectors 24 and 25, which are designed to convert polarization plane azimuth into an electrical signal; narrow-band amplifiers 26 and 27, which are designed for amplifying electrical signal; synchronous detectors 28 and 29, which are designed to increase the measurement sensitivity and provide measurement "on zero signal"; microcontroller 30, which is designed for processing measurement results; storage unit 31, which is designed for collecting and storing measurement results; calculator 32, which is designed to perform mathematical calculations.

5. Result of Research

The paper represents the azimuth-measurement polarimetric method for determining the coordinates of the aircraft and the navigation system that implements the proposed method. Proposed navigation system allows measuring the relative angular position of the aircraft and navigation beacon. At that, aircraft's coordinates are determined by the azimuths of two beacons and altitude. The system consists of two channels. The polarization plane azimuth and incidence angle of radiation are defined in each channel. The value of polarization plane azimuth of incidence radiation in two channels determines the airplane bearing. The measured value of the incidence angle may be used in determining the altitude, pitch angle and station bearing.

In the process of mathematical modeling the dependence graph of the polarization plane azimuth of the refracted beam on incidence angle for different values of polarization plane azimuth of the incident beam (Fig. 4), the dependence graph of the measurement sensitivity on polarization plane azimuth of the incident beam (Fig.5), the dependence graph of the polarization plane azimuth of the refracted beam on the polarization plane azimuth of the incident beam for different values of the incidence angle (Fig. 6) were created. To determine the optimal incidence angle we created the dependence graphs of measurement sensitivity and refracted beam intensity on incidence angle on the same graph (Fig. 7). The incidence angle value of the intersection point is optimal incidence angle.

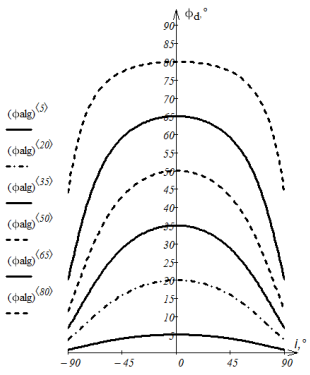


Figure 4.Dependence graph of the polarization plane azimuth of refracted beam on incidence angle

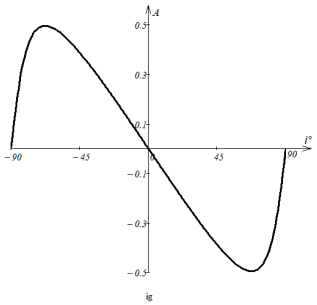


Figure.5 Dependence graph of the measurement sensitivity on polarization plane azimuth of the incident beam

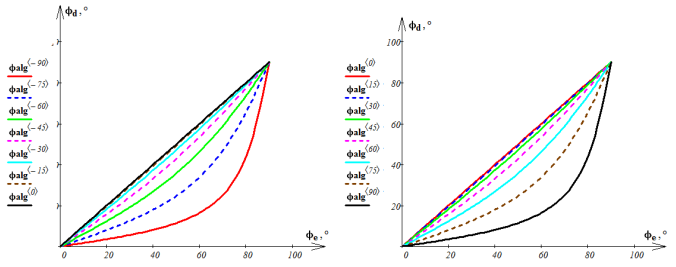


Figure 6. Dependence graphs of the polarization plane azimuth of the refracted beam on polarization plane azimuth of the incident beam

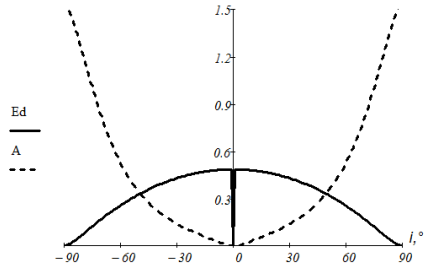


Figure 7 – Dependence graph of measurement sensitivity and refracted beam intensity on incidence angle

After analyzing the dependence graph of the polarization plane azimuth of refracted beam from incidence angle (Fig. 4), we conclude that the relationship is symmetric about the y axis and depends on the polarization plane azimuth of the incident beam. After analyzing the dependence graph of the measurement sensitivity on polarization plane azimuth of the incident beam (Fig. 5), we conclude that the relationship is symmetric about zero and has extremes. The graph shows that the highest sensitivity value provides polarization plane azimuth $\phi_e = \pm 68°$. After analyzing the dependence graph of the polarization plane azimuth of the refracted beam on polarization plane azimuth of the incident beam (Fig. 6), we conclude that the relationship is nonlinear and depends on the incidence angle. At that, the sensitivity of the measurement increases with the increase of the incidence angle. After analyzing dependence graphs of measurement sensitivity and refracted beam intensity on incidence angle on the same graph (Fig. 7), we conclude that the optimal incidence angle (i) equal $i = \pm 52{,}5°$. The optimal incidence angle is provided by installation of dielectric plate at an angle to the horizontal plane. Due to the fact that one measurement channel unit has two dielectric plates, it is proposed to establish one plate at an angle of 52.5 ° to the vertical plane and the second 52.5^0 angle to the horizontal plane.

6. Conclusions

The article deals with the existing methods for determining navigation parameters and the proposed azimuth-measurement polarimetric method for determining the coordinates of the aircraft. This method can potentially increase the accuracy and sensitivity of determining the navigation parameters. Also in the article the measurement system that implements this method is proposed. The proposed method and system can be used with azimuth-measurement and rho-theta navigation system for determining the aircraft's coordinates.

Distinction between the proposed system and optical gyroscopes lie in different principles of action: optic gyroscopes are autonomous means and polarimetric system is non-autonomous means, as well as lie in various physical operation principles: optical gyroscopes are based on the Sagnac effect and polarimetric navigation system is based on the Fresnel's formulas. Integrated utilization of optical gyroscopes and means of polarimetric navigation system will greatly increase the accuracy, sensitivity and reliability of information about aircraft's navigation parameters of flying.

The article also shows the results of mathematical modeling of the measurement channel. Optimal polarization plane azimuth of the incident beam and optimal incidence angle of radiation were determined during mathematical modeling. Calculations can be applied in the development and implementation of polarimetric navigation system. The practical implementation of the proposed method is to create a polarimetric navigation system and requires additional theoretical and practical research. This article reflects results of one theoretical research stage of the possibility to use the polarimetric technology for solving air navigation tasks.

References

[1] Азам Р., Башара Н. Эллипсомерия и поляризованный свет, Мир, 1981. (R. Azzam, N. Bashara, Ellipsometry and Polarized Light, Mir, 1981). (in Russian)

[2] Поль Р. В. Оптика и атомная физика, Физико-математическая литература, 1966. (R. Pohl, Optics and Nuclear Physics, Physical and Mathematical Literature, 1966) (in Russian)

[3] Annual Report of the ICAO Council: 2015

[4] Air Navigation Report 2016 Edition, 2016

[5] Правила організації радіолокаційного забезпечення польотів державної авіації України (Organizational radar support regulations of flight of Civil Aviation of Ukraine.). (in Ukrainian)

[6] В. І. Чепіженко, В. О. Волкогон, “Роль сучасних концепцій трансформації CNS/ATM систем у забезпеченні автономних польотів повітряних кораблів”, Actual Problems of Automation and Information Technologies, **19**, 112–120, 2015. (V. Chepizhenko, V. Volkogon "The role of modern concepts of system transformation CNS / ATM to provide autonomous flight of aircraft", Actual Problems of Automation and Information Technologies, **19**, 112–120, 2015) (in Ukrainian)

[7] M. Asanov, V. Rogozhin, A. Skrypets, V. Tronko “Estimation of accuracy characteristics of onboard vertical meters based on photopolarimetric metod”. Electronics and control systems.. **3 (34)**, 46–50, 2012.

[8] A. Klochan, A. Al-Ammouri V. Romanenko, V. Tronko “Aplication of polarymetry in aviation navigation systems” in IEEE 4th International conference «Methods and Systems of Navigation and Motion Control (MSNMC)», Kyiv, Ukraine, 2016.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 73-81 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Optimization of Multi-standard Transmitter Architecture Using Single-Double Conversion Technique Used for Rescue Operations

Riadh Essaadali[1], Said Aliouane[2], Chokri Jebali[2] and Ammar Kouki[2]

[1]*Aviat Networks, H9H3L4, Canada*

[2]*LACIME, Electrical engineering department, École de technologie supérieure, H3C1K3, Canada*

A R T I C L E I N F O

Article history:
Received: 06 March, 2017
Accepted: 29 March, 2017
Online: 09 April, 2017

Keywords :
Multi-standard, transmitter, optimization, direct conversion, single-double conversion, UAV and rescue operation

A B S T R A C T

A new architecture of multi-standard transmitter that combines single and double frequency conversion is presented. The transmitter is able to generate waveforms of different cellular standards (GSM, UMTS, Wimax), LAN standards (Bluetooth, Zigbee, IEEE802.11 b/g) and Radio navigation standards (GPS, Galileo) as well. Depending on the selected standard and frequency, the transmitter routes the intermediate frequency signal, into direct or double conversion path. Optimization design techniques are introduced to manage multiple standards and efficiently enable them while reducing cost, size and weight of the resulting transmitter. Such a transmitter is aimed to be used as a backup solution for a broken network to reestablish and insure communication links for rescue operations.

1. Introduction

On January 12th 2010, an earthquake has devastated Haiti, causing approximatively 160.000 deaths [1]. One of the main reasons that worsen the disaster was some major failures in the wireless communications towers, located at Port-au-Prince, which hampered the emergency measures.

Disasters have no specific target location and no specific timing. One of the possible solutions to insure the communication between rescue and intervention units is to deploy Unmanned Aerial Vehicles (UAVs, a.k.a Drones) to restore the links with the wearable transmitters in case of emergency and to generate radio navigation signals when satellite signals are lost. As illustrated in Figure 1, an UAV equipped with a multi-standard transceiver will fly over the affected city to take over the communication system.

Existing mobile networks are heterogeneous and people own different devices supporting cellular standards (GSM, UMTS and Wimax), wireless personal and local area network standards (Bluetooth, Zigbee and IEEE802.11 b/g) as well as Radio navigation (GPS and Galileo). In this context, the UAV's transceivers have to support all available wireless communication standards and all operating frequencies. Therefore, there is a real need to a multi-standard transceiver that covers common wireless standards and automatically switches between different standards and frequencies whenever needed.

Some solutions have been already proposed. Indeed, Qualcomm transmitter in [2] serves different GSM variants (GSM850, Primary GSM, EDGE, Railway GSM, and GSM1800 & GSM1900). In [3], another architecture is proposed to generate GSM signals using Phased Locked Loops (PLLs) in the up-converters eliminating the additional filters traditionally used between the up-converter and the amplification bloc. This solution minimizes power consumption but is only applicable to constant envelope signals. Another architecture [4] was proposed for both constant and complex envelope signals. Each waveform is generated by a different configuration of I/Q modulators.

The proposed solution involves a new architecture that generates the RF signal using direct or double conversion, depending on the desired standard. In addition, an optimization

*Riadh Essaadali Aviat Networks and LACIME, 16847 Hymus Boulevard, Kirkland, Quebec, H9H3L4,
Email:Riadh.essaadali@aviatnet.com

technique is introduced to minimize the high number of RF filters and PLLs necessary to cover multiple standards and frequencies. Therefore, cost, size and weight are also minimized.

This paper is an extension of work originally presented in IEEE MILCOM2016 [5] and it is organized as follows: Section II outlines the basics of multistandard architecture optimization. The judicious filtering technique is detailed. After presenting the frequency plan of different standards to be served, two architectures with their different components will be presented. In the last section, both of architectures are validated by carrying out simulations with CAD software: Advanced Design System (ADS).

Figure 1 Artist point of view of a rescue operation using UAVs [1]

2. Multi-standard transmitter architecture definition and optimization

2.1. Basics of architecture optimization

The first step in the transmitter design is defining the architecture. Generally, the transmitter is divided into the following blocks: baseband part, frequency up-converter, RF filter, power amplifier and antenna. Modern architectures could be heterodyne, homodyne, polar and digital-IF. The digital IF which is heterodyne is the most used because of its perfect balance between I and Q signals. Its architecture is illustrated in Figure 2. Digital I and Q signals are combined in the baseband part. The resulting signal is converted by a digital to analog converter.

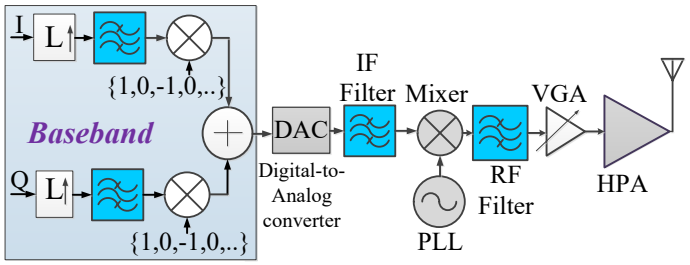


Figure 2 Architecture of digital IF (heterodyne) transmitter

The frequency f_{IF} of the signal at the IF stage is then up-converted to f_{RF}, at the output of the up-converter. In order to reduce the number of required filters, in a multi-standard transmitter, the upper side band signal located in $f_{RF} = f_{LO} + f_{IF}$ and the lower side band signal located in $f_{RF} = f_{LO} - f_{IF}$ are used. The mixing of the IF signal and the carrier produces the desired output signal, the image signal and the intermodulation products. As illustrated in Figure 3, the desired signal can be located either on the upper or lower sideband. Harmonics and inter-modulation products have the form $m*f_{LO}+n*f_{IF}$, and the critical harmonics are: $-f_{LO}+5f_{IF}, f_{LO}-3f_{IF}, f_{IF}, f_{LO}, f_{LO}+3f_{IF}, f_{LO}+5f_{IF}$. The IF-frequency f_{IF} is fixed whereas the PLL-frequency f_{LO} is tuned. If the signal is transmitted in the upper sideband, the filer should reject the nearest unwanted signal components f_{LO} and $f_{LO}+3f_{IF}$. In the same way, if the signal is transmitted in the lower sideband, the rejection of f_{LO} and $f_{LO}-3f_{IF}$ is required. This concept allows increasing the bandwidth of the RF filters, and therefore, it reduces their numbers.

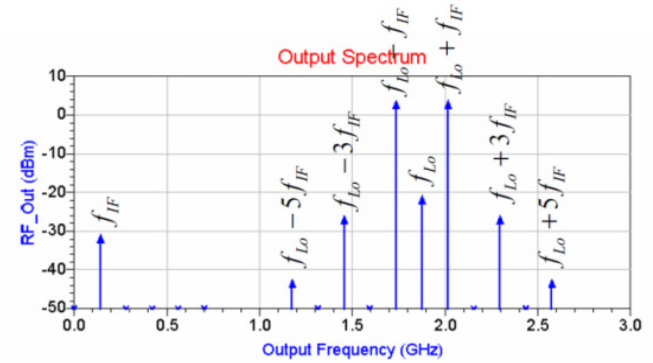


Figure 3 A typical output spectrum of the up-converter

Moreover, it decreases the frequency range that should be covered by the PLLs. Indeed, the sub-frequency band of the PLL that will cover the higher sub-band of the filter will be re-used to cover the lower sub-band of the filter. Figure 4 illustrates this concept. In conventional direct-conversion architecture, the maximum stop bandwidth of a filter is lower than f_{IF}. By using the optimization concept, the stop bandwidth of each RF filter is lower than $2f_{IF}$. An overview of this filtering strategy is given in Figure 5. The difference between the upside ($f_{StopHigh}$) and low side ($f_{StopLow}$) stop band frequencies of the RF filter should be at most equal to $2f_{IF}$. Two scenarios are included in Figure 5, the higher lateral band ($f_{LO1}+f_{IF}$) is filtered where f_{LO1} is the frequency generated by the local oscillator. In the second scenario, the lower lateral band ($f_{LO2}-f_{IF}$) is filtered by the same filter where f_{LO2} is the frequency generated by the local oscillator.

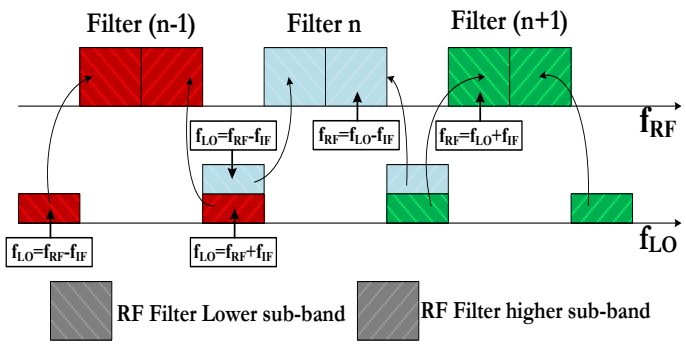


Figure 4 RF filters bands and the corresponding PLL frequency bands.

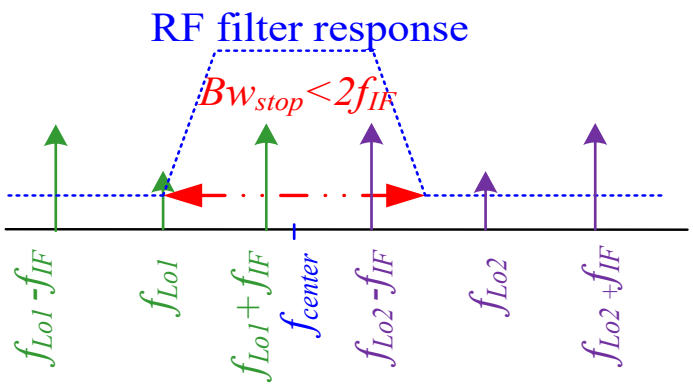


Figure 5 Judicious filtering technique.

2.2. Optimization of a Direct Conversion Multi-standard Transmitter

The frequency plan to be covered is shown in Figure 6. The second generation of a cellular communication standard (GSM) offers voice and low speed data transfer [6]. The need to add multimedia services at a very high-speed in wireless communication infrastructure has led to the development of the third generation such as UMTS and the fourth generation as Wimax [7]. Local and Personal wireless networks such as WLAN, Bluetooth and Zigbee have become very popular [8]. And satellite navigation systems were developed to enable precise and reliable positions for users. European Union launched a new positioning system called Galileo and United States of America have upgraded their satellite navigation system [9], [10].

The method consists on subdividing the whole frequency band into sub-bands, respecting the fact that the rejection bandwidth does not exceed $2f_{IF}$. Another point that should be considered is that closer the pass band is to $2f_{IF}$, the steeper the slope of the filter transfer function has to be. Therefore, the order of the filter increases and the number of resonators increases as well which increases the size of the RF filters. The rejection bandwidth should not exceed $2f_{IF}$. Thus, in order to reduce the number of required RF filters, The IF frequency f_{IF} should be high enough.

The maximum frequency that the afforded digital to analog converter can generate is 140 MHz. Thus, the transmitter requires 8 RF filters, each one having a bandwidth less than 2 fIF =2x140=280 MHz. the standards served by each RF filter are shown in Figure 7. Filter 1 covers the standards GSM, UMTS and Zigbee between 824 MHz and 960 MHz. Filter 2 covers GPS and Galileo signals. It has a bandwidth of 136 MHz centered at 1232 MHz. Filter 3 covers also GPS and Galileo signals between 1559 MHz and 1575 MHz. Filter 4 covers the standards GSM and UMTS between 1710 MHz and 1785 MHz. Filter 5 covers also the standards GSM and UMTS between 1805 MHz and 1990 MHz. Filter 6 covers the UMTS standard between 2010 MHz and 2170 MHz. Filter 7 covers Wimax, Bluetooth, Zigbee and WLAN standards between 2300 MHz and 2497 MHz. Filter 8 covers UMTS and Wimax standards between 2496 MHz and 2690 MHz.

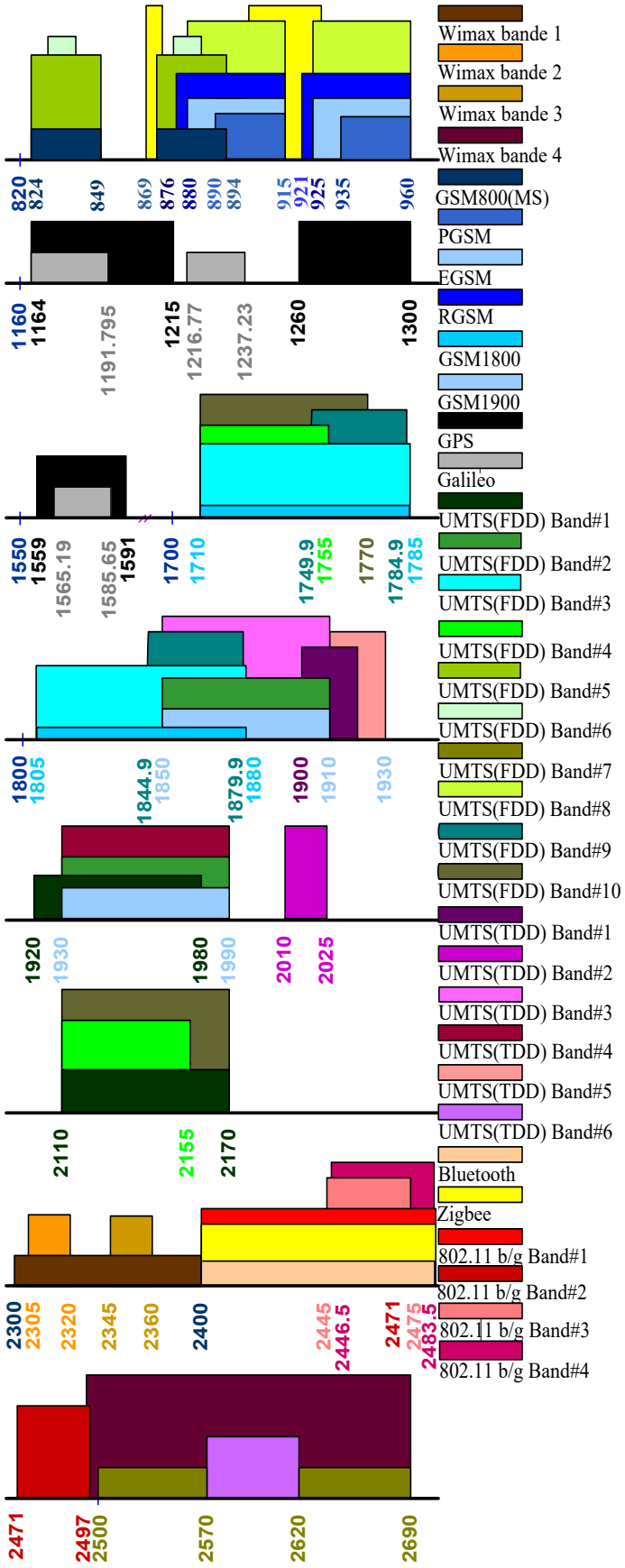
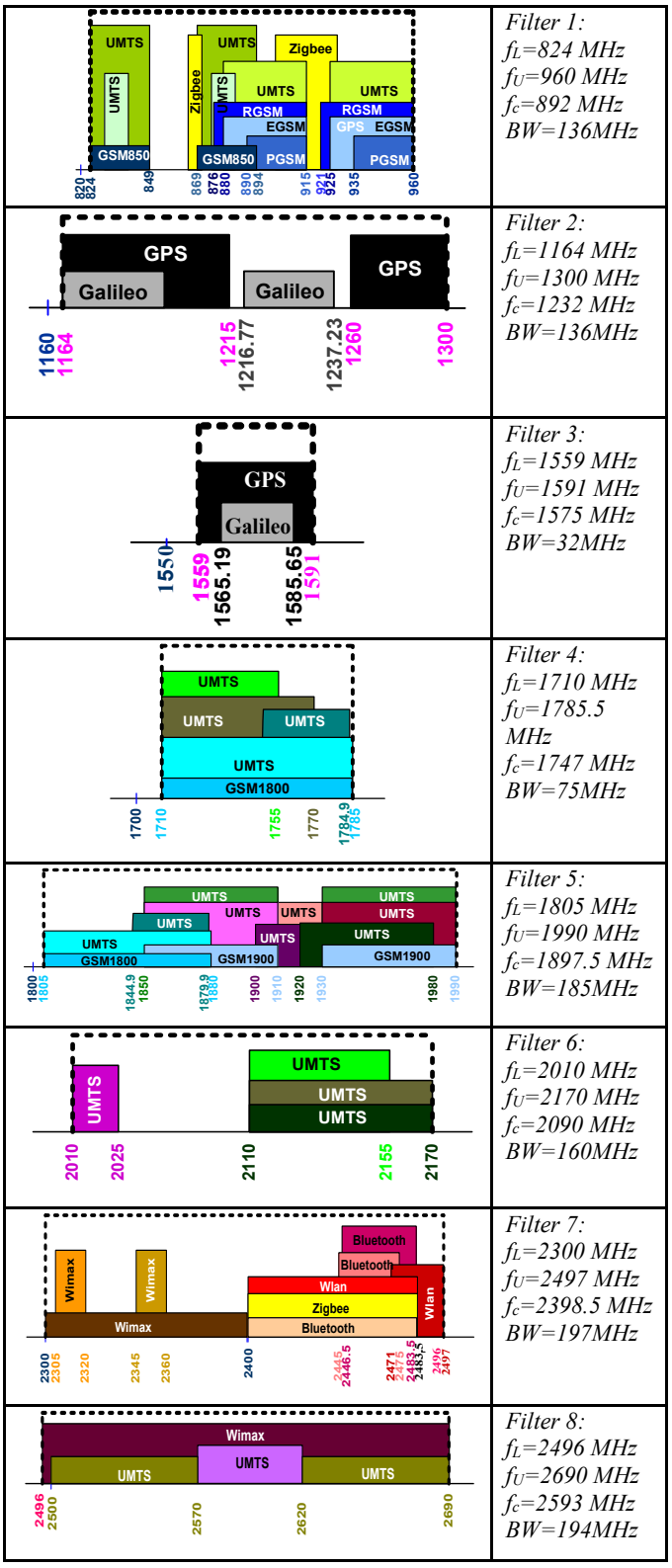


Figure 6 Frequency plan of the considered standards



Filter 1:
f_L=824 MHz
f_U=960 MHz
f_c=892 MHz
BW=136MHz

Filter 2:
f_L=1164 MHz
f_U=1300 MHz
f_c=1232 MHz
BW=136MHz

Filter 3:
f_L=1559 MHz
f_U=1591 MHz
f_c=1575 MHz
BW=32MHz

Filter 4:
f_L=1710 MHz
f_U=1785.5 MHz
f_c=1747 MHz
BW=75MHz

Filter 5:
f_L=1805 MHz
f_U=1990 MHz
f_c=1897.5 MHz
BW=185MHz

Filter 6:
f_L=2010 MHz
f_U=2170 MHz
f_c=2090 MHz
BW=160MHz

Filter 7:
f_L=2300 MHz
f_U=2497 MHz
f_c=2398.5 MHz
BW=197MHz

Filter 8:
f_L=2496 MHz
f_U=2690 MHz
f_c=2593 MHz
BW=194MHz

Figure 7 RF filters specifications.

Figure 8 shows how the multi-standard architecture generates and processes the right signal. GPS signal is taken here as an example to illustrate the process. At the DAC output the FPGA generate the GPS signal that passes through an IF filter of 32 MHz bandwidth, centered at 140 MHz. This corresponds to the maximum pass band allowed by the whole set of standards. The filtered signal is then up-converted to the desired RF frequency using a Phase Locked Loop (PLL). A combination of a two digitally controlled SPDT and SP4T switches insures the routing of the signal to the GPS filtering path which is selected among 8 RF filtering arms. Finally, the filtered signal is amplified before transmission.

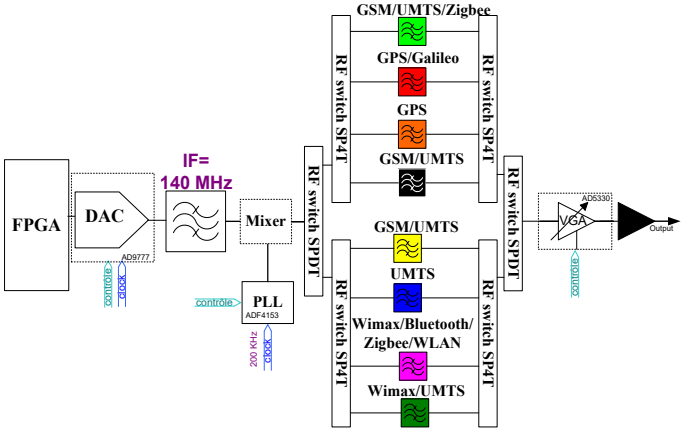


Figure 8 The proposed multi-standard transmitter architecture.

The frequency band covered by the voltage controlled oscillator used in the PLL depends on RF filter bandwidths. For a given filter having a central frequency f_c and a bandwidth BW, the frequency range is:

$[f_c - BW/2 - f_{IF} ; f_c - f_{IF}]$ if we choose $f_{RF} = f_{IF} + f_{LO}$, (1)

$[f_c + f_{IF} ; f_c + f_{IF} + BW/2]$ if we choose $f_{RF} = f_{LO} - f_{IF}$. (2)

Figure 9 shows the obtained frequency ranges to be covered by the frequency synthesizer.

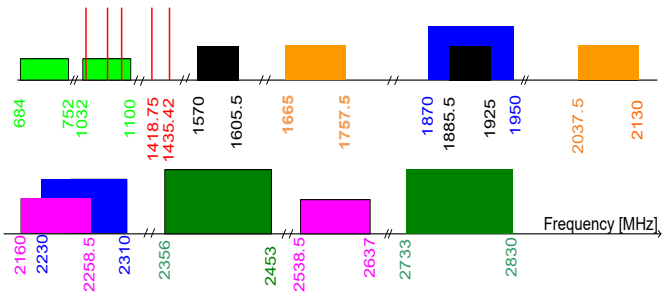


Figure 9 Frequency bands to be covered by the frequency synthesizer.

The mixer is selected primarily depending on its frequency band operation. The characteristics of the mixer are shown in Table 1. The frequency band of the LO port has to cover 824 and 2690 MHz. It has to exhibit a high IIP3 and a high 1dB compression point, as well as low insertion loss. Regarding those conditions, the best candidate is MACOM's ESMD-C50M mixer.

For the SPDT switch, many device candidates are possible. The Analog Devices' ADG918 exhibits the lowest insertion loss and return loss and offers the highest isolation and IIP3. For SP4T,

the Analog Devices'ADG904 fulfil the requirements. The specifications of the RF switches are presented in Table 2.

Table 1 Mixer main characteristics

Mixer Name	MACOM ESMD-C50M
RF band [MHz]	10-2950
IF band [MHz]	30 - 590
LO band [MHz]	10 - 2950
Conversion Loss [dB]	6.21
LO drive level [dBm]	13

Table 2 SPDT & SP4T main characteristics

Name	SPDT-ADG918	SP4T-ADG904
Freq band [MHz]	DC - 4000	DC-2700
Insertion Loss [dB]	0.8	1.1
Isolation [dB]	43	37
Retun Loss [dB]	26	27
IP_{1dB} [dBm]	17	16
IIP_3 [dBm]	36	31

The IF filter has to be centered at 140 MHz with a minimum bandwidth of 32 MHz. Low ripple, low insertion loss and low group delay are required. High stop band rejection and return loss are also needed. Oscilent's 820-IF140.0M is selected as it satisfies all the requirements. Its characteristics are presented in Table 3.

Table 3 IF Filter characteristics

Name	Oscilent820-IF140.0M
Center Freq [MHz]	140
Bandwidh [MHz]	34
Ripple (dB)	0.65
Group delay (us)	0.96
Rejection (dB)	60
Insertion Loss (dB)	22.2

The RF filter reponse depends on the characteristics of the chosen RF components, especially the frequency conversion bloc. The rejection level required to attenuate the level of f_{LO}, $f_{LO} \pm 3f_{IF}$ and of the other remaining intermodulation products should be determined to respect the out-of-band emission mask. A software simulation (ADS) is used to obtain the level of different harmonics and intermodulation products at the input of RF filters and at the output of the first stage of RF switches.

The common RF filter response is plotted in Figure 10. The attenuation A_{min} should be low in the filter passband between F_c-F_p and F_c+F_p. The term $A_{f_{Lo}}$ is the attenuation level needed to reject the oscillator signal. The frequency bands between F_c-F_p-f_{IF} and F_c- f_{IF} and between F_c+ f_{IF} and F_c+F_p+ f_{IF} correspond to the frequency range of the PLL frequency f_{LO}. The term A_{max} is the attenuation level to be achieved for the image frequency. The frequency bands between F_c-F_p-$2f_{IF}$ and F_c-$2f_{IF}$ and between F_c+$2f_{IF}$ and F_c+F_p+$2f_{IF}$ correspond to the frequency range of the image signal. The term $A_{f_{Lo}\pm 3f_{IF}}$ is the minimum level required to attenuate the intermodulation product located in $f_{Lo}-3f_{IF}$ and $f_{Lo}+3f_{IF}$. The frequency bands between F_c-F_p-$4f_{IF}$ and F_c-$4f_{IF}$ and between F_c+$4f_{IF}$ and F_c+F_p+$4f_{IF}$ correspond to the frequency range of the harmonic components $f_{Lo}-3f_{IF}$ and $f_{Lo}+3f_{IF}$ respectively. The term $A_{f_{Lo}\pm 5f_{IF}}$ is the minimum level required to attenuate the harmonics located in $f_{Lo}-5f_{IF}$ and $f_{Lo}+5f_{IF}$. The frequency bands between F_c-F_p-$6f_{IF}$ and F_c-$6f_{IF}$ and between F_c+$6f_{IF}$ and F_c+F_p+$6f_{IF}$ correspond to the frequency range of the harmonic components $f_{Lo}-5f_{IF}$ and $f_{Lo}+5f_{IF}$ respectively. The term $A_{f_{IF}}$ is the attenuation level of the IF component.

Different standards have an attenuation levels depending on the in- and out-of-band emission limits. The level of each harmonic components and undesired signals at the input of the RF filters are designed with a software simulation. The obtained harmonics and intermodulation products rejection levels for each filter are summarized in Table 4. The RF power amplifier is assumed to be linear to guarantee that the levels of the generated harmonics and intermodulation products generated by the up-converter relative to the signal level are kept to be similar after the amplification bloc.

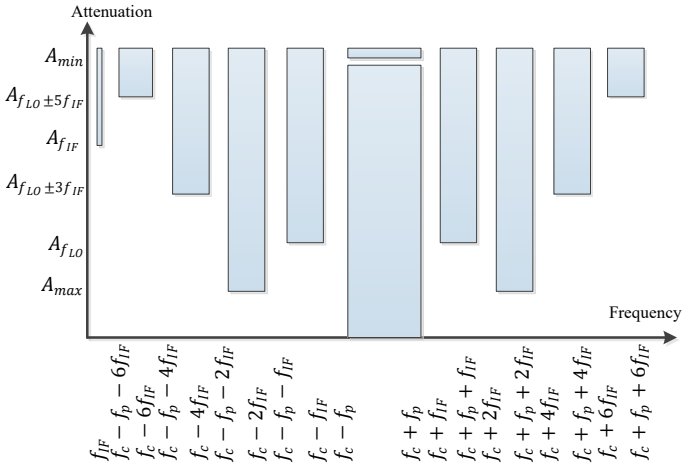


Figure 10 RF Filter response

Table 4 F_{LO} and $F_{LO}\pm 3F_{IF}$ rejections for each channel filter

Specification	Center Freq [MHz]	Pass band Freq [MHz]	$f_{LO}\pm 3f_{IF}$ rejection level [dB]	f_{LO} rejection level [dB]
Filter 1	892	136	80	59
Filter 2	1232	136	40	19
Filter 3	1575	32	40	19
Filter 4	1747.5	75	80	59
Filter 5	1897.5	185	80	59
Filter 6	2090	160	53	32
Filter 7	2398.5	197	52	31
Filter 8	2593	194	53	32

2.3. Optimization of a Multi-standard transmitter architecture using single/double conversion

The first architecture uses eight RF filters, two SPDT switches, four SP4T switches. This architecture is costly in terms of size. For optimization purposes, we have developed another architecture based on double conversion. Two stages are used to transpose the signal from an intermediate frequency to RF frequency.

In order to reduce further the number of required RF filters, The IF frequency fIF should be sufficiently high. The maximum frequency that the afforded DAC can generate is 140 MHz. The purpose is to cover the maximum standards with the minimum number of filters. The judicious filtering technique detailed in the section II.A is re-used to develop this architecture.

The first part of the spectrum is 824 –1300MHz. For this section of the spectrum, the direct conversion method with an IF frequency of 140 MHz (which leads to a maximum channel filter bandwidth of 280 MHz) gives good results in terms of rejections. However, the second section of the spectrum (1559 – 2690 MHz) requires at least six IF-filters to be entirely covered. To reduce the total size and cost of this previous setting, we use instead, a double conversion with a new IF-frequency that will be high enough to require at most two band-pass filters without being too close to the LO frequency. This prevents to generate harmonic interference.

In fact, the intermediate frequency is not fixed for this architecture, it is chosen in such a way that it does not exceed the third of the frequency of the phase-locked loop, $f_{LO}/3$. Indeed, at the output of each mixer at both frequency conversion blocs, the critical harmonics that should be filtered are $-f_{Lo}+5f_{IF}$, f_{IF}, $f_{Lo}-3f_{IF}$, f_{Lo}, $f_{Lo}+3f_{IF}$, $f_{Lo}+5f_{IF}$. These harmonics may disturb the judicious filtering technique when one of them, at least, falls within the frequency ranges between $f_{Lo}-2f_{IF}$ and $f_{Lo}+2f_{IF}$. The RF filter bandwidth is at most equal to twice the intermediate frequency

$$\max(f_{IF}) = \frac{2}{7} f_{LO} = \frac{2}{9} f_{RF} \quad (3)$$

$$\Delta f \le 2 * \min(f_{IF}) \quad (4)$$

Looking at Figure 6, the frequency bands of interest start at $f_{RF} = 824MHz$. According to (3), the maximum IF frequency is therefore 206 MHz. In one hand, if the carriers around 824 MHz are generated by double conversion, the maximum of the frequency at the output of the digital converter must be 51.5 MHz. In the other hand, if the signal is generated by direct conversion, the maximum of the intermediate frequency is 140 MHz and the first filter covers the frequency band between 824 MHz and 960 MHz. The bandwidth of the filter is 136 MHz and the minimum intermediate frequency f_{IF} is 68 MHz. The first filter bandwidth covers GSM, UMTS and Zigbee standards. The second filter starts at $f_{RF} = 1164MHz$. According to (3), the maximum IF frequency f_{IF} is therefore 291 MHz in the case of direct conversion. The bandwidth of the filter is 136 MHz therefore the minimum frequency is 68 MHz. The frequency bands to be covered which come just after the second filter start at frequency 1559 MHz. The corresponding maximum IF frequency is 389.75 MHz. The third filter covers the band 1559-1990 MHz. The maximum frequency and minimum frequency is therefore 215.5 MHz. The fourth filter covers the frequency band between 2010 MHz and 2690 MHz. The maximum intermediate frequency is 502.5 MHz and the minimum intermediate frequency is 340 MHz. Simulations using Advanced Design System (ADS) of Keysight technologies on this set up (see Simulation section) prove that the second stage IF frequency of 357.5 MHz is a good compromise and corresponds to a central frequency of an available intermediate filter. The maximum allowed channel filter bandwidth is 715 MHz. The filters 1 and 2 are used for the direct conversion whereas the filters 3 and 4 are used for the double conversion. Specifications of each RF filter are summarized in Figure 11.

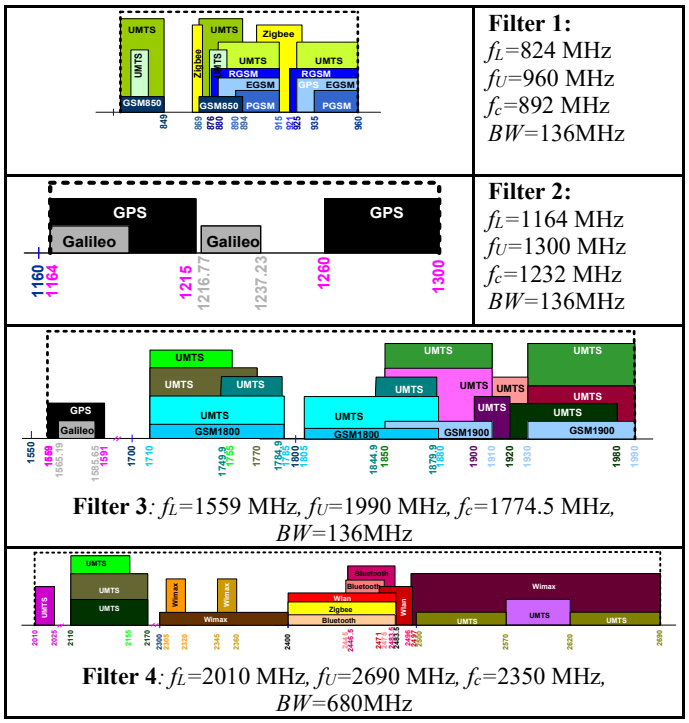


Figure 11 RF filters specifications.

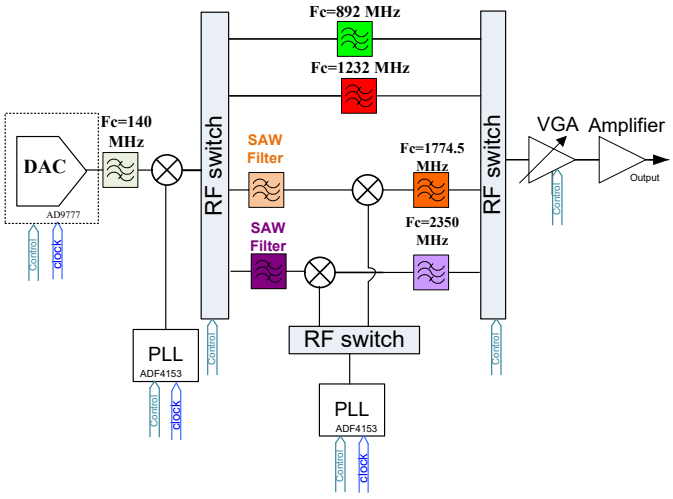


Figure 12 The proposed multi-standard transmitter architecture.

The global transmitter architecture is shown in Figure 12. The DAC output is filtered by an IF-filter with a 32 MHz bandwidth (corresponding to the maximum pass band of the GPS standards) centered at 140 MHz. The signal is then up-converted to a desired

RF frequency. Then, depending on the RF frequency, the signal is routed to either single or double conversion path. For the double conversion, two intermediate surface acoustic wave (SAW) filters centered at 357.5 MHz and having at least 32 MHz bandwidth are to be considered before the second up-conversion stage. The signal is then driven by a variable gain amplifier (VGA) depending on the standard and finally amplified by a high power amplifier (PA).

Figure 13 shows the obtained frequency for the local oscillators. The green bands of 684 – 752 MHz and 1032 –1100 MHz are the LO frequencies corresponding to the direct conversion path (filters 1 and 2). For double conversion path we use the up sideband frequency 497.5 MHz (357.5+140) because it is closer to the other required bands. The orange and purple bands corresponds to the LO frequencies of the second PLL for the 3rd and 4th filters respectively (low side band and higher side band).

In the frequency synthesizer's side, the frequency range to be covered by the Voltage-Controlled Oscillator (VCO) depends on the bandwidth of the filters. For a filter whose center frequency f_c is and the bandwidth is B_W , the frequency range to be covered by the VCO is between $f_c - B_W/2 - f_{IF}$ and $f_c - f_{IF}$ on the one hand and between $f_c + f_{IF}$ and $f_c + B_W/2 + f_{IF}$ on the other hand. Figure 13 shows the total frequency range to be covered by the RF frequency synthesizers. For the signals covered by the first two filters covering the frequency bands 824-960 MHz and 1164-1300 MHz respectively, the intermediate frequency f_{IF} is set to 140 MHz. In the path where the frequency transposition occurs twice, the intermediate frequency at the output of the analog-digital converter is fixed and is equal to 140 MHz. The signal generated by the converter is transposed to 357.5 MHz by the first block of the frequency synthesizer. The first frequency synthesizer should generate a frequency of 497.5 MHz or 217.5 MHz. The frequency 497.5 MHz is chosen because it is closer to the other frequency ranges. The second block of the frequency synthesizer serves to transpose the 357.5 MHz frequency signal to a RF signal that will be filtered by the third and fourth filters.

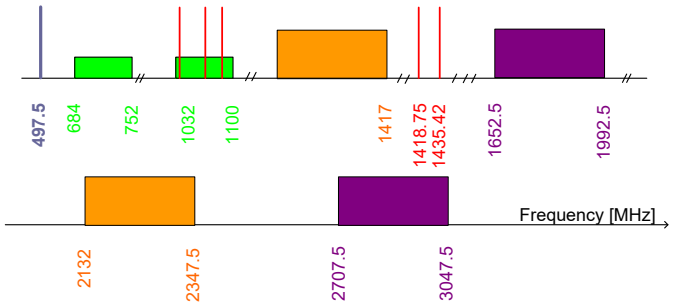


Figure 13 Frequency bands to be covered by the frequency synthesizer

Between the two mixers, a filter should be inserted. In the market, SAW (Surface Acoustic Wave) filters can be used. The third filter covers GPS, Galileo, UMTS, GSM and WLAN standards with a maximum channel width of 32 MHz. The fourth filter covers the UMTS, WIMAX, Bluetooth, Zigbee and WLAN standards with a maximum bandwidth of 28 MHz. The first and the second SAW are pass-band filters. The center frequency of the first SAW is located between 215.5MHz and 389.75 MHz and between 340 MHz and 502.5 MHz, for the second. Their bandwidth should be at least 32 MHz and 20 MHz respectively. Figure 14 illustrates the possible locations of IF channels to be filtered by the third and fourth filters, i.e, at the second frequency conversion bloc.

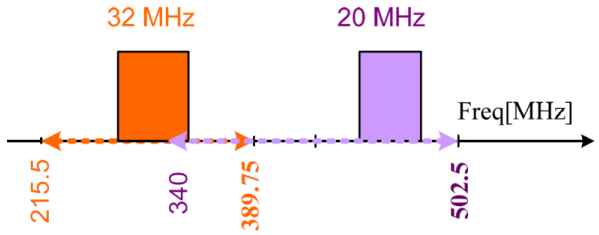


Figure 14 Possible frequencies of intermediate IF channels.

The folding of the two intermediate frequency ranges is considered as an optimization point if a filter that covers the frequency band between 389.75 MHz and 340 MHz is available. From the market, the chosen SAW filter is RFM SF2090C as its ripple, group delay and insertion loss are low and its out-of-band rejection is high. Its characteristics are in Table 5.

The modified transmitter architecture is illustrated in Figure 15. The transmitter includes an analog to digital converter, a mixer, phase locked loops, an SPDT type RF switch, two SP4T type RF switches, a variable gain amplifier and an RF amplifier block. The chosen mixer is MACOM's ESMD-C50M. Its main characteristics are listed in TABLE I. The RF switches ADG918 and ADG904 of analog devices are the best candidates. Their characteristics are in TABLE II. After the DAC, the IF filter (Oscilent's 820-IF140.0M) is inserted. Its characteristics are listed in Table 1.

Table 5 SAW IF filter characteristics

SAW filter Name	RFM SF2090C
Center freq [MHz]	357.5
Bandwidth [MHz]	35
Ripple [dB]	1
Group delay [ns]	55
Rejec-tion [dB]	40
Insertion Loss [dB]	12
Frequency Band [MHz]	340-375

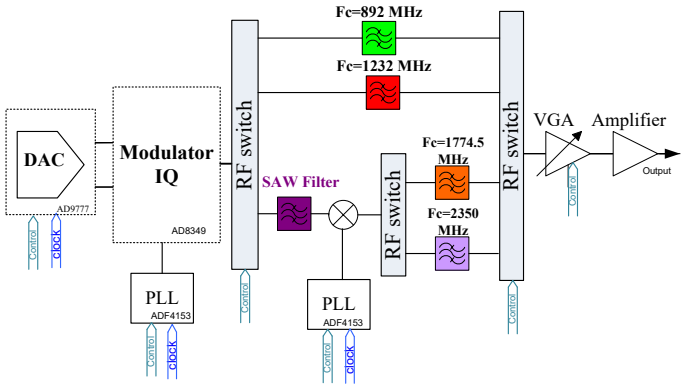


Figure 15 The architecture of single/dual multi-standard transmitter.

The out-of-band attenuations of the first second filters are determined based on the level of the harmonics at the output of the first mixer and the values of the center frequency and bandwidth. The center frequency of the first filter is 892 MHz, and its bandwidth is 136 MHz. The first filter covers GSM, UMTS and Zigbee standards. According to their various masks and spurious emissions cited, the rejection of the image frequency should be 80 dB and of the local oscillator should be 19 dB. The second filter covers the GPS and Galileo systems, the rejection of the image frequency should be 40 dB and of the component is 19 dB. For the third filter and the fourth filter, the levels of harmonics are obtained by carrying out a harmonic balance simulation. The level of harmonic components $f_{Lo}+f_{IF}$ and $f_{Lo}-f_{IF}$ is 10.79 dBm. The level of the local oscillator signal is -19.64. According to the Mask signal and out-of-band emission limits of GSM, UMTS, GPS and Galileo, the third filter should attenuate the image signal by 80 dB and the local oscillator component by 71.15 dB. The stop band of this filter is 720 MHz. The fourth RF filter has a bandwidth of 680 MHz. It covers UMTS, Bluetooth, Zigbee, WLAN and WIMAX standards. It should attenuate the signal at f_{Lo} by 44.15 dB and the image frequency by 53 dB.

3. Simulation

3.1. Simulation of the Direct Conversion Multi-standard Transmitter

To validate the concept, the multi-standard architecture is implemented in ADS, as shown in Figure 16. The RF filters, the RF switches, the mixer as well as the IF filter are implemented by using their models. The simulation allows checking whether the spurious signals are well rejected by the RF filters. All standards generated through the proposed multi-standard architecture are tested. And for each standard, many frequency step simulation runs were carried out. The output spectrum respects the out of band emission limits, in all simulated cases. The following test run is used as an illustration example.

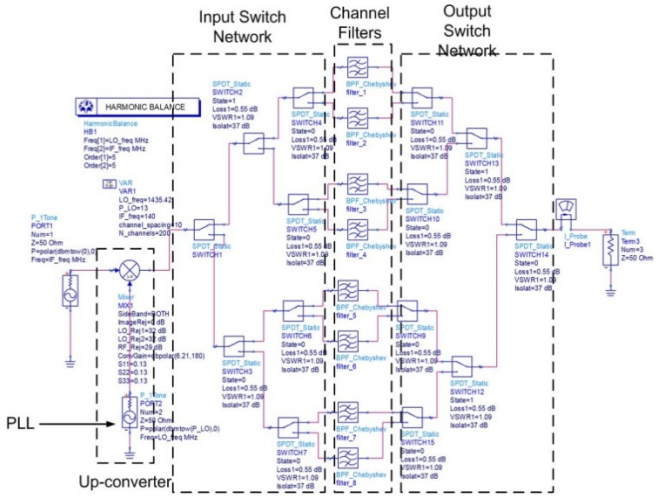


Figure 16 Validation of the direct conversion multi-standard transmitter architecture with ADS.

To generate a WiMAX signal at 2304 MHz, the RF switches are activated in a way that the signal is filtered by filter 7. The frequency of the signal that is generated by the PLL is $f_{Lo}=f_{RF}-f_{IF}=2164MHz$. The spectrum of the signal at the output of the transmitter for this case is illustrated in Figure 17.

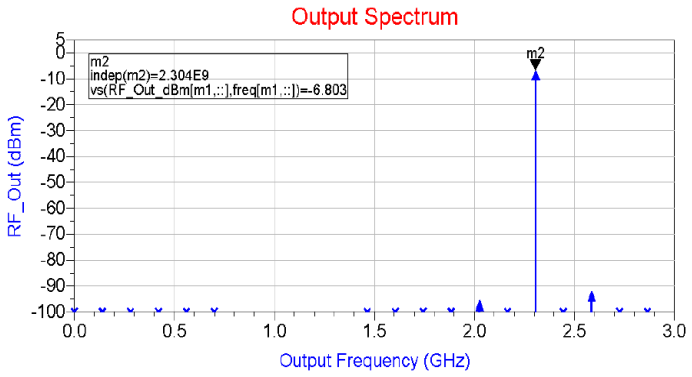


Figure 17 Spectrum of the signal at the output of the transmitter

3.2. Simulation of a multi-standard transmitter using single/ double conversion technique

After the selection of the RF components, the simulation of the second architecture is done using the Keysight–ADS simulator. The circuit diagram useful for completing the simulation in ADS is shown in Figure 18. The simulation consists in carrying out the spectral analysis at the output of the transmitter by varying the frequencies of the frequency synthesizers and adjusting the control parameters of the RF switches.

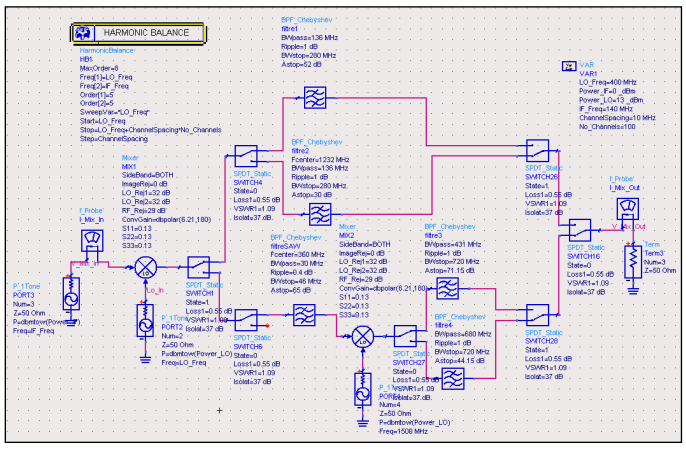


Figure 18 Validation of the multi-standard transmitter architecture using single/double conversion technique with ADS.

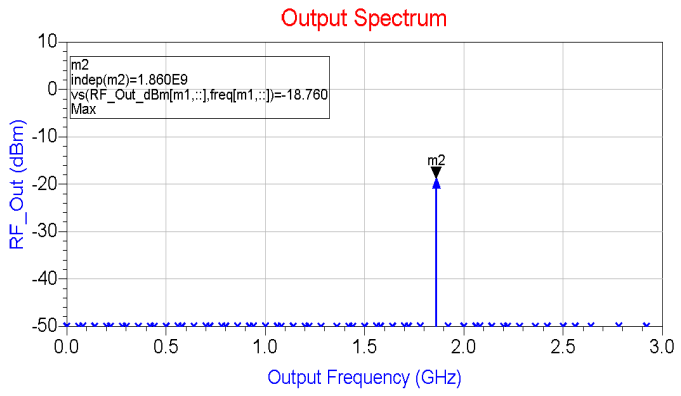


Figure 19 Spectrum of the signal at the output of the transmitter

For example, if the frequency of the first frequency synthesizer and of the second frequency synthesizer are adjusted to $f_{Lo1} = 497.5MHz$ and $f_{Lo2} = 2217.5MHz$. The different RF switches are tuned in a way that the RF signal passes through the third filter. At the output of the transmitter, the spectrum is shown in Figure 19.

4. Conclusion

In this paper, two new optimized architectures of a multi-standard transmitter have been presented. The first one is an heterodyne based architecture. The second combines single and double frequency conversion that allows further reducing the cost, size and weight of the transmitter. This transmitter is a part of a wireless radio that will be carried by an UAV reaching stricken areas after a disaster and therefore enabling and making easier efficient rescue intervention.

It has been shown, through simulation, that both of architectures can generate GPS and Galileo signals which are very critical for rescue operations and waveforms of different cellular wireless communication standards (UMTS, GSM and Wimax) along with waveforms of different local and personal area wireless network standards (Bluetooth, Zigbee and IEE802.11 b/g). The high number of different wireless communication standards and their operating frequency bands requires a high number of RF filters and PLLs. The first architecture is direct conversion. It uses an analog digital converter to convert the digital signal into an analog signal. The analog signal passes through an IF filter with a center frequency of 140 MHz. The filtered signal is mixed with a carrier to be transposed to an RF frequency. The RF signal may be the upper or lower sideband. The RF signal passes through an SP4T switch which points the RF signal to its corresponding RF filter. The first architecture uses eight RF filters. RF filters are partitioned and defined based on the judicious filtering technique. For each RF filter, the lower half of bandwidth serves to filter the upper side signal and the upper half of filter's bandwidth serves to filter the lower side signal. The second architecture is less expensive than the first. It uses fewer components. It uses two SAW IF filters, two mixers, two SP4T RF switches, an SPDT RF switch and four RF filters. The frequency range to be covered by the frequency synthesizers of the second architecture is between 497.5 MHz and 3047.5 MHz. As a logical perspective, the architecture can be further optimized. There is an opportunity to use a frequency doubler that allows using fewer RF filters and fewer RF frequency synthesizers.

References

[1] F. Gagnon, "La radio cognitive, l'avenir de demain," ed, May 2013.

[2] S. M. Mollenkopf, P. H. A. See, and B. C. Walker, "Multi-standard transmitter system and method for a wireless communication system," ed: Google Patents, 2005.

[3] H. Jensen, "Translational Loop RF transmitter architecture for GSM Radio," United States Patent, vol. US007386283B2, no. Broadcom Corporation, p. 14, Jun. 10,2008 2008.

[4] J. Kilpatrick, J. Dawson, J. Bohorquez, and J. Venuti, "Programmable transmitter architecture for non-constant and constant envelope modulation," ed: Google Patents, 2007.

[5] R. Essaadali, T. Hadouej, C. Jebali, and A. Kouki, "Optimization of a direct conversion multi-standard transmitter used for rescue operations," in MILCOM 2016 - 2016 IEEE Military Communications Conference, 2016, pp. 1173-1178.

[6] L. Harte and Books24x7 Inc. (2005). Introduction to global system for mobile communication (GSM) physical channels, logical channels, network, and operation [Text.].

[7] J. G.Andrews, A. Ghish, and R. Muhamed, Fundamentals of WiMAX Prentice Hall, february 2007.

[8] H. A. H. Labiod, C. De Santis. (Springer, 2007). WI-FI TM, BLUETOOTH TM, ZIGBEE TM AND WIMAX TM.

[9] A. R. CORPORATION, Navstar GPS Space Segment/Navigation User Interfaces, 1993.

[10] J. Kreher, "GALILEO Signals : RF Characteristics," presented at the Navigation System Panel, 2004.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 82-87 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Singular Integral Equations in Electromagnetic Waves Reflection Modeling

A. S. Ilinskiy[1], *T. N. Galishnikova*[*,1]

[1] *Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University, 119991, Russia*

ARTICLE INFO

Article history:
Received: 01 March, 2017
Accepted: 29 March, 2017
Online: 11 April, 2017

Keywords:
Reflection of electromagnetic waves
Inhomogeneous medium
Singular integral equations

ABSTRACT

The processes of reflection of three-dimensional electromagnetic waves by locally irregular media interfaces are investigated. The problem under study is mathematically reduced to the solution of a boundary value problem for the Maxwell equations in an infinite space with an irregular boundary. In order to develop a numerical algorithm, the potential theory and a special Green's function are applied to reduce the addressed boundary problem to an equivalent system of two hypersingular integral equations. This system is solved with the use of the approximation and collocation method. Special attention is focused on calculation of the kernels of these equations. Results of simulation of the currents induced on the irregularity and reflected field patterns in the resonance frequency range are presented.

1. Introduction

This paper is an extension of the work originally presented at the URSI Commission B International Symposium on Electromagnetic Theory (EMTS 2016) [1]. In this article we model the scattering of three-dimensional electromagnetic waves on a finite impedance section of a wavy surface separating two media. The scattering of electromagnetic waves by a wavy interface between two different media is an important problem for various applications. In this article we investigate the incidence of an arbitrary plane wave on a local inhomogeneity of the interface assuming sufficiently high conductivity of the underlying medium. The mathematical formulation of such problems reduces to that of solving a system of Maxwell equations in non-regular infinite regions. Depending on the polarization of the incident field, the boundary-value problems are reduced to independent systems of hypersingular integral equations solved by specially developed numerical algorithms utilizing a singularity isolating algorithm.

2. Mathematical model and numerical algorithm

2.1. Mathematical formulation of wave scattering problems

Let us assume that an interface between two media is situated in Cartesian coordinates x, y, z. The interface consists of two half planes Σ that are specified by the equation $y = 0$ and joined by an impedance cylindrical wavy section S that stretches along the z axis. The entire wavy section is located in the half-space $y \geq 0$. The surface S is specified by the equation $y=f(x)$, $0 \leq x \leq a$, $f(0) = f(a) = 0$, where a is the x length of irregularity S. Let $a = 2\pi$. The reflecting surface $S \cup \Sigma$ does not depend on z coordinate. Denote by D_1 the region over interface $S \cup \Sigma$ where the 3D incident electromagnetic field propagates. This region D_1 is characterized by permittivity ε_1, permeability μ_1, and the k_1 wave number $k_1^2 = \omega^2 \varepsilon_1 \mu_1$, where ω is the circular frequency (Figure 1). The time dependence is $\exp(-i\omega t)$.

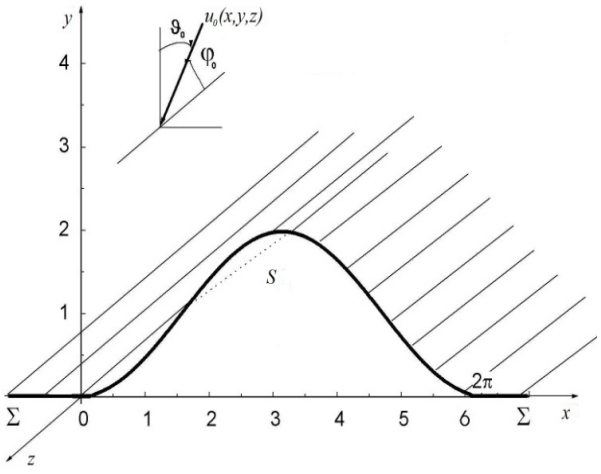


Figure 1: Geometry of perfecting surface

The field of the plane wave

$$\vec{E}_0(x,y,z) = \vec{E}_0 \exp(i\alpha_0 x - i\beta_0 y + i\gamma_0 z),$$
$$\vec{H}_0(x,y,z) = \vec{H}_0 \exp(i\alpha_0 x - i\beta_0 y + i\gamma_0 z),$$

$\alpha_0 = k_1 \sin\vartheta_0 \sin\varphi_0$ □ $\beta_0 = k_1 \cos\vartheta_0 \sin\varphi_0$, and $\gamma_0 = k_1 \cos\varphi_0$ is incident on the surface $S \cup \Sigma$ in region D_1. Here, φ_0 is the angle between axis z and the direction of the incident wave propagation. When $\varphi_0 = \pi/2$, we obtain a plane reflection two-dimensional problem. If $\vartheta_0 = 0$, we have the case of the normal incidence in the plane problem.

In region D_1 we seek for a solution to the homogeneous system of Maxwell equations

$$curl\,\vec{H}(x,y,z) + i\omega\varepsilon_1\,\vec{E}(x,y,z) = \vec{0},$$
$$curl\,\vec{E}(x,y,z) - i\omega\mu_1\,\vec{H}(x,y,z) = \vec{0},$$

which satisfies on the plane section Σ the condition of perfect conduction

$$\left[\vec{n} \times \vec{E}(x,y,z)\right] = \vec{0},$$

Leontovich condition on local section S with a finite conductance

$$\left[\vec{n} \times \vec{E}(x,y,z)\right] = -W_2\left[\vec{n} \times \left[\vec{n} \times \vec{H}(x,y,z)\right]\right],$$

$W_2 = \sqrt{\mu_2/\varepsilon_2}$. Here, ε_2 and μ_2 are the characteristics of the section S, $\vec{n} = (n_x, n_y, 0)$ is the outward normal to region D_1. The reflected field satisfies the radiation conditions at infinity.

We look for the total electromagnetic field in D_1 that has the same dependence on z as the incident field, that is

$$\vec{E}(x,y,z) = \vec{E}(x,y)\exp(i\gamma_0 z),\quad \vec{H}(x,y,z) = \vec{H}(x,y)\exp(i\gamma_0 z).$$

Then, the boundary condition on S can be represented as follows:

$$E_z(x,y) = \frac{W_2}{k_1^2 - \gamma_0^2}\left(-i\gamma_0 \frac{\partial H_z(x,y)}{\partial \tau} - i\omega\varepsilon_1 \frac{\partial E_z(x,y)}{\partial n}\right)$$
$$H_z(x,y) = \frac{1}{W_2(k_1^2 - \gamma_0^2)}\left(-i\omega\mu_1 \frac{\partial H_z(x,y)}{\partial n} + i\gamma_0 \frac{\partial E_z(x,y)}{\partial \tau}\right),$$

$\partial/\partial n = n_x\,\partial/\partial x + n_y\,\partial/\partial y$, $\partial/\partial\tau = n_x\,\partial/\partial y - n_y\,\partial/\partial x$ are the normal and tangent derivatives, respectively. Accordingly, boundary conditions on plane section Σ take the form

$$E_z(x,y) = 0, \qquad \frac{\partial H_z(x,y)}{\partial n} = 0.$$

We denote $u(x,y) = E_z(x,y)$, $v(x,y) = H_z(x,y)$, $\beta_1^2 = k_1^2 - \gamma_0^2$. Then the considered boundary value problem can be reduced to solving in region D_1 of the 2D homogeneous Helmholtz wave equation

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \beta_1^2\right)\left\{\frac{u(x,y)}{v(x,y)}\right\} = 0$$

with appropriate boundary conditions on Σ and S.

2.2. Reduction of the boundary problem for Helmholtz equations to the system of singular equations in case of E-polarization

In order to obtain a system of hypersingular integral equations equivalent to the boundary value problem for Helmholtz equation, we introduce Green's function $g^E(M,P)$, satisfying in region D_1 the inhomogeneous Helmholtz equation

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \beta_1^2\right) g^E(M,P) = -2\pi\,\delta(M,P),$$

the boundary condition for perfect conductor on half-plane Σ

$$g^E(M,P) = 0,$$

that have the form

$$g^E(M,P) = \frac{i\pi}{2}\left[H_0^{(1)}(\beta_1 r_{M,P}) - H_0^{(1)}(\beta_1 \hat{r}_{M,\hat{P}})\right].$$

Here, $H_0^{(1)}(x)$ is the zero-order Hankel function of the first kind; $r_{M,P} = \sqrt{(x_M - x_P)^2 + (y_M - y_P)^2}$, where $(x_{M,P}, y_{M,P})$ are coordinates of M and P points; $\hat{r}_{M,\hat{P}} = \sqrt{(x_M - x_P)^2 + (y_M + y_P)^2}$, $(x_P, -y_P)$ are the coordinates of the $\hat{P}$ point.

Applying in region D_1 the Green's formula successively to functions $u(M)$, $g^E(M,P)$, taking into account the radiation conditions and the properties of surface potentials, we obtain a system of two integral equations for unknown functions $\partial u(P)/\partial n$ and $\partial v(P)/\partial\tau$ on finite section S for E- polarization case:

$$\frac{ik_1 W_2}{2W_1\beta_1^2}\frac{\partial u(M)}{\partial n_M} + \frac{i\gamma_0 W_2}{2\beta_1^2}\frac{\partial v(M)}{\partial \tau_M} +$$
$$\frac{1}{2\pi}\int_S\left[\left(g^E(M,P) + \frac{ik_1 W_2}{W_1\beta_1^2}\frac{\partial g^E(M,P)}{\partial n_P}\right)\frac{\partial u(P)}{\partial n_P} + \right. \qquad (1)$$
$$\left.\frac{i\gamma_0 W_2}{\beta_1^2}\frac{\partial g^E(M,P)}{\partial n_P}\frac{\partial v(P)}{\partial \tau_P}\right] ds_P = -u_0^E(M),$$

$$\frac{1}{2}\frac{\partial u(M)}{\partial n_M}-\frac{1}{2\pi}\int_S\left[\left(\frac{\partial g^E(M,P)}{\partial n_M}+\frac{ik_1W_2}{W_1\beta_1^2}\frac{\partial^2 g^E(M,P)}{\partial n_M\,\partial n_P}\right)\frac{\partial u(P)}{\partial n_P}+\right.$$

$$\left.\frac{i\gamma_0W_2}{\beta_1^2}\frac{\partial^2 g^E(M,P)}{\partial n_M\,\partial n_P}\frac{\partial v(P)}{\partial \tau_P}\right]ds_P=\frac{\partial u_0^E(M)}{\partial n_M},\ M\in S, \qquad (2)$$

where $u_0^E(x,y)=\exp(i\alpha_0 x-i\beta_0 y)-\exp(i\alpha_0 x+i\beta_0 y)$. In the case of the *H*- polarization we can obtain a system of two integral equations for unknown functions $v(P)$ and $\partial u(P)/\partial\tau$ on finite section *S*.

2.3. The numerical algorithm for the solution of singular integral equations for the E-polarization

The systems of integral equations are singular and can be solved through reducing them to the complex system of linear algebraic equation by means of the collocation and approximation method.

To this end, we divide the interval $[0,a]$, on which the equation $y=f(x)$ of the irregular cylindrical surface S is defined, by points $x_i=ia/N,\quad x_0=0,\quad x_N=a$, into N equal parts. The points dividing the curve S have the coordinates $(x_i,f(x_i))$. Then we replace the integrals entering into the systems of equations (1), (2) with the sums of the integrals over the subintervals and set $M=M_{i+1/2},\quad i=1,...,N$, $M_{i+1/2}$ are collocation points. The sought solution on the section $S_{P_j,P_{j+1}}$, $j=0,...,N-1$, is approximated by a zero-order spline. As a result, we obtain a complex system of linear algebraic equations. When calculating the integrals over a section $S_{P_j,P_{j+1}}$, in which the integrand depends on either $g^E(M,P)$ or $\frac{\partial g^E(M,P)}{\partial n_P}$, we replace the integration with respect to s with the integration with respect to x by setting $ds_P=\sqrt{1+(f'(x))^2}\,dx_P$.

Note that the terms that enter the kernels of integral equations and depend on $g^E(M,P)$ have a logarithmic singularity when the arguments coincide. Therefore, when developing the numerical algorithm, this singularity is separated in an explicit form and the integrals containing it are calculated analytically.

The normal derivatives of the function $g^E(M,P)$ have no singularity when the arguments are equal, and

$$\lim_{P\to M}\frac{\partial g^E(M,P)}{\partial n_p}=\frac{1}{2}\kappa_0(M),$$

where $\kappa_0(M)$ is the curvature of the contour S at the point M.

The kernels of the system depending on $\frac{\partial^2 g^E(M,P)}{\partial n_M\,\partial n_P}$ have a strong singularity. In order to calculate the matrix elements that are integrals of the functions with a hypersingularity, it is necessary to develop a numerical algorithm. To this end, we preserve the integration over the arc s in these integrals. We defined the surface S in the form $x=x(s),\quad y=y(s)$. The outward normal to the domain D_1 we write in the form $\vec{n}=-y_s'\,\vec{i}+x_s'\,\vec{j}$, where $\vec{i}$ and $\vec{j}$ are the x and y unit vectors. Note that if the boundary of *S* is specified in the form $y=f(x)$, the relationship between the directional cosines $n_x;n_y$ of the outer normal $\vec{n}$ to D_1 is given by

$$y_S'=-n_x\sqrt{1+(f'(x))^2},\qquad x_S'=n_y\sqrt{1+(f'(x))^2},$$
$$n_x=f'(x)/\sqrt{1+(f'(x))^2},\quad n_y=-1/\sqrt{1+(f'(x))^2}.$$

We have

$$\int_{S_{P_j,P_{j+1}}}\frac{\partial^2}{\partial n_M\partial n_P}g^E(M_{i+1/2},P)ds_P=$$
$$\frac{i\pi}{2}\int_{S_{P_j,P_{j+1}}}\frac{\partial^2}{\partial n_M\partial n_P}H_0^{(1)}(\beta_1 r_{M_{i+1/2},P})ds_P-\qquad (3)$$
$$\frac{i\pi}{2}\int_{S_{P_j,P_{j+1}}}\frac{\partial^2}{\partial n_M\partial n_P}H_0^{(1)}(\beta_1 \hat{r}_{M_{i+1/2},P})ds_P.$$

To develop a special computing algorithm, every integrands is represented by a sum of two terms: the first one is the total differential of some function has been calculated analytically; the function in the second term has a logarithmic singularity and is integrated by separating this singularity in the explicit form, thereafter the corresponding integrals are calculated analytically.

In the first term $\partial/\partial n_M$ let us consider the component $\partial/\partial x_M$ (without the multiplier $i\pi/2$):

$$\int_{S_{P_j,P_{j+1}}}\frac{\partial^2}{\partial x_M\partial n_P}H_0^{(1)}\left(\beta_1 r_{M_{i+1/2},P}\right)ds_P=$$

$$-\int_{S_{P_j,P_{j+1}}}\frac{\partial}{\partial n_P}\frac{\partial}{\partial x_P}H_0^{(1)}\left(\beta_1 r_{M_{i+1/2},P}\right)ds_P=$$

$$-\int_{S_{P_j,P_{j+1}}}d\left(\frac{\partial}{\partial y_P}H_0^{(1)}\left(\beta_1 r_{M_{i+1/2},P}\right)\right)-$$

$$\beta_1^2 \int\limits_{S_{P_j,P_{j+1}}} H_0^{(1)}\left(\beta_1 r_{M_{i+1/2},P}\right) y'_{S_P} d s_P + 4i \int\limits_{S_{P_j,P_{j+1}}} \delta(M,P) y'_{S_P} ds_P = -\frac{\partial}{\partial y_P} H_0^{(1)}\left(\beta_1 r_{M_{i+1/2},P_{j+1}}\right) + \frac{\partial}{\partial y_P} H_0^{(1)}\left(\beta_1 r_{M_{i+1/2},P_j}\right) - \beta_1^2 \int\limits_{S_{P_j,P_{j+1}}} H_0^{(1)}\left(\beta_1 r_{M_{i+1/2},P}\right) y'_{S_P} ds_P + 4i \int\limits_{S_{P_j,P_{j+1}}} \delta(M,P) y'_{S_P} ds_P. \quad (4)$$

The points M and P can coincide, hence, there is an integral of the Dirac delta-function.

Similarly let us consider the component $\partial / \partial y_M$ in the first term $\partial / \partial n_M$:

$$\int\limits_{S_{P_j,P_{j+1}}} \frac{\partial^2}{\partial y_M \partial n_P} H_0^{(1)}\left(\beta_1 r_{M_{i+1/2},P}\right) ds_P = -\int\limits_{S_{P_j,P_{j+1}}} \frac{\partial}{\partial n_P} \frac{\partial}{\partial y_P} H_0^{(1)}\left(\beta_1 r_{M_{i+1/2},P}\right) ds_P = \int\limits_{S_{P_j,P_{j+1}}} d\left(\frac{\partial}{\partial x_P} H_0^{(1)}\left(\beta_1 r_{M_{i+1/2},P}\right)\right) + \beta_1^2 \int\limits_{S_{P_j,P_{j+1}}} H_0^{(1)}\left(\beta_1 r_{M_{i+1/2},P}\right) x'_{S_P} d s_P - 4i \int\limits_{S_{P_j,P_{j+1}}} \delta(M,P) x'_{S_P} ds_P = \frac{\partial}{\partial x_P} H_0^{(1)}\left(\beta_1 r_{M_{i+1/2},P_{j+1}}\right) - \frac{\partial}{\partial x_P} H_0^{(1)}\left(\beta_1 r_{M_{i+1/2},P_j}\right) + \beta_1^2 \int\limits_{S_{P_j,P_{j+1}}} H_0^{(1)}\left(\beta_1 r_{M_{i+1/2},P}\right) x'_{S_P} ds_P - 4i \int\limits_{S_{P_j,P_{j+1}}} \delta(M,P) x'_{S_P} ds_P. \quad (5)$$

Using (4)-(5) and the recurrence relations for Hankel functions

$$\frac{\partial}{\partial x_P} H_0^{(1)}\left(\beta_1 r_{M,P}\right) = \beta_1 H_1^{(1)}\left(\beta_1 r_{M,P}\right) \frac{x_M - x_P}{r_{M,P}};$$

$$\frac{\partial}{\partial y_P} H_0^{(1)}\left(\beta_1 r_{M,P}\right) = \beta_1 H_1^{(1)}\left(\beta_1 r_{M,P}\right) \frac{y_M - y_P}{r_{M,P}},$$

we obtain the final formula for the calculation of the second normal derivative of the Green function [2].

2.4. Calculation of the reflected field pattern for the E-polarization

Using the solution to the system of integral equations (1)-(2) and taking into account the boundary conditions on the section S, we can obtain the integral representation for the field $u_1(M) = u(M) - u_0^E(M)$:

$$u_1(M) = \frac{1}{2\pi} \int\limits_S \left\{ \left[g^E(M,P) + \frac{i k_1 W_2}{\beta_1^2 W_1} \frac{\partial g^E(M,P)}{\partial n_P} \right] \frac{\partial u(P)}{\partial n_P} + \frac{i \gamma_0 W_2}{\beta_1^2} \frac{\partial g^E(M,P)}{\partial n_P} \frac{\partial v(P)}{\partial \tau_P} \right\} d s_P, \quad M \in D_1. \quad (6)$$

We introduce spherical coordinate frame $r, \overline{\vartheta}, \overline{\varphi}$, where angle $\overline{\vartheta}\ (-\pi/2 < \overline{\vartheta} < \pi/2)$ is measured from the y axis and angle $\overline{\varphi}\ (0 < \overline{\varphi} < \pi/2)$ is measured from the z axis. We seek for a representation of the scattered field for $|\vec{r}|_M >> |\vec{r}|_P$, $k_1 |\vec{r}|_M >> 1$. Following that, we introduce the notation $|\vec{r}|_M = \rho$. It can be shown that the following asymptotic relationships $r_{MP} \approx \rho - (\vec{r}_P, \vec{e}_M)$ and $\hat{r}_{M\hat{P}} \approx \rho - (\vec{\hat{r}}_{\hat{P}}, \vec{e}_M)$ are valid. Here, $\vec{e}_M$ is a unit vector oriented in the direction to point M; i.e. $\vec{e}_M = \vec{r}_M / \rho = (\sin\overline{\vartheta}\sin\overline{\varphi}, \cos\overline{\vartheta}\sin\overline{\varphi}, \cos\overline{\varphi})$. Taking into account the above relationships and asymptotic behavior of the Hankel function, we can obtain the asymptotic representations for $g^E(M,P)$, $\frac{\partial g^E(M,P)}{\partial n_P}$ in the form

$$g^E(M,P) \approx \sqrt{\frac{\pi}{2\beta_1\rho}} \exp\left(i\left(\beta_1\rho + \frac{\pi}{4}\right)\right) \times \left[\exp\left(-i\beta_1(\vec{r}_P, \vec{e}_M)\right) - \exp\left(-i\beta_1(\vec{\hat{r}}_{\hat{P}}, \vec{e}_M)\right)\right];$$

$$\frac{\partial g^E(M,P)}{\partial n_P} \approx \sqrt{\frac{\pi}{2\beta_1\rho}} \exp\left(i\left(\beta_1\rho + \frac{\pi}{4}\right)\right) \times \left[-i\beta_1(\vec{n}_P, \vec{e}_M)\exp\left(-i\beta_1(\vec{r}_P, \vec{e}_M)\right) + i\beta_1(\vec{n}_{\hat{P}}, \vec{e}_M)\exp\left(-i\beta_1(\vec{\hat{r}}_{\hat{P}}, \vec{e}_M)\right)\right] \quad (7)$$

and also for $u_1(M)$[3].

We call quantity $F^E(\overline{\vartheta},\overline{\varphi})$ that equals

$$F^E(\overline{\vartheta},\overline{\varphi})=$$

$$\frac{1}{\sqrt{\beta_1}}\int_S\Big[\Big\{\Big[\exp\big(-i\beta_1(\vec{r}_P,\vec{e}_M)\big)-\exp\big(-i\beta_1(\vec{\hat{r}}_{\hat{P}},\vec{e}_M)\big)\Big]+$$

$$\frac{k_1W_2}{\beta_1W_1}\Big[(\vec{n}_P,\vec{e}_M)\exp\big(-i\beta_1(\vec{r}_P,\vec{e}_M)\big)-$$

$$(\vec{n}_{\hat{P}},\vec{e}_M)\exp\big(-i\beta_1(\vec{\hat{r}}_{\hat{P}},\vec{e}_M)\big)\Big]\Big\}\frac{\partial u(P)}{\partial n_P}+$$

$$\frac{\gamma_0W_2}{\beta_1}\Big[(\vec{n}_P,\vec{e}_M)\exp\big(-i\beta_1(\vec{r}_P,\vec{e}_M)\big)-$$

$$(\vec{n}_{\hat{P}},\vec{e}_M)\exp\big(-i\beta_1(\vec{\hat{r}}_{\hat{P}},\vec{e}_M)\big)\Big]\frac{\partial v(P)}{\partial\tau_P}\Big]ds_P$$

the pattern of scattering field $u_1(M)$. Knowing the solution of the system of singular integral equations (1)-(2), we calculate the currents induced on the irregular surface and far-field pattern for the scattered field in the far zone [4].

3. Numerical results

The induced currents and reflected electromagnetic field patterns are calculated for *E*- polarization, $\varepsilon_2=50+4i$, contour S specified in the XOY plane by the equation $f(x)=1-\cos x$, $0\le x\le 2\pi$, $a=2\pi$.

Figure 2 shows the distribution of the calculated function of the induced current $I(x)=|\partial u(P)/\partial n|$ on the boundary S along the x axis for $ak_1=3\pi$, $a/\lambda=1.5$, where λ is the wavelength of the incident field. The results are obtained for various angles ϑ_0 and φ_0 of the electromagnetic wave incidence: (curve *1*) $\vartheta_0=0$, $\varphi_0=85^0$; (curve *2*) $\vartheta_0=30^\circ$, $\varphi_0=60^0$; (curve *3*) $\vartheta_0=45^\circ$, $\varphi_0=45^0$; (curve *4*) $\vartheta_0=60^\circ$, $\varphi_0=30^0$.

Figures 3-4 show for $ak_1=3\pi$, $a/\lambda=1.5$ reflected field pattern $|F^E(\overline{\vartheta},\overline{\varphi})|$ in polar coordinates for the range $-85^\circ<\overline{\vartheta}<85^\circ$ and fixed values of $\overline{\varphi}$. Curves *1-5* correspond to $\overline{\varphi}=85^\circ, 60^\circ, 45^\circ, 30^\circ, 5^\circ$. Figures 3-4 represent the results of calculations for the reflected field when a plane wave incident at the angles $\vartheta_0=0$, $\varphi_0=85^\circ$ (Figure 3) and $\vartheta_0=45^\circ, \varphi_0=45^\circ$ (Figure 4).

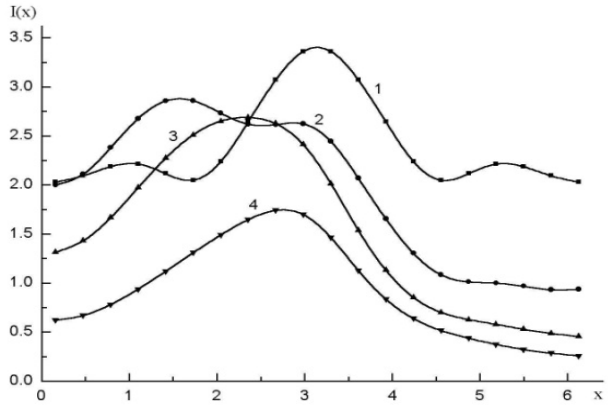


Figure 2

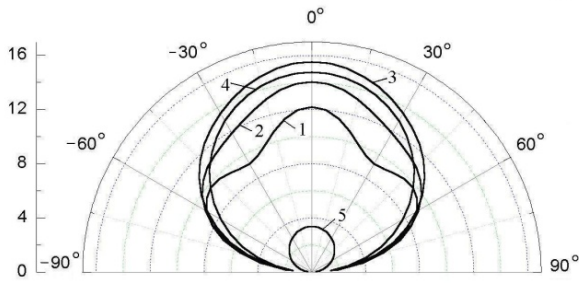


Figure 3

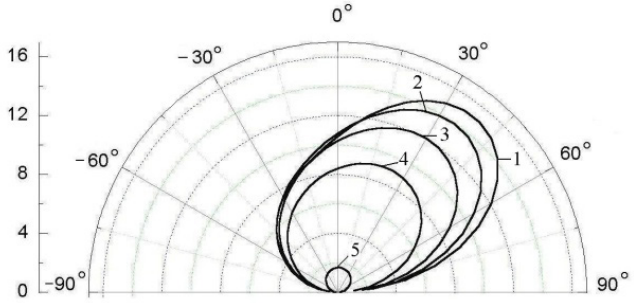


Figure 4

Further, the analogous results are presented on Figures 5-7 for $ak_1=5\pi$, $a/\lambda=2.5$.

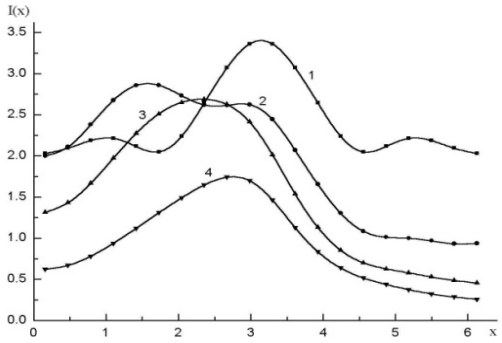


Figure 5

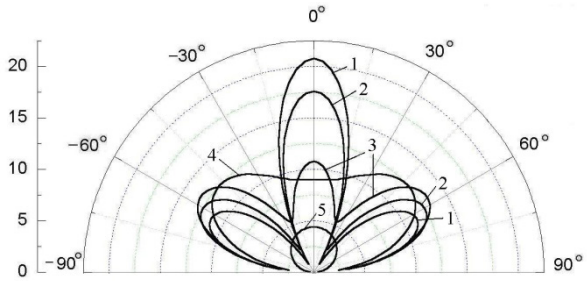


Figure 6

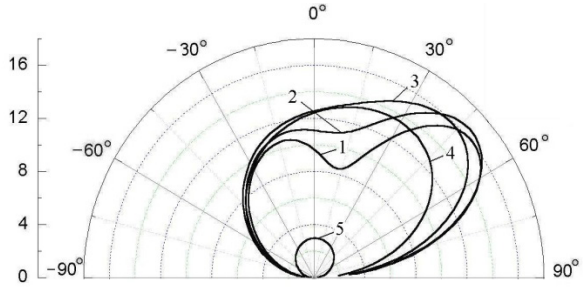


Figure 7

4. Conclusion

1. The exact model of plane electromagnetic wave incident on inhomogeneous conductive surface is investigated.
2. The problem is reduced to the system of singular integral equations. The analytic solution is provided for a singularity in the kernels of integral equations.
3. The numerical algorithm for calculation of scattering characteristics of incident electromagnetic field was developed and calculation examples were presented.

5. References

[1] A. S. Ilinskiy, T. N. Galishnikova, “Singular Integral Equations in the Wave Scattering Problems”. Proc. Of URSI Commission B International Symposium on Electromagnetic Theory (EMTS 2016), 14-18 August, 2016, Espoo, Finland. Pp. 635-637 (electronic publication).

[2] A. S. Il’inskii and T. N. Galishnikova, “The Integral Equation Method in Problems of Diffraction by a Finite Impedance Section of a Medium Interface”, Moscow University Computational Mathematics and Cybernetics (Allerton Press), vol. 32, no. 4, pp.187-193, 2008.

[3] A. S. Il’inskii and T. N. Galishnikova, “Investigation of Diffraction of an Electromagnetic Wave Arbitrarily Incident on a Locally Inhomogeneous Medium Interface”, Journal of Communications Technology and Electronics, vol. 58, no. 1, pp. 40-47, 2013.

[4] A. S. Il’inskii and T. N. Galishnikova, “Integral Equation Method in Problems of Electromagnetic-Wave Reflection from Inhomogeneous Interfaces between Media”, Journal of Communications Technology and Electronics, vol. 61, no. 9, pp. 981-994, 2016.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 2, No. 3, 88-99 (2017)
www.astesj.com

ASTES Journal
ISSN: 2415-6698

Methodology for Management of Information Security in Industrial Control Systems: A Proof of Concept aligned with Enterprise Objectives.

Fabian Bustamante[1], Walter Fuertes[*,1], Paul Diaz[1], Theofilos Toulqueridis[1]

[1]*Universidad de las Fuerzas Armadas ESPE, Department of Computer Sciences, ZIP Code: 17-15-231-B, Sangolqui, Ecuador*

ARTICLEINFO

Article history:
Received: 05 March, 2017
Accepted: 04 April, 2017
Online: 09 April, 2017

Keywords :
Industrial Control Systems,
Information Security,
Automation,
ISMS,
SCADA,
DCS,
PLC,
ISO 27000,
NIST

ABSTRACT

This article is an extended version of the study presented at the IEEE Ecuador Technical Chapters Meeting (ETCM)-2016. At that time, a methodological proposal was designed, implemented, and applied in a group of industrial plants for the management of the information security of the Industrial control systems (ICS). The present study displays an adaptation and improvement of such methodology with the purpose of aligning the proposal for the effective management of information security with the strategic objectives. The development of this study has been divided into three distinctive phases. Firstly, we induced the articulation of PMI-PMBOK v5 and ITIL v3 both for the management of the project and for the verification of risks in the IT services. Second, we applied a set of risk mitigation strategies based on international standards as NIST 800-82 and 800-30. Thirdly, we assembled the two mentioned phases in a Guide for standards-based instructions and security policies, which previously have been encouraged on NIST 800-82, 800-53 and 800-12. Hereby, we observed the reduction of incidents of information security, the correct delimitation of the functions of the direct responsible of the ICS and the improvement of the communication between the operative and technical areas of the involved companies. The results demonstrate the functionality of these improvements, especially in the context of the availability and integrity of information, which generates an added value to the enterprise.

1 Introduction

According to the Guide to Industrial Control Systems (ICS) Security NIST-2nd Revision [1], "threats to control systems may be originated from numerous sources, including hostile governments, terrorist groups, disgruntled employees, malicious intruders, complexities, accidents, and natural disasters as well as malicious or accidental actions by insiders. ICS security objectives typically follow the priority of availability and integrity, followed by confidentiality".

Facing these issues, the industry has proposed several studies as recommended by [2] where best practices and risk assessment of the ICS have been suggested. In [3] authors documented about Supervisory Control and Data Acquisition (SCADA) system security. In [4] enlightens the myths and facts behind Cyber security in ICS, while in [5] [6] [7] [8] [9], authors present issues, methods and countermeasures to the protection of information, control systems as well as automation. In [10] exposes the research about SCADA security in the light of Cyber-Warfare. In recent studies such as [11] examines the State of the Art in Cyber security risk assessment methods for SCADA systems. In [12] [13] proposes a study on efficiency of an Information Security Management System (ISMS) for ICS with compliance. In [14] presents a formal approach that automates the security analysis of ICS. In [15] explains a process of defense of securing ICS. In [16] presents a theory of security metrics in SCADA

[*]Walter Fuertes, Gral. Ruminahui Ave, Contact No:00593-987392793 & Email: wmfuertes@espe.edu.ec

https://dx.doi.org/10.25046/aj020313

and ICS including resilience. In [17] [18] [19] [20] [21] authors explore the ICS cybersecurity landscape including threats, vulnerabilities, intrusions and cyberattacks. In [22] discusses that the wireless systems in all ICS are subject to cybersecurity vulnerabilities. In [23]gives an analysis of behavior, time, and adaptation in the assessment of ICS Cybersecurity Risk. In all mentioned studies, general recommendation and analysis on the ICS were discussed, but unfortunately they missed to propose a management system of information security. The standards listed in the mentioned studies are predominantly: NIST 800-53, ISA 99, NERC, ISO 27001, ISO 27002 [24], ISO 31010, and ISO 27019.

Despite having controls and procedures in the management of information security of its ICS, the manufacturing companies, fail to reduce and resolve the incidents. In addition, there is little or no evidence of a complete vulnerability, threat and risk analysis of their critical assets and services. This may be attributed to the high level of sophistication of the ICS and the poor knowledge of the present personnel that manages and carries out such tasks [10].

Thus, the main contribution of this research has been to provide a methodology to manage information security in ICS. This has been focused on both, Information Technology (IT) professionals and manufacturing automation professionals. Its stages and methods have been consistently combined with the standards outlined above and are internationally used in traditional IT and ICS. Specifically, these contributions are: (1) Novel guides for the management of the project of implementation of the methodology; (2) Updated methodologies for risk assessment in IT services used in ICS; (3) Methodological procedures to encounter new risk mitigation strategies; and finally, (4) Methodological procedures to develop the manual of security standards and policies for ICS.

Based on the findings and conclusions presented by [24] [26], we agree and confirm that NIST fulfills all criteria needed and has been therefore chosen to be ideally for this study, besides the existence of two further interesting options such as the North American Electrical Reliability Corporation / Critical Infrastructure Protection (NERC / CIP). They represent a set of requirements designed to secure the assets required to operate in North America and the International Automation Society (ISA), being is a series of standards, technical reports and related information that defines procedures for Systems Automation and Control Electronics (IACS). Therefore, we have opted for a reliable standard like NIST stands for, which is not exclusively framed in the organization nor in the technical part only. This allows the possibility to manage in an effective and holistic way different areas that coexist in the SCI being: business, computing, electronics, automation, production processes and people. It is also recognized that NIST is internationally highly disseminated and its best practices exist supplementary information that may be available to the professional stakeholders of this project.

This document is an extended version of a previous conference presentation [27], which has been based specially on NIST. In this present article, areas of knowledge of the PMBOK and ITIL standards have been also included, which served to improve the management of the methodology implementation projects and in the risk analysis of the IT services operating in the ICS. Furthermore, we added detailed evidence and discussions of the results presented in [27].

The remainder of this paper is organized as follows: The conceptual framework that describes the theoretical foundation of this study is described in Section 2. Then, the adaptation and improvement of the methodology proposed is explained in section 3. Results and discusses of findings and their implications are displayed in Section 4. Finally, Section 5 closes with conclusions and some future research lines.

2 Theoretical Framework

This section comprises all the important theory and knowledge about the ICS, tools for IT architecture modeling and expert system validation method that have been used in this study:

2.1 Industrial Control Systems (ICS).

As reported by the Guide to Industrial Control Systems (ICS) Security of the National Institute of Standards and Technology (NIST) [1], the term Industrial Control System includes Supervisory Control and Data Acquisition (SCADA), Distributed Control Systems (DCS), as well as programmable logic controllers (PLCs). ICS are used in industries such as electricity, water distribution systems, wastewater, petroleum, natural gas, chemical, transportation, pharmaceutical, pulp, paper, food, automotive, aerospace, among many others. On the other side, SCADA systems are also used to control critical infrastructure. As stated by ENISA in [28] critical infrastructure such as electricity generation plants, transportation systems, oil refineries, chemical factories and manufacturing facilities means an asset, system or part thereof located in a country which is essential for the maintenance of vital societal or strategic functions, such as health, safety, security, economic or social well-being of people, and the disruption or destruction of which would have a significant impact (e.g. disruption to business operations and services but also potential damage and destruction of equipment). As a final point, DCS are used to control production systems within a local area such as a factory using supervisory and regulatory control. PLCs are used for discrete control for specific applications and generally provide regulatory control [1].

2.2 Security incidents in Industrial Control Systems.

The proliferation of connections that use the electromagnetic spectrum, combined with vulnerable source code and errors in configurations of automation systems expose the ICS to potential from cyberspace. Weak IT security policies and weaknesses in the security features of automation systems dramatically increase the risk of a successful cyber-attack [29].

Facing such issues, there are some organizations authorized to investigate and combat cybersecurity incidents. For instance, in the United States are the Cybersecurity Emergency Response Team (ICS-CERT) [30] and the Federal Bureau of Investigation (FBI). ICS-CERT in 2015, received and responded to several incidents reported by owners and manufacturers of industrial products in North America. Other reports considered are also those reported by manufacturers and suppliers of equipment, industrial systems, and information security solutions such as: ABB, Allen Bradley, Verizon [31], among others.

In the same context [27] [28] discloses the current maturity level of ICS-SCADA cybersecurity in Europe and identifies good practices handled by European Member States to improve this area. The first and second part of this study introduces the ICS-SCADA cyber security topic, explains the role of ICS-SCADA in critical sectors and summarizes the methodology of this research.

2.3 Guide for Industrial Control System Security NIST 800-82.

The NIST Industrial Control System Safety Guide 800-82 establishes some guidelines for the implementation of safety in ICS. The main reason for using the regulations for an ICS is that these increasingly adopt computer components in their design and operation. While some features are similar between traditional IT systems and ICS, it is fundamental to understand that they are unable to be handled in a similar procedure. ICS and IT include significant risks such as health, protection of human life, environmental damage, financial impact, production losses, etc., [1].

This document is the second revision to NIST SP 800-82 which has been published in March of 2015. Updates in this revision include: ICS threats and vulnerabilities; ICS risk management, recommended practices and architectures; current activities in ICS security; security capabilities and tools for ICS; Additional alignment with other ICS security standards and guidelines; new tailoring guidance for NIST SP 800-53, among others [1].

Risk is present, when the probability exists of an occurrence of a threat or any vulnerability. Threats are able to take advantage of vulnerabilities. In the case of a computer incident, for example, the magnitude of the potential impact results from a successful exploitation of vulnerability [32] and [33] should be determined. Risk assessment, in turn, is the process of identifying the risks of an organization's operations, assets, and individuals by evaluating the likelihood of an identified vulnerability having a potential impact. The risk assessment also has to compare the costs of such safety with those expenses or financial damages associated with a possible incident.

2.4 Guide to the Project Management Body of Knowledge (PMBOK® Guide)–Fifth Edition.

The Guide to the Project Management Body of Knowledge provides guidelines for the direction of individual projects and defines concepts related to project management. It also describes the life cycle of project managements and related processes, as well as the project life cycle. The PMBOK® Guide contains the globally recognized standard and guidance for the project management profession. By standard means a formal document that describes established standards, methods, processes and practices. Similar to other professions, the knowledge contained in this standard evolved from the recognized good practices of the professionals dedicated to the management of projects that have contributed to its development. Acceptance of project management as a profession indicates that the application of knowledge, processes, skills, tools and techniques may have a considerable impact on the success of a project [34].

2.5 IT Infrastructure Library (ITIL)

ITIL is part of a suite of best-practice publications for IT service management (ITSM). ITIL provides guidance to service providers on the provision of quality IT services, and on the processes, functions and other capabilities needed to support them. Organizations are encouraged to adopt ITIL best practices and to adapt them to work in their specific environments in ways that meet their needs. ITIL is not a standard that is mandatory to be followed, it is more likely a guidance that should be read, understood, and used to create value. The ITIL framework is based on the five stages of the service lifecycle, namely Service Strategy, Service Design, Service Transition, Service Operations and Continual service improvement [34] [35] [36].

2.6 Control Objectives for Information and Related Technology (COBIT).

The COBIT provides a comprehensive framework that supports enterprises to achieve their goals and deliver added value through effective governance as well as IT governance of the Organization. In this way, it supports organizations to create optimal value from IT, by maintaining a balance between realizing benefits and optimizing levels of risk and resource utilization. This allows IT and related organizations to be governed and administered in a holistic manner throughout the Organization. This includes the full scope of

all functional and business areas of responsibility, taking into account the internal and external interests of IT stakeholders.

3 Methodology Proposal.

3.1 Research Methodology

During this project, we have used the Action research Methodology [37], also known as participatory research which is a testing of theories developed to overcome an immediate on the job-difficulty. In short, action research means learning by doing. In this kind of approach first a solution is devised and evaluated, while depending of the results, a new solution may be constructed in order to try to achieve more sophisticated results. These processes are cyclically repeated until an acceptable performance has been obtained. In one of its formal stages it is established the gathering of pertinent data, material, methods, techniques, and so on for setting into testing periods. All data should be relevant to the problem. Therefore, it has been pertinent for us, to use PMBOK v5 and ITIL PRACTITIONER in their respective areas of knowledge called Communication. As collaborative instruments, workshops were developed with ICS experts as a focus group. The next stage has been to concern the developing plans for implementation of a theory.

The research has been carried out under real conditions of four industrial plants of the principal manufacturing enterprises in their markets. These have approximately 600 employees in the manufacturing area and some 400 more, in the administrative area. The four plants were comprised of two production lines containing four distributed control systems (DCS), 120 Allen Bradley and Siemens PLCs, 40 HMIs Allen Bradley, 48 desktop computers, 15 industrial applications using Rockwell Automation and Wonderware software (including 5 critical mission) and a Local and Wide area networks (LAN/WAN) with Cisco equipment and Checkpoint. Reference standards such as: NIST 800-30, NIST 800-82, NIST 800-53, NIST 800-12, PMBOK v5, ITIL v3, interviews, statistical processes and cost-benefit analysis have been used among the most important ones.

3.2 Modeling of IT frameworks & Methodological proposal.

In order to improve the methodology proposed in our previous study [27], we have considered the adaptation of other standards, referential frameworks, and management systems that are used in Information Technology (IT). With the purpose to understand the context, we have started by modeling the indicated in a hierarchical pyramid that is illustrated in Figure 1. In the base stand the standards used in traditional IT: COBIT, ISO 27000, PMBOK and ITIL. The second level of the pyramid holds NIST, NERC and ISA. At the third level is the proposed methodology. From bottom to top, these standards reveal ICS practitioners "how to do it?", while from top to bottom urges the question of: "what to do?"

Firstly, as part of the COBIT IT governance process, the "BAI01" (Building, Acquiring and Implementing, BAI), called Program and Project Management has been included. There, COBIT indicates "what to do?" According to this process, all the investment portfolio programs and projects are strongly recommended to be managed in a coordinated way and in line with the corporate strategy. In a similar case, chapter 13 of [34] indicates "what to do?" to identify project stakeholders. In the case of this proposed management, the "how to do?" procedure is indicated to discover the stakeholders required at each step of the implementation.

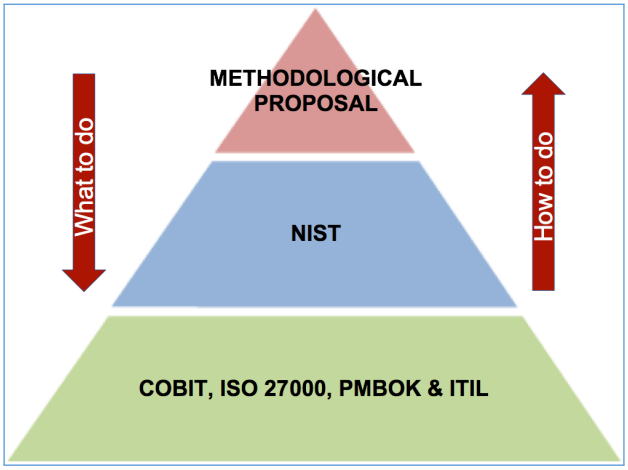


Figure 1: Modeling of IT frameworks & Methodological proposal

Secondly, [1] indicates "what to do?" to evaluate and mitigate ICS risks. In the case of the management proposal for ICS security, it is indicated "how to do it?"

Although COBIT, ISO27000, PMBOK and ITIL are not fully implemented in the case study, it is necessary to use some of their areas of expertise and refer them to fit the methodological proposal into a form, which is already worldwide established. This avoids the occurrence of inaccuracies of concepts that in the future may not articulate in the management of the information security at managerial, operational, and technical level.

3.3 Project to improve the methodological proposal

For the improvement of this methodology and because its implementation requires to perform a considerable amount of tasks with specific deliverables with a defined time, in which the quality of the information is fundamental, having some stakeholders in account, it has become necessary to include in the methodology standards in the field of project management, which help to keep critical success factors controlled (i.e. cost, quality, scope and time). The main

reason for adopting PMBOK [34] has been the better adaptability with NIST, its interdependence and the widespread diffusion that this standard has had in project management.

Prior the implementation of the methodology in an enterprise, several meetings with the responsible of the ICS must be realized. It starts with the collection of information that serves to identify and analyze stakeholders. In line with [34] this point plays a fundamental role in the development and success of a project. Later, a variety of points should be reviewed such as: names and position of who will be the project manager of implementation, identification of the problem to be solved, business need to solve the problem, justifications that lead to implement the methodology in the enterprise, levels of authority of Project Manager, risks that may exist in the implementation, opportunities that the enterprise has with this project, general deliverables, specific deliverables, and pre-allocated resources (i.e., financial, material and human resources). With all this information two documents need to be elaborated to begin the project of implementation of the methodology such as Project Charter and Power-Interest Matrix Stakeholder (see Figure 2). In consonance with [34] the Charter Project must be signed before beginning the implementation of the methodology in order to formalize the support and participation of the entire enterprise. The signers of the charter project are the Project Manager and the Chief Executive Officer (CEO), who is the first and most important stakeholder.

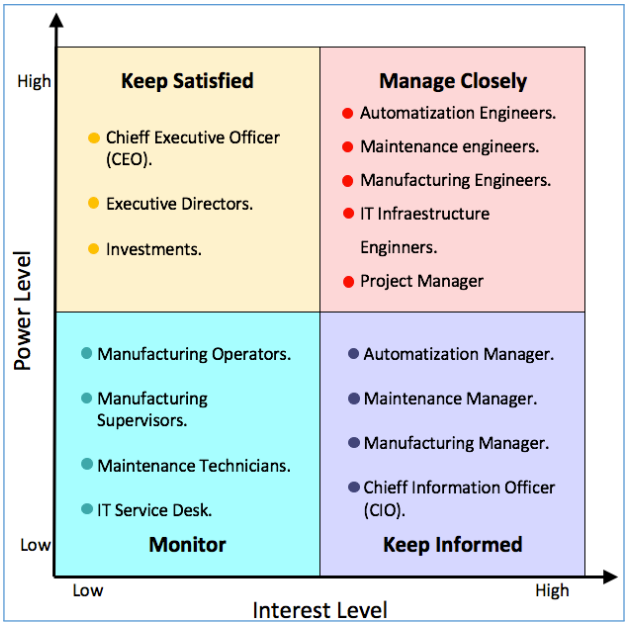


Figure 2: Power-Interest Matrix Stakeholders of the implementation project.

As stated in [34], the analysis of the power-interest matrix stakeholders illustrated in Figure 2, allows identifying who are the experts of the ICS. Moreover, there it's stated that all the workshops and meetings that are needed to obtain the information, which will be vital for the success of the implementation, must be carried out.

Figure 2 illustrates that these experts are those located in the quadrant entitled "Manage Closely" (e.g. Engineers of automation, maintenance, manufacturing, and IT Infrastructure Engineering Engineers). Shown in the "Keep Satisfied" quadrant, investors, executive directors, and the CEO will be placed, because they are only interested in the project for adequate results for the enterprise. Besides, in the "Keep Informed" quadrant, there are all those who are interested in implementing the methodology, but are not directly involved within. Finally, in the "Monitor" quadrant, the managers are placed in charge of monitoring what is being implemented and when the methodology is already in place to notify about the incidents and problems that may appear.

Given the experience gained in the implementations acquired by the research team, it is recommended to meet with the project implementation team and the Project Manager to assess progress and review the project's risk status once a week. As inputs, surveys should be carried out by the experts of each workshop and meeting, with whom the methodology is being implemented. In section 4, evaluation of results and discussion implies such behavior. In order to obtain the results of the surveys carried out by the experts, the following formula has been used (see equation 1):

$$P_m = \frac{\sum_{n=1}^{q} (R_m)_n}{q} \; (m=1,\ 2,\ 3,4) \qquad (1)$$

Where "P" is the average of the answers of question number "m"; "n" is the workshop session number; "q" is the number of the last workshop; and "R" is the qualification the expert gave to each question.

Once the adaptation and improvement of the proposed methodology is planned, starts the implementation of the methodology, being divided into three phases represented in Figure 3. A brief schematic description is provided below:

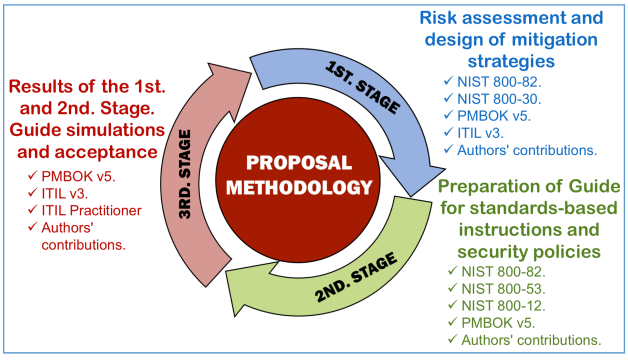


Figure 3: Stages of the methodology proposed, as adapted by [27]

Unlike the previous version presented in [27] and as illustrated in Figure 3, in this extended version, areas of knowledge of both PMBOK and ITIL have been added in blue colors in the three stages of the proposed methodology. In stage 1, the PMBOK, NIST

and ITIL stakeholder management were used to assess risks in ICS components and services, design of mitigation strategies, including the authors' contribution. In the second phase, only PMBOK stakeholder management are incorporated together with what NIST suggests in the preparation of the guide for standards-based instructions and security policies, as well as the contribution of the authors. Stage 3 includes PMBOK stakeholder management, ITIL service strategy catalogs, ITIL service design and communication management suggested by ITIL practitioner, and the authors' contributions.

3.4 Risk assessment and design of mitigation strategies for ICS

The first process in the methodology of risk management is the evaluation of threats and vulnerabilities, according to [1]. To perform such task, the subsequent steps are followed:

Step 1. The characterization of the system. In agreement with [34] it is determined with the whole system of hardware, software, connectivity, data, personnel support and execution processes. They also take into account the revision of existing documentation, the use of exploration tools, questionnaires and stakeholder interviews of the "Monitor" and "Manage closely" quadrants of the Power-Interest matrix (Figure 2).

Step 2. Identification of threats. For this step, we have drawn a table with sources based on threats [25], motivations, actions that would yield these threats and a comment of the stakeholders if necessary. Those responsible for conducting workshop sessions from step 2 through step 8 are the stakeholders of the "Manage closely".

Step 3. Identifying Vulnerabilities: Interviews were conducted, information related to recent incidents has been gathered and subsequently analyzed. Furthermore, documentation risks of the enterprise, I-CAT NIST database vulnerability, and security requirements were check-listed and later analyzed according to [35]. Although other ways to identify vulnerabilities have been available as proposed by [32], we have used a simplified version of the proposed methodology as suggested by [30].

Step 4. Analysis of Controls: A variety of control methods, control categories, and analysis techniques of control according to [24] were taken into account. The results have been documented due to the realization of workshops.

Step 5. Determination of Probability (P): The motivation of the source of threat (TM), the nature of vulnerability and the effectiveness of controls (EC) according to [24] were considered. In this respect:

1. If TM=High and EC=low, then P=1 (High).
2. If TM=medium and EC=medium, then P=0.5 (Medium).
3. If TM=low and EC=high, then P=0.1 (Low).

Step 6. Determination of Impact: The mission of the ICS has been to review the data criticality and to respect the sensitivity of the data and the information security [38] within the enterprise. In the present case study, the mission, the vision, the strategic plan of the enterprise and the documentation ISO 9001: 2015 held by the enterprise were analyzed.

Step 7. Risk Assessment (R): The risk assessment has been estimated according to the probability of threat sources (P) and the corresponding magnitude of the impact (I). Thus: R = P x I. The risk scale has been used by [24]: being HIGH when 100> value> 50, MEDIUM when 50> value> 10 and LOW when 10> value> 1.

Step 8. Recommended Controls: The recommended controls aim to reduce the vulnerability and risk level of the ICS and its data to an acceptable level for the enterprise, which needs to take into account the lowest possible costs and impacts.

Step 9. Documentation Result: A final report about the achieved results has been developed and afterwards addressed to the stakeholders, being rather systematic and analytical than accusatory in any form. However, finally in the present case study an executive summary has been elaborated and addressed to the CEO. Furthermore, the methodology for risk mitigation has been performed according to [24] (Figure 4).

In order to complement and verify that the ICS risk analysis has been accomplished, it is further recommended to use a risk assessment with respect of computer services in the ICS. Therefore, it is necessary to pick up the ICS services catalog as indicated [35] and [36]. To evaluate the risks in the IT services at ICS, a brainstorming session is held with the experts, without losing sight of the services catalog. In this way it is possible to analyze the probability and the impact in each of them, in the same way that it has been performed with the components of the ICS.

Regarding the data characteristics of the experimentation, we may point out, that for the accomplishment of the present study, confidentiality agreements have been signed with the enterprise of the case study. Consequently, this prevents the disclosure of detailed information on the characteristics of the components of their ICS. However, it should be stressed out, that regardless of the technology used, the proposed methodology would work properly as it has been focused on both management and technical criteria. Even so, the characteristics of the components have been better detailed described in the following points.

3.5 Preparation of the Guide for standards-based instructions and security policies

The security of an ICS according to [1] is based on a combination of effective security policies and a set of security controls configured properly. Such consider-

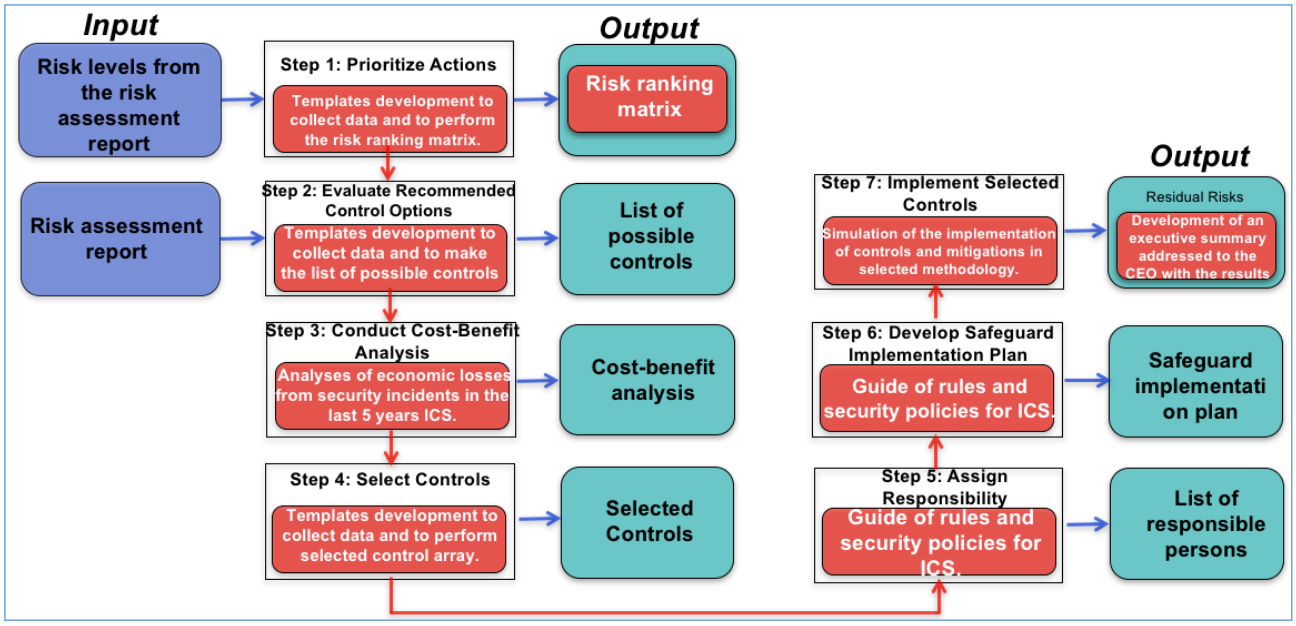


Figure 4: Risk Mitigation methodology conformity with NIST SP 800-30 [24], with Authors' contributions (in red) based on [27]

ations lead to the following five phases:

Phase 1: Meeting with the stakeholders (i.e., the "manage closely", "Keep informed" and "monitor" of the power-interest matrix, Figure 2). The purpose of these meetings is to encounter project requirements, implemented safety controls and other important documentation for the preparation of the manual. For this purpose the basic questions 5W-1H are used.

Phase 2: *Preparation of the guide*. As a reference has been taken of [1] [39] [40] [41], which serve as guidelines for developing security policies. In order to improve to structure the guide, it has been classified into three types of policies: being management, operational, and technical type, operational and technical. The stakeholders of the "manage closely" quadrant (Figure 2) are in charge of conducting the workshop sessions to elaborate phases 3, 4 and 5.

Phase 3: *Management type*. These are according to [1] security countermeasures to an ICS that focus on risk management and information security management: Evaluation of security and Authorization (CA), Planning (PL), Risk Assessment (RA), System and Services Acquisition (SA), and program Management (PM).

Phase 4: Operating Type. According to [1], these are security countermeasures for an ICS, which are executed and implemented mainly by people (i.e., stakeholders of all quadrants of the power-interest matrix, Figure 2): Personal Security (PS), physical and environmental protection (EP), Contingency Plan (CP), Maintenance (MA) and integrity of the information system (SI), media protection (PM), incident response (IR) and awareness training.

Phase 5: *Technical type*. According to [1], the technical types are security countermeasures to the ICS, which are primarily implemented and executed by the system through mechanisms containing hardware, software or firmware of the system: Identification and authentication (IA), Access Control (AC), Audit and Accountability (AU), System Protection and communications (SC).

4 Evaluation of Results and Discussion.

As indicated in section 3, this study includes the results of evaluations of experts' perceptions regarding the methodology implementation project. These were completed after each workshop session lasting between 60 and 120 minutes. This has been achieved by being informed of how the project is being implemented and better managing the risks related to its factors of success. The questions that were given for these evaluations were the following:

Question 1: Do you think that the time you took to collaborate in this workshop add value to the enterprise? Possible answers: Much (100%), Very (75%), little (50%), Very little (25%) and Nothing (0%).

Question 2: Do you think that the team that worked in this workshop provided all the necessary information? Possible answers: Yes (100%), something (50%) and No (0%).

Question 3: Do you agree that this workshop covered all aspects related to the scope of the project? Possible answers: Yes (100%), something (50%) and No (0%).

Question 4: How would you rate the time that has been used for this workshop? Possible answers: Very productive (100%), productive (75%), Normal (50%), Unproductive (25%) and Nothing Productive (0%).

Figure 5 reveals the averages of the answers to the questions asked in each evaluation according to each of the interested areas indicated in the "Manage Closely" quadrant of the Power-interest matrix indicated in section III: Maintenance, Process, Automa-

tion, and IT infrastructure of Figure 2.

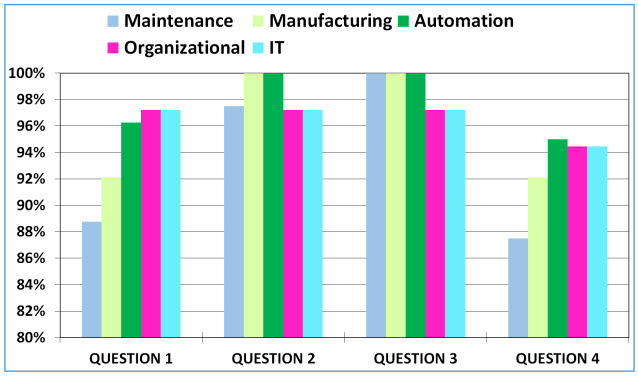


Figure 5: Averages of the scores of the expert's evaluations in the project.

Although the first and the fourth question are similar, it should be noted that the first one assesses whether there is added value to the enterprise or not. In contrast the fourth question assesses the productivity of the expert who has been present. The average answers to question 1 range from 88% to 97%, while the answers to question 4 range from 87% to 95%, respectively. This means that the experts are aware that their time in the implementation of this project gives value to the enterprise. However, it is not very productive or fun for them to spend any given time talking about the subject since their professional profile is more technical than administrative. These results allow important clues about how the project is performing in terms of the cost that the enterprise is investing in implementing such methodology. In addition, it demonstrates if there is any need to motivate the experts with an incentive, which may be analyzed later by the area of human resources and the project manager. In turn, these results indicate that there is a likelihood that the time-cost risks of project experts will be activated, as well as risks due to schedule delays due to fatigue or lack of interest. This information helps the project manager as well as the implementation team of the methodology. In current implementations, the subject of communication with stakeholders has been based on [42] and [34] and is being started in order to mitigate these project risks at the time of its execution.

The averages of the answers in question 2 indicate that the quality of implementation of the methodology is on track, since the corroborating ranks range from 97% to 100%. This question induces to the experts to unholy anything saved for them and to contribute with the vital information required. Also, it indicates to the Project Manager that the risks regarding the quality of the project are unlikely to be activated.

The answers to question 3, range from 97% to 100%, which indicates that the risks in scope are less likely to be activated.

After analyzing the results of risks of the implementation of the methodology, we have proceed to analyze the results of the risk assessment and design of mitigation strategies for ICS. These resulted in five visible threats, 15 vulnerabilities, 17 resulting risk and ten suggested mitigation strategies.

In this study, we assessed the risks from the perspective of ICS components and also of the informatics services. Therefore, the result of lifting the catalog of services according to [35] and [36] of the manufacturing process of the product "P" in the production line "L" is illustrated in Figure 6. There, they found 13 IT Services, 18 Information Systems and 26 informatics components. The risks found in the IT services coincided with items 1,2,6,8 and 9 mentioned in Table 2, and their mitigations would be the same as those analyzed with NIST. For other ICS manufacturing processes the same procedure may be performed.

As for the results obtained in the development of mitigation strategies, as mentioned in section 3, these are directly linked to the cost-benefit analysis presented to the enterprise (case study). Specifically, these indicate a deficiency in the management of the Information Security of this ICS, which forms it adequate to apply the proposed methodology.

Table 1 illustrates the considered variables. These indicate the investment costs of the implementation of the strategies versus the losses if the enterprise avoids the implementation of the methodology. Based on such calculations, the monthly cost of unavailability of the two production lines may reach between $ 109,000.00 in the first production line and $ 226,400.00 in the second production line. These calculations were obtained taking into account an hourly value of $ 5.01, with a total of 160 working hours per month. Total costs of implementation of mitigation strategies have reached up to $ 170,400.00 (i.e. estimated total costs of implementing technology, processes, and personnel requirements for the ten largest mitigations) compared to $ 1,422,760.00, which would be the cost of not implementing them. With these results, we follow that the best decision for the managers of these companies is the application of the proposed mitigation strategies. As mentioned in [38], it is confirmed that availability and integrity are more important than confidentiality in an ICS. Therefore, at this point we may suggest the development of a software application for the management of ICS that is able to evaluate the risks and the cost benefit of information security decisions in a fast and effective way.

The results of the simulations of the application of the above mentioned mitigations are listed in Table 2. The residual risk has been reduced in most cases from 100% to 10%, as well as the probability of execution of the threats. However, the impact remained almost unchanged. Simulations include only risks that would potentially have a high impact on SCI. The other minor risks should not be ruled out, but were carried out in a similar analysis, which produces a complete risk management in the ICS. Analyzing the whole context, we are able to appreciate that the continuous changes in the processes, the behavior of the people and the technology documented in this study, besides the in-

Table 1: Cost–Benefit Analisis

Number	Strategy/Control	Impact of implementation	Aditional costs/ investment	Impact of omited implementation	Aditional costs/ investment
1	Implentation of antivirus and firewall in the network ICS	MEDIUM	$24.000	HIGH	$81.750
2	Implementingan application, whic changes control and manages backups	MEDIUM	$20.000	MEDIUM	$226.400
3	Introduction of a security cameras system in control rooms	LOW	$24.000	HIGH	$60.000
4	Implementation of an application for asset management and inventory	LOW	$15.000	HIGH	$60.000
5	Acquisition of an additional PLC to quality control and later integration into the network	MEDIUM	$16.000	HIGH	$283.000
6	Introduction of an authentication a system (where possible) and daily backups	MEDIUM	$14.000	MEDIUM	$251.550
7	Physical access controls in control rooms	MEDIUM	$15.000	HIGH	$134.160
8	Introduction of a coordination procedure between Automation and IT	LOW	$2.400	HIGH	$45.000
9	Review of software licensing and other major equipment programs for the maintenance of ICS	MEDIUM	$24.000	HIGH	$226.400
10	Documenting algorithms and programs for changes of PLC	LOW	$16.000	MEDIUM	$54.500
			$170.400		**$1.422.760**

dications of [43], demonstrate that it is essential to mitigate and reduce the risks in an ICS. This leads to a risk assessment and mitigation strategy to be performed at least once a year.

For the Guide for standards-based instructions and security policies, the different formats and templates used by the enterprise for this type of documentation have been used. In order to carry out the evaluation of the Guide, a ten-person committee composed of management, technical and operational personnel (stakeholders of the quadrants: "manage closely" and "keep informed") was elected. This group had the task of reviewing, commenting, and proposing changes to improve the content of the policy manual. The collected responses were subsequently evaluated using statistical formulas and graphs (frequency, means, fashion, median, standard deviation and central tendency) as an evaluating tool to visualize the results (Figure 7). From a total of eleven data, the acceptance percentage has a median of 75%, an average of 79.09% and an acceptance for most questions of about 90% (fashion). In general, the Guide lists 18 clear and precise guidelines, which contain the point of view of management and the enterprise, thus considering all aspects of the management of the ICS safety guide. Therefore, it complies with what it stands for [12], where a policy also helps to reduce investment costs in technology control, induced by appropriate management processes and assigned functions.

Due to the Guide evaluation, the low values of about 55.24% may reflect poor judgment or knowledge of the contestants, in which an additional survey may be able to analyze its potential causes. Nonetheless, evidently users accept the Guide, allowing having high expectations that its application will significantly improve management and reduce potential future incidents.

In the Guide-induced simulations, ten stakeholders were consulted (i.e. from the "manage closely", "keep informed" and "monitor" quadrants, Figure 2). Figure 8 illustrates that out of a total of 100 data (ten questions asked to these 10 ICS stakeholders), the percentages of perception of policy compliance vary from

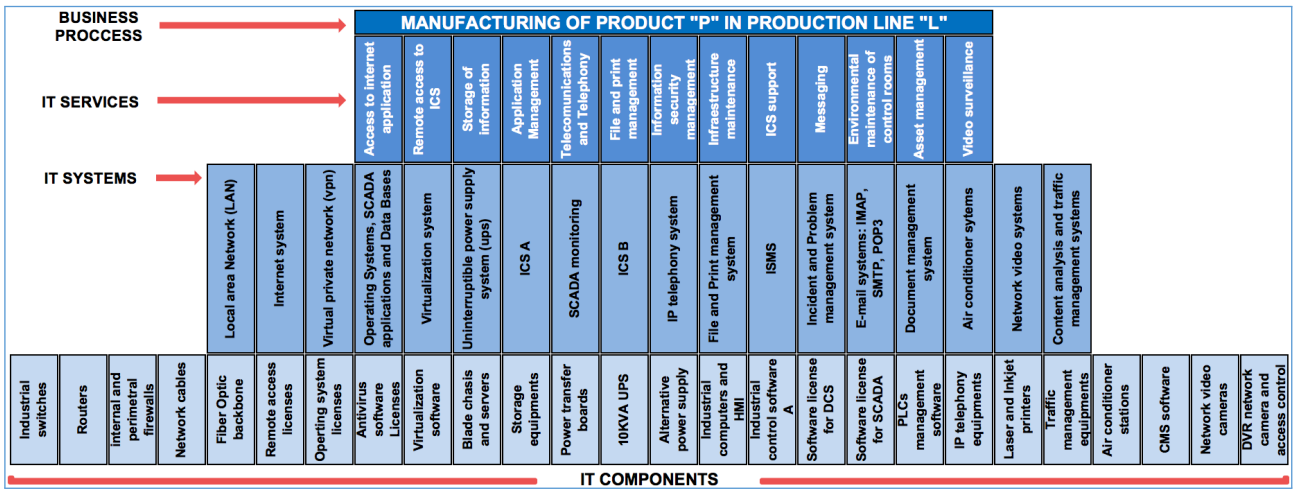


Figure 6: Services catalog of Industrial Control System.

Table 2: Results of The Simulation of Applications of Mitigation Strategies on ICS(The Top 10)

Item	Risk description	Probability	Impact	Risk	Strategy mitigation	Probability	Impact	Residual risk
1	Virus infection and malware network of ICS	1	100	100	Implementation of antivirus and firewall	0,1	100	10
2	Information loss and configuration errors in the ICS	1	100	10	Implementing an application for managing change controls and backups	0,1	100	10
3	Robbery of computers and devices of the ICS backup	1	100	100	Introduction of a camera security system in control rooms and workshops	0,1	100	10
4	Critical infrastructure lacking insurance policy	1	100	100	Implementation of an application for asset management and inventory	0,1	100	10
5	Damage of PLC of Quality Control in the 2nd production line (Outdated equipment)	1	100	100	Acquisition of an additional PLC for quality control, program support, and integration into the control network	0,1	50	5
6	Intentional alteration of information of the ICS	0,5	100	50	Introduction of an authentication system where possible and daily backups	0,1	100	10
7	Failures in ICS caused by unauthorized persons	0,5	100	50	Implementation of physical access controls	0,1	100	10
8	Improper acquisitions by outdated information security technology	5	50	50	Introduction of adequate and contemporary coordination procedure between Automation and IT	0,1	50	5
9	Inadequate sizing of technology resources in ICS maintenance	0,5	100	50	Regular software licensing review and other major equipment programs for the maintenance of ICS	0,1	50	5
10	Allen Bradley PLC Failure (control room 3)	1	50	50	Documenting Algorithms and programs for changes in the PLC	0,1	50	5

15% to 80%, while the average perception of compliance yielded an average of 41.05% and a median of 45%. Despite the negative trend shown in this figure, the slope (about 45 degrees) indicates a reliable distribution of the data collected for this analysis. On the other hand, as illustrated in Figure 8, 60 of the 100 questions answered are between the average (41.05%) and 80% of the fulfillment of perception. This may be considered as satisfactory values of the simulation. It is therefore thought that in the future stakeholders will support the improvement of the proposed methodology for the security of management information in ICS once they become familiar with such a procedure.

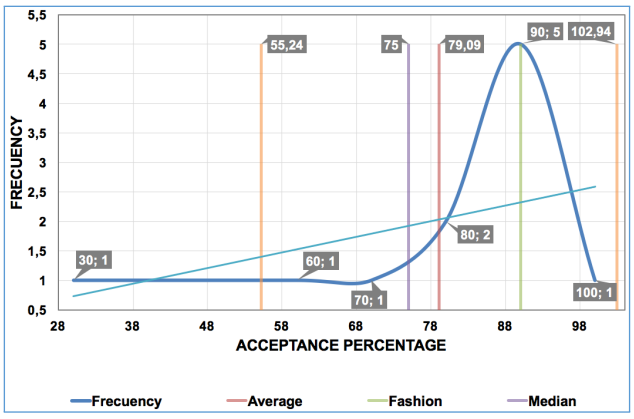


Figure 7: The statistical process, which results by the evaluation of the acceptance of the Guide for standards-based instructions.

In Figure 9 the perception of compliance with the application of the manual that are outside the range between 20.61% and 61.49% may result from the lack of judgment or even ignorance of the respondents regarding the Security of the ICS, since it is a contemporary and new topic for many professionals in the industrial field.

According to enterprise executives in this case study, the percentage of compliance with other policies and standards manuals of the organization vary in the range of 40% to 50%. This suggests that 40% is not a coincidence but the result of the organizational culture of such an enterprise. According to these results a proposal of a follow-up may rise, where the target may be of how an organizational culture has a potential impact or relation with the compliance of the rules and policies of an enterprise. Simultaneously, it should also be considered to propose the development of a software application that simulates the performance of the manual, in order to be able to determine the degree of information security management of an ICS that is a useful tool for the managers and stakeholders of the quadrant "Manage closely".

Finally, the proposed approach and the set of industry standards analyzed in this study have been complemented with each other. Therefore, the substantial difference is situated in the amalgamation and articulation of proven techniques, methods, and the adaptation of other standards, referential frameworks, and management systems that have been used in IT for both project management and industrial security in the same place where the facts are occurring. This allowed significant improvements to effectively and holistically manage the different and complex areas, which coexist in the ICS: business, computing, electronics, automation, production processes and people.

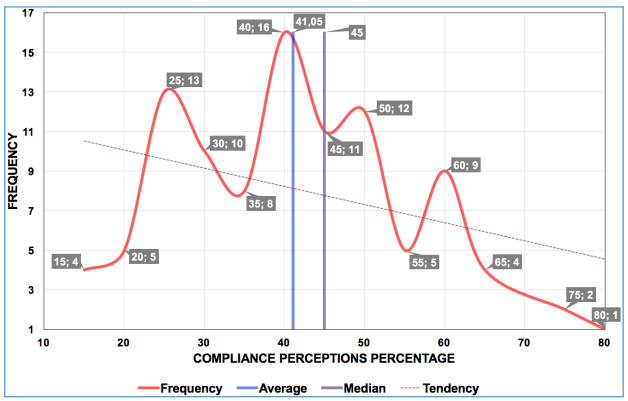


Figure 8: The statistical process, which resulted by the perception of the compliance of the Guide for standards-based instructions.

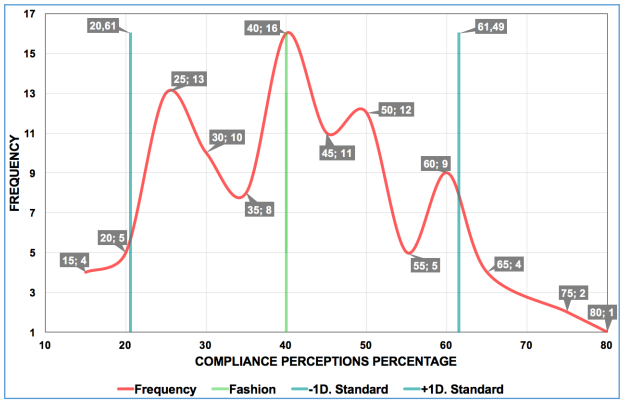


Figure 9: The statistical process, which results by the simulation of the application of the Guide for standards-based instructions.

5 Conclusions and Future Work.

Based on the results obtained, it is concluded that the set of mitigation strategies and the Guide for standards-based instructions and security policies for a manufacturing enterprise may achieve a reduction of 40% of security incidents in its ICS in the contexts of availability, integrity and confidentiality of information. At the same time, it is observed how the project of implementation of the methodology has been favorably accepted by the stakeholders creating motivation, collaboration and synergy in the areas that use, operate and manage the ICS. As for NIST 800, PMBOK, COBIT and ITIL, it has been perceived that they are highly strategic international standards, robust and in line with the management proposal presented for the ICS. That means that in turn, that the management proposal is perfectly articulated with the managerial, operational, technical areas of such enterprise. The proposed methodological process may be compared with ISO 27000 ISMS, being recommended to be used in manufacturing companies to manage information security in industrial control systems.

As future lines of studies, it is planned to combine the proposed methodology with DSS02, DSS03 and DSS04 of COBIT 5, the ITIL service operation stage, and a general guide of information security policies as well as norms throughout the enterprise according to ISO 27000, and NIST 800-82, in order to maintain proper management of traditional IT and ICS. In addition, the design and implementation of an information security incident response team (CSIRT) for ICS, based on NIST 800-61 as a framework and ITIL V3, has also been considered in order to analyze the involved computer services.

Conflict of Interest. The authors declare not to have any conflicts of interest.

Acknowledgment. We would like to thank Comercializadora San Remigio and in particular Engineer Rafael Simon (CEO) for allowing to apply the concept of this study and also the analysis of the required information for this research. This work has been partially funded by the Universidad de las Fuerzas Armadas ESPE within the research project entitled "Training Computational Platforms, Experimentation, Management, and Mitigation of Cyber security Attacks", Code: ESPE-2015-PIC-019.

References.

[1] Stouffer, K., Lightman, S., Pillitteri, V., Abrams, M., Hahn, A., "Guide to Industrial Control Systems (ICS) Security", NIST SP 800-82, Rev. 2, pp. 1-249, May 2015. Last update: March 1st, 2017.

[2] Francia III, G. A., Thornton, D., & Dawson, J., "Security best practices and risk assessment of SCADA and industrial control systems". In Proceedings of the International Conference on Security and Management (SAM) (p. 1), Alabama, January, 2012.

[3] Ning, C., Jidong, W., &Xinghuo Y., "SCADA system security: Complexity, history and new developments," Industrial Informatics, 2008. INDIN 2008. 6th IEEE International Conference on, Daejeon, 2008, pp. 569-574.

[4] Byres, E., & Lowe, J. "The myths and facts behind cyber security risks for industrial control systems". In Proceedings of the VDE Congress (Vol. 116, pp. 213-218), October, 2004.

[5] Tieghi, E. M. "Introduzione alla protezione di reti e sistemi di controllo e automazione (DCS, SCADA, PLC, etc.)". Quaderni CLUSIT-Associazione Italiana per la SicurezzaInformatica. Italia, 2007.

[6] Green, B., Prince, D., Roedig, U., Busby, J., & Hutchison, D., "Socio-technical security analysis of industrial control systems (ICS)". In Proceedings of the 2nd International Symposium on ICS & SCADA Cyber Security Research 2014 (pp. 10-14). BCS. Sept., 2014.

[7] Yang, W., & Zhao, Q., "Cyber security issues of critical components for industrial control system". In Guidance, Navigation and Control Conference (CGNCC), 2014 IEEE Chinese (pp. 2698-2703), Aug. 2014.

[8] Hadziosmanovic, D., Bolzoni, D., Etalle, S., & Hartel, P., "Challenges and opportunities in securing industrial control systems". In Complexity in Engineering (COMPENG), 2012 (pp. 1-6). IEEE, Jun., 2012.

[9] Krotofil, M., & Gollmann, D., "Industrial control systems security: What is happening?" In Industrial Informatics (INDIN), 2013 11th IEEE International Conference on (pp. 670-675). IEEE, Jul. 2013.

[10] Nicholson, A., Webber, S., Dyer, S., Patel, T., and Janicke, H., "SCADA security in the light of Cyber-Warfare", Computer & Security, Volume 31, Issue 4, June 2012, Pages 418-436.

[11] Cherdantseva, Y., Burnap, P., Blyth, A., Eden, P., Jones, K., Soulsby, H., &Stoddart, K. "A review of cyber security risk assessment methods for SCADA systems". Computers & security, 56, 1-27. 2016.

[12] Seo, J., Song, M., and Lee, K., "A Study on Efficiency of ISMS for ICS with Compliance", International Journal of Multimedia and Ubiquitous Engineering", Vol.9 No. 5, pp 301-306, 2014.

[13] Drias, Z., Serhrouchni, A., & Vogel, O. "Analysis of cyber security for industrial control systems". In Cyber Security of Smart Cities, Industrial Control System and Communications (SSIC), 2015 International Conference on (pp. 1-8). IEEE, 2015, Aug.

[14] Lemaire, L., Vossaert, J., Jansen, J., & Naessens, V., "A Logic-Based Framework for the Security Analysis of Industrial Control Systems". Automatic Control and Computer Sciences. 2017

[15] Mansfield-Devine, S., "A process of defense–securing industrial control systems". Network Security, 2017(2), 14-19. 2017.

[16] Collier, Z. A., Panwar, M., Ganin, A. A., Kott, A., & Linkov, I., "Security Metrics in Industrial Control Systems". In Cybersecurity of SCADA and Other Industrial Control Systems (pp. 167-185). Springer International Publishing. 2016.

[17] McLaughlin, S., Konstantinou, C., Wang, X., Davi, L., Sadeghi, A. R., Maniatakos, M., & Karri, R., "The cybersecurity landscape in industrial control systems". Proceedings of the IEEE, 104(5), 1039-1057. 2016.

[18] Keliris, A., Konstantinou, C., Tsoutsos, N. G., Baiad, R., & Maniatakos, M., "Enabling multi-layer cyber-security assessment of Industrial Control Systems through Hardware-In-The-Loop testbeds". In Design Automation Conference (ASP-DAC), 2016 21st Asia and South Pacific (pp. 511-518). IEEE. 2016, Jan.

[19] Graham, J., Hieb, J., & Naber, J., "Improving cybersecurity for Industrial Control Systems". In Industrial Electronics (ISIE), 2016 IEEE 25th International Symposium on (pp. 618-623). IEEE, 2016, Jun.

[20] Abe, S., Fujimoto, M., Horata, S., Uchida, Y., & Mitsunaga, T., "Security threats of Internet-reachable ICS". In Society of Instrument and Control Engineers of Japan (SICE), 2016 55th Annual Conference of the (pp. 750-755). IEEE, 2016, Sept.

[21] Ponomarev, S., & Atkison, T., "Industrial control system network intrusion detection by telemetry analysis". IEEE Transactions on Dependable and Secure Computing, 13(2), 252-260. 2016.

[22] Martellini, M., Abaimov, S., Gaycken, S., & Wilson, C., "Vulnerabilities and Security Issues". In Information Security of Highly Critical Wireless Networks (pp. 11-15). Springer International Publishing. 2017.

[23] Busby, J. S., Green, B., & Hutchison, D. (2017). Analysis of Affordance, Time, and Adaptation in the Assessment of Industrial Control System Cybersecurity Risk. Risk Analysis.

[24] ISO 27000, "Information Technology, Security Techniques, Information Security Management Systems, Overview and Vocabulary," International Organization for Standardization ISO, Geneve, 2013.

[25] Stoneburner, G., Goguen, A., and Feringa, A., "Risk Management Guide for Information Technology Systems", NIST SP 800-30, pp. 1-56, 2002, July.

[26] Konstantinos K. Automated cyber security compliance assessment. 2017. TRITA-EE, 2016:202. Master Thesis.

[27] Bustamante, F., Fuertes, W., Díaz, P., & Toulkeridis, T., "A methodological proposal concerning to the management of information security in Industrial Control Systems". In Ecuador Technical Chapters Meeting (ETCM), IEEE (Vol. 1, pp. 1-6). IEEE, 2016, Oct.

[28] ENISA, "Analysis of ICS-SCADA Cyber-security Maturity Levels in Critical Sectors". ENISA, 11 Dec 2015. Accessed on February 28th, 2017. URL: https://www.enisa.europa.eu/publications/maturity-levels/at_download/fullReport.

[29] Järekallio, T., "Methods for Identification and Classification of Industrial Control Systems in IP Networks", Master's Thesis, School of Electrical Engineering, Aalto University, 2016, Aug.

[30] ICS-CERT, "Industrial Control Systems Cyber Emergency Response Team", U.S Department of Homeland Security, pp. 1-24, 2015.

[31] Vertical Insight, "Data Breach Investigations Report Manufacturing", Verizon, MC1591204/14, pp. 1-60, 2014.

[32] Blank R. M. and Gallagher, P. D., "Guide for Conducting Risk Assessments/Information security ", NIST Special Publication 800-30 Revision 1, pp. 1-95, 2012, Sept.

[33] Locke, G., and Gallagher, P., "Managing Information Security Risk", NIST SP 800-39, pp. 1-88, 2011, March.

[34] Rose, K. H., "A Guide to the Project Management Body of Knowledge (PMBOK® Guide)—Fifth Edition". Project Management Journal, 44: e1. doi:10.1002/pmj.21345. 2013.

[35] Cannon, D., Cannon, D., Wheeldon, D., Lacy, S., & Hanna, A. "ITIL service strategy". TSO. 2011.

[36] Hunnebeck, L., Rudd, C., Lacy, S., & Hanna, A. "ITIL service design". TSO. 2011.

[37] Coughlan, P., & Coghlan, D. (2002). "Action research for operations management". International journal of operations & production management, 22(2), 220-240.

[38] Cheminod, M., Durante, L., &Valenzano, A., "Review of security issues in industrial networks". Industrial Informatics, IEEE Transactions on, 9(1), 277-293. 2013.

[39] Blank, R., and Gallagher, P., "Security and Privacy Controls for Federal Information Systems and Organizations", NIST SP 800-53, Rev. 4, pp. 1-460, April, 2013 updates as of 01-22-2015

[40] Gallagher, P., "Recommended Security Controls for Federal Information Systems and Organizations", NIST SP 800-53, Rev. 3, pp. 1-11, 2009, August.

[41] Guttmann, B., and Roback, E., "An Introduction to Computer Security: The NIST Handbook", NIST SP 800-12, pp. 1-296, October, 1995.

[42] Karu, K., Ferris, K., Hunebeck, L., Rae, B., & Rance, S. (2016). ITIL Practitioner. The Stationery Office.

[43] Talib, M., Barchi, M., Khelifi, A., and Ormandjieva, O., "Guide to ISO 27001: UAE Case Study", Issues in informing Science and Information Technology, 7(2012), pp. 331-349, 2012.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 2, No. 3, 100-110 (2017)
www.astesj.com

ASTES Journal
ISSN: 2415-6698

Dependence-Based Segmentation Approach for Detecting Morpheme Boundaries

Ahmed Khorsi, Abeer Alsheddi*

Computer Science Department, Al-Imam Mohammad Ibn Saud Islamic University, Kingdom of Saudi Arabia

ARTICLEINFO

Article history:
Received: 04 March, 2017
Accepted: 01 April, 2017
Online: 10 April, 2017

Keywords :
Classical Arabic language
Computer science
Corpora
Morpheme boundaries
Unsupervised learning
Word segmentation

ABSTRACT

The unsupervised morphology processing in the emerging mutant languages has the advantage over the human/supervised processing of being more agiler. The main drawback is, however, their accuracy. This article describes an unsupervised morphemes identification approach based on an intuitive and formal definition of event dependence. The input is no more than a plain text of the targeted language. Although the original objective of this work was classical Arabic, the test was conducted on an English set as well. Tests on these two languages show a very acceptable precision and recall. A deeper refinement of the output allowed 89% precision and 78% recall on Arabic.

1 Introduction

This article is an extension of work originally presented in Proceedings of 4th Saudi International Conference on Information Technology (Big Data Analysis) [1]. The article discusses one of the main challenges in computational linguistics is the identification of morpheme boundaries [2, 3, 4]. Boundary detection process plays a central role in extracting stems from words [5]. Moreover, the noisy infixes affect the performance of some tasks such as similarity computation [6, 7, 8]. The fascinating results of this process performed easily on specific examples made computer learning of language morphology a desirable approached for a long time; as are simple problems hiding dissuasive details. As a result of that, Zelig Harris [9] way of segmenting words into morphemes by only counting letters that may occur after a prefix and marking the picks seems to be the first approach published. Since that, investigations did not cease on building a model easily maintainable to grasp how words are coined in a human language. Benefits of unsupervised approaches step beyond avoiding costly human efforts on known languages. These approaches also have the advantage of being fast enough and to some extent language independent.

In Semitic languages [10, 11] such as Arabic, the challenge is even harder because of the irregularities, like mutations and diphthongs situations. Mutation means that one letter can be replaced by another to fit the pronunciation. Take the Arabic word "اضطجع" ([ADoTajaE]: lie-down)[1] comes from applying the pattern "افتعل" and then mutates the third letter "ت" ([t]: 3rd Arabic letter) to "ط" ([T]: 16th Arabic letter). This mutation also can be shown with a long vowel[2] deleted or replaced by another. For example, the past tense of the verb "قول" ([qawl]: say) is "قال" ([qaAl]: said) by mutating "و" ([w]: 27th Arabic letter) into "ا" ([A]: 1st Arabic letter), whereas the imperative form of the same verb is "قل" ([qul]: say) with the letter "و" ([w]: 27th Arabic letter) deleted. The second situation, diphthong; means that two consecutive identical letters are merged into one letter. For example, the word "شد" ([$ad~]: pull) is originally "شدد" ([$adoda]).

The proposed approach is developed to the design of a fully computerized (i.e., unsupervised) extraction of the morphemes. In light of this goal, the approach uses simple and intuitive statistical feature extracted by reading a corpus of plain text with no tag to become able to identify the morphemes boundaries. The approach also was compared with an existed approach named Morfessor [13]. Although the original goal of this work was to address the much

*Corresponding Author Email: abeer.alsheddi@gmail.com (A. Alsheddi)

[1] In this article, Arabic is represented in some or all of three variants according to context: "Arabic word" ([Buckwalter Arabic transliteration] [12]: English translation).

[2] Three Arabic letters have long vowels: "ى" or "ا" ([A]: 1st Arabic letter), "و" ([w]: 27th Arabic letter) and "ي" ([y]: 28th Arabic letter)

richer vocabulary of the classical Arabic, the test was conducted on an English set as well. The results are a proof of concept that unsupervised techniques can accurately handle a morphology as complex as the one of a legacy Semitic language.

The study brings the following values:

1. A model: An intuitive but formal model was developed for morphology learning based on a simple interpretation of the probabilistic dependence [14] to segment a word into morphemes.

2. Languages

   - Classical Arabic: The performance of the model was reported on the morphology of the classical Arabic, which is the language of a huge legacy literature. Arabic morphology is more complicated than simple languages [15]. Because of these complex features, some of the methods widely used in many languages cannot be applied to Arabic [6, 11]. On the other hand, two known varieties of Arabic are classical Arabic and Modern Standard Arabic (MSA). MSA is based on the classical Arabic. However, MSA is used more in informal speech and can combine words borrowed from other languages that may not use Arabic rules. In contrast, the classical Arabic has been used to write literary legacy and traditional vocabulary. It contains purer Arabic words than MSA. Moreover, Arabic morphological and semantic rules are derived from the classical Arabic. At the same time, the researchers on Arabic are paying most of their attention to MSA. To the best of our knowledge, the only studies taking into account the classical Arabic are limited to religious texts.
   - Other Languages: The performance of the model was tested on a concatenative language namely English. It has a totally different degree of morphological complexity from Arabic [16].

3. Comparison: The performance of another word segmentation approach named Morfessor was reported on the same dataset.

4. Corpus: It is a language resource defined as a collection of texts stored in a machine-readable format. To assess the proposed approach, we had to build a corpus of classical Arabic texts authored in the period from 431 to 1104 (in Hijri between 130 b.h and 498 h) counting around 122M words in total and 1M distinct words.

The remaining parts of this article are organized as follows: Section 2 reviews existing methodologies on detecting morpheme boundaries. Section 3 introduces the proposed approach and explains its algorithm. Section 4 discusses the tests and the results it carried out. Finally, section 5 concludes with a summary of contributions and makes suggestions for future research work.

2 Applied Methodologies

One of the oldest published work on unsupervised word segmentation is due to Harris who suggested a process to break a phonetic text into morphemes [9] using only successors numbering. Bordag [17] partially relies on Letter Successor Variety (LSV) due to Zelig Harris [9, 2, 3] and combines it with a trie based classifier [18]. The approach is claimed to reach 67.48% F-measure. Bernhard [19] too combined the same principle of successors variation with a set of heuristics to filter out the less plausible segments. The F-measure is claimed to reach 60.81%.

One of the well-known systems in this area is Morfessor, developed by Creutz and Lagus [13]. Morfessor tries to capture the morpheme boundaries in a probabilistic model with simplistic features such as the frequency and the length of the morpheme. For comparison, we investigate this approach deeper in Section 4.1.4. Eroglu, Kardes, and Torun [20] tried to make Morfessor 1.0 accommodate the phonetic features of the Turkish language. By adding phone-based features the tried to measure the phonemic confusability between the morphemes. Their results are claimed to be better than Morfessor's for Turkish, Finnish and English. Pitler and Keshava [5] use the forward and backward conditional probability to build a list of substantial affixes. They report a 52% F-measure on their test set and 75% on the gold-standard.

A number of works depend on Markov Models (MM).Melucci and Orio [21] investigated the use of the Hidden Markov Models (HMMs). The evaluation though is done on a whole information retrieval system (extrinsic) [22]. Naradowsky and Toutanova [23] designed a model, which tries to exploit the alignment of morphemes in a source language to the morphemes in a target (machine translated to) language. With different settings of the model, the approach is claimed to reach 84.6 F-measure for Arabic.

Peng and Schuurmans [24] proposed a two-level hierarchical Expectation Maximization (EM) model. Their test set holds 10K words and their results are claimed to reach 69% F-measure. Poon, Cherry and Toutanova [25] use a log-linear model [26, 27] to capture the features inherent to the morpheme itself and its surrounding letters. The authors tested the approach on supervised and semi-supervised settings as well. The unsupervised settings reached a 78.1% F-measure for Arabic and 70% for Hebrew.

It is worth mentioning that, the actual list of similar works is longer. Thus, we limited the overview to the previous works only. Some of the extra works are in the surveys [28] and [29] that provided for the interested reader. In light of the proposed goal in this article, this section intentionally avoids supervised approaches [30].

On the other hand, other works in real environments use the segmentation technique in their approaches, such as document image retrieval systems [31, 32].

3 The Proposed Approach

The word *"unbreakable"* is coined by concatenating the prefix *"un"*, the stem *"break"* and the suffix *"able"*. The vocabulary, which suggests such segmentation, should contain other words with different combinations of prefixes, stems and suffixes (e.g. *"rebreakable"*, *"unbreaking"*...etc), which makes the occurrence of *"un"* have a weak dependence to the occurrence of *"break"* whose occurrence is also relatively independent of the occurrence of the suffix *"able"*. On the other hand, each morpheme is supposed to be inseparable from neither its first nor its last letter. The proposed approach (Dependence-Based Segmentation) is all about exploiting these two facts. A good formal modeling of these facts is the definition of the probabilistic dependence [14].

3.1 Segmentation Algorithm

Algorithm 1 iterates over the word W letter by letter and computes for every letter w_i two dependences: 1. the dependence of the prefix on w_i as its last letter, 2. the dependence of the suffix on w_i as its first letter where the prefix ends (inclusive) at w_i and the suffix starts (inclusive) at it. The difference then points to which of the two (i.e. the prefix or the suffix) is more attached to the current letter w_i. The algorithm keeps going until it encounters a change of the direction of the attachment. The change is done when the prefix depends on the preceding letter w_{i-1} more than the suffix does, and at the same time the suffix depends on the current letter w_i more than the prefix does. Then between the previous and the current letters a cutting point (CP) is marked.

3.2 Computation of the Dependence

The concept of dependence we use is symmetric [14]. In this context: *"the string depends on the letter"* means *"the letter depends on the string"* and vice versa.

We will call the dependence of a letter a on the prefix α: the forward dependence, and we denote FD(u) where u=αa is the prefix tailed by the letter under processing a. We call the dependence of the letter a on the suffix β: the backward dependence, and we denote BD(v) where v=aβ is the suffix headed by the letter under processing a. We also mark the beginning and the end of a word by respectively # and $. The forward dependence is then:

$$FD(u) = \frac{P(\alpha \rightarrow a)}{P(a)} \tag{1}$$

$$BD(u) = \frac{P(a \leftarrow \beta)}{P(a)} \tag{2}$$

Where $P(\alpha \rightarrow a)$ is the conditional probability:

$$P(\alpha \rightarrow a) = \frac{Count(\alpha a)}{Count(\alpha)} \tag{3}$$

and $P(a \leftarrow \beta)$ is the conditional probability:

$$P(a \leftarrow \beta) = \frac{Count(a\beta)}{Count(\beta)} \tag{4}$$

$$FD(u) = \frac{Count(\alpha a)}{Count(\alpha)P(a)} \tag{5}$$

$$BD(u) = \frac{Count(a\beta)}{Count(\beta)P(a)} \tag{6}$$

Where the probability of a letter a: P(a) is estimated by its normalized frequency in the corpus.

$$P(a) = \frac{Count(a)}{\sum_{b \in \mathcal{A}} Count(b)} \tag{7}$$

Where $\mathcal{A}$ is the alphabet. Count(α) expresses how often an n-gram α occurs in the corpus.

By the definition of the dependence measure [14], when both forward and backward dependencies at the same letter are lower than one, this means that the letter does not depend on the prefix or the suffix. Thus, we assigned zero to the difference's value and discarded this position from the segmentation process. This dependence is called a *negative dependence* [14].

After segmenting a word using the cutting points, we supposed that a morpheme with the least frequency among the other morphemes is a stem and the other morphemes are affixes.

Arabic example: The Arabic word "الكتابان" ([Al◦kitaAbaAk]: the two books) contains two cutting points as shown in Table 1. The first one occurs when the second letter "ل" is more dependence on the prefix "#ال" than the suffix "لكتابان$, while the third letter "ك" is more dependent on the suffix "كتابان$" than the prefix "#الك". The second position is the sixth letter "ب" when the dependence tendency is changed. Thus, these two cutting points "ال|كتاب|ان" produce the following morphemes: "ال", "كتاب" and "ان". Consequently, the morpheme "كتاب" is the stem depending on the frequency list in Table 3.

Negative dependence example: The Arabic word "كتيب" ([kutayb]: small book) does not contain any cutting point as the table 2 shows it. Consequently, the morpheme "كتيب" is the stem of the word. At the same time, we can note the second letter "ت" ([t]: 3^{ed} Arabic letter) has zero as its difference value. That means both forward and backward dependencies are less than one.

English example: Table 4 is a simulation of the algorithm on the word *"unbreakable"*. The difference at each letter ($FD(u)$-$BD(v)$) points a dependence direction of that letter. In other words, the positive difference means that the current letter w_i is more attached to the prefix $w_{1...i}$ than suffix $w_{i...n}$ and vice versa. The second letter *"n"* depends on the prefix *"#u"* more than it does on the suffix *"breakable$"*, where the third letter *"b"* depends on the suffix *"reakable$"* more than it does on the prefix *"#un"*. This change of the dependence direction makes the point *"un|breakable"* a cutting point. The same logic applies to the seventh letter

Algorithm 1: Word Morphological Segmentation

input : A word $W = w_0 w_1 \dots w_n$
output: Segments of word W
Cutting points: CP
foreach w_i *where* $0 < i \leqslant n$ **do**
  if $FD(\#\dots w_i) < 1$ **and** $BD(w_i \dots w_n\$) < 1$ **then**
    discard this point
  else
    if $FD(\#\dots w_{i-1})$-$BD(w_{i-1} \dots w_n\$) > 0$
    and $FD(\#w_0 \dots w_i)$-$BD(w_i \dots w_n\$) < 0$ **then**
      $CP \leftarrow CP \cup i$
$seg \leftarrow$ segment W at each i in CP
return seg

Algorithm 2: N-grams model construction algorithm

Input : A string of words W
Output: n-grams mapped to their frequency
foreach W *in string* **do**
  W ← '#'+W+'$'
  for $i \leftarrow 1$ *to* $|w|-1$ **do**
    for $j \leftarrow i+1$ *to* $|w|-1$ **do**
      if $w_{i\dots j} \in n\text{-}grams$ **then**
        $freq(w_{i\dots j}) \leftarrow freq(w_{i\dots j}) + 1$
      else
        $n\text{-}grams \leftarrow n\text{-}grams \cup \{w_{i\dots j}\}$
        $freq(w_{i\dots j}) \leftarrow 1$
return n-*grams* and $freq$

Table 1: Arabic example of the segmentation

i	u	FD(u)	v	BD(v)	Difference	Direction
1	ا#	0.991	الكتابان$	7.174	-6.182	↓
2	ال#	9.349	لكتابان$	9.153	0.195	↑
3	الك#	1.017	كتابان$	17.174	-16.157	↓
4	الكت#	0.716	تابان$	1.210	-0.494	↓
5	الكتا#	2.532	ابان$	0.544	1.987	↑
6	الكتاب#	12.067	بان$	1.552	10.514	↑
7	الكتابا#	1.366	ان$	1.865	-0.498	↓
8	الكتابان#	4.560	ن$	1.792	2.767	↑

Table 2: Example of the negative dependence

i	u	FD(u)	v	BD(v)	Difference	Direction
1	ك#	1.125	كتيب$	5.631	-4.505	↓
2	كت#	0.646	تيب$	0.349	0	0
3	كتي#	0.685	يب$	2.030	-1.344	↓
4	كتيب#	6.675	ب$	0.256	6.419	↑

Table 3: A frequency list of chosen words

Morpheme	Frequency	Morpheme	Frequency
un	1,986	ال	72,845
break	46	كتاب	47
able	1,426	ان	10,047

Table 4: English example of the segmentation

i	u	FD(u)	v	BD(v)	Difference	Direction
1	#**u**	0. 46	**u**nbreakable$	30.97	-30. 51	↓
2	#u**n**	6. 36	**n**breakable$	5. 84	0. 52	↑
3	#un**b**	1. 35	**b**reakable$	46. 70	-45. 35	↓
4	#unb**r**	1. 94	**r**eakable$	10. 45	-8. 51	↓
5	#unbr**e**	3. 11	**e**akable$	5. 01	-1. 90	↓
6	#unbre**a**	6. 79	**a**kable$	1. 68	5. 11	↑
7	#unbrea**k**	34. 71	**k**able$	1. 50	33. 21	↑
8	#unbreak**a**	5. 09	**a**ble$	7. 38	-2. 29	↓
9	#unbreaka**b**	46. 70	**b**le$	9. 89	36. 81	↑
10	#unbreakab**l**	9. 72	**l**e$	2. 76	6. 96	↑
11	#unbreakabl**e**	10. 02	**e**$	1. 15	8. 87	↑

"*k*" and the eighth letter "*a*". Therefore, the resulting morphemes will be: "*un*", "*break*" and "*able*". The least frequent morpheme is "*break*" by using Table 3. Consequently, the morpheme "*break*" is the stem.

4 Tests and Results

Before describing the test settings, we will list main steps, that we followed in the study:

1. Build the corpus. (see Section 4.1.1).
2. Make the normalization (see Section 4.1.2).
3. Generate the language model (see Section 4.1.3).
4. Test the proposed approach (see Section 4.1.4) using materials (see Section4.1.5).
5. Evaluate the performance by using the metrics (see Section 4.1.6).

4.1 Test Settings

4.1.1 Test Dataset

Arabic The purpose of the research, which drew our efforts to design this approach, was to address challenges in processing the morphology of the classical Arabic. Works we are aware of try to handle MSA. Actually, MSA is known to include a mixture of words borrowed from other languages and words just coined for the convenience, which usually breaks the classical Arabic rules. Correctly handling the classical Arabic brings three interrelated values: 1. The classical Arabic vocabulary is much richer in terms of pure Arabic words. 2. The Arabic rules in the morphological, grammatical and semantic levels amply documented since the eighth century have been designed on the basis of the classical Arabic. 3. A huge legacy literature is written in the classical Arabic, and most of the time standing beyond the scope of the supervised approaches designed for the MSA. The present corpus[3] gathers Arabic texts dating back from 431 to 1104 (130 b.h and 498 h). It contains around 122M words in total and 1M distinct words with an average size of 6.22 letters per word. Table A.1 in the appendix shows a sample of the index in Arabic that was published with our corpus. Such dataset is by itself a valuable resource for researches in NLP field.

English We obtained the English text by merging the two corpora "wiki" and "news"[4]. This resulted in more than 1M unique words and more than 8M words in total, having an average size of 8.00 letters per word.

4.1.2 Normalization

The normalization process is frequently used to transform text into an approved form, which aids in reducing the noise and sparsity in the text. In present work, the following normalization was applied to the Arabic words:

1. The removal of diacritics.
2. The substitution of all variants of Hamza "ء" ([ˊ]) with the form "ء" ([ˊ]).
3. The substitution of the letter "آ" ([|]) with "ءا" ([ˊA]).

4.1.3 Language Model

The proposed language model extracts all possible q-grams from a word of n length, where q rises from one to n. Then it assigns each gram to its occurrence in the corpus. Algorithm 2 explains the extraction. Table A.2 in the appendix shows the n-grams of the two example words "الكتابان" and "*unbreakable*".

4.1.4 Test Process

Morfessor 2.0[5] and the proposed approach (Dependence-Based Segmentation) are run over the same corpus, then three samples of 100 words each are randomly picked out of the whole segmented corpus. The results of these approaches are evaluated manually. Precision, recall and F-measure of the cutting points are then recorded. This is repeated on each of the four settings combinations where each setting combines one of two occurrence settings with one of two affix settings.

Occurrence Settings

- Distinct: To build the language model, the approach used distinct words, i.e. each word in the corpus occurs one time.
- Plain: The number of occurrences is taken into account during building the model, i.e. a word may occur more than once.

Affix Settings

- Raw: Results sample is picked randomly with no restriction.
- Non-empty affix: Samples are picked randomly only among words for which the proposed approach has carried out at least one cutting point. For a number of words, the segmentation simply did not identify any cutting point. Most of them were because of the scarcity of the stem or stem.affix remaining combination. We then tried to assess the impact of such cases on the performance and how accurate the identified cutting was.

It is worth noting that our objective was to address the challenge of segmenting the classical Arabic words. To the best of our knowledge, there is no suitable gold standard for TA. We had to build ourselves the set of words. Then we proceeded to the manual segmentation of three different randomly picked samples for each setting.

[3] https://sourceforge.net/projects/classical-arabic-corpus/

[4] http://corpora.informatik.uni-leipzig.de/download.html

[5] One of our objectives is to keep the process entirely unsupervised. Thus, we used the latest version without using any optional parameters to discard any semi-supervised extension.

4.1.5 Test Platform

We used the following materials to test the proposed algorithm:

- Operating system: Ubuntu 12.04, 64-bit.
- Programming language: C++.
- Compiler: GCC 4.9.0.
- Database management system: MySQL Server 5.5.43.
- C++ external Library: libmysqlcppconn-dev 1.1.0-3build1 to connect C++ with MySQL server.
- CPU: Intel(R) Xeon(R) CPU E5530 @ 2.40GHz * 16.
- RAM: 6 GB for RAM and 12 GB for Swap.

4.1.6 Performance Metrics

For evaluating, we used three metrics: recall, precision and F-measure [33], where we took into account a number of cutting points during the evaluation.

1. Recall: This metric measures how many correct positions are found among all the existing correct positions. The higher the recall, the more correct positions are found and returned.

$$Recall = \frac{correct\ cutting\ points\ in\ the\ result}{all\ correct\ cutting\ points\ in\ the\ sample}$$

2. Precision: Precision measures how many positions are actually correct among all the positions that the algorithm found. The high precision indicates that the algorithm found significantly the correct positions more than incorrect positions.

$$Precision = \frac{correct\ cutting\ points\ in\ the\ result}{all\ found\ cutting\ points\ in\ the\ result}$$

3. F-measure: A weighted average of the recall and the precision is measured.

$$F - measure = 2 * \frac{Recall * Precision}{Recall + Precision}$$

4.2 Result

4.2.1 Arabic Results

Compared to the English, the results of the proposed approach (Dependence-Based Segmentation)are clearly lower on Table 5. The two obvious causes might be the irregularities in Arabic morphology and the typos in the test set. The latter is confirmed by results obtained when the sample is restricted to the words with a relatively high frequency (thresholded) as shown in Figure 1. Thresholded denotes an additional filter applied to *non-empty affix* sample, where a segmentation is picked only if the affix reappears in more than 1K other segmentations. It is worth noting that we omitted this filter in the English tests because we observed a rise in values of the standard deviation between the three samples. It has fluctuated around 0 to 29 words. Reasons for this deviation need to be investigated more closely in a separate work, and we paid our attention in this work to Arabic results.

Another reason affects the performance which is the lack of a dataset. The proposed approach works totally with the unsupervised method and depends only on the count of words in the corpus. However, it is difficult to include all possible derivatives in the corpus. For example, the derivative "الإجحافات" ([Al◦<ij◦HaAfaAt]: the prejudices) appears one time in the dataset; that is, there is no other derivative that appears in the dataset without the prefix "ال", for instance. Indeed, the word "إجحافات" would be more dependent on the prefix "ال" and the proposed approach does not learn that the prefix "ال" can be cut off from the word "الإجحافات" as shown at the top of Table 6. This problem will be removed if we add the word "إجحافات" into the dataset. The approach then finds the segmentation position "ال|إجحافات" as it appears at the bottom of Table 6.

The segmentation seems to be the best when the word counts are omitted (distinct). This seems to be due to the diversity of affixed in the classical Arabic. Indeed some stems/affixes may not appear more than once in the whole corpus (e.g. "كجخف" ([kajux◦f])as a deep sleep).

Surprisingly, Morfessor was unable to detect any boundary. The expected reason is that Morfessor may not be able to deal effectively with non-Latin letters, such as Arabic letters. We then evaluated Morfessor using Buckwalter Arabic transliteration [12]. Despite the simplicity of the proposed approach, the results of Morfessor is close to the results of the proposed approach especially after applying the filter.

Running the proposed approach over the whole corpus resulted in around 1,805,231 morphemes and 596,358 distinct morphemes with an average size of 5.94 letters per morpheme.

4.2.2 English Results

With no restriction on the results sample, Morfessor outperforms the proposed approach (Dependence-Based Segmentation) in terms of precision where the latter has a better recall as shown in Table 7 . This matches well the results given in [34]. The low recall of Morfessor explains the worsening of its performance with affixed words while the proposed approach improves. The F-measure shows that the overall performance of the proposed approach is better as shown in Figure 2.

When the counts of the words are involved in the computation of the dependence, Morfessor improves. The dependence-base's precision increases too, where its recall worsens.

Table 5: Arabic results

		Distinct		Plain	
		Dependence	Morfessor	Dependence	Morfessor
Raw	Precision	81.21±2.76	86.0215±0.61	78.66±6.73	83.6547±4.32
	Recall	48.11±1.32	76.0147±1.94	44.88±3.98	75.9473±1.74
	F-measure	60.62±1.47	80.7091±1.08	57.14±4.96	79.6149±1.67
Non-empty affix	Precision	78.73±4.97	83.2916±3.37	78.02±1.96	83.9520±5.34
	Recall	74.96±2.44	75.1702±4.66	69.26±1.90	72.0343±5.03
	F-measure	76.80±2.54	79.0228±4.08	73.38±1.90	77.5379±5.19
Thresholded	Precision	89.36±5.18	90.1493±4.60	85.80±8.13	89.9908±4.07
	Recall	78.69±2.30	73.5789±1.29	72.78±7.89	73.0709±2.74
	F-measure	83.69±3.49	81.0256±2.57	78.76±7.90	80.6530±2.43

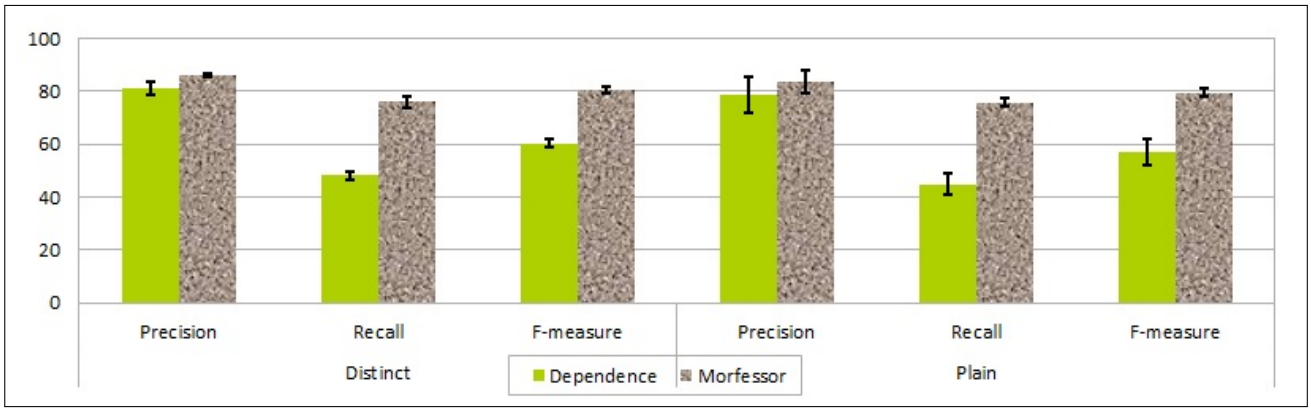
Table 6: Example of the lack of a dataset

i	u	FD(u)	v	BD(v)	Difference	Direction
1	#ا	0.991	$الإجحافات	7.174	-6.182	↓
2	#ال	9.349	$لإجحافات	12.204	-2.855	↓
3	#الإ	2.637	$إجحافات	26.555	-23.917	↓
4	#الإج	2.096	$جحافات	5.499	-3.402	↓
5	#الإجح	1.019	$حافات	3.443	-2.423	↓
6	#الإجحا	4.484	$افات	1.404	3.079	↑
7	#الإجحاف	11.911	$فات	1.296	10.614	↑
8	#الإجحافا	3.587	$ات	2.422	1.164	↑
9	#الإجحافات	17.184	$ت	0.885	16.298	↑
i	u	FD(u)	v	BD(v)	Difference	Direction
1	#ا	0.991	$الإجحافات	7.174	-6.182	↓
2	#ال	9.349	$لإجحافات	6.102	3.246	↑
3	#الإ	2.637	$إجحافات	26.555	-23.917	↓
4	#الإج	2.096	$جحافات	10.081	-7.984	↓
5	#الإجح	1.019	$حافات	3.374	-2.714	↓
6	#الإجحا	4.484	$افات	1.411	3.072	↑
7	#الإجحاف	11.911	$فات	1.298	10.612	↑
8	#الإجحافا	3.587	$ات	2.422	1.164	↑
9	#الإجحافات	17.184	$ت	0.885	16.298	↑

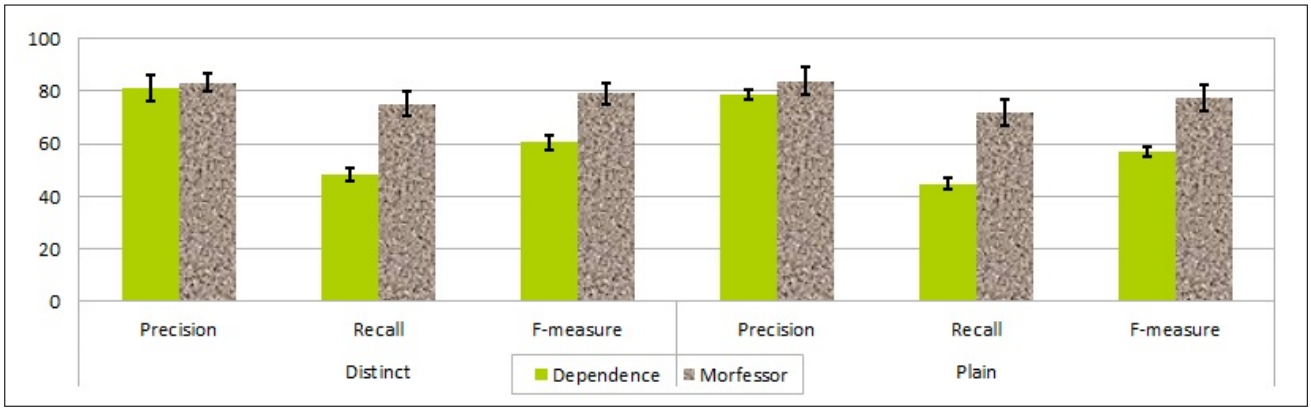
Table 7: English results

		Distinct		Plain	
		Dependence	Morfessor	Dependence	Morfessor
Raw	Precision	71.60±1.68	87.36±2.64	74.99±1.14	90.72±4.83
	Recall	78.11±3.98	70.90±4.84	79.66±1.88	71.98±4.73
	F-measure	74.71±2.17	78.27±3.57	77.25±0.73	80.27±2.52
Non-empty affix	Precision	75.82±1.98	89.73±1.74	76.26±2.34	91.37±1.24
	Recall	88.10±5.73	67.11±6.12	84.19±4.61	68.77±3.40
	F-measure	81.50±3.43	76.79±4.64	80.03±2.81	78.47±1.88

(A) Raw sample



(B) Non-empty affix sample



(C) Thresholded sample

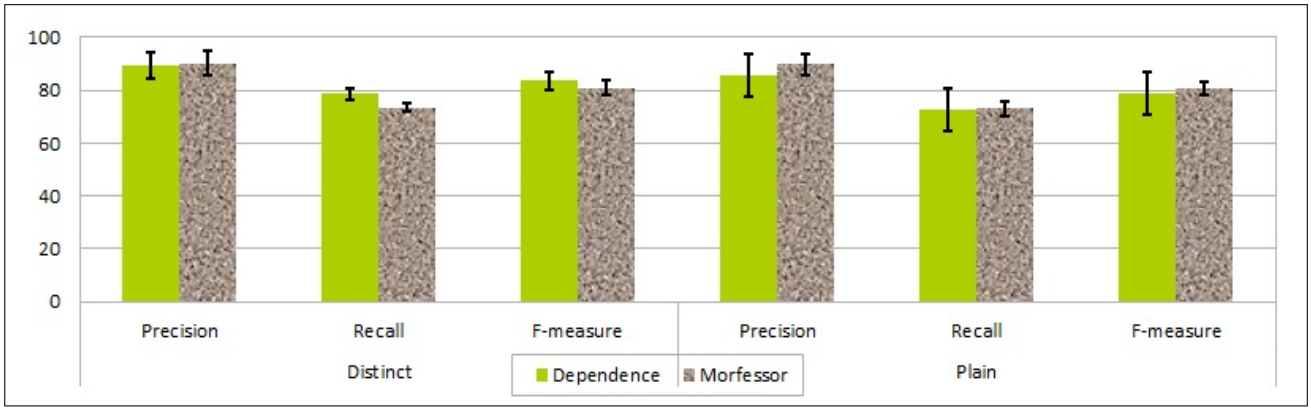


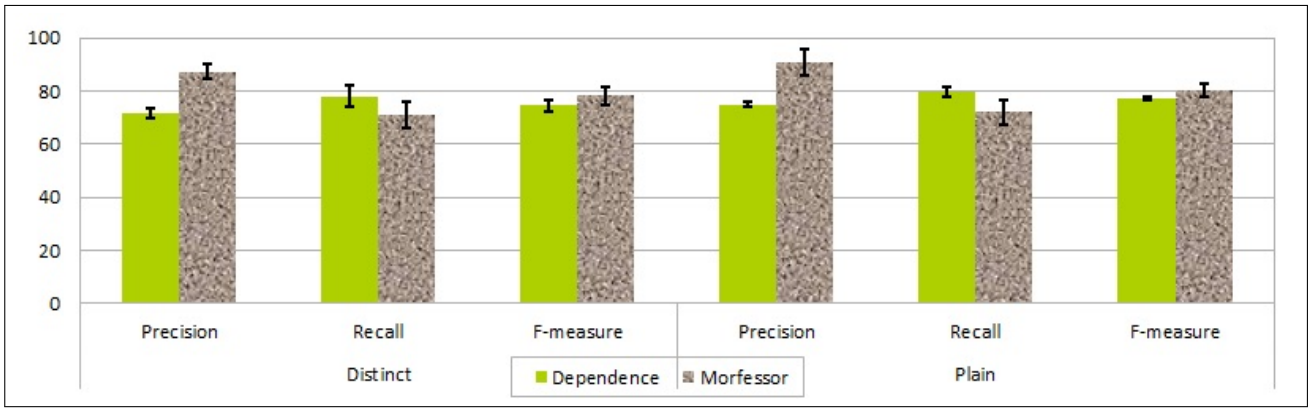
Figure 1: Arabic results of the two approaches

5 Conclusion and Future Work

We investigated the use of an intuitive yet formal definition of event dependence in the detection of morpheme boundaries. The initial target of our work was the classical Arabic, which is a Semitic language recognized to have a complex morphology and a vocabulary richer in pure Arabic words than the Modern Standard Arabic (MSA). The test also was conducted on an English set as well. Results on the classical Arabic and on English have been compared to the results of another unsupervised complicated segmentation approach so-called Morfessor that initially designed for European languages. As simple and elegant as is the proposed approach (Dependence-Based Segmentation) , it seems to be more committed to the recall where the latter's precision is better on English words. This has been confirmed when the evaluation focused on affixed words. On Arabic, Morfessor could not identify boundary. The results are a proof of concept that unsupervised techniques can decently handle a morphology as complex as the one of a legacy Semitic language like Arabic.

In addition, required for the evaluation, we carefully built a pure classical Arabic dataset. Where such linguistic resource is abundant in languages like En-

(A) Raw sample



(B) Non-empty affix sample

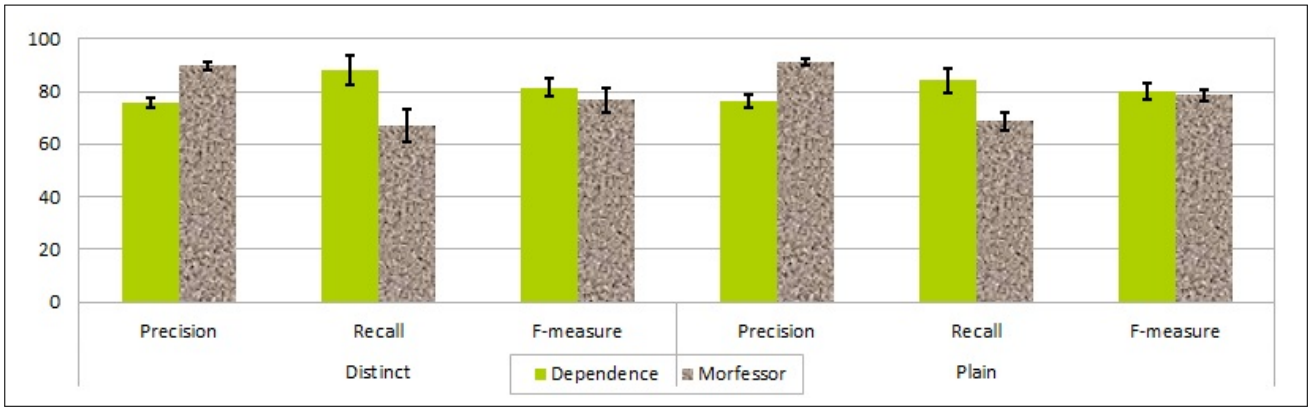


Figure 2: English results of the two approaches

glish, we are not aware of any other classical Arabic dataset.

One of the things made possible by morpheme boundaries identification is the automated word generation. The segmentation of a small set of well-chosen words will result in a set of diverse morphemes. The generation of new words is then a proper concatenation of the resulting morphemes. The challenge is then two folds: to build a suitable word-set of a decent size, which allows an accurate segmentation and a wide coverage of the different morphemes, then to find the appropriate concatenation strategy to minimize the number of irrelevant words.

References

[1] A. Khorsi and A. Alsheddi, "Unsupervised detection of morpheme boundaries," in *2016 4th Saudi International Conference on Information Technology (Big Data Analysis) (KACSTIT)*, Nov 2016, pp. 1–7.

[2] Z. Harris, *Morpheme Boundaries Within Words: Report on a Computer Test*, ser. Pennsylvania. University. Dept. of Linguistics. Transformations and discourse analysis papers. University of Pennsylvania, 1967.

[3] M. A. Hafer and S. F. Weiss, "Word segmentation by letter successor varieties," *Information Storage and Retrieval*, vol. 10, no. 11–12, pp. 371 – 385, 1974.

[4] J. R. Saffran, E. L. Newport, and R. N. Aslin, "Word segmentation: The role of distributional cues," *Journal of Memory and Language*, vol. 35, no. 4, pp. 606 – 621, 1996.

[5] S. Keshava and E. Pitler, "A segmentation approach to morpheme analysis," in *Working Notes for the CLEF Workshop 2007*, Hungary, 2007, pp. 1–4.

[6] S. H. Mustafa, "Character contiguity in n-gram-based word matching: The case for Arabic text searching," *Information Processing & Management*, vol. 41, no. 4, pp. 819–827, Jul. 2005.

[7] S. H. Mustafa, "Word-oriented approximate string matching using occurrence heuristic tables: A heuristic for searching Arabic text," *Journal of the American Society for Information Science and Technology*, vol. 56, no. 14, pp. 1504–1511, Dec. 2005.

[8] A. Khorsi, "On morphological relatedness," *Natural Language Engineering*, vol. 19, no. 4, pp. 537–555, Oct. 2013.

[9] Z. S. Harris, "From phoneme to morpheme," *Language*, vol. 31, no. 2, pp. 190–222, Jun. 1955.

[10] G. A. Kiraz, *Computational Nonlinear Morphology with Emphasis on Semitic Languages*. Cambridge, England: Cambridge University Press, 2001.

[11] S. Wintner, "Morphological processing of semitic languages," in *Natural Language Processing of Semitic Languages*, ser. Theory and Applications of Natural Language Processing, I. Zitouni, Ed. Springer Berlin Heidelberg, 2014, pp. 43–66.

[12] N. Habash, A. Soudi, and T. Buckwalter, "On Arabic transliteration," in *Arabic Computational Morphology*, ser. Text, Speech and Language Technology, A. Soudi, A. v. d. Bosch, and G. Neumann, Eds. Springer Netherlands, 2007, no. 38, pp. 15–22.

[13] M. Creutz and K. Lagus, "Unsupervised models for morpheme segmentation and morphology learning," *ACM Transactions on Speech and Language Processing*, vol. 4, no. 1, pp. 1–34, Jan. 2007.

[14] R. Falk and M. Bar-Hillel, "Probabilistic dependence between events," *The Two-Year College Mathematics Journal*, vol. 14, no. 3, pp. 240–247, Jun. 1983.

[15] D. Jurafsky and J. H. Martin, *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*. Prentice Hall, 2009.

[16] I. Al-Sughaiyer and I. Al-Kharashi, "Arabic morphological analysis techniques: A comprehensive survey," *Journal of the American Society for Information Science and Technology*, vol. 55, no. 3, pp. 189–213, Feb. 2004.

[17] S. Bordag, "Unsupervised and knowledge-free morpheme segmentation and analysis," in *Advances in Multilingual and Multimodal Information Retrieval*. Springer, 2008, pp. 881–891.

[18] D. Morrison, "Patricia—practical algorithm to retrieve information coded in alphanumeric," *JACM: Journal of the ACM*, vol. 15, no. 4, pp. 514–534, Oct. 1968.

[19] D. Bernhard, "Simple morpheme labelling in unsupervised morpheme analysis," in *Advances in Multilingual and Multimodal Information Retrieval*. Springer, 2008, pp. 873–880.

[20] O. Eroglu, H. Kardes, and M. Torun, "Unsupervised segmentation of words into morphemes," 2009.

[21] M. Melucci and N. Orio, "A novel method for stemmer generation based on hidden markov models," in *Proceedings of the twelfth international conference on Information and knowledge management*. ACM, 2003, pp. 131–138.

[22] A. Clark, C. Fox, and S. Lappin, *The Handbook of Computational Linguistics and Natural Language Processing*. John Wiley & Sons, 2013.

[23] J. Naradowsky and K. Toutanova, "Unsupervised bilingual morpheme segmentation and alignment with context-rich hidden semi-Markov models," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 2011, pp. 895–904.

[24] F. Peng and D. Schuurmans, "A hierarchical EM approach to word segmentation," in *In 6th Natural Language Processing Pacific Rim Symposium (NLPRS2001) Shai Fine, Yoram Singer, and Naftali Tishby.1998*, 2001, pp. 475–480.

[25] H. Poon, C. Cherry, and K. Toutanova, "Unsupervised morphological segmentation with log-linear models," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2009, pp. 209–217.

[26] R. Christensen, *Log-Linear Models and Logistic Regression*, ser. Statistics. Springer Verlag, 1997.

[27] S. Haberman, *The Analysis of Frequency Data*. Chicago: The University of Chicago Press, 1974.

[28] B. Can and S. Manandhar, "Methods and algorithms for unsupervised learning of morphology," in *Computational Linguistics and Intelligent Text Processing*. Springer, 2014, pp. 177–205.

[29] H. Hammarström and L. Borin, "Unsupervised learning of morphology," *Comput. Linguist.*, vol. 37, no. 2, pp. 309–350, Jun. 2011.

[30] J. Goldsmith, "Unsupervised learning of the morphology of a natural language," *Comput. Linguist.*, vol. 27, no. 2, pp. 153–198, Jun. 2001.

[31] R. Tavoli, E. Kozegar, M. Shojafar, H. Soleimani, and Z. Pooranian, "Weighted pca for improving document image retrieval system based on keyword spotting accuracy," in *2013 36th International Conference on Telecommunications and Signal Processing (TSP)*, July 2013, pp. 773–777.

[32] M. Keyvanpour, R. Tavoli, and S. Mozaffari, "Document image retrieval based on keyword spotting using relevance feedback," *International Journal of Engineering-Transactions A: Basics*, vol. 27, no. 1, p. 7, 2014.

[33] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to information retrieval*. Cambridge University Press, 2008.

[34] M. Creutz and K. Lagus, "Unsupervised models for morpheme segmentation and morphology learning," *ACM Transactions on Speech and Language Processing (TSLP)*, vol. 4, no. 1, 2007.

Appendix

Table 1: Sample of index of the corpus

اسم الملف	اسم الكتاب	المؤلف	سنة الوفاة	مكان الميلاد	مكان الوفاة	عدد الكلمات	المصدر
0.1.txt	القرآن الكريم	-	-	-	-	78,245	http://ksucorpus.ksu.edu.sa
44.1.txt	معاني القرآن وإعرابه	أبو إسحاق الزجاج، إبراهيم بن السري بن سهل	311 AH	بغداد	بغداد	558,238	http://ksucorpus.ksu.edu.sa
52.4.txt	المغازي	أبو عبد الله الواقدي، محمد الأسلمي	207 AH	المدينة المنورة	بغداد	279,632	http://shamela.ws
198.1.txt	فضائل القرآن وتلاوته	أبو الفضل، عبد الرحمن بن الرازي	454 AH	مكة المكرمة	نيسابور	12,245	http://shamela.ws
198.2.txt	أحاديث في ذم الكلام وأهله	أبو الفضل، عبد الرحمن بن الرازي	454 AH	مكة المكرمة	نيسابور	80,945	http://sourceforge.net/projects/newarabiccorpus

Table 2: Sample of *N*-grams table

N-gram	Frequency
#	1,138,616
#ا	157,361
#ال	120,542
#الك	2,856
#الكت	119
#الكتا	42
#الكتاب	21
#الكتابا	4
#الكتابان	1
الكتابان$	3
لكتابان$	3
كتابان$	4
تابان$	10
ابان$	142
بان$	1,871
ان$	29,092
ن$	111,887
$	1,138,616

#	1,180,304
# u	17,387
#un	7,583
#unb	219
#unbr	29
#unbre	9
#unbrea	6
#unbreak	4
#unbreaka	2
#unbreakab	2
#unbreakabl	1
#unbreakable	1
unbreakable$	2
nbreakable$	2
breakable$	5
reakable$	5
eakable$	7
akable$	14
kable$	85
able$	2,949
ble$	4,069
le$	19,211
e$	135,377
$	1,180,304

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 111-115 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Paper Improving Rule Based Stemmers to Solve Some Special Cases of Arabic Language

Soufiane Farrah*, Hanane El Manssouri, Ziyati Elhoussaine, Mohamed Ouzzif

RITM Laboratory, IT Department, EST - ENSEM, Casablanca, Morocco

ARTICLE INFO

Article history:
Received: 24 February, 2017
Accepted: 03 April, 2017
Online: 15 April, 2017

Keywords :
Rule Based
Arabic text
Stemming

ABSTRACT

Analysis of Arabic language has become a necessity because of its big evolution; we propose in this paper a rule based extraction method of Arabic text to solve some weaknesses founded on previous research works. Our approach is divided on preprocessing phase, on which we proceed to the tokenization of the text, and formatting it by removing any punctuation, diacritics and non-letter characters. Treatment phase based on the elimination of several sets of affixes (diacritics, prefixes, and suffixes), and on the application of several patterns. A check phase that verifies if the root extracted is correct, by searching the result in root dictionaries.

1. Introduction and Related Works

Arabic is a Semitic language spoken by more than 400 million people as a native language and ranked at the seventh position of Internet users in 2010. However, the task of performing the retrieve of information of Arabic language is very problematic, because of many aspects, such as: polysemy, irregular and inflected derived forms, various spelling of certain words, various writing of certain combination character, short vowels (diacritics) and long vowels, and the spectacular availability of affixes in the Arabic words [1, 2]. Different methods and approaches have been introduced to retrieve Arabic information [2, 3, 4, 5, 6].

To study Arabic morphology effectively, we divide words in Arabic into three self-contained categories as follows:

- ✓ اِسْم: It includes nouns, pronouns, adjectives, adverbs, etc
- ✓ فِعْل: Verbs
- ✓ حَرْف: Particles, articles, and conjunctions

Particles are completely unpredictable; they don't fall into the templatic system (i.e. they have no patterns) nor do they undergo any morphophonemic changes. They are what they are and must be memorized. The up side is that there are relatively few of them in the language – within one hundred.

Nouns and verbs do fall into the templatic system and have very systematic morphophonemic rules that govern them. This includes the study of how verbs are conjugated, how they move from pattern to pattern to enhance their meanings, how the participles and other nouns are derived, how nouns pluralize, etc.

Each declinable noun and each verb is made up of a certain set of base letters, called its root: جذر. (Nouns that are always indeclinable (such as pronouns) usually don't follow this system).

Verbs can have either 3 base letters, or 4. Nouns can have 3, 4 or 5. Now these base letters can be augmented with extra letters, and they can be dropped or changed due to morphophonemic rules as well.

Particles - The third part of speech in Arabic mentioned above is the particle. The meaning of a particle is often understood in the context of the sentence and words before and after the particle. The sign of the particle is that it does not accept the signs of nouns or verbs.

Analyzing Arabic text was treated by many researchers, all of them tried to extract an exact root or stem from a word, there is two ways to treat a text; morphological analyze, which consist to find roots, and there are statistical stemmers that group word variants using clustering techniques.

The first approach of morphological analyze is manually constructed dictionaries based on roots, Kharashi and Evans worked with small text collections, for which they manually built dictionaries of roots for each word to be indexed [7]. Tim

*Corresponding Author : Soufiane FARRAH, 46 Rue 17 Hay Anigret, No +212663782432, Email: f.soufiane@gmail.com

Buckwalter developed a set of lexicons of Arabic stems, prefixes, and suffixes, with truth tables indicating legal combination [1].

Nehar [5] and Taghva [8] introduces new stemming techniques that do not rely on any dictionary, the first one is based on the use of transducers. Nehar [5] proposed also a heavy stemmer that does not use any dictionary of roots. Khoja and R. Garside [2] developed a dictionary-based stemmer, and Larkey [5] developed a Java program based on their own Arabic stemmer that will develop and evolve to take in count some nouns and verbs categories described in the previous paragraph. Taghva [8] proposed IRSI Arabic Stemmer Algorithm, which does not use a Root Dictionary. ISRI stemmer per-forms better than the other approaches on the shorter title queries. For the long texts and narrative queries, stemming made a difference: the Khoja, ISRI, and Light stemmers were significantly better than not stemming. Ghwanmeh [9] presents an Arabic root-based algorithm based on patterns. This stemmer is restricted to native Arabic words that consist of four or more Arabic alphabets.

All algorithms mentioned before have some weaknesses. In this paper, we will prove that the best way of stemming is the one that have a strong preprocessing phase, and it is based on both "patterns check" and "root list".

This paper is an extension of work originally presented in conference [10].

In fact, we present the weakness in Heavy and Light Stemming Algorithms and we try to propose some new solutions for each point treated, then we will compare results of our new stemmer with other ones.

In section II, we present the different areas for improvement in Arabic text classification, in section III we present our approach, and in section IV we present some tested examples and, compare our solution with others.

2. Improvement Areas in Arabic Classification

Light stemming algorithms removes suffixes, and prefixes from words, producing a form of word called "stem" [11], there was many versions of the light stemming algorithms and the last one is light10 [12]. This algorithm after removing punctuation and non-letters, diacritics, Hamza from letter "أ", he replaces final letter "ة" with letter "ه" and then replace final letter "ى" with letter "ي". After that, the algorithm search in irregular word list to find out if the word exists on this table or not. Then the algorithm removes the letter "و" from the beginning of the words if the length of the word is more than three characters, because it considers that this letter is usually a conjunction.

This step generates several errors on stem extraction, I give below some examples:

Word	Stem extracted by Light10
وَبِيل	بيل
وُجُوب	جوب
وَرِيث	ريث

Table 1. Morphology of Arabic Word

As we can see, when removing the letter "و" from the beginning of those words, we change the meaning of the word, for the first word "وَبِيل" it means calamitous; disastrous, and when we remove the letter "و" the word means torch.

In Khoja's Approach, and TC system proposed by M.Hadni, A.Lachkar and S. Alaoui Ouatik in [13], and also in Mohammed N. Al-Kabi who proposed evolution of Khoja's algorithm [14], we find this same issue, so in our algorithm we will take care of this point and we propose to check if the word doesn't exist in the list of words that begins by 'و', and then remove diacritic 'و' (primarily weak vowels), this list is constituted manually and must be maintained regarding the evolution of Arabic language.

The second point we have improved is removing the letter "أ", in Light Approach, Khoja's Approach and M.Hadni's one, this letter is deleted because it is considered as a prefix. The issue is when this letter is a part of word as for the word "أباح" which means "permit", and when we delete it the word means "confide", to solve that, we built a list of words that starts with letter "أ".

The third point we involve in this paper is the stemming of five nouns (أب, أخ, حمو, فو, ذُو), those nouns are excluded from the other single nouns per the syntactic case. They have other marks to indicate them syntactic cases that the other doesn't have. The single noun always depends on rules to indicate its syntactic cases but the five nouns are contradicting those rules. The five nouns aren't depending on al Harakat (vowelization on system) rather than the letters. They have preconditions to be different from other single nouns:

- It has to be adjunct to another noun in other words there must be a noun after it that is genitive noun.
- The noun after must not be (ي) that indicates the speaker [15].

Therefore, for our algorithm, we handle the fives nouns separately.

It is true the orthography in Arabic is less ambiguous and more phonetic with the use of diacritics. For example, a word can be written using the same characters and be pronounced differently. The main purpose of diacritics including vowel marks, known as Harakat a phonetic is to provide, "حركات" aid to show the correct pronunciation. Arabic vowel marks include Fatha "فتحة", Kasra "كسرة", Damma "ضمة", Sukun "سكون", Shadda "شدة" and Tanwin "تنوين". The pronunciation of these vowel marks are represented in Table 2 below:

Double Constant	No Vowel	Tanwin			Vowel		
ـّ shadda	ـْ Sukun	ـٍ in	ـٌ un	ـً an	ـِ Kasrah	ـُ Ḍammah	ـَ Fatḥah

Table 2. Arabic Diacritics

However, in Modem Standard Arabic (MSA), vowel marks are not usually included in printed and electronic text, and the understanding and correct pronunciation of the word is determined within its context by the reader, so we decide not to remove (if it exists), the vowels as a step on preprocessing phase.

3. Important Steps

The method we propose is based on preprocessing step, treatment and check steps, here is a description of each one:

3.1. Preprocessing

In this step we proceed to:

- Divide text into words
- Format the word by removing any punctuation, diacritics and non-letter characters
- Check if the word is a stop word (3)

(3) This step consists on eliminating (very frequent) words that contain no or little information to help discriminate the text they occur in. A large list of stop words are used (Table 3).

إن	أنما	إنما	بعد	ضد
أحد	إلى	في	من	حتى
تلك	كذلك	التي	وهو	به
لكن	عن	منذ	على	وكان

Table 3 Stop Word

3.2. Searching in strange words list

In this step our algorithm will check if the word is a part of strange words (it is a word that comes from another language than Arabic, and used in the modern Arabic language especially), those words exists in a list of Strange words constituted manually (Table 4).

ألمانيا	فرنسا	ديسمبر	أوروبا	إفريقيا
اكلنيكي	ميكانيك	ناهوند	أوباما	بنكيران

Table 4 Strange Words

If the word exists the algorithm returns the word, otherwise the treatment continues.

3.3. Check if the word exists in the list of words that begins by "و":

The stemmer removes letter "و" ("and") from the beginning of the words if the length of the word is more than three characters, and if the word doesn't exist in the list of "Words_begins_by_Waw.txt", because many common Arabic words begin with this character.

ومي	وقر	وصب	وخط	وخض
وهث	وكز	وعع	وغي	وغف

Table 5 Words starting with *"و"*

3.4. Check if the word exists in the list of words that begins by "ال"

The stemmer removes letter "ال" from the beginning of the word if the length of the word is more than three characters, and if the word doesn't exist in the list of "Words_begins_by_AL" (table 6), because many common Arabic words begin with this character.

أتم	أبط	أبض	أبر	أبب
أول	أوض	أوش	أمس	أذي

Table 6 Words starting with "ال"

3.5. Normalization

The third step in the stemmer is normalization of the words. Normalization process in the proposed stemmer is the similar to the normalization process in Light10 stemmer which runs as following:

- Remove Hamza from letter "أ" (Replace " أ إ آ" with "ا")
- Replace final letter "ة" with "ه".
- Replace final letter "ى" with "ي".

3.6. Removing prefixes and suffixes

This step consists of removing the prefixes and suffixes from the words

س	و	ب	ل	لل
وا	ها	كما	تما	هما
نا	كن	تن	كم	تم
هم	هن	ين	ون	تا

Table 7 Prefixes and Suffixes

3.7. Check if the word matches any of the patterns

The last step after deleting the prefixes and suffixes of the words is correcting any word that its meaning changed. In some cases, a letter in the pattern of the word is deleted affecting the process of root extraction, like in the word.

In (فعل) which is the pattern of (رأى), the letter "أ" is deleted, and in the present tense of pattern (فعل) is (يفعل) which is introduced to (يرأى) and not (يرى). If we take (يرى) as a present verb, the past will be (رى) so the letter ' ع ' is deleted for this reason becoming (يفل) instead of (يفعل).

There are three rules which apply in the stemmer for correcting some words their meaning was affected:

1. Adding "ي" to the end of the word if the suffix "يه" is deleted
2. Adding "ه" to the end of the word if the suffix "ته" is deleted
3. Replacing the letter "ئ" to the end of the word by "ء" if the suffix of the word is deleted.

4. Detailed Algorithm

To implement our algorithm, we have used Java code program, based on Khoja's one.

These are the schema of our algorithm:

```
{
    // check if the word consists of two letters:
    if ( word.length ( ) == 2 )
        if the word consists of two letters, we treat two cases:
```

- A root consisting of two letters (though I can't think of any!)
- A letter was deleted as it is duplicated or a weak middle or last letter.

```
    // if the word consists of three letters
    if( word.length ( ) == 3 && !rootFound )
```

- If the last letter is a weak letter or a hamza, then check for last week words list.
- If the second letter is a weak letter then check for second week words list.

```
// if the word consists of four letters
if( word.length ( ) == 4 )
   // check if it's a root
```

- Check on the list of four letters root.

```
// if the root hasn't yet been found
if( !rootFound )
{
   // check if the word is a pattern
```

- Try and find a pattern that matches the word

```
}
// if the root still hasn't been found
if ( !rootFound )
{
   // check for a definite article, and remove it
   word = checkDefiniteArticle ( word );
```

- look through the vector of definite articles search through each definite article, and try and find a match
- Check to see if the word is a root of three or four letters
- If the word has only two letters, test to see if one was removed
- If the root hasn't been found, check for patterns/ check for suffixes/ prefixes

```
}
// if the root still hasn't been found
if ( !rootFound && !stopwordFound )
{
   // check for the prefix waw
   word = checkPrefixWaw ( word );
```

- Check to see if the word is a stopword
- Check to see if the word is a root of three or four letters, that begin by 'waw'
- If the word has only two letters, test to see if one was removed
- if the root hasn't been found, check for patterns
- Check for suffixes
- check for prefixes

```
}
// if the root STILL hasnt' been found
if ( !rootFound && !stopwordFound )
{
   // check for suffixes
   word = checkForSuffixes ( word );
}
// if the root STILL hasn't been found
if ( !rootFound && !stopwordFound )
{
   // check for prefixes
   word = checkForPrefixes ( word );
```

- Check to see if the word is a stopword
- Check to see if the word is a root of three or four letters.
- If the word has only two letters, test to see if one was removed
- if the root hasn't been found, check for patterns
- Check for suffixes
- check for prefixes

```
}
return word;
}
```

5. Results

The stemming result of the word will be correct, if the output form of the word is the same as the target form of the word. Otherwise, the result of the word will be incorrect. We have used Khoja's java code and we applied our approach on it. This program take in input a text file and returns in output a list of words theirs stems and the type of the each word. We have used as a first test a list of 524 words that begins by "waw" letter, we give below results of our stemmer comparing it to Khoja's one.

	Khoja's Stemmer	Our Stemmer : EST.Stemmer
Number of words used	524	534
Stemmed words	96,75 %	98,85 %
Not Stemmed words	3,25 %	1,15 %

Table 8 Results Comparison

Figures below shows the running result of both stemmers:

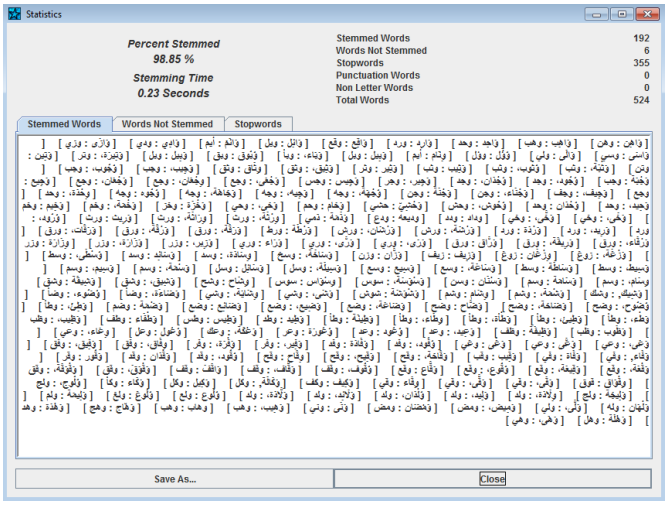


Figure 1 EST.Stemmer

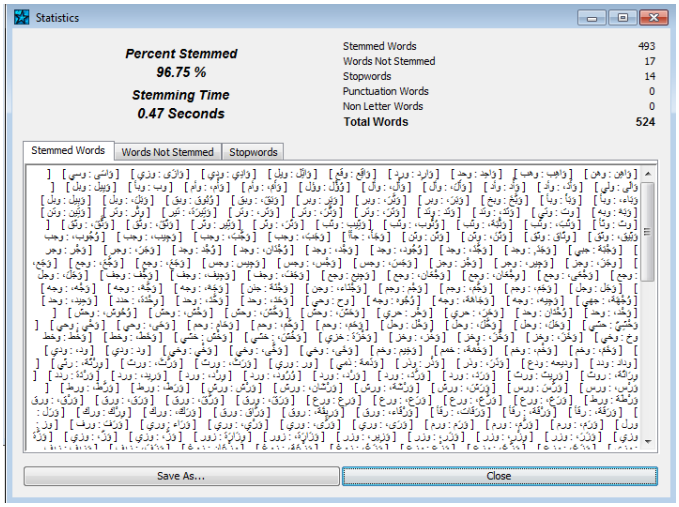


Figure 2 Khoja's stemmer

A second test that we have done with an article that contained 1418 words, results are presented in the table below:

	Khoja's Stemmer	Our Stemmer: EST.Stemmer
Number of words used	1418	1418
Stemmed words	93,16 %	93,23 %
Not Stemmed words	6,84 %	6,77 %

Table 9 Results Comparison

Since the text didn't contain words that begins by "waw", some strange words, and five nouns, the difference between the two results is not large.

EST.Stemmer is able solve some cases, for example if we take وَسَمَة ، وَسَامَة.

With Khoja's stemmer we have as stemmed text:

Word	Stem	Type
وَسْمَة	سمي	ROOT
وَسَامَة	سوم	ROOT

Table 10 Results Example

With our algorithm EST.Stemmer:

Word	Stem	Type
وَسْمَة	وسم	ROOT
وَسَامَة	وسم	ROOT

Table 11 Results Example

We have also review and modify stop word list to solve some issues detected is our tests, for example we have added 'ففي' and 'منهم' in this list.

6. Conclusion

In this paper we proposed new methods to improve the detection of the stem for Arabic language. Indeed, specific cases related to the five nouns and words starting with a vowels are processed successfully by the algorithm.

In the future work we will test the accuracy of our algorithm and compare it with Light and Heavy stemmers.

References

[1] Buckwalter,T.Qamus: Arabic lexicography http://members.aol.com/ArabicLexicons/

[2] Khoja, S.and Garside, R. Stemming Arabic Text. Computing Department, Lancaster University, Lancaster. Internet Home Page, 1-7 (1999). [Available at:http://www.comp.lancs. ac.uk/computing/users/khoja/stemer.ps].

[3] H. AlSerhan, R. A. Shalabi, and G. Kannan, "New approach for extracting arabic roots," in Proceedings of The 2003 Arab conference on Information Technology (ACIT 2003), Alexandria, Egypt, Dec. 2003, pp. 42–59.

[4] Nehar, Attia, Djelloul Ziadi, Hadda Cherroun, and Younes Guellouma. "An Efficient Stemming for Arabic Text Classification." In *Innovations in Information Technology (IIT), 2012 International Conference on*, 328–32. IEEE, 2012.

[5] Larkey, Leah S., et Margaret E. Connell. « Arabic Information Retrieval at UMass in TREC-10. » In *TREC*, 2001.

[6] Saudagar, Abdul Khader Jilani, Habeeb Vulla Mohammed, Kamran Iqbal, et Yasir Javed Gyani. « Efficient Arabic text extraction and recognition using thinning and dataset comparison technique ». In *Communication, Information & Computing Technology (ICCICT), 2015 International Conference on*, 1–5. IEEE, 2015.

[7] Al-Kharashi, I. A. and Evans, M. W. (1994) Comparing words, stems, and roots as index terms in an Arabic information retrieval system. Journal of the American Society for Information Science (JASIS) 45(8), pp 548-560.

[8] Taghva, Kazem, Rania Elkhoury, et Jeffrey S. Coombs. « Arabic Stemming Without A Root Dictionary. » In *ITCC (1)*, 152–157, 2005.

[9] Ghwanmeh, S., Kanaan, G., Al-Shalabi, R., Rabab'ah, 2009. "Enhanced Algorithm for Extracting the Root of Arabic Words". In: the Sixth International Conference on Computer Graphics, Imaging and Visualization, (CGIV '09). pp. 388–391.

[10] Soufiane Farrah, Hanane El Manssouri, El Houssaine Ziyati. « Proposition of improvement areas in most heavy an light stemmer algorithms novel stemmer: EST.Stemmer». In *Information Science and Technology (CIST), 4th IEEE CiSt'16 Program*. IEEE, 2016.

[11] M. A. H. Omer and S. Ma, Stemming Algorithm to Classify Arabic Documents , Journal of Communication and Computer, Vol. 7, No. 9, Sep. 2010.

[12] Elrajubi, Osama Mohamed. « An improved Arabic light stemmer ». In *Research and Innovation in Information Systems (ICRIIS), 2013 International Conference on*, 33–38. IEEE, 2013.

[13] Hadni, Meryeme, Abdelmonaime Lachkar, et S. Alaoui Ouatik. « A new and efficient stemming technique for Arabic Text Categorization ». In *Multimedia Computing and Systems (ICMCS), 2012 International Conference on*, 791–796. IEEE, 2012.

[14] Mohammed N. Al-Kabi, Saif A. Kazakzeh, Bilal M. Abu Ata, Saif A. Al-Rababahc, Izzat M. Alsmadi« A novel root based Arabic stemmer ». In *Journal of King Saud University - Computer and Information Science,* 94–103, Elsevier, April 2015.

[15] The Five Nouns In Arabic : https://fr.scribd.com/doc/93320581/The-Five-Nouns-In-Arabic.

[16] Chen, A., Gey, F., 2002. "Building an Arabic stemmer for information retrieval". In: Proceedings of the Eleventh Text REtrieval Conference (TREC 2002). National Institute of Standards and Technology, pp. 631–639.

[17] A.F.A. Nwesri, S.M.M. Tahaghoghi, F. Scholer "Stemming Arabic Conjunctions and Prepositions". Lect. Notes Comput. Sc., 3772 (2005), pp. 206–217 http://dx.doi.org/10.1007/11575832_23.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 116-124 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Medical imbalanced data classification

Sara Belarouci*, Mohammed Amine Chikh

Biomedical Engineering Laboratory, Department of Biomedical Engineering, Tlemcen University, 13000, Algeria

ARTICLE INFO

Article history:
Received: 19 March, 2017
Accepted: 04 April, 2017
Online: 15 April, 2017

Keywords :
Multilayer neural networks
k- nearest neighbor
Support vector machine
Imbalanced medical database
Least Mean Square algorithm

ABSTRACT

In general, the imbalanced dataset is a problem often found in health applications. In medical data classification, we often face the imbalanced number of data samples where at least one of the classes constitutes only a very small minority of the data. In the same time, it represent a difficult problem in most of machine learning algorithms. There have been many works dealing with classification of imbalanced dataset. In this paper, we proposed a learning method based on a cost sensitive extension of Least Mean Square (LMS) algorithm that penalizes errors of different samples with different weights and some rules of thumb to determine those weights. After the balancing phase, we apply the different techniques (Support Vector Machine [SVM], K- Nearest Neighbor [K-NN] and Multilayer perceptron [MLP]) for the balanced datasets. We have also compared the obtained results before and after balancing method. We have obtained best results compared to literature with a classification accuracy of 100%.

1. Introduction

Learning algorithms from imbalanced data has attracted a significant amount of interest in recent years. This is because in real world, imbalanced data exist in many applications, such as fault diagnosis [1], medical diagnosis [2], intrusion detection [3,4], text classification [5,6], financial fraud detection [7], data stream classification [8], and soon. In those applications, there are often one or some minority classes possessing very few samples compared with the other classes. And most of time, the "small" classes are more important than those "large" ones. Because of the unbalance data distribution of imbalanced learning problems, it is often difficult to obtain good performance for most cases by using traditional classifiers where a balanced distribution of classes is assumed and an equal misclassification cost for each class is assigned. As a result, traditional classifiers tend to be overwhelmed by the majority classes and ignore the minority ones, which is not acceptable in many real applications [9, 10].

Most previous works focused on the binary classification problems [11]. The others [12, 13] also tried to employ the multi-class data and define the class with a small number of data as the minority class while the other data are merged in to the majority class. Al though the minority class can be recognized by classifiers, the artificial majority class might be more likely to be misclassified. The knowledge of imbalanced data is complex especially when we solve the multi-class problems, since the amounts of some data classes are the same or similar to each other, which increases the difficulty to artificially select the minority class. The imbalanced learning problems can be summarized as two categories: absolute imbalance and relative imbalance [14]. The absolute imbalance occurs in the situation when the minority instances are significantly scarce and implicit, whereas the dataset with relative imbalance can show explicit data distribution but still rare quantity for minority examples. The characteristic of rare instances exists in the typical imbalance where the limited representative data lead to difficult learning regard less of between class imbalances. The other form of imbalance is within-class imbalance. It concentrates on the representative data distribution for the sub- concepts with in a class. The within-class imbalance problem seems to be more difficult than the datasets with the concepts in a similar characteristic [15, 16].

The works to be cited in Section 2 clearly show that most techniques cited in the literature were not able to find the best effective ways to address minority data.

In this paper, the learning method based on a cost-sensitive extension of Least Mean Square (LMS) algorithm is proposed to solve the imbalanced learning problems, and that penalizes errors of different samples with different weights, which increase the

*Corresponding Author: Sara Belarouci, Biomedical Engineering Laboratory, Department of Biomedical Engineering, Tlemcen University, 13000, Algeria
Email: belaroucisara@gmail.com

classification rate. In order to validate our empirical study, we have chosen three different algorithms from different paradigms of data mining, including Multilayer Perceptron (MLP), Support Vector Machines (SVMs) and the K-Nearest Neighbour (K-NN) as an Instance-Based Learning approach. We have also compared the results obtained before and after balancing the different datasets by the adopted LMS.

The rest of the paper is organized as follows. State of Art is presented in Section 2, it reviews several techniques applied for problems with imbalanced datasets. Next, section 3 presents the different techniques (MLP, SVM and K-NN) and our proposed method (LMS). In Section 4, the experimental work is presented also the obtained results are discussed and compared to other works in literature. Finally, Section 5 concerns conclusions and outlines possible directions for future research.

2. State of art

A variety of solutions has been proposed to address the imbalanced learning. To understand this issue comprehensively, most of the state of the art methods are generalized as the following categories. A critical and comprehensive survey on imbalanced learning can be found in [17].

Random oversampling for minority instances and under sampling for majority instances can facilitate change of the distribution for original dataset [18]. The data formed under sampling using K- Nearest Neighbor (K-NN) [19] is also presented. To overcome the disadvantages of the basic sampling methods, Such as, risk of overfitting for oversampling approach and risk of a loss of information for undersampling method, the Synthetic Minority Oversampling Technique (SMOTE) [20] is used. It selects one from the nearest neighbors for each original minority example, and generate synthetic minority data, based on the linear interpolations between the original examples and randomly selected nearest neighbors. Borderline Synthetic Minority Oversampling Technique (Borderline-SMOTE) [21] generates only synthetic data for the minority instances near the border rather than every original minority instance. Adaptive Synthetic (ADASYN) [22] is proposed to adaptively create the different quantities of synthetic data corresponding to the density distribution. Parallel Selective Sampling (PSS) technique [23] is proposed to select data from the majority class to reduce imbalance in large datasets. The PSS is a filter method, which can be combined with the Support Vector Machine (SVM) classification. The PSS-SVM showed excellent performances on synthetic datasets, much better than SVM. Other sampling strategies are integrated with ensemble learning techniques [24, 25] to address the imbalanced learning issue. The Synthetic Minority Oversampling Technique in boosting (SMOTEBoost) [26] algorithm is achieved via combining SMOTE with Adaboost.M2 (Adaptive Boosting.M.2.). Ranked Minority Oversampling in boosting (RAMOBoost) [27] adjusts the sampling weights of minority class examples based on the data distributions [16]. Other weighting approaches are proposed to overcome the problem of imbalanced datasets. The Least Mean Square (LMS) [28] algorithm is proposed to penalise errors of different samples with different weights and some rules of thumb to determine those weights. After the balancing phase, different classifiers (Support Vector Machine [SVM], K-Nearest Neighbour [K-NN] and Multilayer Perceptron [MLP]) are applied for the new balanced dataset. In addition, the results obtained by the LMS method are compared with the results obtained by the sampling methods (Under-sampling, Oversampling and SMOTE). Other local strategies are proposed to address the within-class imbalance issue of positive data sparsity, by directly adjusting the induction bias of specificity-oriented learning algorithms. The k Rare-class Nearest Neighbour (KRNN) algorithm [29] is proposed, where dynamic local query neighbourhoods are formed that contain at least k positive nearest neighbours and the positive posterior probability estimation is biased towards the rare class based on the size and positive distribution in local regions.

The goal of cost-sensitive learning [30- 33] is to calculate the costs for misclassification through different cost matrices. The Adaptive Cost sensitive boosting (AdaCost) [34] adopts the cost-sensitive learning with boosting. Cost-sensitive decision tree [35] can prune the scheme for imbalanced data with misclassification costs through specifying decision threshold. Cost-sensitive neural network models [36, 37] are also widely applied for imbalanced learning [16].

The kernel-based learning approaches include many state-of-the-art techniques for the application of data mining domain [38-41]. A Granular Support Vector Machines-Repetitive Undersampling (GSVM-RU) algorithm [42] carries out the iterative learning procedure based on GSVM. Kernel-Boundary Alignment (KBA) [43] is proposed to modify the kernel matrix via a kernel function based on the distribution of imbalanced data. There is another typical kernel-based learning algorithm for maximizing Area Under Curve (AUC) of the Receiver Operating Characteristic (ROC) graph [16, 44].

The active learning methods [45- 47] are traditionally adopted to handle the special issues relevant to training data without class labels (unlabeled data). As mentioned in [48], the criteria of termination for active learning methods are investigated to apply for the class imbalance issues on Word Sense Disambiguation (WSD) through maximal confidence and minimal error [16].

3. Materials and Methods

A brief description of the used algorithms is reported below:

3.1. Classification techniques used

In this work, we have used a K-Nearest Neighbor (K-NN) as a statistical machine, a Support Vector Machine (SVM) as a kernel machine, and a Multi-Layer Perceptron (MLP) as a neural network. Brief descriptions of these algorithms are already reported in literature [49].

3.2. Least Mean Square algorithm

The Least Mean Square (LMS) algorithm which is also called the stochastic gradient algorithm is relatively easy to implement and is based on a simple concept, it was introduced by Widrow and Hoff in 1960 [50].

The LMS algorithm is an adaptive algorithm, which uses a gradient-based method of steepest decent. LMS algorithm uses the estimates of the gradient vector from the available data. LMS incorporates an iterative procedure that makes successive corrections to the weight vector in the direction of the negative of the gradient vector, which eventually leads to the minimum mean square error.

Compared to other algorithms LMS algorithm is relatively simple; it does not require correlation function calculation nor does it require matrix inversions [51].

In the LMS algorithm, the squares of mean quadratic errors are minimized by solving a system of linear equations. In this paper, to remedy the problem of the learning of the imbalanced dataset, we used a cost-sensitive extension of Least Mean Square algorithm that penalizes errors of different samples with different weights.

LMS algorithm formulation

We are given a training set which consists of n independent identically distributed samples which is described by [50- 53]:

$$S = \{(x_1, y_1), \dots (x_n, y_n)\} \quad (1)$$

Where $x_i \in R^d$ and $y_i \in \{-1, 1\}$, $i = 1,2 \dots, n$. The classification consists of finding a hyper-plane $w.x + b = 0$ where $W \in \mathbb{R}^d$ and $b \in \mathbb{R}$.

The solution for the least mean square algorithm classification can be found by solving the following constrained minimization problem:

$$min_w \frac{1}{n}\sum_{i=1}^{n}(Y_i - W.X_i)^2 \quad (2)$$

The LMS algorithm is probably the most popular adaptive algorithm that exists because of its simplicity.

From the method of steepest descent, the weight vector equation is given by:

$$w_{i+1} = w_i + \frac{1}{2}\mu\,[-\nabla(E\{e_i^2\})] \quad (3)$$

Where μ is the step-size parameter and controls the convergence characteristics of the LMS algorithm; e_i^2 is the mean square error between the beam former output Y_i and the reference vector which is given by,

$$e_i = y_i - x_i^T w_i \quad (4)$$

The gradient vector in the above weight update equation can be computed as

$$\begin{aligned}\nabla(E\{e_i^2\}) &= \partial J\,[w_i]/\partial w_i \\ &= -2E\{x_i\, e_i\} \\ &= -2P + 2R\, w_i\end{aligned} \quad (5)$$

In the method of steepest descent, the biggest problem is the computation involved in finding the values P and R matrices in real time. The LMS algorithm on the other hand simplifies this by using the instantaneous values of covariance matrices P and R instead of their actual values i.e.

$$\begin{aligned}\widehat{R_\iota} &= x_i x_i^T \\ \widehat{P_\iota} &= x_i y_i\end{aligned} \quad (6)$$

These are simply the estimated instantaneous correlations.

Therefore, the weight update can be given by the following equation,

$$w_{i+1} = w_i + \mu\,[\,\widehat{P_\iota} - \widehat{R_\iota} w_i\,] \quad (7)$$

$$= w_i + \mu\, x_i\,[\,y_i - x_i^T w_i\,] = w_i + \mu\, x_i\, e_i \quad (8)$$

Note that w_i is a random variable [since each new iteration i, w_i depends random processes of x_i and y_i.

Therefore, the LMS algorithm can be summarized in following equations [53]:

- Filter output: $Y_i = w_i^T x_i$
- Error: $e_i = y_i - x_i^T w_i$
- Update filter Weight: $w_{i+1} = w_i + \mu\, x_i e_i$

The LMS algorithm is initiated with an arbitrary value $w(0)$ for the weight vector at $i = 0$. The successive corrections of the weight vector eventually leads to the minimum value of the mean squared error.

μ is the step-size parameter and controls the convergence characteristics of the LMS algorithm :

- If μ is chosen to be very small then the algorithm converges very slowly.
- A large value of μ may lead to a faster convergence but may be less stable around the minimum value.

The LMS algorithm is very simple: it requires only 2L + 1 multiplications and 2L additions by iteration, where L is the number of filter coefficients.

4. Results and discussions

4.1. Medical datasets

We have used five medical datasets from UCI database [54]. In order to validate the proposed methods on each one, we chose a subset of these datasets providing a heterogeneous test bench. These five datasets are Pima Indian Diabetes, Wisconsin Breast Cancer (WBC), Wisconsin Diagnostic Breast Cancer (WDBC), Liver disorder and Appendicitis. The main characteristics of these datasets are depicted in Table 1.

Table 1. Characteristics of the medical datasets.

Dataset	Classes	Attributes	Instances	Min. class	Maj. class
Pima	2	8	768	268	500
WBC	2	9	683	137	546
WDBC	2	31	569	212	357
Liver disorder	2	6	345	145	200
Appendicitis	2	7	106	21	85

4.2. Employed classifiers

In this subsection, we describe how we adjust some parameters of these techniques and how we estimate the classification reliabilities.

The K-NN algorithm requires no specific set-up. We test values of k $\{1, 3, 5, 7\}$ and we choose the value providing the best performances on a validation set according to a fivefold cross validation. We estimate the reliability of each classification act on the basis of information directly derived from the output of the expert and analyzing also the reasons in the feature space giving rise to unreliable classification. For further details, we may refer to [49, 55].

We test a SVM algorithm with a Gaussian radial basis kernel. Values of regularization parameter C and scaling factor σ are selected within intervals $[1;\ 10^4]$ and $[10^{-4};\ 10]$, adopting a log scale to sample the two intervals. The value of each parameter is

tuned using a fivefold cross-validation on a validation set. The reliability of a SVM classification is estimated as proposed in [56], where the decision value of the classifier is transformed in a posterior probability [49].

We use a MLP algorithm with a number of hidden layers equal to half of the sum of features number plus class number. The number of neurons in the input layer is fixed by the number of the features whereas we chose two neurons in the output layer. The reliability is a function of the values provided by neurons in the output layer [49, 55].

4.3. Statistical metrics

To assess the predictive ability of constructed models, five statistical evaluation methods were employed and they are defined as follows:

1) The correct classification rate (CC %): is the recognition rate;

$$CC = (TP + TN)/(TP + TN + FP + FN) * 100 \quad (9)$$

Where true positives (TP) denote the correct classifications of positive samples; true negatives (TN) denote the correct classifications of negative samples; false positives (FP) denote the incorrect classifications of negative samples into the positive samples; and false negatives (FN) denote the positive samples incorrectly classified into the negative samples.

2) Error rate (E %):

$$E = (FP + FN)/(TP + TN + FP + FN) * 100 \quad (10)$$

3) Sensitivity (SE %): the percentage of positive samples which are correctly classified;

$$\mathrm{SE} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN}) * 100 \quad (11)$$

4) *S*pecificity (SP %): the percentage of negative samples which are correctly classified;

$$SP = TN/(TN + FP) * 100 \quad (12)$$

5) Gmean (%): It provides a simple way to evaluate the model's ability to correctly classify the minority and majority class by the combination of Sensitivity and Specificity into a single metric. Gmean is considered as a measure of the balanced accuracy and is defined as:

$$\mathrm{Gmean} = \sqrt{\mathrm{Sensitivity} \times \mathrm{Specificity}} \quad (13)$$

4.4. Statistical performance of different classifiers combined with LMS

In order to validate the influence of LMS on the different classifiers (MLP, SVM, K-NN), it is interesting to compare the performances of MLP, SVM and K-NN techniques in the two cases (with and without LMS). However, selecting objective statistical metrics are used to estimate the performance of different classifiers. Indeed, for the imbalanced classification problem, the overall classification accuracy is often not an appropriate measure of performance given that a trivial classifier that predicts every sample as the majority class could achieve very high accuracy in extremely skewed domains. In the present work, instead of the complicated metrics, five intuitive and practical measures (correct classification rate, error rate, Sensitivity, Specificity and Gmean) were adopted to estimate the current classifiers based on the following reasons: first, both Sensitivity and Specificity provide a class-by-class performance estimate, making one easily investigation on the predictive ability of a classification method for each sample class, especially the predictive ability for the interesting minority classes; second, Gmean is a combination of both Sensitivity and Specificity, which indicates the balance between classification performance on the majority and minority classes. A poor performance in prediction of the positive (interesting) samples still leads to a low Gmean value, even if the negative samples are classified with high accuracy, which is a common case for imbalanced dataset. The comparative study results are summarized in table 2.

The classification of the different imbalanced databases used in this work involves four steps:

- Step 1: application of the different techniques (SVM, K-NN and MLP) on imbalanced data.
- Step 2: application of the LMS algorithm to remedy the imbalance of data.
- Step 3: application of the different techniques (SVM, K-NN and MLP) on obtained balanced data.
- Step 4: comparison between results obtain successively in first and second step.

We notice from these experiments that the classification performances (CC, SE, SP, and Gmean) increase after balancing databases by using the least mean square algorithm.

We remark that before balancing the different datasets, the minority class is hardly recognized by the different classifiers (MLP, SVM and K-NN). However, after balancing these imbalanced databases, the performance is improved significantly with the employment of LMS algorithm as illustrated in table 4; by increasing the Sensitivity, the specificity, the correct classification rate, and the Gmean. Therefore, we have obtained the best classification performances. We can say that the classifiers have a good recognition of the minority classes and the majority classes, since in our experimentations the samples of the minority classes and the majority classes are correctly classified (TP and TN will increase / FN and FP will decrease after balancing).

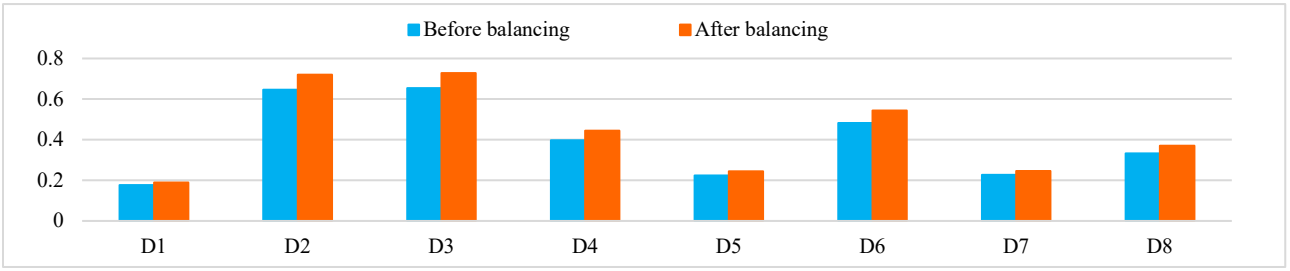
Behavior of descriptors before and after balancing approach

To validate the influence of LMS on the different techniques (MLP, SVM and K-NN), we compare the values of descriptors before and after balancing. So we take a miss-classified case from the minority class before balancing the different databases (PIMA, WBC, WDBC, liver disorder and Appendicitis); and we apply the LMS algorithm, where each descriptor is weighted by a coefficient, and the same case was correctly classified (see figure1).

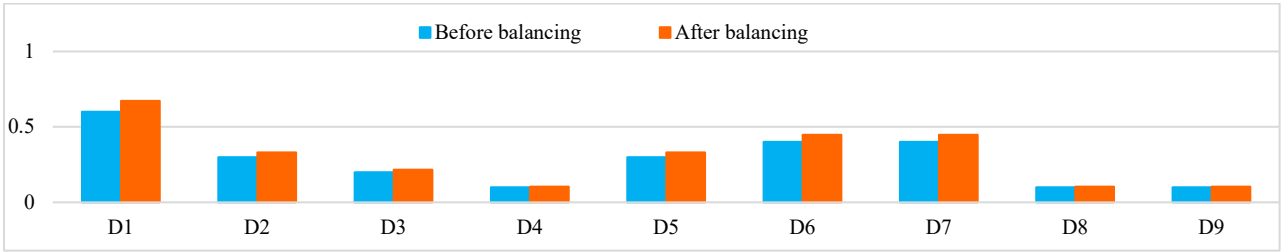
We notice also from this figure1 (a) that some descriptors in PIMA dataset remain unchanged (D1, D5, D7) however the rest has changed in a certain percentage which enhance the importance of the attributes. Also in the other used databases (WBC, WDBC, liver disorder and Appendicitis) we have obtained some changes in the different descriptors (figure1 (b), (c), (d), (e)).

Table 2. The obtained performances before and after balancing of the different imbalanced dataset on the testing base

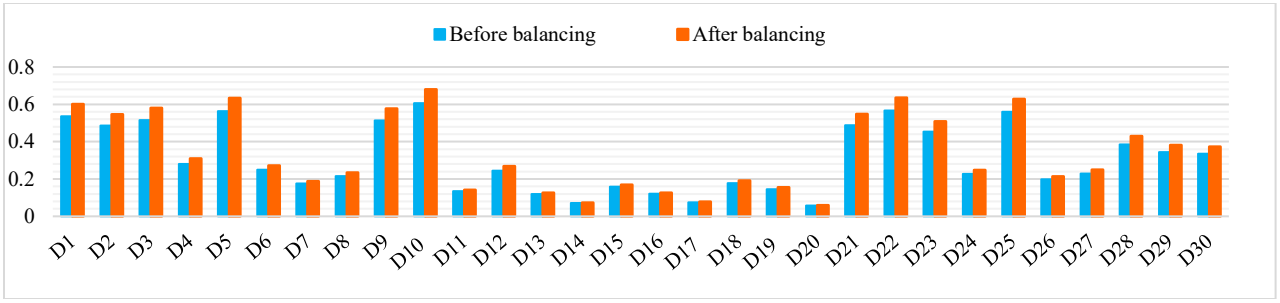
Datasets	Classifiers		CC (%)	E (%)	SE (%)	SP (%)	Gmean (%)
PIMA	MLP Classifier	MLP	73.85	26.15	51.28	83.52	65.44
		LMS	**99.24**	**0.76**	**97.44**	**100**	**98.71**
	SVM Classifier	SVM	83.85	16.15	87.18	82.42	84.77
		LMS	**99.23**	**0.76**	**100**	**98.90**	**99.45**
	K-NN Classifier	K-NN	80	20	61.54	87.91	73.55
		LMS	**100**	**0**	**100**	**100**	**100**
WBC	MLP Classifier	MLP	90.79	9.21	65.38	98.30	80.17
		LMS	**99.56**	**0.44**	**98.08**	**100**	**99.03**
	SVM Classifier	SVM	97.81	2.19	98.08	97.73	97.90
		LMS	**99.12**	**0.88**	**100**	**98.86**	**99.43**
	K-NN Classifier	K-NN	98.68	1.32	96.15	99.43	97.78
		LMS	**100**	**0**	**100**	**100**	**100**
WDBC	MLP Classifier	MLP	96.32	3.68	97.73	95.89	96.81
		LMS	**100**	**0**	**100**	**100**	**100**
	SVM Classifier	SVM	97.37	2.63	97.73	97.26	97.49
		LMS	**100**	**0**	**100**	**100**	**100**
	K-NN Classifier	K-NN	96.32	3.68	97.73	95.89	96.81
		LMS	**100**	**0**	**100**	**100**	**100**
Liver disorder	MLP Classifier	MLP	77.39	22.61	65.85	83.78	74.28
		LMS	**100**	**0**	**100**	**100**	**100**
	SVM Classifier	SVM	66.96	33.04	65.85	67.57	66.70
		LMS	**100**	**0**	**100**	**100**	**100**
	K-NN Classifier	K-NN	61.74	38.26	41.46	72.97	55.00
		LMS	**99.13**	**0.87**	**97.56**	**100**	**98.77**
Appendicitis	MLP Classifier	MLP	71.43	28.57	77.78	50.00	62.36
		LMS	**100**	**0**	**100**	**100**	**100**
	SVM Classifier	SVM	74.29	25.71	74.07	75.00	74.53
		LMS	**100**	**0**	**100**	**100**	**100**
	K-NN Classifier	K-NN	82.86	17.14	88.89	62.50	74.54
		LMS	**94.29**	**5.71**	**96.30**	**87.50**	**91.79**



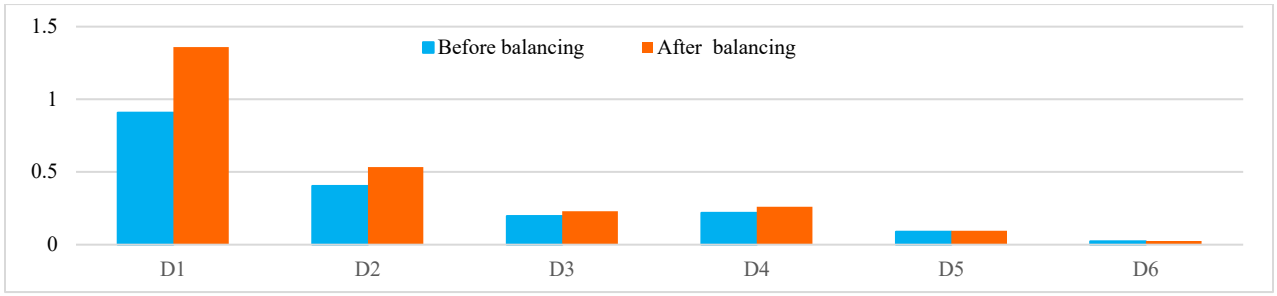
(a) PIMA (FN)



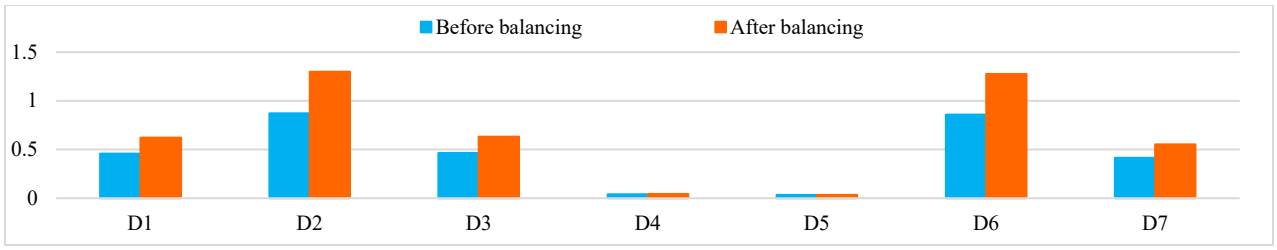
(b) WBC (FN)



(c) WDBC (FN)



(d) Liver disorder (FN)



(e) Appendicitis (FP)

Figure 1. The obtained results before and after balancing the different databases of a miss-classified case from the minority class.

4.5. Comparative study with related works

In this section, we have compared the classification accuracies of our method with other methods applied to the same database:

Works tested on PIMA database

Table 3 gives the classification accuracies of our method and other methods applied on the PIMA database.

Table.3 classification accuracies obtained with our method and other classifiers in literature (PIMA)

Method	Classification Accuracy (%)
GSVM [57]	74.15
WLTSVM [58]	76.78 ±0.35
MLP with LMS	**99.24**
SVM with LMS	**99.23**
K-NN with LMS	**100**

L. Gonzalez-Abril and al. have proposed a new Support Vector Machine method (called GSVM), which is specially designed for bi-classification problems its objective was balanced accuracy between classes [57]. For the evaluation of the results, (L. Gonzalez-Abril and al.) have used many databases (23) and obtained an accuracy of 74.15% for Pima dataset. Y.Shao and al. proposed an efficient Weighted Lagrangian Twin Support Vector Machine (WLTSVM) for the imbalanced data classification, they use different training points for constructing the two proximal hyperplanes [58], they achieve 76.78 ±0.35% of accuracy. In this work, as can be seen from the results (Table. 3), our method (MLP with LMS, SVM with LMS and K-NN with LMS) gave excellent classification accuracy.

Works tested on WBC database

Table 4 gives the classification accuracies of our method and other methods applied on the WBC database.

Table.4 classification accuracies obtained with our method and other classifiers in literature (WBC)

Method	Classification Accuracy (%)
S- AIRS [59]	96.91
WLTSVM [58]	96.30±0.31
MLP with LMS	**99.56**
SVM with LMS	**99.12**
K-NN with LMS	**100**

Wang and Adrian proposed a hybrid method by combining Synthetic Minority Over-Sampling Technique (SMOTE) and Artificial Immune Recognition System (AIRS) to handle the imbalanced data problem that are prominent in medical data . This approach denoted as S- AIRS [59]. They obtain 96.91% accuracy. Y. Shao and al. proposed WLTSVM [58] and they achieve 96.30± 0.31% of accuracy. In this study, as can be seen from the results (Table. 4), our approach obtain the best classification accuracy with the different classifiers.

Works tested on WDBC database

Table 5 gives the classification accuracies of our method and other methods applied on the WDBC database.

Table.5 classification accuracies obtained with our method and other classifiers in literature (WDBC)

Method	Classification Accuracy (%)
S- AIRS [59]	96.52
K-NN with resampling [60]	98.42
MLP with LMS	**100**
SVM with LMS	**100**
K-NN with LMS	**100**

Wang and Adrian proposed a hybrid method S- AIRS [59]. Their approach obtained 96.52% accuracy. G. NAGA RAMADEVI and al. applied the five classifiers K-NN, SVM, Logistic Regression, C 4.5 and Random Forest on original four breast cancer datasets with and without resampling technique, they compare the obtained performances before and after resampling datasets [60]. They obtain the best accuracy with 98.42% by using K-NN and resampling method. In this work, as can be seen from the results (Table. 5), our approach obtain the best classification accuracy.

Works tested on Liver disorder database

Table 6 gives the classification accuracies of our method and other methods applied on the Liver disorder database.

Table.6 classification accuracies obtained with our method and other classifiers in literature (Liver disorder)

Method	Classification Accuracy (%)
DGC+ [61]	67.44
GSVM [57]	71.07
MLP with LMS	**100**
SVM with LMS	**100**
K-NN with LMS	**99.13**

Alberto Cano and al. proposed an algorithm called weighted Data Gravitation Classification (DGC+) that compares the gravitational field for the different data classes to predict the class with the highest magnitude. The proposal improves previous data gravitation algorithms by learning the optimal weights of the attributes for each class and solves some of their issues such as nominal attributes handling, imbalanced data performance, and noisy data filtering [61]. They achieve 67.44% of accuracy. L. Gonzalez-Abril and al. proposed GSVM method [57]; they obtained an accuracy of 71.07%. In this work, as can be seen from the results (Table. 6), our approach obtain the best classification accuracy with the different classifiers.

Works tested on Appendicitis database

Table 7 gives the classification accuracies of our method and other methods applied on the appendicitis database.

Table.7 classification accuracies obtained with our method and other classifiers in literature (Appendicitis)

Method	Classification Accuracy (%)
DGC+ [61]	84.09
BSMAIRS [62]	92.5926
MLP with LMS	**100**
SVM with LMS	**100**
K-NN with LMS	**94.29**

Alberto Cano and al. proposed a DGC+ algorithm [61] and they achieve 84.09% of accuracy. Kung-Jeng and al. developed a hybrid classifier approach, they combine Borderline Synthetic Minority oversampling technique (BSM) and Artificial Immune Recognition System (AIRS) as global optimization searcher with the nearest neighbor algorithm used as a local classifier. This approach denoted as BSMAIRS. For the evaluation of the results, Kung-Jeng and al. have used a fivefold cross validation strategy and they have obtained five accuracies; the best one obtained with 92.5926% [62]. In this study, as can be seen from the results (Table.7), our approach obtains an excellent classification accuracy.

5. Conclusion

In this paper, we proposed a learning method based on a cost sensitive extension of least mean square algorithm that penalizes errors of different samples with different weights. This approach is used to overcome the problem of imbalanced data, it gives high weights for the samples of the minority classes.

The proposed approach was applied on five medical datasets from UCI database to assess its performance. Experimental results revealed that LMS algorithm performed better (achieved higher performance values) than the other balancing methods. It shows clearly the advantage of LMS when we handle the imbalanced data. Moreover, the results showed that the combination of LMS with different techniques (MLP, SVM and K-NN) can enhance classifier performance, particularly in terms of accuracy.

We propose that our approach will be applied for multiclass datasets, also the LMS algorithm will be tested on other intelligent methods based techniques such as fuzzy logic in order to increase the interpretability of the results. Also we can extend the ratio of the minority class to majority class in order to do a study in this situation. Our approach will be proposed to overcome the disadvantages of the basic sampling methods. Because the LMS algorithm that penalizes errors of different samples with different

weights. However, this approach does not eliminate the instances of the majority class and does not add the instances of the minority classes. We can conclude that, this method keeps the same database.

References

[1] D.Y. Liu, "Feature selection based on mutual information for gear faulty diagnosis on imbalanced dataset" J. Comput. Inf. Syst., Vol. 8, No.18, pp.7831–7838, 2012.

[2] L. Mena, J.A. Gonzalez, "Machine learning for imbalanced datasets: application in medical diagnostic" in: Proceedings of the 19th International FLAIRS Conference, May 11–13, Melbourne Beach, Florida, 2006.

[3] D.A. Cieslak, N.V. Chawla A. Striegel, "Combating imbalance in network intrusion datasets" in: Proceedings of 2006 IEEE International Conference on Granular Computing, May 10-12, pp. 732–737, Atlanta, USA, 2006.

[4] C. Thomas, "Improving intrusion detection for imbalanced network traffic" Security and Communication Networks, Vol.6, No.3, pp.309-324, 2013.

[5] Z. H. Zheng, X.Y. Wu, R. Srihari, "Feature selection for text categorization on imbalanced data" ACM SIGKDD Explor. Newslett, Vol. 6, No.1, pp. 80–89, 2004.

[6] Y. L. Li, G.S. Sun, Yehang Zhu, "Data imbalance problem in text classification" in: Proceedings of the Third International Symposium on Information Processing, October 15-17, pp. 301–305, Qingdao, China, 2010.

[7] J. Perols, "Financial statement fraud detection: an analysis of statistical and machine learning algorithms" AUDITING: J. Pract. Theory, Vol. 30, No.2, pp.19–50, 2011.

[8] A. Ghazikhani, R. Monsefi, H.S. Yazdi, "Ensemble of online neural networks for non-stationary and imbalanced data streams" Neurocomputing, Vol.122, pp. 535–544, 2013.

[9] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince and F. Herrera, "A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches" IEEE Transactions on systems, man, and cybernetics—part c: applications and reviews, Vol. 42, No. 4, pp. 463–484, 2012.

[10] Yun Qian, Yanchun Liang, Mu Li, Guoxiang Feng and Xiaohu Shi, "A resampling ensemble algorithm for classification of imbalance problems" Neurocomputing, Vol. 143, pp. 57–67, 2014.

[11] R. Pearson, G. Goney, and J. Shwaber, "Imbalanced clustering for microarray time-series" in Proceedings of the International Conference on Machine Learning, Workshop Learning from Imbalanced Data Sets II, Washington, USA, 2003.

[12] Y. Sun, M. Kamel, and Y. Wang, "Boosting for learning multiple classes with imbalanced class distribution" in Proceedings of 6th International Conference on Data Mining, December 18-22, Washington, USA, 2006.

[13] K. Chen, B. Lu, and J. Kwok, "Efficient classification of multi- label and imbalanced data using min-max modular classifiers" in Proceedings of the International Joint Conference on Neural Networks, July 16-21, Vancoudar, Canada, 2006.

[14] G. Weiss, "Mining with rarity: a unifying framework" ACM SIGKDD Explor. Newslet, Vol. 6, pp.7–19, 2004.

[15] T. Jo, N. Japkowicz, "Class imbalances versus small disjuncts" ACM SIGKDD Explor. Newslett, Vol. 6, pp.40– 49, 2004.

[16] Qiao Cai, Haibo He, and Hong Man, "Imbalanced evolving self-organizing learning, Neurocomputing" Vol. 133, pp.258–270, 2014.

[17] H. He, E. Garcia, "Learning from imbalanced data" IEEE Trans. Knowl. Data Eng, Vol. 21, pp.1263–1284, 2009.

[18] X. Liu, J. Wu and Z. Zhou, "Exploratory under sampling for class imbalance learning" in Proceedings of the 6th IEEE International Conference on Data Mining, Hong Kong, China, 2006.

[19] J. Zhang, I. Mani, "K-NN approach to unbalanced data distributions: a case study involving in formation extraction" in proceedings of the International Conference on Machine Learning, Workshop Learning from Imbalanced Data Set, Washington, USA, 2003.

[20] N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique" J. Artif. Intell. Res, Vol. 16, pp. 321– 357, 2002.

[21] H. Han, W. Wang, and B. Mao, "Borderline- SMOTE: a new over-sampling method in imbalanced datasets learning" in proceedings of the International Conference on Intelligent Computing, August 23-26, Hefei, China, 2005.

[22] H. He, Y. Bai, E. Garcia, and S. Li, "ADASYN: adaptive synthetic sampling approach for imbalanced learning" in proceedings of the International Conference on Neural Networks, June 1-8, Hong Kong, China, 2008.

[23] A. D'Addabbo and R. Maglietta, "Parallel selective sampling method for imbalanced and large data classification" Pattern Recognition Letters, Vol. 62, pp. 61–67, 2015.

[24] J. Caballero, F. Martinze, C. Hervas, and P. Gutierrez, "Sensitivity vs. accuracy in multiclass problems using memetic Pareto evolutionary neural networks" IEEE Trans. Neural Netw, Vol. 21, pp.750–770, 2010.

[25] N. Garcia-Pedrajas, "Constructing ensembles of classifiers by means of weighted instance selection" IEEE Trans. Neural Netw, Vol. 20, pp.258– 277, 2009.

[26] N. Chawla, A. Lazarevic, L. Hall and K. Bowyer, "SMOTEBoost: improving prediction of the minority class in boosting" in proceedings of the European Conference on Principles and Practice of Knowledge Discovery in Databases, September 22-26, Cavtat Dubrovnik, Croatia, 2003.

[27] S. Chen, H. He, and E. Garcia, "RAMOBoost: ranked minority over sampling in boosting" IEEE Trans. Neural Netw, Vol. 21, pp.1624–1642, 2010.

[28] Belarouci, S., Bouchikhi, S. and Chikh, M.A., "Comparative study of balancing methods: case of imbalanced medical data" Int. J. Biomedical Engineering and Technology, Vol. 21, No. 3, pp.247–263, 2016.

[29] X. Zhang, Y. Li, R. Kotagiri, L. Wu, Z. Tari, and M. Cheriet, "KRNN: k Rare-class Nearest Neighbour classification" Pattern Recognition, Vol. 62, pp. 33–44, 2017.

[30] Y.Sun, M. Kamel, A. Wong, and Y. Wang, "Cost- sensitive boosting for classification of imbalanced data" Pattern Recognit, Vol. 40, pp.3358–3378, 2007.

[31] K. Ting, "A comparative study of cost- sensitive Boosting algorithms" in proceedings of the International Conference on Machine Learning, San Francisco, CA, USA, 2000.

[32] F. Cheng, J. Zhang, and C. Wen, "Cost-Sensitive Large margin Distribution Machine for classification of imbalanced data" Pattern Recognition Letters, Vol. 80, pp. 107–112, 2016.

[33] S. Ali, A. Majid, S. G. Javed, and M. Sattar, "Can-CSC-GBE: Developing Cost-sensitive Classifier with Gentleboost Ensemble for breast cancer classification using protein amino acids and imbalanced data" Computers in Biology and Medicine, Vol. 73, pp. 38–46, 2016.

[34] W. Fan, S. Stolfo, J. Zhang and P. Chan, "AdaCost: misclassification cost-sensitive boosting" in proceedings of the 6th International Conference on Machine Learning, Slovenia, 1999.

[35] M. Maloof, "Learning when datasets are imbalanced and when costs are unequal and unknown" in proceedings of the International Conference on Machine Learning, Workshop Learning from Imbalanced Data Sets II, Washington, USA, 2003.

[36] Z. Zhou, X. Liu, "Training cost- sensitive neural networks with methods addressing the class imbalance problem" IEEE Trans. Knowl. Data Eng., Vol. 18, pp.63– 77, 2006.

[37] M. Kukar, I. Kononenko, "Cost- sensitive learning with neural networks" in proceedings of the European Conference on Artificial Intelligence, Wiley, New York, 1998.

[38] K. Muller, S. Mika, G. Ratsch, K. Tsuda, B. Scholkopf, "An introduction to kernel- based learning algorithms" IEEE Trans. Neural Netw, Vol. 12, pp.181– 201, 2001.

[39] X. Hong, S. Chen, C. Harris, "A kernel- based two-class classifier for imbalanced data sets" IEEE Trans. Neural Netw, Vol. 18, pp.28– 41, 2007.

[40] M. Pérez-Ortiz, P. A. Gutiérrez, and C. Hervás-Martinez, "Borderline Kernel Based Over-Sampling" J.-S. Pan et al. (Eds.): HAIS 2013, LNAI 8073, Springer-Verlag Berlin Heidelberg, pp. 472–481, 2013.

[41] P. Li, K. L. Chan, S. Fu, and S. M. Krishnan, "Kernel Machines for Imbalanced Data Problem in Biomedical Applications" Y. Ma and G. Guo (eds.), Support Vector Machines Applications, Springer International Publishing Switzerland, 2014.

[42] Y. Tang, Y. Zhang, "Granular SVM with repetitive under sampling for highly imbalanced protein homology prediction" in proceedings of the International Conference on Granular Computing, May 10-12, Atlanta, USA, 2006.

[43] G. Wu, E. Chang, "KBA: kernel boundary alignment considering imbalanced data distribution" IEEE Trans. Knowl. Data Eng, Vol. 17, pp.786 –795, 2005.

[44] T. Fawcett, "An introduction to ROC analysis", Pattern Recognit. Lett. Vol. 27, pp.861–874, 2006.

[45] S. Ertekin, J. Huang, L. Bottou and C. Giles, "Learning on the border: active learning in imbalanced data classification" in proceedings of the ACM Conference on Information and Knowledge Management, Lisbon, Portugal, 2007.

[46] A. Bordes, S. Ertekin, J. Weston and L. Bottou, "Fast kernel classifiers with online and active learning" J. Mach. Learn. Res, Vol. 6, pp.1579– 1619, 2005.

[47] M. Zieba and J. M. Tomczak, "Boosted SVM with active learning strategy for imbalanced data" Soft Comput, Vol. 19, pp.3357–3368, 2015.

[48] J. Zhu, E. Hovy, "Active learning for word sense disambiguation with methods for addressing the class imbalanced problem" in proceedings of the International Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Prague, Czech Republic, 2007.

[49] Roberto D'Ambrosio, "Handling Imbalanced Datasets by Reconstruction Rules in Decomposition Schemes" Ph.D Thesis, Campus Bio-Medico University, Roma, Italy, 2014.

[50] Christian JUTTEN, "FILTRAGE LINEAIRE OPTIMAL" During the fifth year of the department 3i options Images and Signals and Automatic to Joseph Fourier – Polytech' Grenoble University, 2010.

[51] Thomas FILLON, "Traitement Numérique du Signal Acoustique pour une Aide aux Malentendants" Ph.D Thesis, Paris University, Paris, France, 2004.

[52] Yilun Chen, Yuantao Gu, and Alfred O. Hero, "Regularized Least Mean Square Algorithms" In IEEE, 2010.

[53] J.-F. Bercher and P. Jardin, "Introduction au filtrage adaptatif" ESIEE Paris, 2003.

[54] UCI MACHINE LEARNING REPOSITORY, http://archive.ics.uci.edu/ml/data / (last update 01/03/2015).

[55] L. P. Cordella and al., "Reliability parameters to improve combination strategies in multi-expert systems" Pattern Analysis & Applications, Vol. 2, No. 3, pp. 205–214, 1999.

[56] J.-C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods" Advances in Large Margin Classifiers, pp. 61–74, Cambridge, MA, 1999.

[57] L. Gonzalez-Abril, H. Nuñez, C. Angulo and F. Velasco, "GSVM: An SVM for handling imbalanced accuracy between classes inbi-classification problems" Applied Soft Computing, Vol.17, pp. 23–31, 2014.

[58] Yuan-Hai Shao, Wei-Jie Chen, Jing-Jing Zhang, Zhen Wang and Nai-Yang Deng, "An efficient weighted Lagrangian twin support vector machine for imbalanced data classification" Pattern Recognition, 2014.

[59] Kung Jeng Wang and Angelia Melani Adrian, "Breast Cancer Classification Using Hybrid Synthetic Minority Over-Sampling Technique and Artificial Immune Recognition System Algorithm" International Journal of Computer Science and Electronics Engineering (IJCSEE), Vol. 1, No. 3, 2013.

[60] G. NAGA RAMADEVI, Dr. K. USHA RANI and Dr. D. LAVANYA, "Evaluation of Classifiers Performance using Resampling on Breast cancer Data" International Journal of Scientific & Engineering Research, Vol. 6, No. 2, 2015.

[61] Alberto Cano, Amelia Zafra, and Sebastián Ventura, "Weighted Data Gravitation Classification for Standard and Imbalanced Data" IEEE Transactions on cybernetics, Vol. 43, No. 6, 2013.

[62] Kung-Jeng Wang, Angelia Melani Adrian, Kun-Huang Chen and Kung-Min Wang, "A hybrid classifier combining Borderline-SMOTE with AIRS algorithm for estimating brain metastasis from lung cancer: A case study in Taiwan" computer methods and programs in biomedicine, Vol.119, pp. 63–76, 2015.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 125-136 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

ADOxx Modelling Method Conceptualization Environment

Nesat Efendioglu*, Robert Woitsch, Wilfrid Utz, Damiano Falcioni

BOC Asset Management GmbH, Innovation Group, AT-1040, Austria

A R T I C L E I N F O

Article history:
Received: 03 March, 2017
Accepted: 07 April, 2017
Online: 18 April, 2017

Keywords :
Meta-modelling, Modelling Method Design, Agile Modelling Method Engineering Conceptualization

A B S T R A C T

The importance of Modelling Methods Engineering is equally rising with the importance of domain specific languages (DSL) and individual modelling approaches. In order to capture the relevant semantic primitives for a particular domain, it is necessary to involve both, (a) domain experts, who identify relevant concepts as well as (b) method engineers who compose a valid and applicable modelling approach. This process consists of a conceptual design of formal or semi-formal of modelling method as well as a reliable, migratable, maintainable and user friendly software development of the resulting modelling tool. Modelling Method Engineering cycle is often under-estimated as both the conceptual architecture requires formal verification and the tool implementation requires practical usability, hence we propose a guideline and corresponding tools to support actors with different background along this complex engineering process. Based on practical experience in business, more than twenty research projects within the EU frame programmes and a number of bilateral research initiatives, this paper introduces the phases, corresponding a toolbox and lessons learned with the aim to support the engineering of a modelling method. "The proposed approach is illustrated and validated within use cases from three different EU-funded research projects in the fields of (1) Industry 4.0, (2) e-learning and (3) cloud computing. The paper discusses the approach, the evaluation results and derived outlooks.

1. Introduction

This paper is an extension of work originally published in 2016 IEEE 20th International Enterprise Distributed Object Computing Workshop (EDOCW)[1]

The importance of Modelling Method Engineering is equally rising with the importance of Domain Specific Conceptual Modelling Methods and individual modelling approaches. In addition to existing (de-facto-) standards (e.g. Business Process Model and Notation (BPMN)[2], Decision Model and Notation (DMN)[3], Case Management Model and Notation (CMMN)[4]), a growing number of groups around the world design their individual modelling-methods (in accordance with the definition of such a method by [5], [6]for a variety of application domains.).

This is often seen as necessary, when model-based approaches are transferred in new application domains and hence require adaptations for modelling methods. A simple sample can demonstrated using the well-known standard for business process BPMN. Although BPMN can be used to design a administrative process, such as sending an invoice, it cannot be used to design a simple production process like producing a chair. The successor relation that indicates that one activity follows the other does not have properties like distance to the station, which is not necessary when sending an invoice, but is of utmost importance, when producing a chair. When analysing more complex scenarios like a car manufacturer shop floor (we faced in projects DISRUPT [60], GO0DMAN [65]), the adaptation requirements for a modelling language like BPMN becomes quite complex. Hence, providing well-known model-based approaches requires the adaptation by e.g. introducing the concept “distance” between two activities.

Challenging question is, how to support the generation of modelling tools that can range from a minor adaptation like the one introduced above, till the complete realisation of totally new modelling approach like a cyber threat modelling for cloud computing .

The authors of [8] believe that supporting the automatic generation of modelling tools can open a new quality in

*Corresponding Author: Nesat Efendioglu BOC Asset Management GmbH, Innovation Group, AT-1040, Austria
Email: nesat.efendioglu@boc-eu.com

information systems development for engineers and customized design as well as encourage the use of modelling languages that are fitting to the custom needs. Often customer needs are defined by desired features like visualisation, querying, simulation or configuration and transformation, which are applied onto the model. Individual solutions enable the generation of light-weighted solutions with targeted provision of features that reduce the developers' aversion against overloaded modelling languages and inflexible or expensive modelling tools.

The engineering of such modelling tools is a result of the so called "conceptualization process of modelling methods", which is disseminated by the world-wide community of OMiLAB[10]. The complexity lies in the mapping of language artefacts and their corresponding functionality to concrete implementable and deployable modelling tools. In addition, knowledge domains are divided into more refined and specialized subdomains, where each domain needs to be characterized by its own abstraction and refinement of concepts from the so-called "real world" objects from the "subject under study". Hence, in order to capture the relevant semantic primitives that address domain specific needs, it is necessary to involve both the method engineers as well as domain experts. Today, there are different approaches, guidelines and practices for the development of modelling tools available that do not consider the full lifecycle from the design and collaborative development of a modelling tool, which unavoidably leads to limitations and inconsistencies in the conceptualization [7]. We, hence, propose a guideline and corresponding tools supporting method engineers along the conceptualization process supporting all phases and ensuring collaboration among involved stakeholders.

Karagiannis proposes in [6] the Agile Modelling Method Engineering (AMME) framework. Authors of [9] propose the Modelling Method Conceptualization Process that based on AMME and guides the method engineers during conceptualization. The same authors in [35] propose a toolbox that supports this process. The work at the hand introduces an extended version of this toolbox so-called "Modelling Method Conceptualization Environment" as well as introduces corresponding services. Hence the paper introduces a product-service-system proposal for modelling method conceptualisation. Moreover the paper evaluates this product-service system in three European Research projects, and one additional in the context of an in-house research project, and discusses evaluation results.

In this respect, the remainder of the paper is structured as follows: Section 2 briefly presents results of our state of the art analysis from [9], revisits AMME, the Modelling Method Conceptualization Process. Section 3 outlines the toolbox Modelling Method Conceptualization Environment with an emphasis on new extension to ADOxx Library Development Environment, so-called ADOxx-JAVA-MM-DSL. Section 4 presents Modelling Method Conceptualization Services, Section 5 introduces evaluation cases and discusses the evaluation results, while Section 6 concludes the paper and gives an outlook on future work.

2. State of the Art in Modelling Method Definition and Development Approaches

We published the initial version of this state of the art analysis in [9]. In this paper, we revisit shortly it for completeness reasons.

We analyse modelling standards from five well-known standardization organizations, which provide intensively used industry modelling standards in last decade; (1) OASIS (www.oasis-open.org), (2) OMG (www.omg.org), (3) Open Group (www.opengroup.org), (4) W3C(www.w3.org) and (5) WfMC(www.wfmc.org). It is highly possible that each organization follows same or similar approach and technologies to specify standards within its organization. Therefore, we assume that, it would be enough selecting control samples from each standardization organization in order to understand, which approach they are following, which tools they are using to define modelling standards. As the control samples we select UBL [39], ebXML BPSS [40] and WS-BPEL [53] from OASIS; BPMN [2] CMMN [4] and DMN [3] from OMG; ArchiMate® [42] from Open Group; OWL [44], SPARQL [45] and WSDL [46] from W3C, finally XPDL [47] from WfMC.

We investigate, how the modelling language of modelling standards has been specified. We focus on four aspects specification of (1) abstract syntax-, (2) semantic- and (3) notation (concrete syntax) of the modelling languages as well as (4) samples provided to ease to understand modelling language specification and its usage.

The results of the analysis are presented in Table 1. In respect of abstract syntax specification organization used (1) graphical approaches and/or, (2) formal textual approaches and/or textual informal approaches. All of organizations utilize UML-Class Diagram [41] to specify abstract syntax of most of their modelling languages. In addition to UML-Class Diagram all standardization institute use natural language to specify the abstract syntax. Additionally, in order to ensure interoperability between systems using the given modelling standard and collaboration among parties utilizing the standard, most of the organizations attach importance to formal textual specification of abstract syntax either using with BNF [48] or XSD [49].

It seems that, regarding to specification of semantics there is no standard/approach used commonly by all organization. W3C uses formal languages such as RDF [50], Z Notation [54] [51] and standard mathematical notation. However, most of the organizations utilize semi-formal keywords defined in RFC-2119 [52] in order to define requirement level and constraints. Additions to these, all organizations use natural language to specify semantics of all modelling standards.

Regarding to specification of notation (concrete syntax), obviously the modelling standards, which have been developed preferentially in the first place for human interpretation, precise has concrete syntax that is illustrated with images and described with natural language. On the other side, the standards, which have been developed for machine interpretation, have concrete syntax specified with formal textual approaches. Mostly, XSD is utilized for specification of concrete syntax of these standards. Finally, all the organizations choose to introduce features of their modelling standards with samples. The samples can be found in either in form of graphical or textual models.

According to analysis results, for the specification of selected modelling standards, the UML (UML Class Diagram) has been mostly utilized to specify the meta-model of the modelling language.

Table 1 Figure 1 Approach, standards and artifacts used to specify modelling languages in control samples

		Abstract Syntax			Semantics			Notation			Sample
		Graphical	Textual (formal)	Textual informal)	Formal Languages	Semi-formal Languages	textual (informal)	Graphical	textual (formal)	textual (informal)	
OASIS	Universal Business Language (UBL)	UML-Class Diagram	BNF	Natural Language			Natural Language		XSD	Natural Language	Sample Graphical Models
	The eBusiness eXtensible Markup Language (ebXML) Business Process Specification Schema (BPSS)		XSD	Natural Langauge		RFC-2119			XSD	Natural Language	Sample Textual Models
	Web Services Business Process Execution Language (WS-BPEL)	UML-Class Diagram	XSD	Natural Language		RFC-2119			XSD	Natural Language	Sample Graphical Models
OMG	Business Process Model and Notation (BPMN)	UML-Class Diagram	XSD	Natural Language		RFC-2119; WS-BPEL	Natural Language	Image	XSD	Natural Language	Sample Graphical Models
	Case Management Model and Notation (CMMN)	UML-Class Diagram	XSD	Natural Language		RFC-2119	Natural Language	Image		Natural Language	Sample Graphical Models
	Decision Model and Notation (DMN)	UML-Class Diagram	XSD; BNF	Natural Language			Natural Language	Image		Natural Language	Sample Graphical Models
Open Group	ArchiMate	UML-Class Diagram		Natural Langauge			Natural Language	Image		Natural Language	Sample Graphical Models
W3C	Web Ontology Language (OWL)	UML-Class Diagram	BNF	Natural Langauge	RDF	RFC-2119			XSD ;BNF; RDF	Natural Language	Sample Textual Models
	SPARQL Protocol and RDF Query Language (SPARQL)		BNF	Natural Langauge	Standard mathematical notation		Natural Language		BNF;RFC-3986; RFC-3987	Natural Language	Sample Textual Models
	Web Services Description Language (WSDL)		XSD; BNF		Z Notation	RFC-2119	Natural Language		XSD	Natural Language	Sample Textual Models
WfMC	XML Process Definition Language (XPDL)	UML-Class Diagram	XSD	Natural Langauge			Natural Language		XSD	Natural Language	Sample Textual Models

2.1. UML as a Modelling Method Design Approach

As UML is utilized for specification of most of modelling standards in our set, we extend our literature research to analyze UML as a prominent modelling method, which is used during create and design phases of the conceptualization process. The pragmatic objective of this analysis is to understand if UML - in the context of modelling method design - covers all requirements for creating and designing a modelling method, or if there are any shortcomings that we have to consider during the development of our approach.

Glinz evaluates UML in [56] as a requirement specification language. He states that definition of requirements with UML Use Case Diagrams is possible but Use-Case-Diagrams alone are not sufficient. Definitions of concrete functional and non-functional requirements as well as definition of relation between the concepts in domain specific modelling method and requirements are not possible. The authors of [58] indicate the same problem and propose again a UML-based solution to define non-functional requirements and relation between the components in system by combining constructs from UML Class Diagram and UML Component Diagram. Hence the first shortcoming needs to be considered in our approach is;

1. Definition of functional and non-functional requirements and their relation between the concepts in modelling methods.

According to Selic [59] the domain specific modelling languages specify different viewpoints of a complex system. Hence, the complex system should be represented from different viewpoints and some features will be presented in several viewpoints. Moreover the language, presumably, will support multiple viewpoints for different sub-domains, which means language should allow use of different abstract and concrete syntax. With UML Class Diagrams the whole meta-model (abstract syntax) of language can be fragmented. But depicting all alternatives of abstract syntax together can make the class diagram very complex and hard to read. To the best of our knowledge the UML does not offer definitions of different concrete syntax for same concept of modelling language. Hence the next shortcomings are

2. Fragmenting whole meta-model into individual meta-models composing concepts for different sub-domains and still be able to define link between concepts in different individual meta-models
3. Having another abstraction layer to represent modules and layers of modelling language as well as relation among them without representing complexity of abstract and concrete syntax
4. Assigning different concrete syntax to the concepts in modelling language.

According to authors of [55], nowadays, models are used to elaborate design decisions, sort out different concepts and exchange the ideas in mainstream software development. They state the importance of traceability during the transformation of information, communication of decisions etc. from design into implementation and vice versa. They indicate that in order to establish the traceability the support of modelling environment is as essential as the approach itself. To the best of our knowledge UML itself does not support such traceability. Then the next shortcoming is:

5. Traceability of information (e.g design decisions, changes, suggestions etc.) during the knowledge exchange among experts within design phase and also from design to implementation vice versa.

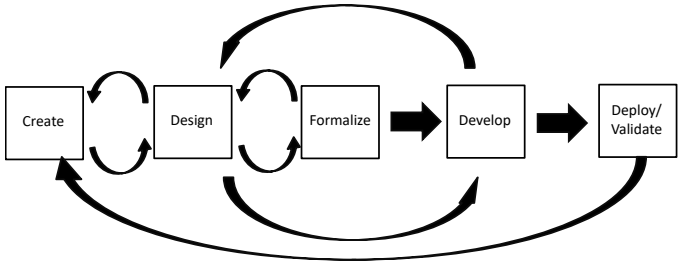
Karagiannis and Kühn in[5][57] argue that modelling methods have three major components (1) Modelling Language, (2) Modelling Procedure and (3) Mechanisms & Algorithms. Hence besides the modelling language of domain specific modelling method, we have to consider also designing modeling procedure and mechanisms & algorithms during the design of a modelling method. The design of modelling procedure and mechanisms & algorithms can be possible with using UML (e.g UML Activity diagram for design of the modelling procedure and if we consider description of mechanism and algorithms as sequence of object interactions and message exchange, the UML Sequence diagram can be used for description of mechanism and algorithms). Hence we would not see any shortcoming in UML for this issue. But in order to emphasize requirement of ability to design modelling procedure and mechanisms & algorithms besides the modelling language of a domain specific modelling method and requirement, and requirement of having corresponding supportive modelling environment, we would list this issue here rather as a general requirement than shortcoming of UML that we have to consider in our approach. Hence last but not least;

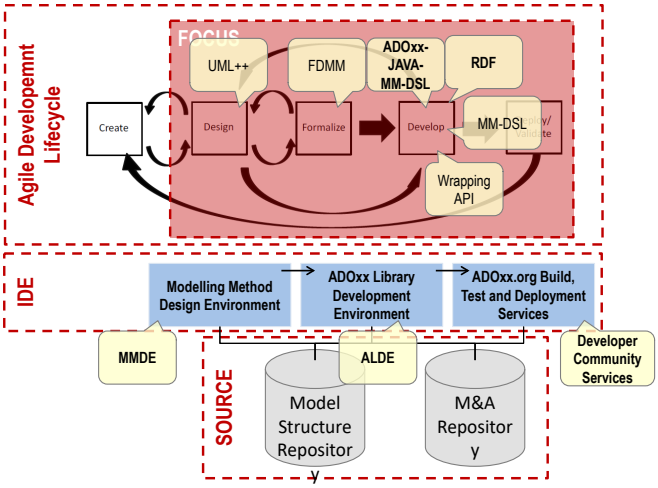
6. Possibility to design modelling procedure and mechanisms & algorithms of a domain specific modelling language.

2.2. Agile Modelling Method Engineering and Conceptualization Process

Having roots in software engineering, as it is in agile software development, during the modelling method engineering, involved stakeholders need procedures, structures and supportive tools allows high iterative process with as less as possible bureaucracy, and offers agile value and follows principles in Agile Manifesto [37]

AMME is proposed in [6] to support modelling method engineering during propagation and evolution of modelling requirements. The OMiLab Lifecycle [10] instantiates AMME and defines the internal cycle of a modelling method conceptualization with five phases; (1) “Create” as a mix of goal definition, knowledge acquisition and requirements elicitation activities that capture and represent the modelling requirements; (2) “Design” specifies the meta-model, language grammar, notation and functionality as model processing mechanisms and algorithms; (3) “Formalize” aims to describe the outcome of the previous phase in non-ambiguous, formal representations with the purpose of sharing results within a scientific community; (4) “Develop” produces concrete modelling prototypes and finally (5) “Deploy/Validate” involves the stakeholders in hands-on experience and the evaluation process as input for upcoming iterations.





Due to the involvement of several stakeholders with varying knowledge backgrounds, perspectives and different objectives, in the conceptualization of a modelling method, the authors of [9] propose so-called Modelling Method Conceptualization Process (as depicted inFigure 1) by adding additional feedback channels into the OMiLab Lifecycle between: (1) Create and Design, to prove, if the designed modelling language covers the identified application scenarios and considers the identified requirements; (2) Design and Formalize to ensure formal approval of modelling language, as well as (3) Design and Develop - to improve modelling language in earlier stages before it is released and deployed.

3. Modelling Method Conceptualization Environment

The work at hand introduces the "Modelling Method Conceptualization Environment from ADOxx.org. A toolbox (as its high-level architecture depicted in Figure 2) that initially has been introduced in [35] and that instantiates the above-mentioned conceptualization process and supports method engineers during each phase. The only exception is that of the "Create" phase, as this part is regarded as the most creative phase and standard tools and methods (also in some cases pen and paper can be the most appropriate tools) shall be freely selected. Modelling Method Conceptualization Environment proposes a combination of tools in sense of Integrated Development Environment (IDE), such as the Modelling Method Design Environment (MMDE, available to download and install at [11]) for the Design, the ADOxx Library Development Environment (ALDE) for Formalize and Develop, Adoxx.org Build, Test and Deployment Services (available at [22]) for Deploy/Validate Phases.

As depicted in Figure 3, typical application scenario would be; during the create phase domain experts and method engineers come together, define goal of modelling method, acquire and elicit requirements, in design phase method engineers with tight collaboration of domain experts specifies the meta-model, language grammar, notation and processing functions on MMDE, as method engineers formalize design of modelling method collaboratively and commit on ALDE, developer(s) based on that formalization implements concrete modelling toolkit prototype within ALDE and ADOxx Development Toolkit. Developer(s) uploads the prototype into ADOxx.org build server, semi-automatic service behind starts with completeness check, building installation package, testing of installation package and optionally deploy it on selected developer's space to allow to download the toolkit, to be tested and validated by community members, so to get feedback from them or the build services simply sends a link to corresponding owner to download and/or share the modelling toolkit.

It is worth to mention that one of the objectives is to provide loosely coupled tools, so involved actors have the flexibility to decide to use one, a combination of tools from the toolbox or even use other appropriate tools of their choice, (e.g. method engineer uses MMDE during the Design Phase, but formalize the modelling

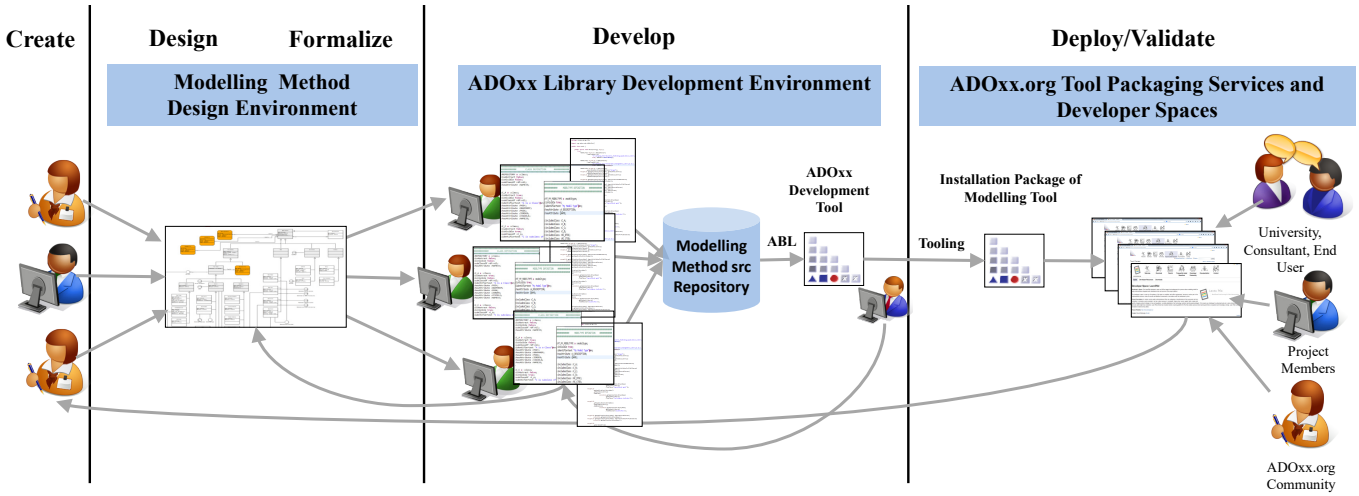


Figure 3 Life-cycle within Modelling Method Conceptualization Environment from Users' Perspective

method design with mathematical models or use another development tool during the Develop Phase and deploys them at the Developer Spaces and enable validation).

In the following sub-sections current abilities of the tools from the environment are shortly presented.

3.1. Modelling Method Design Environment

The Modelling Method Design Environment (MMDE) is itself a modelling tool to design other modelling methods. MMDE has been implemented based on lessons learned from results of the state of the analysis, which is discussed in pervious chapter, and from the experience of the authors, who have been involved in more than 20 modelling method/tool development projects for varying domains. Based on these lessons learned, UML [12] has been identified as a starting point. Hence, the MMDE takes a subset of UML and extends it with required concepts and functionalities in order to overcome the following challenges (**Ch**), which are introduced after the state of the art analysis: **Ch1**-Definition of functional and non-functional requirements and their relation between the concepts in modelling methods; **Ch2**-Fragmenting the whole meta-model into individual meta-models composing concepts for different sub-domains and still be able to define links between concepts in different individual meta-models; **Ch3**-Having another abstraction layer to represent modules and layers of modelling language as well as relation among them without representing the complexity of abstract and concrete syntax; **Ch4**-Assigning different concrete syntax to the concepts in modelling language; **Ch5**-Possibility to design modelling procedure and mechanisms & algorithms of a domain specific modelling language.

To overcome **Ch1**, "Requirements" model type (as depicted in Figure 4) is implemented that allows the elicitation of requirements, specifying their status as well as dependencies among them. The described requirements in this model type can be referenced to (a) all the modelling classes modelled in the related model type "Meta-Model" classes, (b) graphical notation (concrete syntax) definitions modelled in the "Graphical Notation" model type, (c) the "Modelling Stack" definition and (d) to the functionalities modelled in "Mechanisms & Algorithms" models.

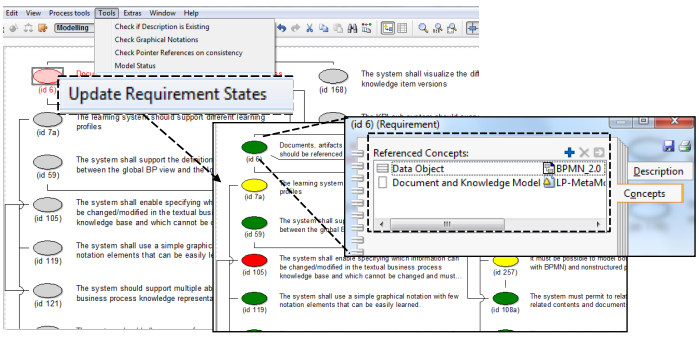


Figure 4 . Example: Requirement Model before and after Updating the Status of Requirements

For **Ch2** and **Ch3,** we extend the class diagram from UML (as depicted in Figure 5) with concepts, so method engineers can differentiate between class and relation class as well as relate different meta-models (-fragments) with each other using "Weaving" techniques as they are introduced in [8] [9].

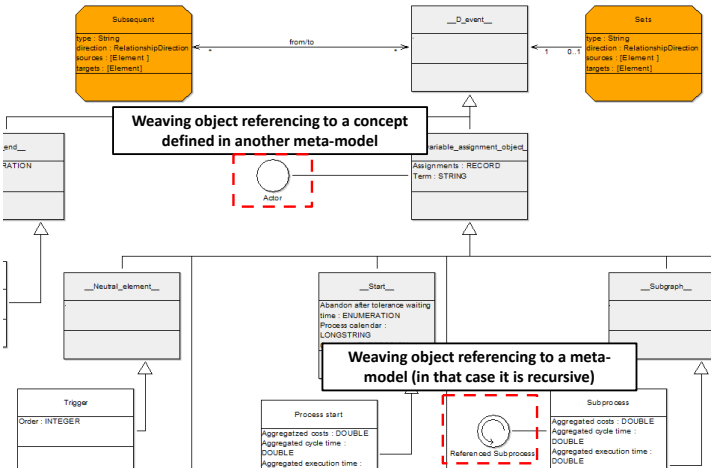


Figure 5 A sample meta-model showing usage of weaving concept

The modularization and layering of modelling language is essential to avoid complexities during the design of domain specific modelling methods [15][16] Hence, we propose representation of the Modelling Stack as the "Meta-models Stack model type (as depicted in Figure 6) allowing method engineers to differentiate meta-models in sense of different model types that target different fragments of the system.

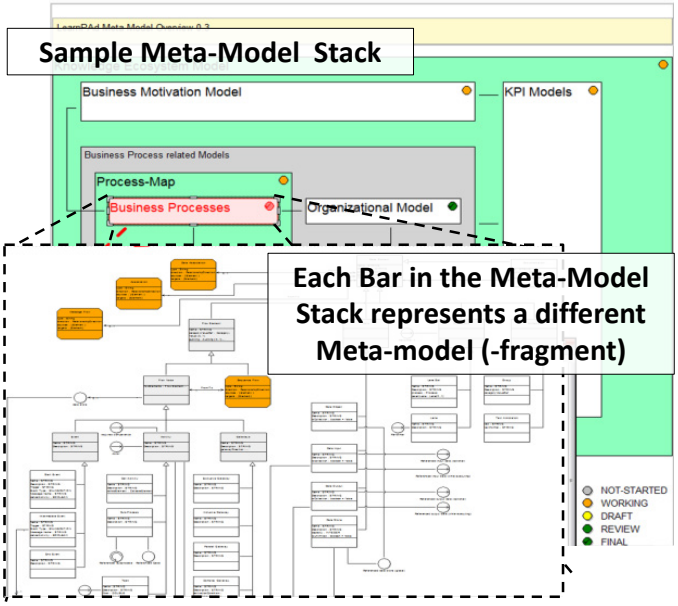


Figure 6 A Sample Meta-model Stack Model

In order to target Ch4 and specify a proper graphical representation (concrete syntax) of each concept in a meta-model, we introduce another model type called "Graphical Notation" model type (as depicted in

Figure 7) allows definition of concrete syntax of model types with specifying graphical representations for each constructs in meta-models. This model type allows the description of graphical representations with the assignment of concrete images in PNG, JPG or Bitmap format including a description of the functionalities in the notation (e.g. attribute-value dependent visualization, context related views)

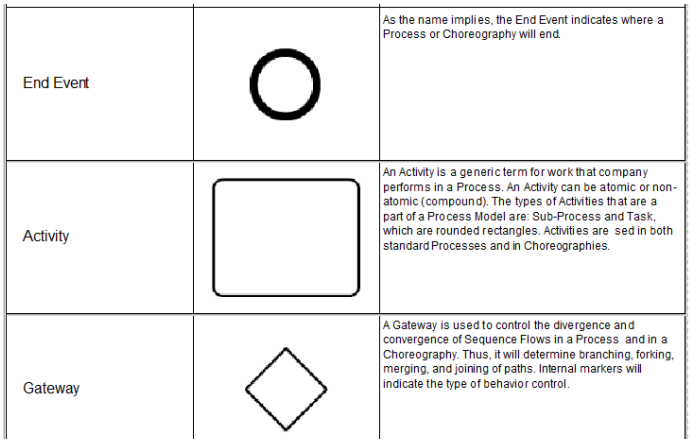


Figure 7 Sample Graphical Notation Model

In order to target Ch5 to define the applicable modelling technique as steps and corresponding results we propose a model type called "Modelling Procedure" model type". The Modelling Procedure Model Type (as depicted in

Figure 8) allows method engineers to define the steps with their required inputs and produced outputs, as well as the sequence of steps based on the input – output relationships, in order to introduce case specific proper usage of their modelling method.

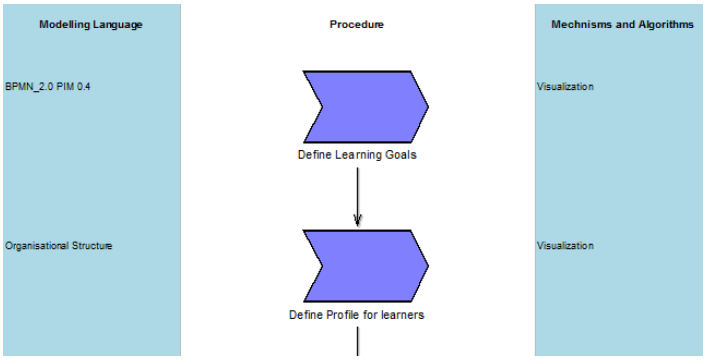


Figure 8 A Sample Modelling Procedure Model

Based on this procedural view, concrete Mechanisms and Algorithms, can be derived and depicted as Sequence and Component Diagrams from UML (therefore these diagram types has been implemented as model types in MMDE). Further details about MMDE can be found in [9]

3.2. ADOxx Library Development Environment

The ADOxx Library Development Environment (ALDE) aims to enable parallel development of modelling tools libraries based on the designs deriving from Design Phase, merging different libraries and ensuring maintainability. As an experimental prototype ALDE is uses the Resource Description Framework (RDF) as a format for data interchange [17]

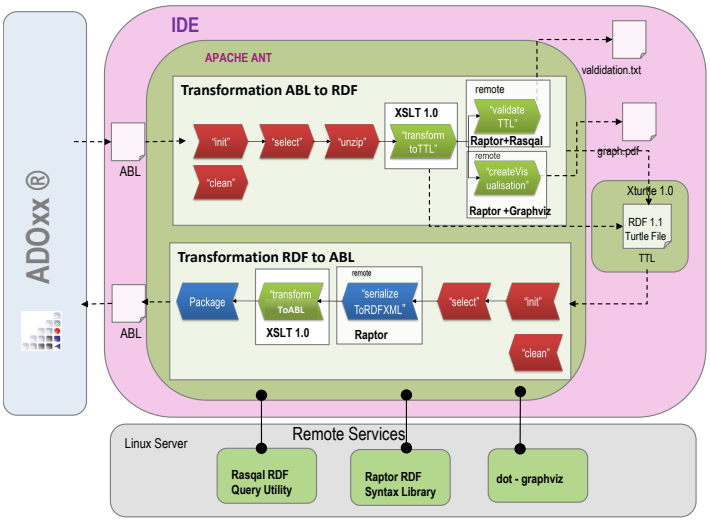


Figure 9 Architecture of ADOxx Library Development Environment (ALDE)

The Figure 9 depicts the architecture of ALDE. It is a development environment based on the Eclipse IDE [18] and includes a meta-meta-model defined in RDFS, the ALDE vocabulary. Having the vocabulary and utilizing Apache Ant® [19] as a build mechanism, the environment enables the definition of the transformation processes from ADOxx Library Languages to RDF and vice versa. Moreover ALDE serializes libraries in an arbitrary RDF format; for the prototypical realization RDF Turtle [20] has been used (the Figure 10 depicts code snippets produced by the environment) and includes the RDF XTurtle Editor developed by [21]. Having libraries in RDF Turtle format and a RDF Turtle Editor available, method engineers can adapt declaratively and script libraries collaboratively using standard functionalities of source-code management systems. Merging several libraries or integration of parts of libraries in each other becomes possible. On the other hand, ALDE includes verification services to ensure that the newly developed, edited or merged libraries are consistent with ADOxx platform requirements.

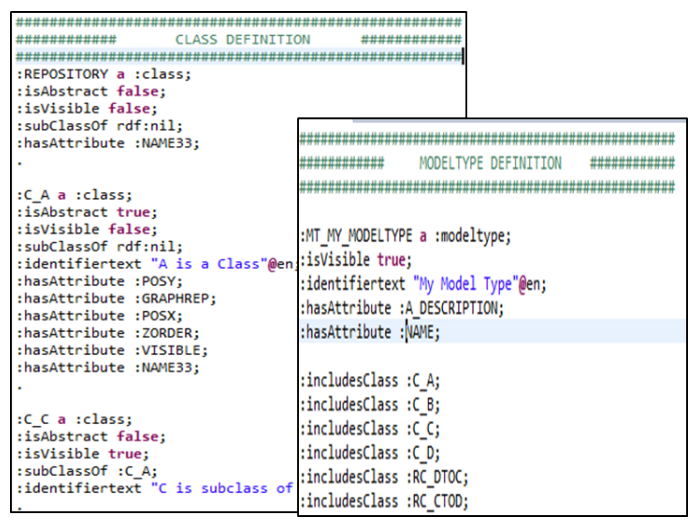


Figure 10 Snipper of Class and Model Type Definitions in ALDE in RDF

The new extension to ALDE is a new DSL based on Java, which has a working title of "ADOxx-JAVA-MM-DSL". The ADOxx-JAVA-MM-DSL is developed according to feedback on

previous version of toolbox, which are presented and discussed in [35].

The ADOxx-JAVA-MM-DSL is a framework that creates several abstraction layers over the ADOxx Library Language (ALL) format, the ADOxx internal language that describes a meta-model [64]. Each layer simplifies and adds features to the bottom one. The framework gives, the possibility to operate and easily perform modification on a meta-model without dealing with its complexity. In order to assure that, an internal structure is managed that represents the ALL structure. This internal structure can be imported from an existing ALL meta-model.

All the constraints and rules present in the ALL syntax are managed, so the framework can guarantee that only syntactically valid ALL conform meta-models can be loaded and generated. The whole internal structure is an instance of java classes, so operations and definitions on the meta-model's concepts can be done using the java features. Parallel to that, additional utilities are integrated in order to bridge the gap between the development phases:

- The compilation to the ALL meta-model to its binary format ABL: The ALL is a text based language used to describe a meta-model, in order to be imported and be usable in ADOxx it need to be compiled in its binary format ABL. A conversion engine is integrated into the framework. In such a way it is possible to automatize a previously manual step, required in order to create an automatized flow between formalization and development phase.
- The importing and deployment of the compiled meta-model in the form of ABL file, into a ADOxx Database: In order to create the respective modelling environment based on the created meta-model, is required to import the ABL compiled meta-model into the Database used by ADOxx. After this step the modelling environment relative to that meta-model can be executed and opened simply specifying at the launch time, the newly created database. The framework contains a feature that automatizes such a process, generating the Database and launching the modelling environment relative to the created meta-model. This step bridges the gap between formalization and development phase.

As depicted in Figure 11, the framework can be divided into three abstraction layers and two transversal layers of primitives and general features:

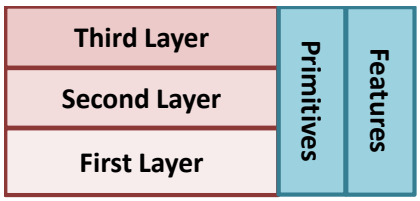


Figure 11 Layers of ADOxx-JAVA-MM-DSL

The first abstraction layer is a package of java classes that exactly reflect the structure of the ALL syntax. This layer is responsible for converting each class in its ALL representation and applies syntax rules. Due to its closeness to the ALL syntax, working directly with this layer is difficult. The advanced of creating a meta-model using this layer instead using directly the ALL syntax is that all possible errors derived from miss-spelling and unhallowed sequences will be avoided, but the complexity in the definition of the meta-model remain the same as the ALL.

The second abstraction layer abstracts all the concepts of the First layer to a more design friendly way, providing the possibility to work directly with the concepts of Libraries, Classes, Relations and Attributes. These concepts are more familiar to method engineers that are familiar with the interactive development approach directly in the ADOxx Development Toolkit. This abstraction layer defines also some semantic rules over the syntactic rules provided by the first layer. The entire concepts created at this level will be mapped to the relative concepts at the first layer in order to reflect the ALL syntax. At this level, certain helpful features are also implemented, such as the possibility to merge two or more libraries or to import one or several classes from a library to another.

The third abstraction layer is a factory layer and contains predefined methods that generate empty libraries as a starting point for the definition of specific meta-models with classes and relations. The main entry point of the framework is also at this layer with the possibility to explore all the features of the development environment.

The transversal primitives layer gives a formal representation of all the primitives allowed in the ALL syntax. The primitives are used at every level in order to define data such as Identifiers or Attribute Values;

The transversal features layer contains functions used at every level and the ALL parser. The parser is the component that allows an ALL to be read from a file and being instantiated and loaded into the framework in order to be extended or modified.

The ADOxx-JAVA-MM-DSL supports following three scenarios:

Definition of a meta-model from scratch: Using the third abstraction layer, it is possible to generate an empty ADOxx library that is the best starting point to create own specific meta-model. Starting from that object the method engineer can add Classes and Relations and their respective Attributes to the meta-model by writing pure java code. Once the java code is executed, it generates the ALL file and/or compile it in the ABL file and/or directly create the ADOxx Database and execute the prototype of newly developed modelling toolkit based on defined meta-model.

Extension or modification of an existing meta-model: Using the parsing module of the transversal layer, it is possible to load an ALL file into an instance of the java classes present in the second and First layer of the developer environment. After the ALL have been loaded is possible to extend, modify or manage it in the same way as creating from scratch. Certain methods to find specific Libraries, Classes, Relations and Attributes that the method engineer wants to work with are available. As in the previous scenario, after the java code is executed generates the ALL file and/or compile it in the ABL file and/or directly create the ADOxx Database and execute the prototype of newly developed modelling toolkit based on defined meta-model

Merging of two or more meta-model in one: This scenario is a particular case of the first two. In particular, with using the framework, it is possible to merge two or more meta-model. As it is in a sample snippet depicted in

Figure 12, thanks to the features present in the second layer is possible to import the whole concepts from a meta-model to

another or do a fine-grained import of specific concepts from each meta-model to another one. It is possible to import a class from a meta-model to another managing automatically all its dependencies like the presence of Super Classes. The merging scenario is supported in order to minimize the conflict that may rise in the merging of two incompatible meta-models, providing useful information to method engineer regarding how to correct occurred conflicts.

```
package caxman_merge_lib;

import org.adoxx.ado.ADOUtils;

public class Main {

    public static void main(String[] args) {
        try {
            ADOAllFile mainLib = ADOLibFactory
                    .loadFromAblFile(
                            "D:\\merging_test\\Business_Modelling_Application_Li
                            true, ADOUtils.adonisAblKey);

            ADOAllFile secondLib = ADOLibFactory
                    .loadFromAblFile(
                            "D:\\merging_test\\Performance_Management_Library_0.
                            true, ADOUtils.adoxxAblKey);

            ADOAllFile mergedLib = ADOLibFactory
                    .generateADOxxTotalEmptyLibrary("Merged");
            mergedLib.getApplicationLibraryNew().getDynamicLibrary()
                    .addADOxxEmptyDefaultClasses();
            mergedLib.getApplicationLibraryNew().getStaticLibrary()
                    .addADOxxEmptyDefaultClasses();

            mergedLib.getApplicationLibraryNew().importClassesAndRelations(
                    mainLib.getApplicationLibraryNew());

            mergedLib.getApplicationLibraryNew().importAttributeProfileClasses(
                    mainLib.getApplicationLibraryNew());
            mergedLib.getApplicationLibraryNew().importRecordClasses(
                    mainLib.getApplicationLibraryNew());
            mergedLib.getApplicationLibraryNew().importFiles(
                    mainLib.getApplicationLibraryNew());
            mergedLib
                    .getApplicationLibraryNew()
                    .getDynamicLibrary()
                    .addClass(
                            secondLib.getApplicationLibraryNew()
                                    .getDynamicLibrary()
                                    .findClass("Operational goal"));
            mergedLib
                    .getApplicationLibraryNew()
                    .getDynamicLibrary()
                    .addClass(
                            secondLib.getApplicationLibraryNew()
                                    .getDynamicLibrary()
                                    .findClass("Performance indicator"));

            mergedLib
                    .getApplicationLibraryNew()
                    .getDynamicLibrary()
                    .addRelation(
                            secondLib.getApplicationLibraryNew()
                                    .getDynamicLibrary()
                                    .findRelation("quantifies"));


            mergedLib.getApplicationLibraryNew().importRecordClasses(
                    secondLib.getApplicationLibraryNew());
            mergedLib.getApplicationLibraryNew().importAttributeProfileClasses(
                    secondLib.getApplicationLibraryNew());

            mergedLib
                    .generateABLFile(
```

Figure 12 Snippet of a Selective Merge of two Meta-models

3.3. Adoxx.org Build, Test and Deployment Services.

Adoxx.org Build, Test and Deployment Services [38] are web-based services that allow method engineers of the ADOxx community to build verified, professional and installable distribution packages that can be distribute to interested stakeholders. The service combines and validates all available inputs, integrates all elements, compiles the necessary artefacts and signs the outcomes and creates the actual installer as a download archive.

As a collaboration space for the development and deployment phases, the concept of "Developer Spaces" has been introduced in ADOxx.org [23]. These spaces enable sharing of intermediate/release results, discussing development resources from all pre/past phases in the form of source code, snippet, examples and distribution packages with the community.

4. Modelling Method Conceptualization Services

In addition to the development tools described in the previous chapter, an appropriate service support is foreseen to support the modelling method engineers. The services are provided on the ADOxx.org portal, supporting a community of more than 1.300 modelling method engineers world-wide.

1. **Download:** For the download, ADOxx.org provides the Meta Modelling Platform ADOxx in combination, Installation Instructions, Frequently Asked Questions, Startup-Package as well as a set of more than 30 available application libraries, which can be used to start with.
2. **Get Started:** For getting started, ADOxx.org provides important readings, provides a Forum that is structured according active communities, lists tutorial and training events that are offered free of charge, provides tutorial material for both the students – in form of guide samples and slides – as well as for the trainer – in form of a trainer handbook and offers tutorial videos and webinars.
3. **Development and Support:** For the development, ADOxx.org provides aforementioned tools and additional developer utilities, 3rd parties add on like but not limited to simulation, documentation, dashboards or Web-APIs. A collection of 200 graphical representations that introduce the major elements conclude the development support.
4. **Community:** For collaborative development within the ADOxx.org community a map is provided indicating the ten laboratories – nine in Europe, one in Asia, indicating the hot spots of developers, the participating research institutes, a set of 24 modelling tools as a result of [66], and development spaces that enable a collaborative development and enable the use from solutions and tools from other projects.
5. **Documentation:** A complete specification and documentation is offered, where each relevant element of the modelling method is (a) explained based on the corresponding theory, (b) introduced with hand-on samples, (c) demonstrated with real-world scenarios, (d) mapped to forum entries of the community and finally (e) supported with tools where possible.

The operation of this service centre is provided via the portal, social media like Twitter, Facebook and LinkedIn, or via email. In justified cases an onsite support is possible, where either the method engineer is trained, supported or critical implementation steps are performed by the ADOxx.org service centre.

5. Evaluation

Given that usually each modelling method engineering case differs from each other in sense of complexity of domain, variety of aspects to be targeted, number of involved actors, to calculate quantified evaluation means is difficult, and – to best of our knowledge - there is no similar conceptualization environment, hence, it is difficult to bench-mark our proposal and quantify the

evaluation and provide statistically objective results. On the other hand, the most important tangible and objective evaluation result would be deployed and ready to use modelling toolkits, specification of modelling methods and communication of community members as proofs of concept. Those proofs of concepts for each are online and freely accessible (with exception of in-house project case). The links to access those proofs of concepts for each case are provided under regarding sub-section below.

The conceptualization environment introduced above has been applied in four different cases for evaluation: three EU-funded research projects in the domain of multi-stage manufacturing, eLearning and cloud computing and additionally in an in-house development project, in the area of decision modelling extensions. These cases have been selected since the involved partners have varying profiles and expertise in given domains, in development and in modelling method engineering. In the following sub-section we introduce the cases and their requirements in method engineering manner.

Case 1: Conceptualization of a Modelling Method for E-Learning: The FP7 project Learn PAd [24] proposes a process-driven-knowledge management approach based on conceptual and semantic models for transformation of public administration organizations into learning organizations. Learn PAd proposes a model-driven collaborative learning environment. In this case, 4 domain experts and method engineers have been involved. In addition, two developer teams, each consisting of 4 developers worked on the implementation of the tool. The results of the conceptualization process of this modelling method can be found at Learn PAd Developer Space [25], as well as the developed prototypes [26] can be downloaded and feedback can be given.

Case 2: Conceptualization of Modelling Method for Cloud Computing: The H2020 project CloudSocket [27] introduces the idea of Business Processes as a Service (BPaaS), where conceptual models and semantics are applied to align business processes with Cloud-deployed workflows [28]. In this case, 6 domain experts and method engineers have been involved, as well as two developer teams, one with 5 developers, the other one with 2 members. The results of the conceptualization process of this modelling method can be found at CloudSocket Developer Space [29], as well as developed prototypes [30] can be downloaded and feedback can be given.

Case 3: Conceptualization of Modelling Method for holistic Manufacturing System Management:

The H2020 project DISRUPT [60] deals with the integration of innovative technologies into a holistic manufacturing system and optimization of production flow. The DISRUPT projects needs a modelling method to describe manufacturing system from supply-chain level down to shop-floor level. In this case 2 domain experts, one requirement engineer and one method engineer have been involved. Preliminary results can be found on DISRUPT Developers Space [61].

Case 4: Integration of existing BPMN and DMN Modelling Methods: The in-house project requires integration of an already implemented DMN Modelling Method into existing BPMN 2.0 realization as part of a commercial product. In this case, 3 domain experts and method engineers, and a team of two developers have been involved.

The evaluation process was enacted in the following steps: (1) Provisioning: the tools -of the toolbox have been provided to the stakeholders in the involved cases. (2) Team Formation: representatives, - of the stakeholders in the project created development teams consisting of domain experts and method engineers following the conceptualization process and developing tools individually. (3) Feedback Phase: individual results have been consolidated periodically through video conferences and workshops, constituting the evaluation results.

Feedback on MMDE.

Pro: It is possible to specify requirements and dependencies among them as well as tracing them; (2) to define modelling language fragments and modules, (3) layering the modelling language with navigational constructs; (4) definition of syntax, semantic and assignment of notation (concrete syntax); (5) definition of weaving among construct in different meta-models; (6) assignment of (multiple-) graphical notation (concrete syntax); (7) explicit definition of modelling procedure;

Contra: It is not possible to define application scenarios and use cases, and design results can be exchanged solely using ADOxx specific formats or as static content (image, PDF or HTML). Hence, double effort in the design and in the formalisation and or development is currently necessary.

Outlook: The MMDE is currently updated, to offer an XML export, which then can be transformed into different formats like the one that is used for the ADOxx-Java-MM-DSL.

In addition, several improvements on the modelling language are implemented to (a) enable the design of user scenarios, (b) better describe the features of the modelling method and their corresponding components as well as (c) enable a more detailed representation of the method procedure enabling the mapping from components and the corresponding elements of the modelling method.

Feedback on ALDE.

Pro: it is possible to transform libraries in a machine as well as human interpretable format, ability to use reasoning algorithms, due to standard semantic formats; reduces complexity to edit, merge and maintain libraries.

Contra: To take over results from Design Phase require manual steps.; it re-quires different transformation scripts for different meta-modelling technologies (such as ADOxx, EMF).

Outlook: The semantic-based verification of meta model is seen as a useful extension of the ADOxx-Java-MM-DSL, hence an integration will be experimented. However, we see the necessary skill level for the meta model developer currently as inappropriate and tend not to follow this path.

Feedback on ADOxx-Java-MM-DSL:

Pro: It is possible to merge libraries and start libraries from scratch. Furthermore, the code base can be stored and versioned in a versioning system enabling several developers in parallel to work on one library. Built scripts enable the automatic generation and deployment of the tool.

Contra: The current code maturity needs improvement and documentation, enabling also non specialists to handle the tool.

Outlook: This tool will be further improved and tested in two EU-H2020 research projects and will consequently be taught at the

ADOxx.org Training Days and Webinars to achieve the required maturity.

Feedback on ADOxx.org Tool Packing Services and Developer Spaces.

Pro: It is possible to have an installation package to distribute to interested stake-holders, building your own community around the modelling method, and get feed-back from them.

Contra: Setting up and handling issues of a certain Developer Space involves certain manual steps, such, as the interested stakeholder has to send an e-mail to the administrator with a request of an own Developer Space.

Outlook: This tool packaging service will be stepwise opened, so that the developer can also include own software components, which are then composed into a single tool package.

6. Conclusion and Outlook

In this paper we introduce a toolbox instantiating the Modelling Method Conceptualization Process, which supports agile modelling method engineering. The toolbox has been evaluated through an analysis of four different cases: three EU research projects and one in-house project. The evaluation results put forward that having an approach and a corresponding toolbox following the idea of model-driven engineering approach is effective in terms of transferring knowledge from the analysis of requirements up to the development of solutions. Being three main tools, MMDE, ALDE and ADOxx-Java-MM-DSL, prototypes that are at about Technology Readiness Level 5, hence lack of full integration or automatic data exchange ability, and the need of manual steps building Developer Spaces came out as major limitations of the toolbox. As an outlook the following items derived from the evaluation as future work: (1) currently we are evaluating development alternatives of DSLs with using Xtend [62] or RDF; building on existing work [63] in the field and integrating it into ADOxx-Java-MM-DSL, (2) enabling graphical modelling method design to transform into machine understandable format, (3) formalization of modelling method design using mathematical models such as FDMM [33] or Proof of Concept prototyping, (4) automatization of tooling services and deployment onto developer spaces, (5) full integration of tools within a holistic development environment.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This work has been partly supported by European Union's Horizon 2020 research and innovation programme within the projects DISRUPT(www.disrupt-project.eu) under grant agreement No: 723541 and CloudSocket (www.cloudsocket.eu) under grand agreement no: 644690.

References

[1] N. Efendioglu, R., Woitsch, "Modelling Method Design: An ADOxx Realisation" in 20th IEEE International Enterprise Distributed Object Computing Workshop, EDOC Workshops 2016, Vienna, Austria, 2016

[2] Object Management Group (OMG), "Business Model and Notation Version 2.0," 2011. [Online]. Available: http://www.omg.org/spec/BPMN/2.0/. [Accessed 15 July 2016].

[3] Object Management Group (OMG), "Decision Model and Notation, Version 1.0," 2015. [Online]. Available: http://www.omg.org/spec/DMN/1.0/. [Accessed 15 July 2016].

[4] Object Management Group (OMG), "Case Management Model and Notation", [Online] Available: http://www.omg.org/spec/CMMN [Accessed 22 February 2017]

[5] D. Karagiannis and H. Kühn, "Metamodelling platforms," in In Proceedings of the 3rd International Conference EC-Web 2002, Dexa 2002, France, Springer-Verlag, 2002, p. 182.

[6] D. Karagiannis, "Agile Modeling Method Engineering," in Proceedings of the 19th Panhellenic Conference on Informatics, Athens, Greece, ACM, 2015, pp. 5-10.

[7] V. Hrgovcic, D. Karagiannis and R. Woitsch, "Conceptual Modeling of the Organisa-tional Aspects for Distributed Applications: The Semantic Lifting Approach," in COMPSACW, IEEE, 2013.

[8] J. Michael, F. Al Machot, M. C. Heinrich "ADOxx Based Tool Support for a Behavior Centered Modeling Approach" in Proceedings of the 8th ACM International Conference on PErvasive Technologies Related, PETRA'15, Corfu, Greece, 2015

[9] N. Efendioglu, R. Woitsch and D. Karagiannis, "Modelling Method Design: A Model-Drive Approach," in IIWAS '15: Proceedings of the 17th International Conference on Information Integration and Web-based Applications, Brussels, Belgium, ACM, 2015.

[10] Open Models Laboratory (OMILab), "Idea and Objectives," 2015. [Online]. Available: http://austria.omilab.org/psm/about. [Accessed 15 July 2016].

[11] ADOxx.org, "LearnPAd Developer Space," 2015. [Online]. Available: https://www.adoxx.org/live/web/learnpad-developer-space/design-environment. [Accessed 07 July 2016].

[12] Object Management Group (OMG), "Documents Associated With UML Version 2.0," 2005. [Online]. Available: http://www.omg.org/spec/UML/2.0/. [Accessed 12 July 2015].

[13] H. Kühn, "Methodenintegration im Business Engineering, PhD Thesis (in German)," University of Vienna, 2004.

[14] R. Woitsch, "Hybrid Modeling: An Instrument for Conceptual Interoperability," in Revolutionizing Enterprise Interoperability through Scientific Foundations , Hershey, 2014, pp. 97-118.

[15] B. Selic, "The Theory and Practice of Modeling Language Design for Model-Based Software Engineering—A Personal Perspective," in Generative and Transformational Techniques in Software Engineering III, Springer Berlin Heidelberg, 2011, pp. 290-321.

[16] D. Karagiannis, V. Hrgovcic and R. Woitsch, "Model Driven Design for e-Applications: The Meta Model Approach," in Proceedings of the 13th International Conference on Information Integration and Web-based Applications and Services, iiWAS11, Ho Chi Minh City, Vietnam, ACM, 2011, pp. 451-454.

[17] W3C, "RDF-Resource Description Framework," 2014. [Online]. Available: https://www.w3.org/RDF/. [Accessed 14 July 2016].

[18] Eclipse Foundation, "Eclipse IDE for Java EE Developers," 2016. [Online]. Available: http://www.eclipse.org/downloads/packages/. [Accessed 14 July 2016].

[19] The Apache Software Foundation, "Apache Ant Download," 2016. [Online]. Available: https://www.apache.org/dist/ant/binaries/. [Accessed 14 July 2016].

[20] W3C, "RDF 1.1 Turtle Terse RDF Triple Language,," 2014. [Online]. Available: https://www.w3.org/TR/2014/REC-turtle-20140225/. [Accessed 14 July 2016].

[21] The Research Group Agile Knowledge Engineering and Semantic Web (AKSW), University of Leipzig, "Xturtle," 2015. [Online]. Available: http://aksw.org/Projects/Xturtle.html. [Accessed 15 July 2016].

[22] ADOxx.org, "AutoPDP Tool Packaging Service," 2016. [Online]. Available: https://www.adoxx.org/live/autopdp-packaging-service. [Accessed 03 March 2017].

[23] ADOxx.org, "ADOxx.org Developer Spaces," 2016. [Online]. Available: https://www.adoxx.org/live/development-spaces. Accessed 03 March 2017].

[24] Learn PAd Consortium, "The EU Project Learn PAd," 2014. [Online]. Available: http://www.learnpad.eu/. [Accessed 03 March 2017].

[25] LearnPAd Consortium, "LearnPAd Developer Space," 2015. [Online]. Available: https://www.adoxx.org/live/web/learnpad-developer-space. [Accessed 03 March 2017].

[26] LearnPAd Consortium, "LearnPAd Developer Space - Downloads," 2015. [Online]. Available: https://www.adoxx.org/live/web/learnpad-developer-space/downloads. [Accessed 03 March 2017].

[27] CloudSocket Consortium, "CloudSocket Project," 2016. [Online]. Available: https://www.cloudsocket.eu/. [Accessed 03 March 2017].

[28] R. Woitsch and W. Utz, "Business Process as a Service, Model Based Business and IT Cloud Alignment as a Cloud Offering," in ES 2015, Third International Conference on Enterprise Systems, Basel, Switzerland, 2015.

[29] CloudSocket Consortium, "CloudSocket Developer Space," 2015. [Online]. Available: https://www.adoxx.org/live/web/cloudsocket-developer-space/. [Accessed 15 July 2016].

[30] CloudSocket Consortium, "CloudSocket Developer Space - Downloads," 2015. [Online]. Available: https://www.adoxx.org/live/web/cloudsocket-developer-space/downloads. [Accessed 03 March 2017].

[31] Eclipse Fundation, "Xtend," 2015. [Online]. Available: https://eclipse.org/xtend. [Accessed 15 July 2016].

[32] N. Visic and D. Karagiannis, "Developing Conceptual Modeling Tools Using a DSL," in Knowledge Science, Engineering and Management, Sibiu, Romania, Springer, 2014, pp. 162-173.

[33] H.-G. Fill, T. Redmond and D. Karagiannis, "FDMM: A Formalism for Describing ADOxx Meta Models and Models," in Proceedings of ICEIS 2012, Wroclaw, Poland, Vol. 3, Wroclaw, 2012, pp. 133-144.

[34] ADOxx.org "GraphRep" 2016 [Online. Available https://www.adoxx.org/live/graphrep [Accessed 31.08.2016]

[35] N. Efendioglu,, R. Woitsch, W. Utz, "A Toolbox Supporting Agile Modelling Method Engineering: ADOxx.org Modelling Method Conceptualization Environment". In J. Horkoff, M. A. Jeusfeld, & A. Persson, The Practice of Enterprise Modeling (pp. 317-325), 9th IFIP WG 8.1. Working Conference, PoEM 2016, Skövde, Sweden, November 8-10, 2016, Proceedings, Springer

[36] Principles behind the Agile Manifesto [Online. Available http://agilemanifesto.org/iso/en/principles.html] [Accessed 24 January.2017]

[37] Cees Van Halen; Carlo Vezzoli; Robert Wimmer (2005). Methodology for Product Service System Innovation. Assen: Uitgeverij Van Gorcum. p. 21. ISBN 90-232-4143-6.

[38] ADOxx.org, "Developer Community," 2017. [Online]. Available: https://www.adoxx.org/live/community. [Accessed 23.January.2017].

[39] OASIS, "Universal Business Language Version 2.1". 04.November 2013, OASIS Standard [Online] Available at "http://docs.oasis-open.org/ubl/os-UBL-2.1/UBL-2.1.html [Accessed 13 July 2015]

[40] OASISI, "ebXML Business Process Specification Schema Technical Specification v2.0.4" 21 December 2006, OASIS Standard Available at http://docs.oasis-open.org/ebxml-bp/2.0.4/OS/spec/ebxmlbp-v2.0.4-Spec-os-en-html/ebxmlbp-v2.0.4-Spec-os-en.htm [Accessed 13.July 2015]

[41] OMG, 2005. Documents Associated With UML Version 2.0. [Online] Available at: http://www.omg.org/spec/UML/2.0/ [Accessed 12 July 2015]

[42] The Open Group, 2013. ArchiMate® 2.1 Specification, [Online] Available at http://pubs.opengroup.org/architecture/archimate2-doc/toc.html [Accessed 13 July 2015]

[43] W3C, 2007, Web Services Description Language (WSDL) Version 2.0 Part 1: Core Language [Online] Available at: http://www.w3.org/TR/wsdl20/wsdl20.pdf [Accessed 13.July 2015]

[44] W3C, 2012 OWL 2 Web Ontology Language Structural Specification and Functional-Style Syntax (Second Edition) [Online]. Available at http://www.w3.org/TR/2012/REC-owl2-syntax-20121211/ [Accessed 13 July 2015]

[45] W3C, 2012, SPARQL 1.1 Query Language [Online] Available at: http://www.w3.org/TR/owl2-direct-semantics/ [Accessed 13.July 2015]

[46] Web Services Business Process Execution Language Version 2.0 11 April 2007, OASIS Standard http://docs.oasis-open.org/wsbpel/2.0/OS/wsbpel-v2.0-OS.html [Accessed 13 July 2015]

[47] WfMC, 2005, Process Definition Interface – XML Process Definition Language [Online] Available at: http://www.xpdl.org/standards/xpdl-2.2/XPDL%202.2%20(2012-08-30).pdf [Accessed 13. July 2015]

[48] ISO/IEC 14977:1996, Information technology -- Syntactic metalanguage -- Extended BNF, International Organization for Standardization, 1996 http://standards.iso.org/ittf/PubliclyAvailableStandards/s026153_ISO_IEC_14977_1996(E).zip

[49] W3C, 2012 W3C XML Schema Definition Language (XSD) 1.1 Part 1: Structures [Online] Available at: http://www.w3.org/TR/xmlschema11-1/ [Accessed 13.July 2015]

[50] W3C, Resource Description Framework (RDF): Concepts and Abstract Syntax. Graham Klyne and Jeremy J. Carroll, eds. W3C Recommendation, 10 February2004, [Online]Available at http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/. [Accessed 20 July 2015]

[51] Spivey, J. M., 1992, The Z Notation: A Reference Manual, Second Edition, Prentice Hall.

[52] Internet Engineering Task IETF, 1997, Key words for use in RFCs to Indicate Requirement Levels, [Online] Available at http://www.ietf.org/rfc/rfc2119.txt. [Accesed 20 July 2015]

[53] OASIS, "Web Services Business Process Execution Language Version 2.0 11" April 2007, OASIS Standard [Online] Available at http://docs.oasis-open.org/wsbpel/2.0/OS/wsbpel-v2.0-OS.html [Accessed 13 July 2015]

[54] ISO/IEC, 2002 Z formal specification notation –Syntax, type system and semantics First Edition ISO/IEC 13568

[55] Aagedal, J. Ø., Bézivin,J. & Linington, P. F. 2004. Model-Driven Development (WMDD 2004) in Object-Oriented Technology: (ECOOP) 2004 Workshop Reader, (ECOOP) 2004 Workshops, Oslo, Norway, June 14-18, 2004, Final Reports pp. 148-157

[56] Glinz, Martin 2000. Problems and Deficiencies of UML as a Requirements Specification Language in Proceedings of the 10th International Workshop on Software Specification and Design IWSSD'00 IEEE Computer Society

[57] H. Kühn, "Methodenintegration im Business Engineering"[in German] PhD Thesis, University of Vienna, April 2004

[58] Saadatmand, M., Cicchetti, A. & Sjödin, M. 2011, UML-Based Modeling of Non-Functional Requirements in Telecommunication Systems in The Sixth International Conference on Software Engineering Advances , ICSEA 2011

[59] Selic, B., 2011. The Theory and Practice of Modeling Language Design for Model-Based Software Engineering—A Personal Perspective. In: Generative and Transformational Techniques in Software Engineering III. s.l.:Springer Berlin Heidelberg, pp. 290-321.

[60] DISRUPT Consortium, "Project Overview", 2017 [Online] Available at http://disrupt-project.eu/about/overview [Accessed 24 January 2017]

[61] DISRUPT Consortium, "DISRUPT Developers Space" [Online] Available at https://www.adoxx.org/live/web/disrupt/ [Accessed 24 January 2017]

[62] Eclipse Fundation, "Xtend," 2015. [Online]. Available at https://eclipse.org/xtend. [Accessed 15 July 2016].

[63] N. Visic and D. Karagiannis, "Developing Conceptual Modeling Tools Using a DSL," in Knowledge Science, Engineering and Management, Sibiu, Romania, Springer, 2014, pp. 162-173

[64] ADOxx.org ADOxx Library Language (ALL), [Online] Available at: https://www.adoxx.org/live/library-language-all/abl, [Accessed 03 March 2017]

[65] GO0D MAN Consortium,, The project "GO0D MAN: Agent Oriented Zero Defect Multi-Stage Manufacturing" [Online] Available at: http://go0dman-project.eu/ [Accessed 03 March 2017]

[66] D. Karagiannis, H. C. Mayr, J. Mylopoulos, Domain-Specific Conceptual Modelling, Springer International Publishing, 2016

[67] D. Karagiannis, H. C. Mayr, J. Mylopoulos, Domain-Specific Conceptual Modelling, Springer International Publishing, 2016

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 137-146 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

GPSR+Predict: An Enhancement for GPSR to Make Smart Routing Decision by Anticipating Movement of Vehicles in VANETs

Zineb Squalli Houssaini*[1], Imane Zaimi[2], Mohammed Oumsis[2, 3], Saïd El Alaoui Ouatik[1]

[1] *IT laboratory and Modeling (LIM) Faculty of Science at Dhar El Mahraz (FSDM) Sidi Mohammed Ben Abdellah University (USMBA), Fez, Morocco.*

[2] *LRIT, Associated Unit to CNRST (URAC 29) Faculty of Science, Mohammed V University, Rabat, Morocco.*

[3] *High School of Technology, Mohammed V University, Sale, Morocco.*

ARTICLEINFO

Article history:
Received: 06 March, 2017
Accepted: 07 April, 2017
Online: 18 April, 2017

Keywords:
VANET
GPSR
Estimation
Routing
NS2
VanetMobiSim
QoS
Highway and urban environments

ABSTRACT

Vehicular ad-hoc networks (VANETs) are still more challenging to overcome even if they have been widely studied during the last decades. The routing mechanism is the essentially relevant issue in this field. Indeed, it must strictly to be adapted to specific and unique characteristics such as the high mobility of the vehicles, the dynamic nature of network topology as well as the high link breakage probability. In this paper, our objective is to improve the greedy perimeter stateless routing protocol (GPSR) as being the most promising position-based mechanism. However, according to the impact of position information on routing decision, our proposed approach defined by GPSR+PRedict protocol ensures that each vehicle estimates its own position for the near future. Afterwards, through extensive experiments, we show the ability of the GPSR+PRedict to overcome the observed problems and to enhance the overall performance of the traditional GPSR approach. The simulations are carried out on both highway and urban scenarios by using NS-2 and VanetMobiSim simulators.

1. Introduction

Vehicular ad-hoc network (VANET) is a new class of mobile ad-hoc network (MANET) that has emerged through the advances of wireless technologies and automotive industry [1]. This application field has become the core of the intelligent transport systems (ITS) in order to enhance road safety and improve comfort of road users [2, 3]. In addition, VANET's main features are [4]: traffic monitoring, traffic control, increasing visibility in dangerous intersections, detection of collisions and real time traffic calculation. Moreover, other applications apart from intelligent transport reside on providing connectivity to network or any communication between vehicles (e.g. games, chats or file exchange).

VANET behaves distinctly when compared to the other infrastructureless networks. Indeed, it is distinguished by specific and unique characteristics, especially, the huge mobility of the vehicles and the instability of network topology [5-7]. Consequently, a series of constraints and challenges including security, quality of service (QoS), and routing mechanism should be raised to meet the needs of aforementioned VANET features [8]. For that reason, different proposals and research studies are booming in order to achieve better results than the previously proposed solutions.

To be part of such network and be able to exchange messages smoothly either between vehicles (vehicle-to-vehicle communication (V2V)) or between vehicles and roadside infrastructure (vehicle-to-roadside communication (V2R)) [9], the intelligent vehicle represents the main application. Nevertheless, it must be equipped with some electronic tools namely, wireless communication devices, environmental perception devices (e.g. radars, cameras…), geographic positioning systems (e.g. GPS) [10, 11], and a platform for processing incoming and outgoing information. Figure 1 shows some of these intrinsic equipment [12, 13]. This paper is based

*Zineb Squalli Houssaini, Faculty of science at dhar el mahraz (FSDM) sidi mohammed ben abdellah university (USMBA), Fez, Morocco, Contact No: +212676468098 & Email: zinebsqualli@gmail.com.

on purely V2Vad-hoc communication architecture, but it could be easily adapted to V2R communication with only simple changes.

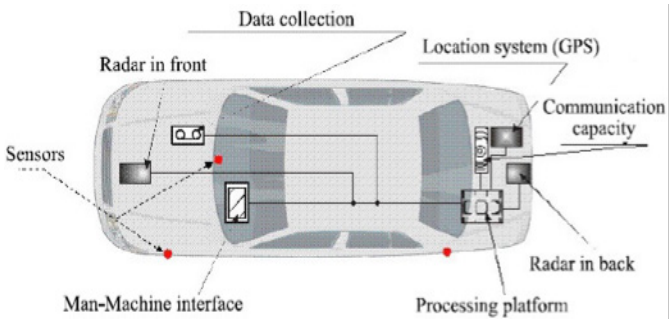


Figure 1: Illustration of different components of an intelligent vehicle [13]

Within this ever-changing vehicular ad-hoc network, messages must be transmitted smoothly from a vehicle to another in order to reach their final destinations. Hence, a routing protocol should be used to ensure inter-vehicular communications in an efficient manner. Exposed to the difficult constraints of vehicular networks, the obvious question is how to ensure routing with a satisfying quality of service (i.e., short delay, minimum lost packets and low overhead).

VANET routing protocols history starts with MANET protocols [14-18]. Unfortunately, these latter cannot achieve a good performance when they are used directly in high challenging environment like VANET. Therefore, many researches focus on adapting these protocols in order to be suitable for VANETs [19-22]. The geographical routing protocols [23-25] are considered as the most stable and functional schemes for VANET's large-scale areas compared to the topology-based routing protocols [26]. The Greedy Perimeter Stateless Routing (GPSR) is a typical example of protocols based on geographical location[27], where each node uses the current information of its position, the position of its neighbors and that of the final destination to make the routing decision.

Considering three significant points, namely, 1) the interest that presents the location's information and vehicles' movement on the both vehicular networks and geographic protocols, 2) the unemployment of amount information, namely, speed and direction given by GPS, 3) the problems observed in geographic protocols as link breakage, it would be advantageous to exploit all these information together to estimate future positions and to consider them during decision process of routing protocols.

In this context, a new geographical routing protocol based on GPSR is proposed. The improvement is mainly defined by using supplementary information to estimate the location of vehicles in the near future. This estimated future position will be included in the periodic message (hello packet) in order to be taken into account in the decision of the best next hop.

The rest of the paper is organized as follows: Section 2 summarizes the related works of routing protocols over VANETs. In Section 3, we provide a brief overview of GPSR protocol. Section 4 describes fully the design of our proposed GPSR+PRedict routing protocol. Simulation settings are presented in section 5, while section 6 shows the effectiveness of GPSR+PRedict via simulation experiments. Finally, the section 7 concludes the study and provides a number of recommendations.

2. Related Work

Actually, there are a lot of studies that are interested to the MANET routing protocols adjustment for being used in VANET environments effectively [28]. Additionally, most of researchers are gathering on the point that the geographical routing protocols are the most suitable for vehicular network [26, 29, 30]. Hence, our attention was directed towards one of the protocols belonging to this class: the famous GPSR protocol. GPSR protocol selects the shortest possible route to destination; however, it may suffer from a higher packet error rate due to the poor link quality or the high break link probability.

To this end, many improvements have been recommended. The authors of [31] proposed the cross-layer weighted position-based routing (CLWPR) to improve the efficiency of geographical routing protocol in vehicular network. The proposed algorithm utilizes information about link layer quality and positioning from navigation to anticipate the characteristics of an urban environment. Compared to the default GPSR, CLWPR demonstrates significantly better performance in terms of packet delivery ratio and end-to-end delay metrics. In the same context, authors of [32], designed an enhanced GPSR protocol and presented its performance for both urban and highway scenarios. The novelty resides on the fact that they used weight calculation to select best neighbor (best next hop). To compute this weight, they used location and velocity information as well as link quality metrics extracted during the hello messages receipt to produce routes that improve the performance of the network. Their proposed protocol achieves higher packet delivery ratio for the network, lower end-to-end delay, while keeping the energy consumption at the same low levels of GPSR. Furthermore, a grid-based predictive geographical routing protocol (GPGR) has been proposed by the researchers of [33]. Indeed, they used a map data to generate the road grid and to predict the moving position during the relay node selection process. Thus, it was noticed that the GPGR protocol reduces the possibility of link breakage and avoids falling on local maximum situation in urban scenarios. Moreover, authors of [33] suggested a new concept based on vehicles movement prediction. In fact, they estimated the Link Stability (LS) for each neighboring vehicle before selecting the next hop for data forwarding/sending. In this way, movement prediction-based routing (MOPR) protocol determines the most stable communication link of the network in terms of communication lifetime with respect to the movement of vehicles. Afterward, they selected the most stable route from the source till the destination.

To the best of our knowledge, all these approaches have been beneficial in most scenarios, but they concentrate all the required calculation on the forwarder during best next hop process or at the moment of receiving beacon packets, which can affect desperately the network's performance. In addition, in most of these aforementioned enhancements, there is a radical change in the process of choosing the best next hop of the default GPSR using more complicated calculations.

For that reason, we propose in our study to filling the observed gaps and to enhance even more the routing process while retaining the basic mechanism of GPSR protocol. Thus, our new GPSR+PRedict routing protocol does not include direction and speed knowledge in the hello packet, but it includes the estimated future position calculated according to this knowledge. Thereby, each vehicle is able to anticipate the state of its neighborhood in the near future. Our idea is directed regarding both highway and urban areas.

3. Overview of GPSR

GPSR was originally created for MANET, but quickly adapted to VANET [27]. It is one of the most known examples that approve the concept of position based routing. GPSR protocol switches between two forwarding strategies depending on the situation of communicating nodes. Usually, when node needs to send packets, it uses the first forwarding strategy defined by greedy forwarding strategy (GFS). In case of failure (i.e., there is no neighbor with least distance to the destination except the transmitter node itself), GPSR bypasses the problem by using the second strategy known by perimeter forwarding strategy (PFS). Figure 2 illustrates the behavior of GPSR.

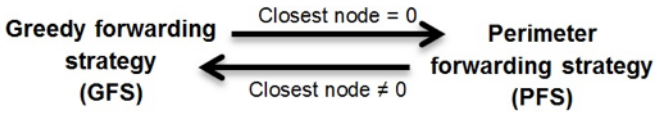


Figure 2: Communication strategies of GPSR

3.1. Maintain strategy

In GPSR, before choosing the best path to route a data packet, each node in the network must know its neighbors and their positions. Hence, the nodes flood the network by sending hello packets (beacon packets) containing current position and an identifier. Indeed, each node periodically shares a hello packet with the existing nodes within its own range radio, and simultaneously collects the information sent by these nodes. Due to the beacon packets exchanges, each node can build its own table of neighbors thanks to the collected information, and can maintain it according to the next beacon packets.

3.2. Routing strategy

In greedy forwarding strategy (see figure 3), the best next hop is selected based on the optimal path, i.e., it always selects the progressively closer node to the destination. This process is repeated at each intermediate neighbor until the intended destination is reached. With this routing decision, the packet can reach the destination through the optimal path in the aspect of distance. However, sometimes the forwarder vehicle has the shortest distance to destination compared to all its neighbors. In this context, there are two cases. In the first one, the final destination is under the radio range of transmitter vehicle; therefore, packet is forwarded directly. The second case is when the final destination is not accessible in a one hop, so it is impossible to use the GFS. Thus, the PFS is used in order to solve this problem (see figure 4). The PFS algorithm recovers by routing around the perimeter of the region until arriving at a node closer to the destination. At this moment, the GFS takes over.

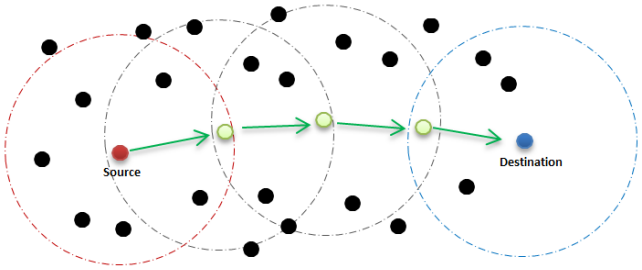


Figure 3: Greedy forwarding packets of GPSR

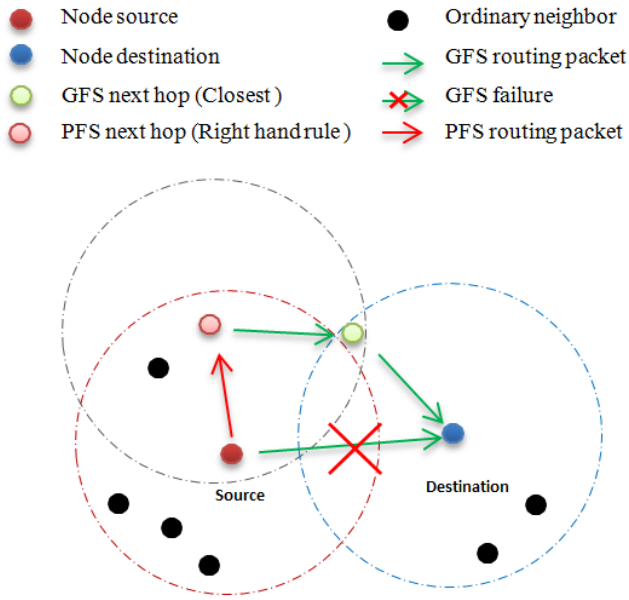


Figure 4: Coping with gap encountered thanks to perimeter forwarding strategy in GPSR

3.3. Benefits and Drawbacks

The most important advantage of GPSR, distinctly from the other proactive and reactive protocols, is the fact that the vehicles can get the exact movement information (e.g. geographic position, speed and movement direction) according to a positioning system like GPS or GALILEO. In addition, the one-hop propagated information over the topology, leading to know the geographic position of every single neighbor, minimizes more or less the network overload which is advantageous especially in dense networks. However, regarding the greedy routing strategy, the vehicle often chooses the best next hop in the distance aspect. Nevertheless, without considering speed or moving direction while researching the best next hop, the GPSR protocol could lead to wrong packet forwarding decisions, which may cause packet loss. Moreover, the link breakage problem can also affect the quantity of delivered packets, particularly in highway environment. Actually, when node receives a beacon packet to update its list of neighbors, it computes the best next hop, then it forwards packet. So, during this process, the selected neighbor can be out of the transmission range and the packet cannot reach it because of high network's mobility. In order to solve these problems that lead to performances degradation, we thought that it would be appealing, in terms of QoS, to improve the routing strategy. The enhancements are mainly based on the estimation of the future position of all participating nodes in routing operation. In the next section, a deep description of all proposals is provided.

4. GPSR+Predict

4.1. Motivation

GPSR+Predict has been conceived on the basis of three significant aspects: First, the importance of node's localization in geographic protocols. Second, the information available in GPS but not exploited by GPSR. Third, the various problems

encountered, for example, high breakage link probability due to the great mobility with different speeds, bad routing decisions due to opposite directions of vehicles, or inaccuracies in the prediction of the driver's direction at intersections.

In our point of view, it is constructive to propose a new greedy routing mechanism that estimates positions of the vehicles in the near future and takes this information into consideration to make smart routing decision.

4.2. Prediction the near future position

With the proposed protocol, each node uses its geographical position, its speed and its direction to estimate its position in a near future. This future position is attached to the hello packets and recalculated at each sent of periodic hello packet. In this way, node continuously has a prediction of future state of the network.

Moreover, unlike the basic GPSR, where the sender node simply searches the nearest neighbor to the destination when it needs to forward a data packet, that of the "GPSR+PRedict" searches firstly if the destination is in its list of neighbors. If it is the case, the packet data is sent directly to the destination, if not, then the sender searches the best next hop in terms of distance by using the estimated future position of neighbors. In fact, GPSR+PRedict practices this strategy in order to know the neighbors that will still or not in sender's neighborhood. Thus, it will make a smarter routing decision by avoiding several bad situations such as link breakage.

The choice of the prediction model is fundamental because it does not affect only the quality of prediction but also the speed and time of the calculation. In fact, there are different mathematical extrapolation methods with different levels of complexity allowing the estimation of future positions of vehicles [35-37]. For this study, we have opted to use a method based on previous positions, direction and speed. This technique is basic with very little intricacy so that the hello packet size does not involve the network overhead (i.e., it does not incur much bandwidth or computational power).

As mentioned above, this model exploits previous positions $\boldsymbol{X_{t0}}$ and velocity $\vec{\boldsymbol{V}}(\boldsymbol{X_{t0}})$. The velocity vector is calculated by using the current speed and heading direction of a vehicle (see equations (1) and (2)).

$$\vec{\mathbf{V}}(\mathbf{X_{t0}}) = \mathbf{dX_{t0}} * \mathbf{Speed} \quad (1)$$

$$\vec{\mathbf{V}}(\mathbf{Y_{t0}}) = \mathbf{dY_{t0}} * \mathbf{Speed} \quad (2)$$

Where dX_{t0} and dY_{t0} are a unit vector that indicate the direction of the vehicle at time t_0.

Therefore, the future position from any node can be calculated with the following equations (3) and (4):

$$\mathbf{fut(X_{t0})} = \mathbf{X_{t0}} + \vec{\mathbf{V}}(\mathbf{X_{t0}}) * \Delta\mathbf{T} \quad (3)$$

$$\mathbf{fut(Y_{t0})} = \mathbf{Y_{t0}} + \vec{\mathbf{V}}(\mathbf{Y_{t0}}) * \Delta\mathbf{T} \quad (4)$$

Where ΔT is the time between two hello packets.

4.3. Optimization

The highlight is to improve routing mechanism by assisting GPSR protocol to make smarter routing decisions and to better choose the suitable neighbor as next hop while maintaining the basic search process of best next hop. Also, we aim to optimize the calculation processing of GPSR and to not concentrate all the processing on the router of the data packet.

From most existing prediction approaches applied to the geographic protocols, the direction and the speed are added then sent in the hello packets. Hence, the nodes do not predict the future positions only after receiving these hello packets or during the research of the next best hop process, i.e., each node must traverse its entire neighbors list and make the required calculation to estimate the future position of each neighbor. This procedure causes two mainly drawbacks: the first is that each node has a lot of calculations to perform, while the second resides on the fact that the same calculation is repeated by several nodes (shared neighbors). As consequence, this strategy of prediction is not optimal and can affect the performance and the QoS especially in dense networks. We have addressed this problem by proposing the idea that each node in the network exploits its information (position and velocity) and estimates its future position in advance, then includes the result in the hello messages (see table 1). In this way, the calculations will be dispersed over the network and will be performed only once by the concerned node.

Through GPSR+PRedict protocol, when a node broadcasts the beacon packet to the other vehicles within its radio range, each vehicle that receive this hello packet, stores all information in its neighbors list. Therefore, the current and the future positions information are available to be used while computing best next hop. The illustration of both hello packet and neighbors list used in GPSR+PRedict is shown respectively in tables 1 and 2 taking the node1 at time t_0 as example.

Table 1: Format of hello packet

ID node	Geo-coordinates of node		Future position estimated of node	
1	X_{t0}	Y_{t0}	$fut(X_{t0})$	$fut(Y_{t0})$

Table 2: Format of neighbors list

ID neighbors	Geo-coordinates of neighbors		Future position estimated of neighbors	
1	X_{n1_t0}	Y_{n1_t0}	$fut(X_{n1\_t0})$	$fut(Y_{n1\_t0})$

Figure 5 summarizes all that has been explained concerning our improved greedy routing strategy. It represents the sending process of the hello packets and that of the data packets as well.

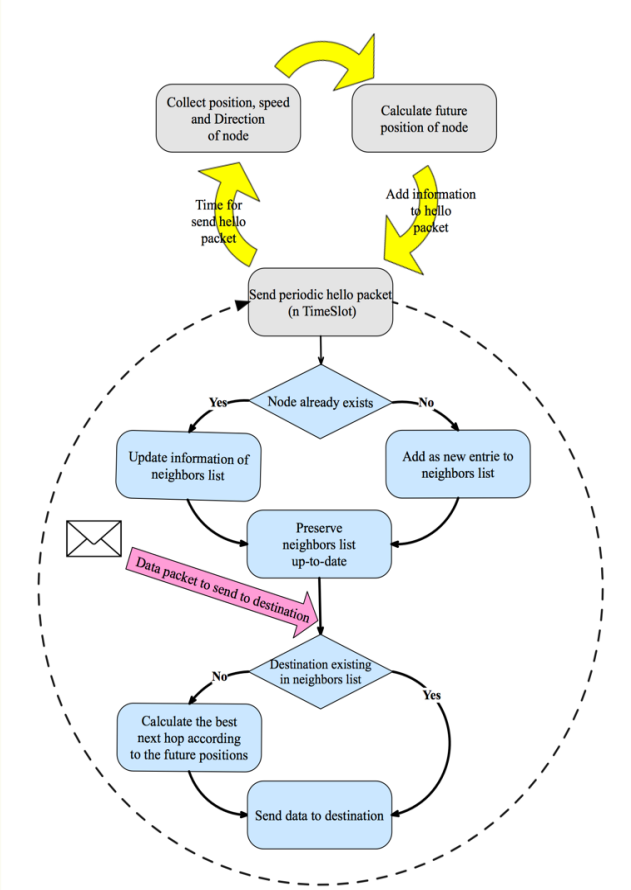


Figure 5: Greedy strategy of new GPSR+PRedict protocol

Our interest to improve the GPSR's routing mechanism is clearly motivated by the three cases presented respectively in figures 6, 7 and 8. It should be noticed that the default GPSR routing mechanism forms the red route (red arrows) where there is often breakage links, while the GPSR+PRedict routing mechanism forms the optimal green route (green arrows).

From figure 6, according to the default greedy strategy, the source node chooses the node A as the next hop since it is the closest one to the final destination. Nevertheless, the packet will be lost since the direction of A is different from that of the destination. The proposed mechanism forms the green route and avoids the problem of opposite directions that will occur in the red route in a very short amount of time.

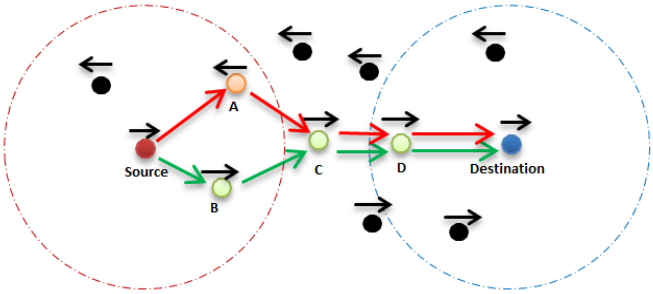


Figure 6: Problem of opposite direction in GPSR

Another problem is overcome as shown in figure 7: the problem of out of the radio range. In fact, according to the default GPSR forwarding process, when the highly mobile node marked by A is chosen as the next hop to send data packet to the final destination, it will be out of the range before the starting of transmission. In regards to GPSR+PRedict, it anticipates link breakage problem, which certainly will occur, and avoids it by selecting the node B as the next hop.

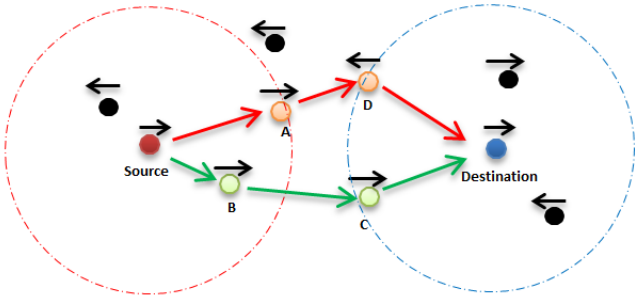


Figure 7: Problem of out of radio range in GPSR

In figure 8, considering that the vehicle B is much faster than A, so the vehicle B will exceed the vehicle A in a very short period of time. With GPSR+PRedict forwarding process, the source expects this situation since it has a future position estimated of its neighbors A and B. Consequently, the source chooses the node B as next hop instead of node A in case of classical GPSR. Thus, the speed problem is bypassed.

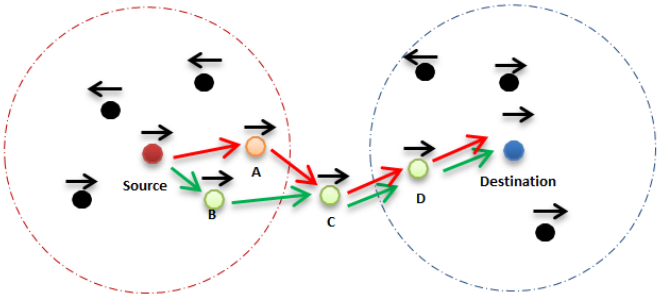


Figure 8: Problem of speed in GPSR

5. Simulations settings

5.1. Parameters of simulation

In the literature, various tools have been implemented to perform experimental studies. In this paper, we used NS-2 software which is the most employed simulator to study the protocol's performance [38]. Additionally, VanetMobiSim is used as a realistic mobility traffic generator [39, 40], where vehicles travel according to an Intelligent Driver Model with Lane Changes (IDM-LC). VanetMobiSim simulator is considered as the most efficient mobility modeling framework. Indeed, itis widely used to simulate realistic automotive motions at a macroscopic and microscopic level. Here, we have used this software to generate both of realistic mobility models and realistic urban/highway traffic environments. The output mobility trace obtained from VanetMobiSim is directly used as input in NS-2.

In order to evaluate the effectiveness of our enhancement, we have opted to compare our GPSR+PRedict and the default GPSR protocol in highway and urban scenarios since they are two

different milieus with a very distinct infrastructures and traffic rules. Each scenario is designed by varying the number of vehicles from 50 to 350 vehicles and keeping data traffic sources number at15 connections. The "TwoRayGround" model is employed as propagation model. The protocol MAC sets according to the IEEE 802.11p standard, and operates at 5.9 GHz on a 10 MHz control channel (CCH). The PHY data rate is configured to 6 Mbps, which is the minimum value to safe the communication [41]. The couples' source/destination were randomly selected and all nodes exchange beacon packets every 5 seconds (i.e., hello period) according to those protocols. Table 3 shows the common configuration parameters for the following set of experiments.

Table 3: Network parameters for both highway and urban scenarios

NS2.33 network simulator	
Parameter	**value**
MAC Layer	802.11p
Propagation Model	Two-Ray-Ground
Radio range	250 meters
Traffic model	CBR/UDP
Hello Period	5s
Hello Timeout	50 s
Time Simulation	500 s
Packet Interval	1 pkt/s
Packet Size	512 octets
Number of traffic	15 connections
Number of vehicles	50, 100, 150, 200, 250, 300 and 350
Routing Protocols	GPSR+PRedict and GPSR

5.1.1. *Urban scenario*

For urban scenario, a space of 1000*1000 meters^2 is randomly generated. It consists of various intersections, two-way road, and different density of vehicles in motion with variation speed between 20 and 60 km/h. Figure 9 illustrates the urban topology used for our experiments, while table 4 shows the parameters of the mobility scenario.

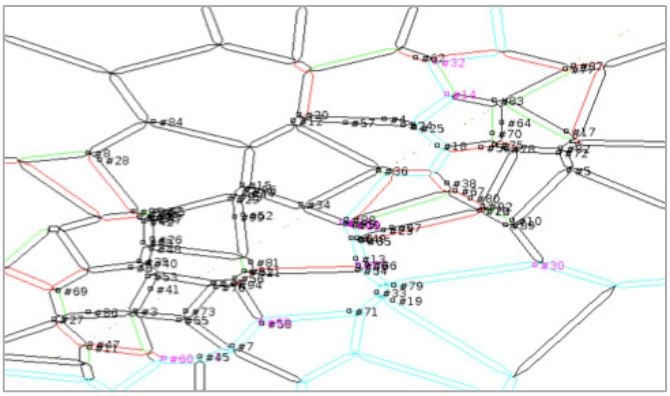


Figure 9: Urban topology randomly generated by VanetMobiSim

Table 4: Urban mobility model parameters

VanetMobiSim mobility generator: urban scenario	
Parameter	**value**
Mobility model	IDM-LC
Space area	1000*1000 meters^2
Vehicle speed	20 to 60 km/h
Traffic light	10
Number of lanes per direction	2

5.1.2. *Highway scenario*

In this subsection, a highway scenario over 700*5000 meters^2 is randomly generated. It consists of two-way road, i.e., two lanes per direction, without any traffic lights. Different density of vehicles is driving at speeds varying between 80 and 120 km/h. Figure 9 illustrates an example of the highway topology used for our simulations, while table 5 shows the parameters of the mobility scenario.

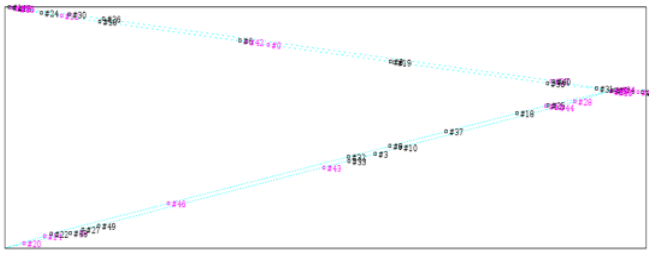


Figure 10: Highway topology randomly generated by VanetMobiSim

Table 5: Highway mobility model parameters

VanetMobiSim mobility generator: highway scenario	
Parameter	**value**
Mobility model	IDM-LC
Space area	700*5000 meters^2
Vehicle speed	80 to 120 km/h
Traffic light	0
Number of lanes per direction	2

5.2. *Metrics*

In order to evaluate and assess the performance of the simulated routing protocols, we use the following metrics:

Packet Delivery Ration (PDR): corresponds to the ratio between the number of packets of data correctly received by the destinations and the number of packets sent by the sources; this latter allows revealing whether a protocol is able to send all outgoing data packets. The final equation of PDR shown in (5):

$$\boldsymbol{PDR = \frac{\sum Received\ Packets}{\sum Sent\ Packets}} \quad (5)$$

End-to-End delay average (E2E) is defined as the average time that a data packet takes to travel from the source to the destination vehicle. It represents the main parameter that must be improved during the evaluation of a routing protocol. A good protocol must have an end-to-end delay average as low as possible. Its calculation is as follows in equation (6):

$$E2E = \frac{PacketTransmissionTime}{\sum Received\ Packets} \quad (6)$$

Throughput: it measures the flux or the quantity of data traveling over a communication channel per unit time, usually expressed in bits per second or packets per second. We use the following equation (7) to calculate the average throughput:

$$Throughput = \frac{Size\ of\ ReceivedPacket}{TimeReception - TimeSending} \quad (7)$$

Route-costs is calculated by dividing the total number of bytes of the routing packets that (includes forwarded routing packets and control packets as well) and the total amount of the received data (see equation (8)).

$$Routecosts = \frac{\sum Size\ of\ TransmittedRoutingPackets}{\sum Size\ of\ ReceivedDataPackets} \quad (8)$$

Normalized routing Load (NRL) defined as the total number of routing packets transmitted per data packet. It calculated by dividing the total sum of routing packets that sent (includes forwarded routing packets as well) by the total number of data packets received as follows in below equation (9):

$$NRL = \frac{\sum TransmittedRoutingPackets}{\sum ReceivedDataPackets} \quad (9)$$

6. Simulation Experiments

Initially, the objective of our experiments is to evaluate the effectiveness of our improved protocol, taking into account the impact of node density regarding routing quality of service. Therefore, we varied the number of vehicles to 50, 100, 150, 200, 250, 300 and 350 respectively, while keeping number of connections fixed at 15 connections. In order to guarantee the credibility and reliability of the experimental results, we respectively made 10 time experiments during 500 seconds for each simulation. The simulations result in both highway and urban are presented in figures 11 to 20, in which we illustrate the packet delivery ratio, the average end-to-end delay, the throughput, the route-costs and the normalized routing load for GPSR+PRedict and default GPSR.

In figures 11 and 12, default GPSR seems to have worst performance in both scenarios, while our proposed GPSR+PRedict gives good values. This is because GPSR+PRedict anticipates in advance the changes of vehicles localizations. Thus, it avoids the problems caused by the high mobility of the vehicles and the dynamic nature of the topology. In the highway scenario, our approach remarkably improves the PDR compared to the default GPSR since the GPSR mismanages the very high mobility of vehicles.

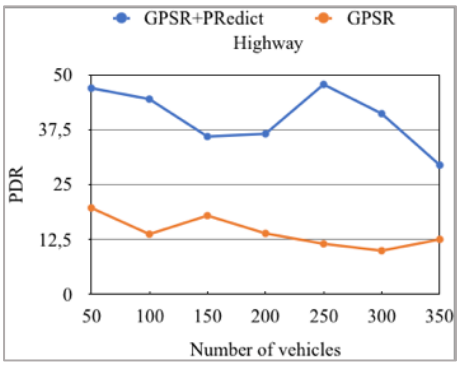


Figure 11: Packet Delivery ratio VS nodes density in highway scenario

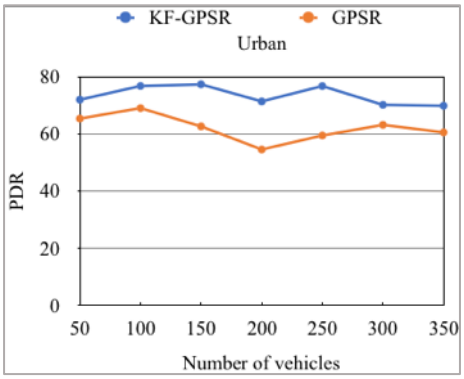


Figure 12: Packet Delivery ratio VS nodes density in urban scenario

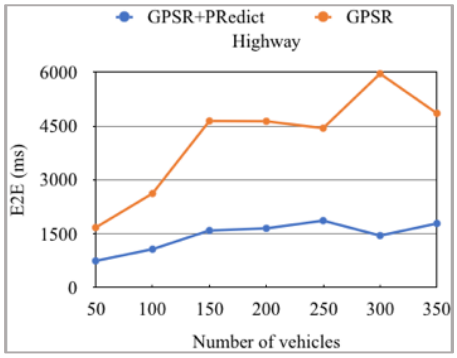


Figure 13: End-to-end delay VS nodes density in highway scenario

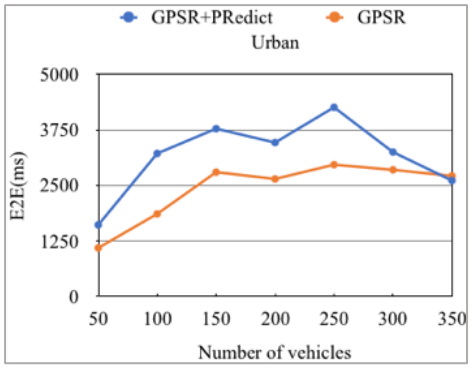


Figure 14: End-to-end delay VS nodes density in urban scenario

In figure 13, regarding the highway scenario, it can be noticed that the classical GPSR has the longer average end-to-end delay due to the outdated location information that results very probably data loss. While, GPSR+PRedict decreases End-to-end delay since it avoids the wrong routing decisions by using future position estimated, hence, the routed data packet will have less chance to be lost. In contrast to the urban scenario (see figure 14), we note that our enhanced protocol has the worst end-to-end delay. This is probably because of two things: its higher PDR compared to the default GPSR, and the mobility of vehicles, which is relatively inferior in this environment than in the highway.

Figures 15 and 16 present the flow in function of density. Both figures show that our improved GPSR+PRedict has higher flow than default GPSR. This can be explained by the fact of the good receipt of data packets. As we see, the performance of default GPSR is very low compared to our approach especially in the highway scenario. This is probably because of these bad routing decisions since it does not take into account the high mobility of vehicles, and this is clearly seen from the significant drop shown in figure 16.

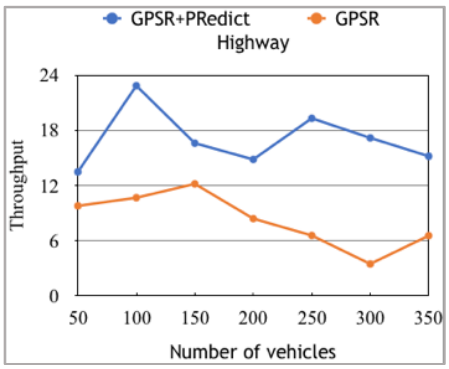


Figure 15: Throughput VS nodes density in highway scenario

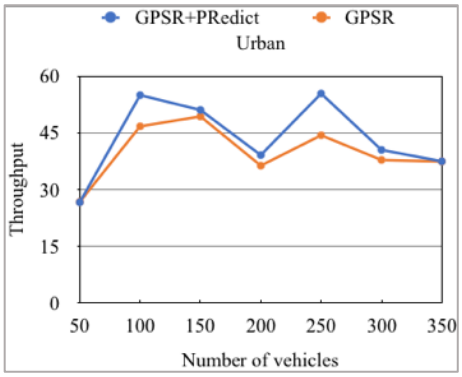


Figure 16: Throughput VS nodes density in urban scenario

Figures 17 and 18 illustrate the cost of routing according to the number of moving vehicles respectively in both highway and urban scenarios. The simulation results show that GPSR+PRedict guarantees low cost of routing compared to the default GPSR in highway scenario (see figure 17). The improved protocol suffers less from link breakages, which conduct to less packet loss and less routing cost. As for results of urban scenario presented in figure 18, GPSR+PRedict has a very slight rise in cost of routing compared with default GPSR. These values can be justified by the high PDR of GPSR+PRedict in the same scenario.

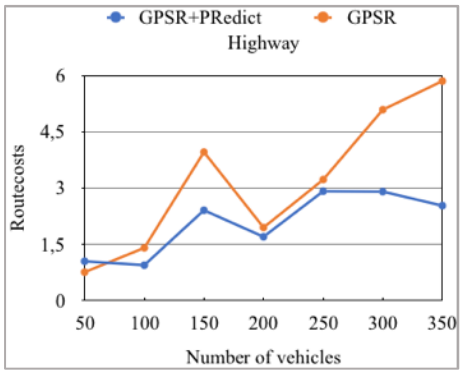


Figure 17: Route-cost VS nodes density in highway scenario

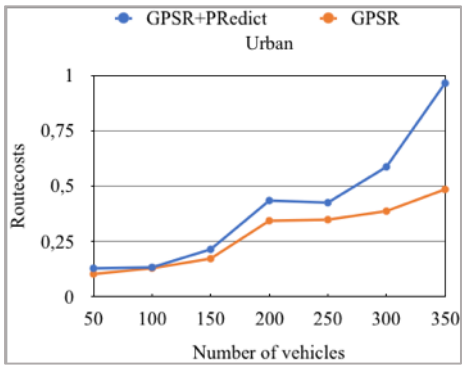


Figure 18: Route-cost VS nodes density in urban scenario

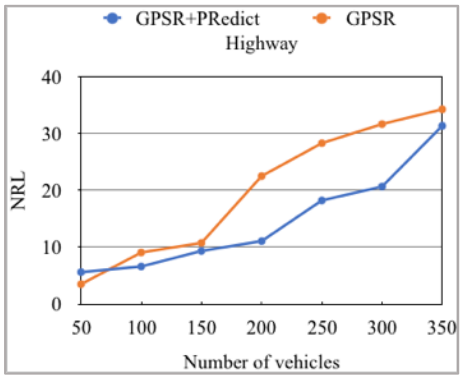


Figure 19: Normalized routing load VS nodes density in highway scenario

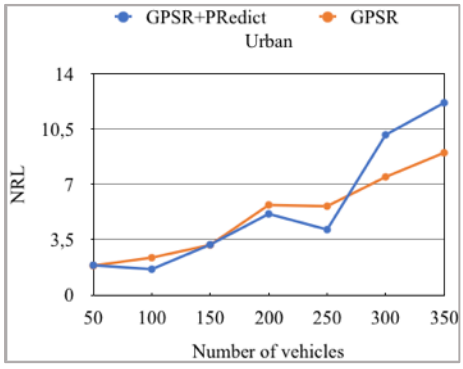


Figure 20: Normalized routing load VS nodes density in urban scenario

In regards of density, the normalized routing load increases when there are more router packets in the network. As shown in figures 19 and 20, the NRL proportionally increases with the number of vehicles. In The highway environment (figure 19) the performance evaluation shows that GPSR+PRedict outperforms GPSR in term of the normalized routing overload rate. As for the urban environment, presented in figure 20, our enhanced GPSR+PRedict improves the NRL for almost all the tested cases. In fact, default GPSR marks slight results than the GPSR+PRedict, especially in very dense network.

Regarding comparison of the obtained results for both routing protocols, it is clear that the proposed enhancement GPSR+PRedict practically guarantees better performance in all simulations compared to default GPSR.

The enhancements of GPSR protocol that are made in order to adapt it to the specific characteristics of VANET, have shown their gains especially in highway environment where mobility and topology are instable. Indeed, we notice a significant progress with all metrics in all cases compared to default GPSR. Meanwhile, in urban environment where mobility and topology are relatively less dynamic, our approach improves quietly packet delivery ratio, flow and normalized routing load. However, for average end-to-end delay and route-costs, the default GPSR is better, but the deficit is quite small and tolerable. Therefore, the proposed enhancement is a strong candidate to implement with GPSR.

7. Conclusion and Perspectives

In this paper, a new routing approach is developed in order to guarantee a high quality of service in VANET areas. The novelty resides on the simplicity of design and identity preserved of the original GPSR protocol. In fact, GPSR+Predict routing protocol exploits in a simple and optimized way node movement information, namely the position, the direction and the speed, to estimate vehicles localization in the near future. This predicted future position will be included in the hello packet and diffused periodically among neighbors, hence, it will greatly contribute to make a smart routing decision and to better select the appropriate next hop.

Our simulations using different frameworks (NS-2, VanetMobiSim) are directed respectively in both highway and urban environments with taking into account the impact of density. The results show that GPSR+PRedict has the ability to preserve and to increase the performance of the vehicular network. Moreover, it indicates that the proposed technique avoids relatively various problems resulting from the delicate properties of this type of networks. Indeed, it achieves a higher packet delivery ratio, a low end-to-end delay, a low routing cost, a low routing overhead and a high throughput, practically in all scenarios.

For future work, we plan to enhance the accuracy of prediction to be more near to reality, and to carry out more enhancement considering QoS on large environments. Furthermore, it should be noted that the proposed approach is appropriate for constant bit rate traffic (CBR); we plan to extend the approach to other types of applications, for example, streaming in multimedia application.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] Lee, Kevin C., Uichin Lee, and Mario Gerla. "Survey of routing protocols in vehicular ad hoc networks." Advances in vehicular ad-hoc networks: Developments and challenges (2010): 149-170.

[2] Ferdinands, A. "Intelligent transport systems and road safety." INSURANCE COMMISSION OF WESTERN AUSTRALIA CONFERENCE ON ROAD SAFETY, 1999, PERTH, WESTERN AUSTRALIA. 1999.

[3] Rumar, K., et al. "Intelligent transport systems and road safety." (1999).

[4] Li, Fan, and Yu Wang. "Routing in vehicular ad hoc networks: A survey." IEEE Vehicular technology magazine 2.2 -2007.

[5] Berbineau, Marion, et al. "Communication technologies for vehicles." NewYork: Springer. doi 10.1007 -2013-: 978-3

[6] Xin Wang, editor. Mobile Ad-Hoc Networks: Applications. INTECH, January 30, 2011. DOI: 10.5772/882

[7] Moustafa, Hassnaa, and Yan Zhang. Vehicular networks: techniques, standards, and applications. Auerbach publications, 2009.

[8] Blum, Jeremy J., Azim Eskandarian, and Lance J. Hoffman. "Challenges of intervehicle ad hoc networks." IEEE transactions on intelligent transportation systems 5.4 -2004-: 347-351.

[9] WAN, Jiafu, YAN, Hehua, SUO, Hui, et al. Advances in Cyber-Physical Systems Research. TIIS, 2011, vol. 5, no 11, p. 1891-1908.

[10] BH-Wellenhof, H. Lichtenegger, and J. Collins. "Global Positioning System: theory and practice." 4th ed., Springer-Verlag, 1997.

[11] Elliott, D. Kaplan, and J. Hegarty Christopher. "Understanding GPS: principles and applications." Edition Kaplan, 1996.

[12] Kumar, Rakesh, and Mayank Dave. "Mobile agent as an approach to improve qos in vehicular ad hoc network." arXiv preprint arXiv:1108.2095, 2011.

[13] J. P. Hubaux, S. Capkun and Jun Luo, "The security and privacy of smart vehicles," in *IEEE Security & Privacy*, vol. 2, no. 3, pp. 49-55, May-June 2004. doi: 10.1109/MSP.2004.26

[14] Li, Huaizhi, and Mukesh Singhal. "A scalable routing protocol for ad hoc networks." Vehicular Technology Conference, 2005. VTC 2005-Spring. 2005 IEEE 61st. Vol. 4. IEEE, 2005.

[15] Giruka, Venkata C., and Mukesh Singhal. "A self-healing on-demand geographic path routing protocol for mobile ad-hoc networks." Ad Hoc Networks 5.7 -2007-: 1113-1128.

[16] De, Swades, and Sajal K. Das. "Dynamic multipath routing -DMPR-: An approach to improve resource utilization in networks for real-time traffic." Modeling, Analysis and Simulation of Computer and Telecommunication Systems, 2001. Proceedings. Ninth International Symposium on. IEEE, 2001.

[17] Haas, Zygmunt J., and Marc R. Pearlman. "The performance of query control schemes for the zone routing protocol." IEEE/ACM Transactions on Networking -TON- 9.4 -2001-: 427-438.

[18] Boleng, Jeff, and Tracy Camp. "Adaptive location aided mobile ad hoc network routing." Performance, Computing, and Communications, 2004 IEEE International Conference on. IEEE, 2004.

[19] Saeed Ahmadi Arzil, Majid Hosseinpour Aghdam and M. A. J. Jamali, "Adaptive routing protocol for VANETs in city environments using real-time traffic information," *2010 International Conference on Information, Networking and Automation -ICINA-*, Kunming, 2010, pp. V2-132-V2-136. doi: 10.1109/ICINA.2010.5636966

[20] Lochert, Christian, et al. "A routing strategy for vehicular ad hoc networks in city environments." Intelligent Vehicles Symposium, 2003. Proceedings. IEEE. IEEE, 2003.

[21] Viriyasitavat, Wantanee, Fan Bai, and Ozan K. Tonguz. "Dynamics of network connectivity in urban vehicular networks." IEEE Journal on Selected Areas in Communications 29.3 -2011-: 515-533.

[22] Blum, Jeremy J., Azim Eskandarian, and Lance J. Hoffman. "Challenges of intervehicle ad hoc networks." IEEE transactions on intelligent transportation systems 5.4 -2004-: 347-351.

[23] Leontiadis, Ilias, and Cecilia Mascolo. "Geopps: Geographical opportunistic routing for vehicular networks." World of Wireless, Mobile and Multimedia Networks, 2007. WoWMoM 2007. IEEE International Symposium on a. Ieee, 2007.

[24] Mo, Zhaomin, et al. "MURU: A multi-hop routing protocol for urban vehicular ad hoc networks." Mobile and Ubiquitous Systems: Networking & Services, 2006 Third Annual International Conference on. IEEE, 2006.
[25] Ding, Yong, Chen Wang, and Li Xiao. "A static-node assisted adaptive routing protocol in vehicular networks." Proceedings of the fourth ACM international workshop on Vehicular ad hoc networks. ACM, 2007
[26] Altayeb, Marwa, and Imad Mahgoub. "A survey of vehicular ad hoc networks routing protocols." International Journal of Innovation and Applied Studies 3.3 2013-: 829-846.
[27] Karp, Brad, and Hsiang-Tsung Kung. "GPSR: Greedy perimeter stateless routing for wireless networks." Proceedings of the 6th annual international conference on Mobile computing and networking. ACM, 2000.
[28] Fazio, Peppino, et al. "A new channel assignment scheme for interference-aware routing in vehicular networks." Vehicular Technology Conference -VTC Spring-, 2011 IEEE 73rd. IEEE, 2011.
[29] Liu, J., Wan, J., Wang, Q. et al. TelecommunSyst -2016- 62: 15. doi:10.1007/s11235-015-9979-7
[30] Cho, Kuk-Hyun, and Min-Woo Ryu. "A survey of greedy routing protocols for vehicular ad hoc networks." SmartCR 2.2 -2012-: 125-137.
[31] Katsaros, Konstantinos, et al. "CLWPR—a novel cross-layer optimized position based routing protocol for VANETs." Vehicular Networking Conference -VNC-, 2011 IEEE. IEEE, 2011.
[32] Bouras, Christos, Vaggelis Kapoulas, and Enea Tsanai. "A GPSR enhancement mechanism for routing in VANETs." International Conference on Wired/Wireless Internet Communication. Springer International Publishing, 2015.
[33] Cha, Si-Ho, Keun-Wang Lee, and Hyun-Seob Cho. "Grid-based predictive geographical routing for inter-vehicle communication in urban areas." International Journal of Distributed Sensor Networks (2012).
[34] Menouar, Hamid, Massimiliano Lenardi, and FethiFilali. "Movement prediction-based routing -MOPR- concept for position-based routing in vehicular networks." Vehicular Technology Conference, 2007. VTC-2007 fall. 2007 IEEE 66th. IEEE, 2007.
[35] D. Katsaros and Y. Manolopoulos, "Prediction in wireless networks by Markov chains," in IEEE Wireless Communications, vol. 16, no. 2, pp. 56-64, April 2009. doi: 10.1109/MWC.2009.4907561
[36] Feng, Huifang, et al. "Location prediction of vehicles in VANETs using a Kalman filter." Wireless personal communications 80.2 (2015): 543-559.
[37] Coupelon, Olivier. "Neural network modelling for stock movement prediction A state of the art." Blaise Pascal University (2007).
[38] "The network simulator ns-2," 2010. [Online]. Available :http://www.isi.edu/nsnam/ns.
[39] Härri, Jérôme, et al. "VanetMobiSim: generating realistic mobility patterns for VANETs." Proceedings of the 3rd international workshop on Vehicular ad hoc networks. ACM, 2006.
[40] Harri, J., and Marco Fiore. "VanetMobiSim–Vehicular Ad hoc Network mobility extension to the CanuMobiSim framework." Institute Eurecom Department of Mobile Commu 6904 (2006): 1-19.
[41] Jiang, Daniel; Chen, Qi; Delgrossi, Luca; Optimal data rate selection for vehicle safety communications, Proceedings of the fifth ACM international workshop on Vehicular Inter-NETworking,30-38,2008

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 147-152 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Optimal Synthesis of Universal Space Vector Digital Algorithm for Matrix Converters

Adrian Popovici*, Mircea Băbăiţă[1], Petru Papazian[2]

Applied Electronics Department, Politehnica University Timisoara, 300223, Romania

ARTICLE INFO

Article history:
Received : 04 March, 2017
Revised : 07 April, 2017
Accepted : 20 April, 2017

Keywords :
Space vector
Digital algorithm
Modulation
Matrix converter
Transfer function
Switching function

ABSTRACT

This paper presents the synthesis of a dynamic space vector modulation for power matrix converters so that it is possible to implement a universal modulator that provides a control algorithm dynamically optimized according to the requirements for the matrix converter.

1. Introduction

In recent years the three-phase to three-phase matrix converter has received considerable attention as an alternative to hard and soft switching dc link converters. This paper is an extension of work originally presented in International Symposium on Electronics and Telecommunications ISETC '16 Twelfth Edition [1]. The SVM (space vector modulation) control strategy for matrix converter offers the best performance, maximal voltage gain and controllable input displacement factor. There are several possible SVM algorithms focused first on reducing energy losses or optimizing the parameters for output voltage and input current [2]. The optimum choice of one of the SVM techniques is a complex task because these requirements are contradictory since reduced power losses mean a reduced sampling frequency leading to a degree of large distortion of waveforms generated by the matrix converter. So for choose correctly a control technique it is required a prior accurate knowledge of the specific conditions which the matrix converter will operate and the characteristics of semiconductor devices which implement the bidirectional switches [3]. Therefore, depending by the operation mode of the power circuit and the conditions imposed by the nature of the load it is useful an implementation of an universal algorithm enabling the transition from an reducing energy losses mode to a very low distortion of the generated waveforms mode and possibility to operation intermediate regimes between them. For this reason it is necessary to implement a digital modulator circuit that allows switching from one SVM algorithm to another without major modifications of the hardware control circuit.

*Corresponding Author: Applied Electronics Department, Politehnica University Timisoara, Romania, Email: adrian.popovici@upt.ro

2. Space Vector Modulation

The matrix converter consists of nine bidirectional switches that connect the three outputs with each of the three inputs, as shown in Figure 1. The converter being supplied by a voltage source, the three input phases can't be short-circuited, and the load having generally a mainly inductive character, the load current can't be interrupted. Because of these constraints the nine switches can generate only 27 permitted combinations. The SVM control algorithm uses 21 of these combinations, a number of 18 combinations, A122- A133; B212-B313 and C221-C331 which connect two output phases at the same input, called active states and a number of 3 combinations Z111-Z333 with all three outputs connected at the same input, called passive states of the matrix converter. To analyze the SVM modulation applied to a three-phase system x1, x2, x3, the space vector X of the system can be used, represented in the complex plan according to the transformation:

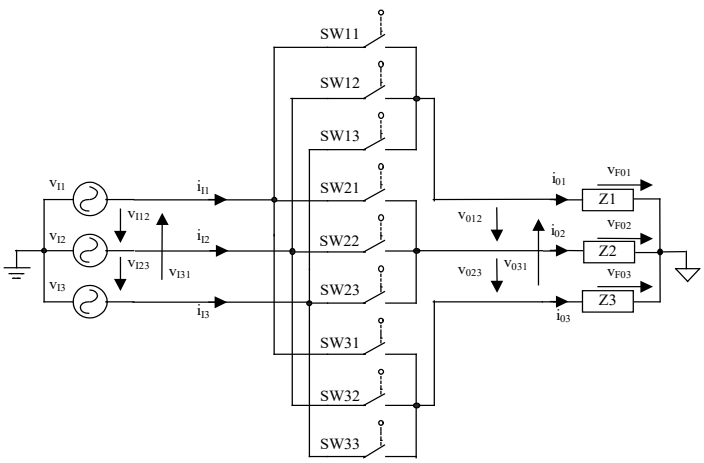


Figure 1: The matrix converter switch layout

$$\underline{X} = Re(\underline{X}) + j\, IM(\underline{X}) = 2/3\,(x_1 + \underline{a}\,x_2 + \underline{a}^2 x_3) \quad (1)$$

where $\underline{a} = e^{j\,2\pi/3}$.

According to this transformation the desired output line voltages space vector is $\underline{V}_{0LW}$ which rotates with the ω_0 angular speed, and the desired input phase currents space vector is $\underline{I}_{IW}$ which rotates with ω_I angular speed as shown in Figure 2 and Figure 3 respectively.

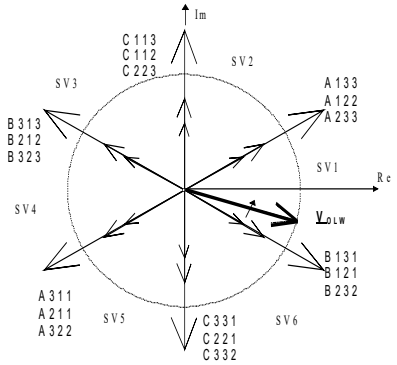


Figure 2: Output voltage SVM vectors diagram

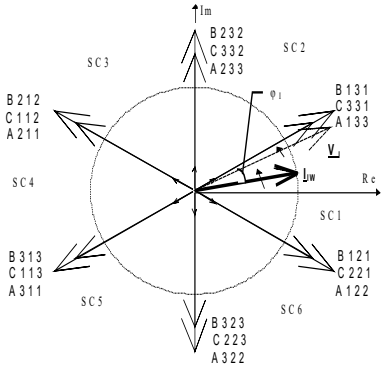


Figure 3: Input current SVM vectors diagram

The circular trajectories of these vectors can be divided in six voltage and current sectors, *SV* and *SC* respectively. If *Ts* time interval is short enough so that the reference vectors $\underline{V}_{OLW}$ and $\underline{I}_{IW}$ do not change significantly, the mean vector space result from different combinations of stationary space vector and null corresponding input voltage line will be equal with the reference vector in this time. Simultaneously it must be performed the desired phase difference between the input current reference space vector and the input voltage space vector, through various combinations of stationary and null space vector corresponding to input currents. The best approximation of spatial reference vector at a time, can be done using vectors that have direction adjacent with the reference space vector [2]. The principle is shown in Figure 4a and Figure 4b, where *a,b* and *m,n* are adjacent directions for output voltage reference space vector and input current reference space vector, respectively. In Figure 4c and Figure 4d it is shown how to choose the vectors for SV1 and SC1 active sectors. In this way from the 21 matrix converter permitted states, in any moment only four active states uniquely established and one of the passive states can simultaneously synthesise the two reference vectors [4]. The order of these active states and the passive state required in a sampling period are chosen so that the number of switches is minimum. Using the trigonometric laws, the normalised time intervals in a sampling period *Ts*, corresponding to the four active states can be established:

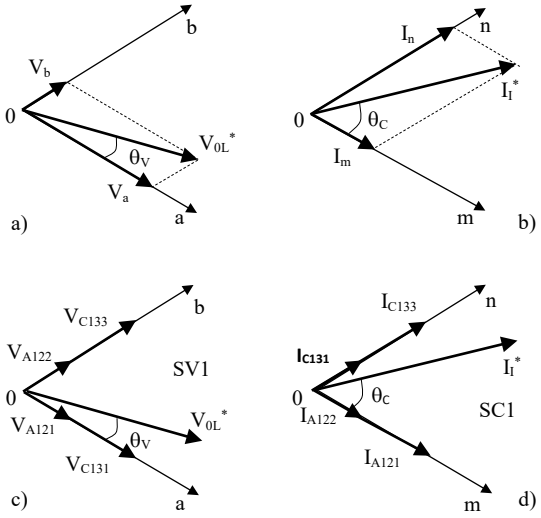


Figure 4: The space vector modulation principle

$$h_{am} = \frac{2}{\sqrt{3}} * (\frac{V_{OL}}{V_{IL}}) * \frac{\sin(\pi/3 - \theta_V) * \sin(\pi/3 - \theta_C)}{\cos(\varphi_I)} \quad (2)$$

$$h_{an} = \frac{2}{\sqrt{3}} * (\frac{V_{OL}}{V_{IL}}) * \frac{\sin(\pi/3 - \theta_V) * \sin(\theta_C)}{\cos(\varphi_I)} \quad (3)$$

$$h_{bm} = \frac{2}{\sqrt{3}} * (\frac{V_{OL}}{V_{IL}}) * \frac{\sin(\theta_V) * \sin(\pi/3 - \theta_C)}{\cos(\varphi_I)} \quad (4)$$

$$h_{bn} = \frac{2}{\sqrt{3}} * (\frac{V_{OL}}{V_{IL}}) * \frac{\sin(\theta_V) * \sin(\theta_C)}{\cos(\varphi_I)} \quad (5)$$

where $h_{am}=T_{am}/T_S$, $h_{an}=T_{an}/T_S$, $h_{bm}=T_{bm}/T_S$, $h_{bn}=T_{bn}/T_S$ an and $1-h_{am}-h_{an}-h_{bm}-h_{bn}=h_z$ [2].

For example, the relation between output and input voltages is

$$\left[v_0^*\right]=\begin{bmatrix} v_{01}^* \\ v_{02}^* \\ v_{03}^* \end{bmatrix}=\begin{bmatrix} 1 & 0 & 0 \\ h_z & h_{am}+h_{bm} & h_{an}+h_{bn} \\ h_{am}+h_{an}+h_z & h_{bm} & h_{bn} \end{bmatrix}\cdot\begin{bmatrix} v_{11} \\ v_{12} \\ v_{13} \end{bmatrix} \quad (6)$$

for *SV1* and *SC1* active sectors. This fact means that the normalised time intervals h_{am}, h_{an}, h_{bm}, h_{bn}, h_z represent the transfer functions for the SVM control strategy.

3. SVM1-SVM3 Versions of Space Vector Modulation

For a three-phase voltage supply for a given three phase output voltages and input phase shift between current and voltage, the association between active SVM and permitted matrix converter states is unique (Table 1). If you know the T_S, then periods T_{am}, T_{bm}, T_{an}, T_{bn}, can be calculated with relationships (2)-(5). In the time interval T_z the matrix converter can pass in any of the passive states. The order in which matrix converter switches from one state to another, within the time T_S, does not affect the local average value of voltages and currents [2],[5]. Basically, choosing between passive states and the order in which the converter switches from one state to another, does not affect the fundamental components of generated waveforms, but it determines the number of commutations required per unit time and the spectral characteristics of generated waveforms. Generally it is intended that the arrangement of the matrix converter states matrix in a sampling period T_S is made such that are necessary as few commutations when passing from one state to another. For example if the active sectors are *SV1* and *SC1* respectively, then from Table 1, the correlation between active states and switching states is *am, bm, an, bn* and *A121, A122, C131, C133* respectively. If a sampling period will use the SVM sequence *am - bm - bn - an -z* , then this corresponds to sequence *A121 -(1)- A122 -(1)- C133 -(2)- C131 -(1)- Z111-(1)-...*, if the z state is associated with Z111state. It is noted that in this case 6 switching are required within the sampling period T_S. If it is kept the same SVM sequence and combination of active sectors is SV1-SC2 then the sequence becomes *C131 -(1)- C133 -(1)- B233 -(1)- B232 -(3)- Z111-(1)-...* which means that required 7switching/T_S. But if the SVM combination is changed as follows *bm - am - an - bn -z* , then we have the next switching sequence: *C133 -(1)- C131 -(2)- B232 -(1)- B233 -(1)- Z333-(1)-...*, if zero state is associated with *Z333* state. It is noted that in this case it is required 6switching/T_S too. Generalizing for all possible combinations it is obtained the Table 2 association which can be easily implemented in a digital control system. Depending on the combination of the active sectors at a time, we define the complementary logical variables *Pa* and *Pb*, as follows:

$$Pa=(SV1+SV3+SV5)(SC1+SC3+SC5)+(SV2+SV4+SV6)(SC2+SC4+SC6) \quad (7)$$

$$Pb=(SV1+SV3+SV5)(SC2+SC4+SC6)+(SV2+SV4+SV6)(SC1+SC3+SC5) \quad (8)$$

Table 1: Matrix converter active states and SV,SC sectors association

Voltage Sector SV	Current Sector SC	Active State am	Active State bm	Active State an	Active State bn
1	1	A121	A122	C131	C133
1	2	C131	C133	B232	B233
1	3	B232	B233	A212	A211
1	4	A212	A211	C313	C311
1	5	C313	C311	B323	B322
1	6	B323	B322	A121	A122
2	1	A122	A112	C133	C113
2	2	C133	C113	B233	B223
2	3	B233	B223	A211	A221
2	4	A211	A221	C311	C331
2	5	C311	C331	B322	B332
2	6	B322	B332	A122	A112
3	1	A112	A212	C113	C313
3	2	C113	C313	B223	B323
3	3	B223	B323	A221	A121
3	4	A221	A121	C331	C131
3	5	C331	C131	B332	B232
3	6	B332	B232	A112	A212
4	1	A212	A211	C313	C311
4	2	C313	C311	B323	B322
4	3	B323	B322	A121	A122
4	4	A121	A122	C131	C133
4	5	C131	C133	B232	B233
4	6	B232	B233	A212	A211
5	1	A211	A221	C311	C331
5	2	C311	C331	B322	B332
5	3	B322	B332	A122	A112
5	4	A122	A112	C133	C113
5	5	C133	C113	B233	B223
5	6	B233	B223	A211	A221
6	1	A221	A121	C331	C131
6	2	C331	C131	B332	B232
6	3	B332	B232	A112	A212
6	4	A112	A212	C113	C313
6	5	C113	C313	B223	B323
6	6	B223	B323	A221	A121

Table2: Sequence SVM states and voltage current sectors association for 6switchings/T_S

Active Sectors	SVM States Sequence
SV1-SC1 SV1-SC3 SV1-SC5 SV2-SC2 SV2-SC4 SV2-SC6 SV3-SC1 SV3-SC3 SV3-SC5 SV4-SC2 SV4-SC4 SV4-SC6 SV5-SC1 SV5-SC3 SV5-SC5 SV6-SC2 SV6-SC4 SV6-SC6	*am - bm - bn - an -z*
SV1-SC2 SV1-SC4 SV1-SC6 SV2-SC1 SV2-SC3 SV2-SC5 SV3-SC2 SV3-SC4 SV3-SC6 SV4-SC1 SV4-SC3 SV4-SC5 SV5-SC2 SV5-SC4 SV5-SC6 SV6-SC1 SV6-SC3 SV6-SC5	*bm - am - an - bn -z*

This SVM version can be changed if the passive state *z* is placed at the beginning of the sampling period *z- am- bm- bn -an* for *Pa =1* and *z- bm- am- an- bn* for *Pa =0* or *z- an- bn- bm- am* and *z- bn- an- am- bm* for *Pa=1* and *Pa=0*, respectively. The combination of the SVM sequences and the matrix converter states can be easily implemented by some memory circuits. The variant thus obtained will be referred to as SVM1. This principle

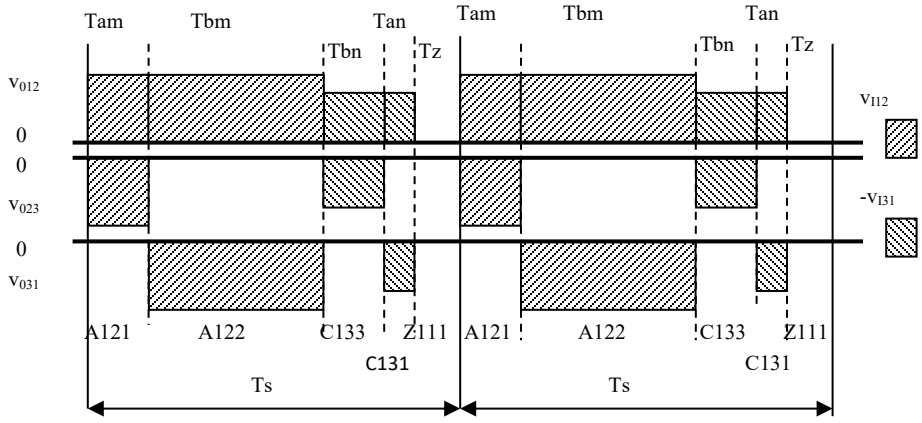


Figure 5: Line output voltage generation in the sampling period for SVM1 algorithm

is shown in Figure 5 for a sampling period. Again 6 switchings/T_S are necessary if for *Pa =1* the SVM sequence is *an- bn- bm- am - z* and for *Pa =0* it is *bn- an- am- bm- z* . Tis s SVM1a algorithm. The SVM1 version can be modified if the zero state is at the beginning of the sequence *z- am – bm –bn – an* for Pa=1 and *z- bm – am – an - bn*. for Pa=0. This is SVM1b algorithm.

By modifying the SVM1 algorithm in the same way we can get the algorithm SVM1c ie *z- an – bn –bm – am* for Pa=1 and *z- bn – an – am - bm*. for Pa=0. Again 6 switchings/T_S are necessary if the passive state is placed between active states*: bm- am- z- an- bn* and *am- bm- z- bn- an* for *Pa =1* and *Pa =0* or *bn- an- z- am- bm* and *an- bn- z- bm- am* for *Pa =1* and *Pa =0* synthesizing another algorithms named SVM1d and SVM1e respectively. A further reduction in the number of commutations can be made by combining certain type of previous algorithms. These new combinations will be named SVM2. A first additional reduction in the number of commutations can be done if for three consecutive sampling periods the SVM sequences will be *SVM1b - SVM1e - SVM1* or *SVM1c - SVM1d - SVM1a*, versions thus obtained are called SVM2a and SVM2b. For example, for SVM2a and *SC1-SV1* active sectors the SVM sequence will be *z- am- bm- bn- an- bn- an- z- am-bm- am- bm- bn- an- z* for *Pa =1*, which means that the number of commutation will be reduced with 5.5%. It can also be reduced the number of switching whether to combine two types of SVM1 sequences for two consecutive sampling periods in sequences *SVM1-SVM1c, SVM1a-SVM1b* and *SVM1d-SVM1e,* synthesizing the control techniques SVM2c, SVM2d and SVM2e, respectively. For example for SVM2c and SV1-SC1 active sectors, if Pa =1 the matrix converter sates will be *A121-(1)-A122-(2)-C133-(1)-C131-(1)-Z111-(0)-Z111-(1)-C131-(1)-C133 -(2)-A122-(1)-A121-(0)-* therefore it means that two consecutive sampling periods requires 10 switching so that the average switching requires 5switchings/T_S. This means a 16% reduction in the necessary number of commutations.

Another direction is optimization techniques to control the dominant harmonics, movement of these to high frequencies, for the same sampling frequency, without increasing excessive the switching losses. This can be achieving using symmetrical waveforms to the center of the sampling period [3]. Variants proposed in the literature accomplishes this balancing around the passive state, which is placed in the center of the sampling period, being framed by the active states that are divided into equal time intervals placed at the beginning and the end of the sampling period. To minimize the number of necessary switching sequences the implementation of SVM sequences proposed in this paper are *am/2- bm/2- bn/2- an/2- z- an/2- bn/2- bm/2- am/2* for *Pa =0* and *bm/2- am/2- an/2- bn/2- z- bn/2- an/2- am/2- bm/2* for *Pa=1*. This version is named SVM3. For this type of algorithms, the converter switches between 9 states during a sampling period, so the minimum number of commutations required is 8switching/T_S. If the SVM sequences are *an/2- bn/2- bm/2- am/2- z- am/2- bm/2- bn/2- an/2* for *Pa =0* and *bn/2- an/2- am/2- bm/2- z- bm/2- am/2- an/2- bn/2 for Pa =1* then are necessary 8switching/TS too. This is named SVM3a algorithm. For SVM3 the output voltage is symmetrical around "big" pulses, while for SVM3a it is symmetrical around "low" pulses for angle θ_C=0-30° in each current sector and vice versa for θ_C =30-60°. În [4] it is suggested that maintaining symmetry around the “big” or “small” voltage pulses regardless of the angle θ_C this can improve the quality of spectral waveforms generated at the converter output matrix. If for θ_C =0-30° will be used SVM3a and for θ_C =30-60° will be used SVM3, then at any time the output voltage waveform will have a shape symmetrical around “small” voltage pulses. This is SVM3b technique. If it is used SVM3 for θ_C =0-30° and SVM3a for θ_C =30-60°we will have symmetry around „large” pulses. This is SVM3c. An improvement of the quality of the frequency spectrum for the waveforms generated by the matrix converter can be obtained by alternating "big" voltage pulses with "small" voltage pulses keeping the symmetry of the output voltages [4]. In order to implement this technique it is necessary to control all switchings between nine states that requires 10 switching/T_S which results in an increase of 66% in the number of switching per unit time. To implement this technique it is necessary to control all switching between nine states. But if in a sampling period is chosen the sequence *bn/2- an/2- z/2- am/2- bm- am/2- z/2- an/2- bn/2* for *Pa =1* and *an/2- bn/2- z/2- bm/2- am- bm/2- z/2- bn/2- an/2* for *Pa =0* or *bm/2- am/2- z/2- an/2- bn- an/2- z/2- am/2- bm/2* and *am/2- bm/2- z/2- bn/2- an- bn/2- z/2- bm/2- am/2*, for *Pa =1* and *Pa =0* respectively then it is possible to reduce the

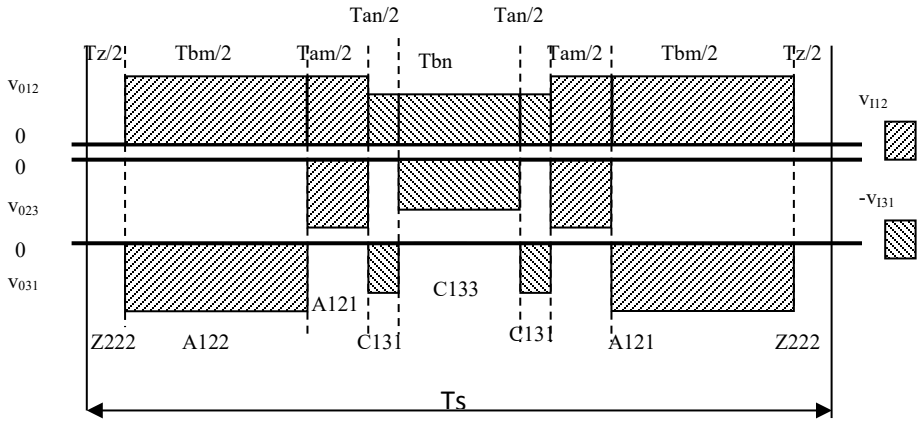


Figure 6: Line output voltage generation in the sampling period for SVM3d algorithm

number of commutation to 8commutations/T_S. This is SVM3d algorithm. If the sequences are *z/2-bn/2- an/2- am/2- bm- am/2- an/2- bn/2- z/2* for Pa=1 and *z/2-an/2- bn/2- bm/2- am- bm/2- bn/2- an/2- z/2* for Pa=0 then we obtain version SVM3e. For angle θ_C=0-30°, SVM3d generates symmetrical waveforms around the „small" voltage pulses like in Figure 6 and SVM3e around "big" voltage pulses, and θ_C =30-60° we have a reversed situation. If for θ_C=0-30° we will use SVM3e and for θ_C =30-60° we will use SVM3d, then at any time the output voltage waveform will be symmetrical around "high" pulses and the algorithm will be named SVM3f. If for θ_C=0-30° we will use SVM3d and for θ_C =30-60° we will use SVM3e, then at any time the output voltage waveform will be symmetrical around "small" pulses and the algorithm will be named SVM3g. If in a sampling period it is chosen the sequence *bn/2- an/2- z/2- am/2- bm- am/2- z/2- an/2- bn/2* for Pa =1 and *an/2- bn/2- z/2- bm/2- am- bm/2- z/2- bn/2- an/2* for Pa =0 we will obtain SVM3h. If for Pa =1 and Pa =0 we will have *bm/2- am/2- z/2- an/2- bn- an/2- z/2- am/2- bm/2* and *am/2- bm/2- z/2- bn/2- an- bn/2- z/2- bm/2- am/2* respectively then we will obtain SVM3i. These algorithms are characterized by 8 switchings/T_S too, the waveforms being symmetrically around the center of tha sampling period. The output voltages consist of alternative "high" and "small" voltage pulses set apart by intervals of "zero" voltage. Combining SVM3i and SVM3h in the same manner we will be obtain SVM3i and SVM3k characterized by waveform that are symmetrical around only the "high" pulses or only around the "small" voltage pulses. A detailed analyses of these algorithms will be presented in a future paper.

4. Implementation of the SVM1-SVM3 Modulator

Whichever you choose to control SVM algorithm, a computer system will determine the value of transfer functions h_{am}, h_{an}, h_{bm}, h_{bn} şi h_z, at T_S time intervals. These transfer functions are then transformed into *S11-S33* switching functions, which controls the ON/OFF states of the 9 bidirectional switches. To convert these transfer functions in PWM switching functions are sufficient four timers TM1-TM4. For this calculus it is necessary to know the modulation index $m=V_{0L}/V_{0L\,max}$ *or* $V_0/V_{0\,max}$, space vector angles θ_C şi θ_V and input phase angle. Analyzing the proposed variants of SVM algorithm previous proposed it is noted that according to the sequence of matrix converter switching states are 12 possible combinations of timer intervals TM1-TM4, according to the timeframes $T_{am}=h_{am}\cdot T_S$, $T_{bm}= h_{bm}\cdot T_S$, $T_{an}= h_{an}\cdot T_S$, $T_{bn}= h_{bn}\cdot T_S$ and $T_z=h_z\cdot T_S$. These combinations, denoted *CTV1-CTV12* are presented in Table 3. For example for *CTV1* combination the logical signals *ST1-ST4* at the outputs of timers will be like in Figure7. The switching functions *S11-S33* associated with switches of the matrix converter are calculated based on the combinations of *CTV, SV, SC* variables and auxiliary variables *Pam, Pbm, Pbn, Pan and Pz* calculated based on the five possible states of the signals *ST1-ST4*. For example switching function formula for *S11* is given by equation (9).

$$S11 = Pam\ ((SC1+SC2)(SV1+SV2+SV3)+(SC4+SC5)(SV4+SV5+SV6))+Pbm\ ((SC1+SC2)(SV1+SV2+SV6)+(SC4+SC5)(SV3+SV4+SV5))+Pbn\ ((SC1+SC6)(SV1+SV2+SV6)+(SC3+SC4)(SV3+SV4+SV5))+Pan\ ((SC1+SC6)(SV1+SV2+SV3)+(SC3+SC4)(SV4+SV5+SV6))+Pz\ (Saz1\cdot(SC1+SC4)+Saz2\cdot(SC3+SC6)+Saz3\cdot(SC2+SC5)) \quad (9)$$

Formula (10) is an example calculation for auxiliary variable *Pam*

$$Pam = ST1\cdot ST2\cdot ST3\cdot ST4\cdot(CTV2\cdot CTV9)+\overline{ST1}\cdot ST2\cdot ST3\cdot ST4\cdot(CTV1\cdot CTV6\cdot CTV10)+\overline{ST1}\cdot\overline{ST2}\cdot ST3\cdot ST4\cdot(CTV3\cdot CTV5)+\overline{ST1}\cdot\overline{ST2}\cdot\overline{ST3}\cdot ST4\cdot(CTV4\cdot CTV7\cdot CTV12)+\overline{ST1}\cdot\overline{ST2}\cdot\overline{ST3}\cdot\overline{ST4}\cdot(CTV8\cdot CTV11) \quad (10)$$

Saz1-Saz3 variables are associated with the passive states *z* of the converter.

Table 3: The time intervals generated by timers

	TM1	TM2	TM3	TM4
CTV1	T_{bm}	$T_{bm}+T_{am}$	$T_{bm}+T_{am}+T_{an}$	$T_{bm}+T_{am}+T_{an}+T_{bn}$
CTV2	T_{am}	$T_{am}+T_{bm}$	$T_{am}+T_{bm}+T_{bn}$	$T_{am}+T_{bm}+T_{bn}+T_{an}$
CTV3	T_{bn}	$T_{bn}+T_{an}$	$T_{bn}+T_{an}+T_{am}$	$T_{bn}+T_{an}+T_{am}+T_{bm}$
CTV4	T_{an}	$T_{an}+T_{bn}$	$T_{an}+T_{bn}+T_{bm}$	$T_{an}+T_{bn}+T_{bm}+T_{am}$
CTV5	T_z	T_z+T_{bm}	$T_z+T_{bm}+T_{am}$	$T_z+T_{bm}+T_{am}+T_{an}$
CTV6	T_z	T_z+T_{am}	$T_z+T_{am}+T_{bm}$	$T_z+T_{am}+T_{bm}+T_{bn}$
CTV7	T_z	T_z+T_{bn}	$T_z+T_{bn}+T_{an}$	$T_z+T_{bn}+T_{an}+T_{am}$
CTV8	T_z	T_z+T_{an}	$T_z+T_{an}+T_{bn}$	$T_z+T_{an}+T_{bn}+T_{bm}$
CTV9	T_{am}	$T_{am}+T_{bm}$	$T_{am}+T_{bm}+T_z$	$T_{am}+T_{bm}+T_z+T_{bn}$

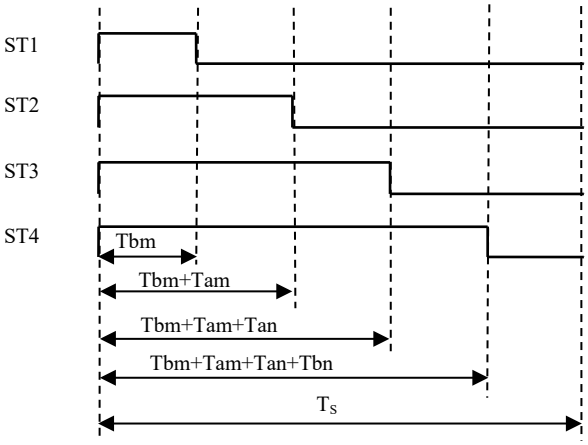


Figure 7: The logical signals at the timers output

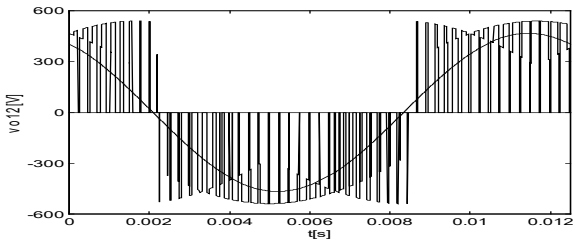


Figure: 8. Output line voltage waveform for SVM1 algorithm

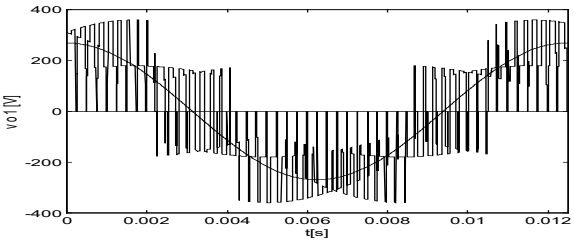


Figure 9: Output phase voltage waveform for SVM1 algorithm

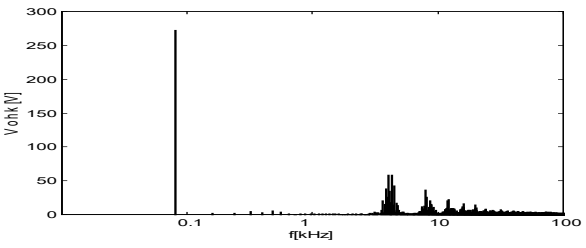


Figure 10: Frequency spectrum of phase output voltage for SVM1

CTV10	T_{bm}	$T_{bm}+T_{am}$	$T_{bm}+T_{am}+T_z$	$T_{bm}+T_{am}+T_z+T_{an}$
CTV11	T_{an}	$T_{an}+T_{bn}$	$T_{an}+T_{bn}+T_z$	$T_{an}+T_{bn}+T_z+T_{bm}$
CTV12	T_{bn}	$T_{bn}+T_{an}$	$T_{bn}+T_{an}+T_z$	$T_{bn}+T_{an}+T_z+T_{am}$

The output voltage waveforms and the frequency spectrum for SVM1 algorithm using S11-S33 switching functions generated with the universal modulator are presented in Figure 8, Figure 9 and Figure 10.

5. Conclusion

In this paper it was presented a dynamic algorithm that enables the implementation of a universal SVM1-SVM3 modulator to control the matrix converters. In this way it is possible to choose different control options that are optimized for reducing energy losses or to control the dominant harmonics. A more detail comparative analysis of these algorithms will be presented in a future work.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] A. Popovici, M. Babaita, P. Papazian "Synthesis of Universal Space Vector Modulator for Matrix Converters", 2016 12th International Symposium on Electronics and Telecommunications (ISETC) October 27-28, Timisoara, pp. 211-214, 2016.

[2] L. Huber, D. Borojevic, "Space vector modulated three-phase to three-phase matrix converter with input power factor correction", IEEE Tran. Ind. Appl., vol.31, no.6, pp. 1234-1246, 1995.

[3] B.K. Bose, "Power Electronics and Variable Frequency Drives - Technology and Applications", IEEE, New York, 1997.

[4] A. Popovici, "Contributii la cercetarea si dezvoltarea convertoarelor matriceale de curent alternativ", Ph.D Thesis, Politehnica University Timisoara, 2002.

[5] S. M. Dabour E.M. Rashad, "Analysis and implementation of space-vector-modulated three-phase matrix converter", IET Power Electronics Journal, Volume 5, Issue 8, pp.1374–1378, September 2012.

[6] A. Popovici, V. Popescu , Simulation and Implementation of Control Strategy for Matrix Converters Using Simulink Software Package, European Power Electronics And Drives Association-PEMC 2004 Riga, Latvia.

[7] A. Popovici, V. Popescu, Simulation and Analyses of Matrix converter circuits, European Power Electronics and Drives Association-PEMC 2004 Riga, Latvia.

[8] A. Popovici, V. Popescu, M. Băbăiță, D. Lascu, D. Negoiţescu, The Analysis Of Input Filter For Matrix Converters, WSEAS Transactions on Signal Processing, Issue 2, Volume 2, ISSN: 1790-5022, pp. 182-189, February 2006.

[9] P. Nielsen, F. Blaabjerg, J. Pedersen; Space vector modulated matrix converter with minimised number of switchings and a feedforward compensation of input voltage unbalance; PEDES'96, vol.2, pg. 833-839; 1996.

[10] A. Popovici, V. Popescu, M. Băbăiță, D. Lascu, D. Negoiţescu, Modeling, Simulation and Design of Input Filter for Matrix Converters, WSEAS International Conference on Dynamical Systems and Control, Venice, Italy, ISBN: 960-8457-37-8, pp 439-444, November 2-4, 2005.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 153-159 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Control design for axial flux permanent magnet synchronous motor which operates above the nominal speed

Xuan Minh Tran*, Nhu Hien Nguyen, Quoc Tuan Duong

Thai Nguyen University of Technology, Vietnam

ARTICLE INFO

Article history:
Received: 22 March, 2017
Accepted: 07 April, 2017
Online: 20 April, 2017

Keywords:
AFPM motor
Optimal control above nominal speed

ABSTRACT

The axial flux permanent magnet synchronous motor (AFPM motor) using magnet bearings instead of ball-bearings at both two shaft ends could allow rotational speed of shaft much greater than nominal speed. One of the solutions to increase motor speed higher than its nameplate speed is reducing rotor's pole magnetic flux of rotor (Ψ_p). This paper proposes a method to boost the speed of AFPM motor above nominal speed by adding a reversed current i_{sd} of (Ψ_p).

1. AFPM Motor

1.1. Structure

In terms of structure, the AFPM motor has its own particular specialists, in details, the stator module may include several types: A single module has one winding set and a dual module has two sets of winding sharing a common core and back-to-back establishment. Similarly, a single rotor module includes only one permanent magnet on one side, and in dual module one, both sides have permanent magnets leaning against each other.

In this research, an object with two single module stator outside and one dual module rotor inside is designated as shown in Figure 1.

1.2. Principle of Operation

When a three-phase voltage is granted to stator coils, different currents are generated (including current i_q) flowing inside, they will interact with the magnetics of rotor to generate torque (M) and the currents in phase windings (component i_d) of stator generate thrust and drag (F) based on the principle of the electromagnet. Thanks to special structure and above-mentioned operating principle, the rotor of the motor will not generate axial displacement although both ends of the shaft have magnetic bearings. It allows the absence of additional axial movement block of the rotor, therefore, the motor structure is being compact. Due to the way of winding roll, the rotational magnetic field generates torques M_1 and M_2 on the same direction on the rotor shaft and generates thrust-drag forces $F1$ and $F2$ between the rotor and the stator on opposite direction. The total torque ($M=M_1+M_2$) is the summation of the torques but the total force is the difference of the axial attractive forces ($F=F_1-F_2$).

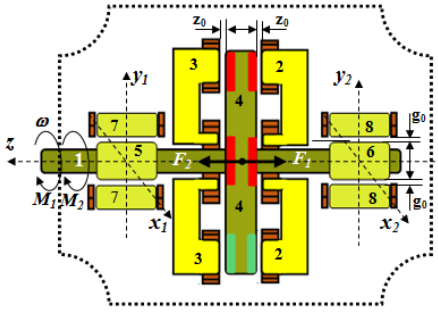


***Figure 1**: The AFPM motor section with magnetic bearing at both ends integrated (1: The shaft; 2, 3: the stator and the winding of left side and right side of motor; 4: the permanent magnetic rotor of motor; 5, 6: the magnetic bearing rotors on left and right sides; 7, 8: the stator and the winding of magnetic bearing on left and right sides; z_0, g_0: the nominal gaps between the rotor and stator of motor with magnetic bearings)*

From the structure and the principle of operation mentioned above, AFPM motor can be considered as two motors that have a common rotor or share a common sharp.

* Xuan Minh Tran, Thai Nguyen University of Technology, Vietnam
E-mail: tranxuanminh@tnut.edu.vn

2. Mathematical Model of the AFPM Motor

According to [1, 2, 3], the mathematical model of AFPM motor was developed in *dq* coordinate system, as presented on (1). The indicator 1 and 2 are present for the left side motor and right side motor, respectively.

$$\left.\begin{aligned}
u_{sd1} &= R_s i_{sd1} + L_{sd1}\frac{di_{sd1}}{dt} - \omega_s L_{sq1} i_{sq1} \\
u_{sq1} &= R_s i_{sq1} + L_{sq1}\frac{di_{sq1}}{dt} + \omega_s L_{sd1} i_{sd1} + \omega_s \psi_p \\
u_{sd2} &= R_s i_{sd2} + L_{sd2}\frac{di_{sd2}}{dt} - \omega_s L_{sq2} i_{sq2} \\
u_{sq2} &= R_s i_{sq2} + L_{sq2}\frac{di_{sq2}}{dt} + \omega_s L_{sd2} i_{sd2} + \omega_s \psi_p \\
m_{M1} &= \frac{3}{2} z_p \left[\psi_p i_{sq1} + i_{sd1} i_{sq1}(L_{sd1} - L_{sq1})\right] \\
m_{M2} &= \frac{3}{2} z_p \left[\psi_p i_{sq2} + i_{sd2} i_{sq2}(L_{sd2} - L_{sq2})\right] \\
m_{\Sigma} &= m_{M1} + m_{M2} = m_m + \frac{J}{z_p}\frac{d\omega}{dt} \\
F_{\Sigma} &= k_1(i_{2d} - i_{1d}) + k_1(i_{2d} - i_{1d})z - k_2 z
\end{aligned}\right\} \quad (1)$$

$$k_1 = 2\frac{\mu_0^2 N^2}{g_0^2}\psi_p; \quad k_2 = 2\frac{\mu_0}{S_p g_0}\psi_p^2.$$

The stucture of AFPM motor as methematical model (1) is presented in Figure 2.

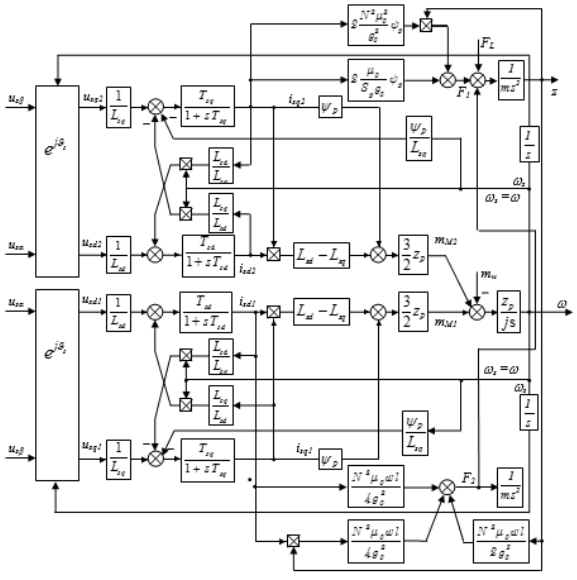


***Figure 2**: Structure diagram of AFPM motor*

3. The voltage, current and magnetic space vectors of APFM motor in two working areas.

The voltage, current, and magnetic space vector of AFPM motor in two working areas are presented in Figure 3 [1, 5, 7]. From (1), it is clearly seen that the torque of AFPM motor includes two parts: the major part with multiplication $\psi_p i_{sq}$ and the reactive part due to the difference of stator inductance (L_{sd}-L_{sq})≠0). In all operating condition, the AFPM motor must produce a sub-torque to compensate the reactive part. The clear existence of the reactive part normally neglected in classical control solutions. Ignoring that component helps to simplify control system and can be accepted in reality within nominal rotational speed, because in that range i_{sd}=0 (Figure 3a). On the contrary, in the speed range above the base speed, in order to speed up, the flux must be reduced by applying a negative current i_{sd}. The motor is now in flux weakening mode, the current is proportional to the rotor's speed (Figure 3b). This leads to the reactive component of the torque is significant and it can't be ignored.

When the motor's speed is adjusted to above base speed, frequency is greater than the nominal frequency $f_{đm}$ but its voltage cannot exceed the nominal voltage $U_{đm}$. The maximum value of voltage is $U_1 = U_{đm}$. The motor's electromotive force expression shows that the flux is inversely proportional to frequency, this is equivalent with the case that the flux of a DC motor is reduced to speed up.

Unlike the synchronous motor with the permanent magnet attached inside the rotor and reluctance synchronous motor, the AFPM motor operating in nominal speed *range* maintains the current i_{sd}=0. Therefore, we only consider the range of above nominal speed.

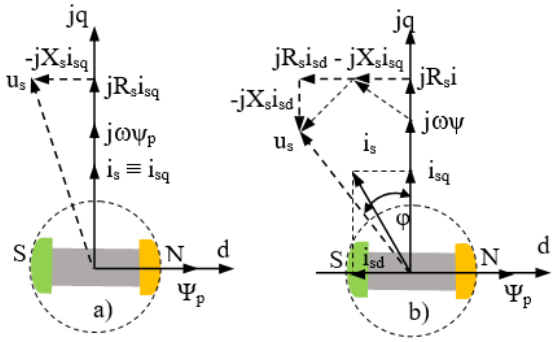


***Figure 3**: State vector of voltage, current and the magnetic flux of AFPM motor in two working areas: a) Under nominal speed; b) Above nominal speed.*

4. Calculating the critical values of voltage and current:

Before introducing a variety of control strategies, we need to consider the limitation of current and voltage [4, 5]. The limitation of current vector:

$$i_s = \sqrt{i_{sd}^2 + i_{sd}^2} < i_{sm} \quad (2)$$

The current limit of (2) is a circular with radius i_{sm} in plane of (i_{sd}, i_{sq})

The voltage limit vector: From the formula of the voltage at steady state neglecting the resistance of stator:

$$u_{sd} = -\omega_s L_{sq} i_{sq}; \; u_{sq} = \omega_s L_{sd} i_{sd} + \omega_s \psi_p$$

The fundamental maximum phase voltage of stator is determined by the DC-link voltage:

$$|u_s| = \sqrt{u_{sd}^2 + u_{sd}^2} < u_{sm} \quad (3)$$

Where: u_{sm} is the maximum DC voltage applied to the inverter.

$$u_s^2 = u_{sd}^2 + u_{sd}^2 = (\omega_s L_{sq} i_{sq})^2 + (\omega_s L_{sd} i_{sd} + \omega_s \psi_p)^2 \quad (4)$$

$$\frac{u_s^2}{\omega_s^2} = (L_{sq} i_{sq})^2 + (L_{sd} i_{sd} + \psi_p)^2 \quad (5)$$

From (3) and (4), ignoring the stator resistances in the steady state, the currents i_{sd}, i_{sq} are defined. Substitute them into (1), the motor torque is obtained:

$$m_M = k_1 \left(\frac{u_s}{\omega_s}\right) cos\varphi - k_2 \left(\frac{u_s}{\omega_s}\right)^2 sin 2\varphi \quad (6)$$

$$k_1 = \frac{3z_p}{2}\psi_p \left(\frac{(L_{sd} - L_{sq})}{L_{sd}} - \frac{1}{L_{sq}}\right); \quad k_2 = \frac{3z_p}{2}\left(\frac{L_{sd} - L_{sq}}{L_{sd} L_{sq}}\right)$$

where: φ is angle between stator voltage vector u_s and the q axis.

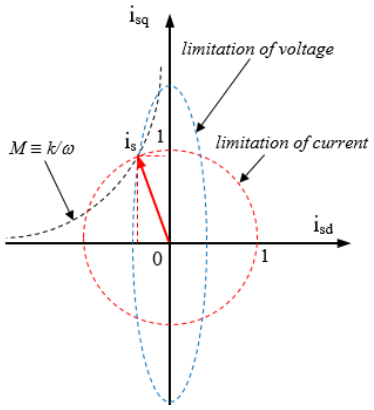


Figure 4: *The limits and the optimal point of torque of the AFPM motor above nominal speed*

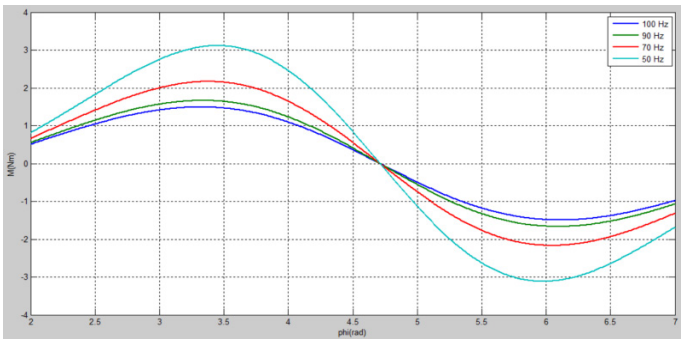


Figure 5*: Simulation result of the motor torque with DC voltage U=400 V;*

Figure 5 shows that if frequency is increased to speed up the motor to above base speed, the flux along with *d* axis will be decreased. However, if φ is maintained at 60° then the motoring torque will be maximum.

5. Control design for AFPM motor

5.1. General control scheme

In term of structure, AFPM motor is completely different from conventional motors; its control scheme includes two different control loops: rotor displacement control loop (attractive force control) and speed control loop (motoring torque control). The general control scheme of AFPM motor is shown in Figure 6 [1, 3].

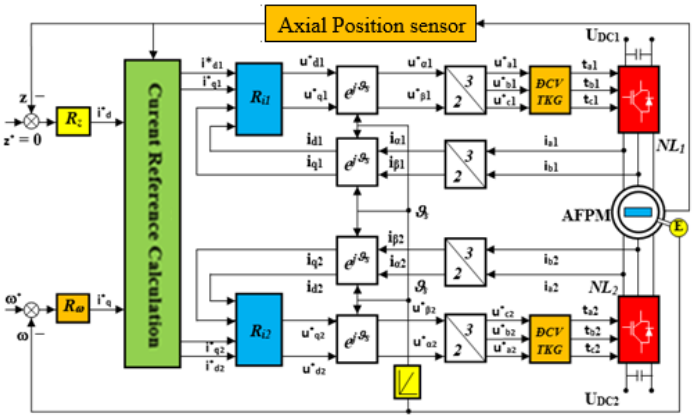


Figure 6: *Vector control structure of AFPM motor*

Vector control of AFPM motor is based on the analysis of the motor's instantaneous currents into two components: the direct component i_d and the quadrature component i_q. The direct component generates the attractive forces and the quadrature component generates the motoring torques. By this way, the control scheme of AFPM motor becomes similar to that of a DC motor.

The rotor displacement along *z* axis from its equilibrium point can be detected by displacement sensor. The detected value is compared to the reference value z^* and the difference is inputted to the R_z controller. The reference z^* is always set to make the rotor in-between the two stators at the equilibrium point. The output of R_z is used to compute the current reference i_d^* . The reference i_{d1}^* and i_{d2}^* of the two stator windings can be calculated by using the compensating current i_{d0}: $i_{d1}^* = i_{d0} + i_d^*$; $i_{d2}^* = i_{d0} - i_d^*$; i_{d0} can be zero or a small value around zero.

The rotor speed detected by the encoder is compared with the speed reference, the difference is fed to the speed controller R_ω. The output of R_ω is used to determine the *q*-axis reference current i_q^*. The *q*-axis reference current of the two stator windings i_{q1}^* and i_{q2}^* is then calculated similar to i_d.

Control design implemented for AFPM motor includes following loops:

5.2. Current control

Most of the modern AC motor drives have a control structure comprising an internal current control loop. Therefore, the drive system characteristic depends on the quality of applied current control strategy [1, 3, 4, 5]

The main task of the current control loop is to make the motor's currents follow the reference signals. By comparing the measured currents with the reference currents, the current control loop generates switching states for the inverter in order to decrease the current errors. In general, the current control loop performs two tasks: error compensation (reduce current error) and modulation (determine switching states).

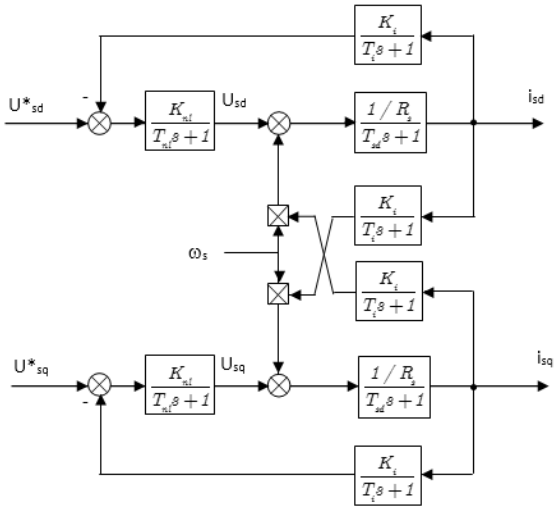


***Figure 7**: The current control loop of AFPM motor*

The current control loop of AFPM motor as mentioned in Figure 7 is extracted from Figure 2 and following formulas:

$$\begin{aligned} i_{sd} &= (u_{sd} + i_{sq}.w_s.L_{sq})W_{sd} \\ i_{sq} &= (u_{sq} - i_{sd}.w_s.L_{sd})W_{sq} \end{aligned} \quad (7)$$

$$\begin{aligned} i_{sd} &= \frac{1}{1 + W_{sd}.W_{sq}.w_s^2.L_{sd}.L_{sq}}(W_{sd}u_{sd} + w_s.L_{sq}.W_{sd}.W_{sq}.u_{sq}) \\ i_{sq} &= \frac{1}{1 + W_{sd}.W_{sq}.w_s^2.L_{sd}.L_{sq}}(-w_s.L_{sd}.W_{sd}.W_{sq}.u_{sd} + W_{sq}u_{sq}) \end{aligned} \quad (8)$$

$$W(s) = \frac{W_i.W_{nl}}{1 + W_{sd}.W_{sq}.\omega_s^2.L_{sd}.L_{sq}}\begin{bmatrix} W_{sd} & \omega_s.L_{sq}.W_{sd}.W_{sq} \\ -\omega_s.L_{sd}.W_{sd}.W_{sq} & W_{sq} \end{bmatrix} \quad (9)$$

Where: $W_{sd} = \frac{1/R_s}{T_{sd}s + 1}$; $W_{sq} = \frac{1/R_s}{T_{sq}s + 1}$;

$$W_{nl} = \frac{K_{nl}}{T_{nl}s + 1}; W_i = \frac{K_i}{T_i s + 1}.$$

The formulas (7), (8), (9) show that interactive actions exist in current control loop. To obtain desire quality of the closed system, the modulus optimum criterion is applied:

$$W_m(s) = \frac{1}{1 + 2T_s s + 2_{sd}.T_s^2 s^2} \quad (10)$$

Before applying the modulus optimum criterion, the current control loop must be decoupled.

Decoupling regulators are defined as follows:

$$W_R(s) = W_m(s)(1 - W_m(s))^{-1}.W(s) \quad (11)$$

$$T_s = T_{si} = 2T_i = 0,002 \quad .$$

Hence, T_σ^2 can be ignored. After that, transfer functions of the decoupling regulators for the current control loop can be determined.

$$\begin{aligned} W_R(s) &= \frac{1}{4T_i.K_{nl}K_i}\begin{bmatrix} L_{sd}(1 + \frac{1}{T_{sd}s}) & \frac{-\omega_s.L_{sq}}{s} \\ \frac{\omega_s L_{sd}}{s} & L_{sq}(1 + \frac{1}{T_{sq}s}) \end{bmatrix} \\ &= \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix} \end{aligned}$$

$$\left.\begin{aligned} R_{11}(s) &= \frac{1}{4T_i.K_{nl}K_i}L_{sd}(1 + \frac{1}{T_{sd}s}) \\ R_{12}(s) &= \frac{1}{4T_i.K_{nl}K_i}\frac{\omega_s L_{sd}}{s} \\ R_{21}(s) &= \frac{1}{4T_i.K_{nl}K_i}\frac{-\omega_s L_{sd}}{s} \\ R_{22}(s) &= \frac{1}{4T_i.K_{nl}K_i}L_{sq}(1 + \frac{1}{T_{sq}s}) \end{aligned}\right\} \quad (13)$$

Insert the decoupling regulators into the current loop diagram in Figure 7 we have the decoupled diagram as shown in Figure 8.

5.3 Axial displacement control

For simplicity, assume that the radial movement of the rotor is supported by two ideal radial magnetic bearings. Thus, the axial displacement and the radial displacement are independent, and can be expressed as follows [1, 2, 3, 4]:

$$F - F_L = m\ddot{z} \quad (14)$$

where m is the weight of the motion part and F is the axial force.

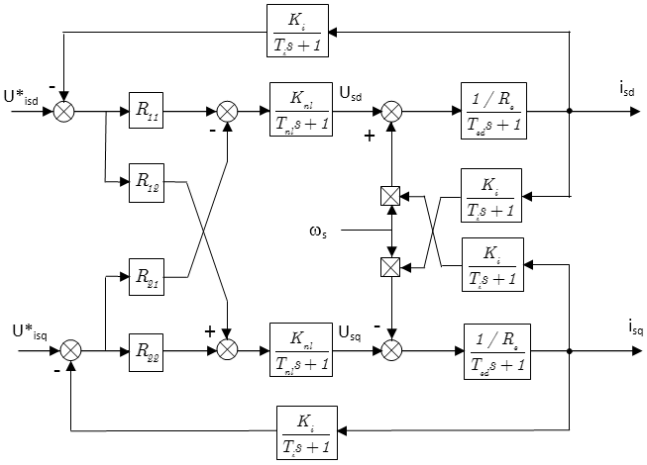


***Figure 8**: Decoupled current control loop*

The equation (15) is obtained by substituting (1) into (14):

$$m\ddot{z} + F_L = k_1(i_{2d} - i_{1d}) + k_1(i_{2d} - i_{1d})z - k_2 z \quad (15)$$

$$k_1 = 2\frac{\mu_0^2 N^2}{g_0^2}\psi_p; \; k_2 = 2\frac{\mu_0}{S_p g_0}\psi_p^2.$$

It is easy to see that this system is unstable. To make the system stable, a controller with derivative component is used. The axial displacement control loop is shown in Figure 9.

The axial displacement control loop contains the transfer function of the current control loop. Since the torque load is usually undefined, so it is regarded as an external interference.

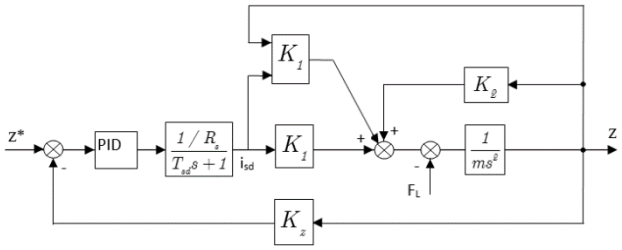


***Figure 9**: Axial displacement control loop*

To eliminate the steady state error, a PID controller is used.

$$G_{cz}(s) = K_P + \frac{K_I}{s} + K_D s \qquad (16)$$

The system will be stable if the controller parameters satisfy the following conditions:

$$\begin{cases} K_P > \dfrac{K_2}{K_1 K_z} \\ K_I < \dfrac{K_D\left(K_1 K_P + K_2\right)}{m} \\ K_I > 0 \\ K_D > 0 \end{cases} \qquad (17)$$

4.4 Speed control

For all types of electric motor, the difference between electromagnetic torque M and torque load M_L causes the acceleration of the rotor according to the mechanical characteristic of the motor drive. The rotational motion equation can be written as follows [1, 3, 4, 5]:

$$M - M_L = J\frac{d\omega}{dt} \qquad \frac{\omega}{M - M_L} = \frac{1}{Js} \qquad (18)$$

M can be controlled by the q-axis current i_q as shown in (1); thus, the speed control loop is depicted in Figure 10.

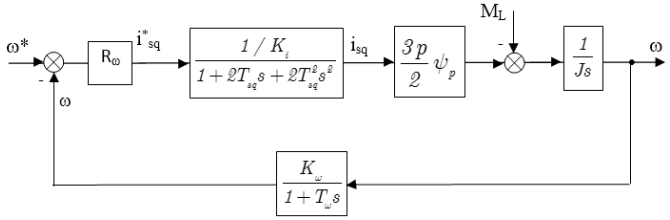


***Figure 10**: Speed control loop*

Similar to the axial displacement control loop, the speed control loop also includes an inner current control loop and a fixed motor transfer function. Due to the uncertain load, in the first step M_L is regarded as an external disturbance of the system. The influence of the speed measurement is usually included in the equivalent time constant of the current control loop.

Hence, the closed-loop transfer function can be expressed as follow:

$$W_K(s) = \frac{R_\omega W_0}{1 + R_\omega W_0} = \frac{KR_\omega}{s(T_s's+1) + KR_\omega} \qquad (19)$$

Apply the symmetrical optimum criterion for the speed control loop:

$$W_C(s) = \frac{1+4T_s s}{1+4T_s s+8T_s^2 s^2+8T_s^3 s^3}$$

Equalizing (19) and (20), we have:

$$\frac{KR_\omega}{s(T_s's+1)+KR_\omega} = \frac{1+4T_s s}{1+4T_s s+8T_s^2 s^2+8T_s^3 s^3}$$

Choose: $2T'_s = T_s$, solve the above equation we have:

$$R_\omega = \frac{1+4T_s s}{K8T_s^2 s} = \frac{1}{8KT_s^2 s} + \frac{1}{2KT_s}$$

6. Quality evaluation by numerical simulation

6.1 Specifications of the motor and simulation parameters

$R_s = 2{,}3\ \Omega$; $L_{sq} = 9{,}6.10^{-6}$ H; $L_{sd} = 8{,}2.10^{-6}$ H;

$T_{sq} = 4{,}2.10^{-6}$; $T_{sd} = 3{,}56.10^{-6}$; $\Psi_p = 0{,}0126$Wb;

$Z_p = 1$; $g_0 = 1{,}7.10^{-3}$ m; $m_{rotor} = 0{,}235$kg;

$J_r = 0{,}0000082$ kgm^2; $\mu_0 = 4\pi$ x 10^{-7} H/m;

Knl = 5; $K_i = 1$; $T_i = 0{,}001$; $K_\omega = 0{,}00417$;

$T_\omega = 0{,}1$; $T_{s\omega} = T_\omega + 2T_i$; $2T'_s = T_{s\omega}$.

6.2 The simulation diagram of AFPM motor

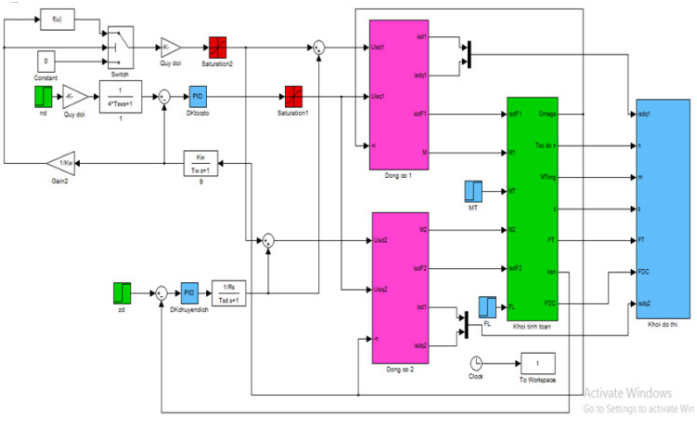


***Figure 11:** Simulink model of AFPM motor*

6.3. Simulation results

+ When the motor operates at the nominal speed: $n = 3000$ rpm; $m = 0{,}08$ Nm; displacement $z = 0$

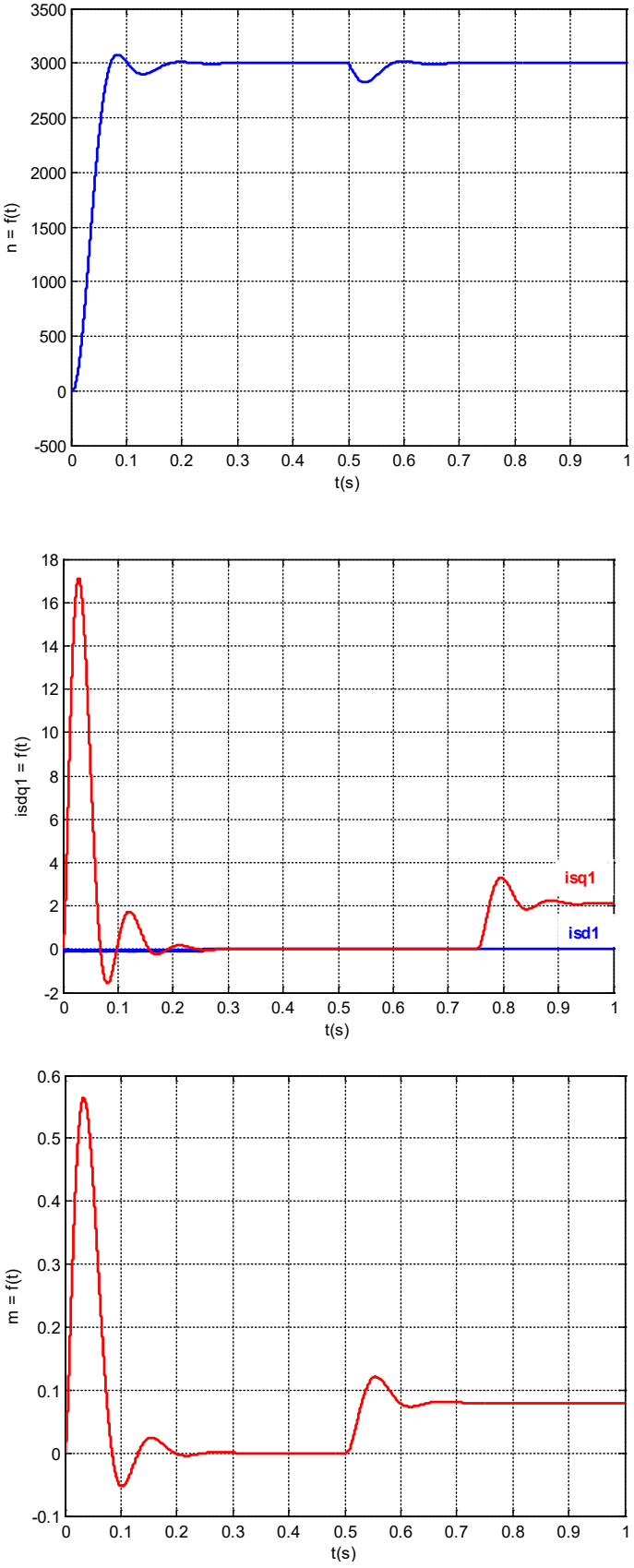


***Figure 12**: Characteristics of AFPM motor when it operates at the nominal speed and no axial force: n = 3200 rpm, displacement z = 0; the torque m (has the same shape as the current i_{sq}); i_{sd} = 0*

+ When the motor operates above the nominal speed: n = 4000rpm; m = 0,08Nm; $z \neq 0$

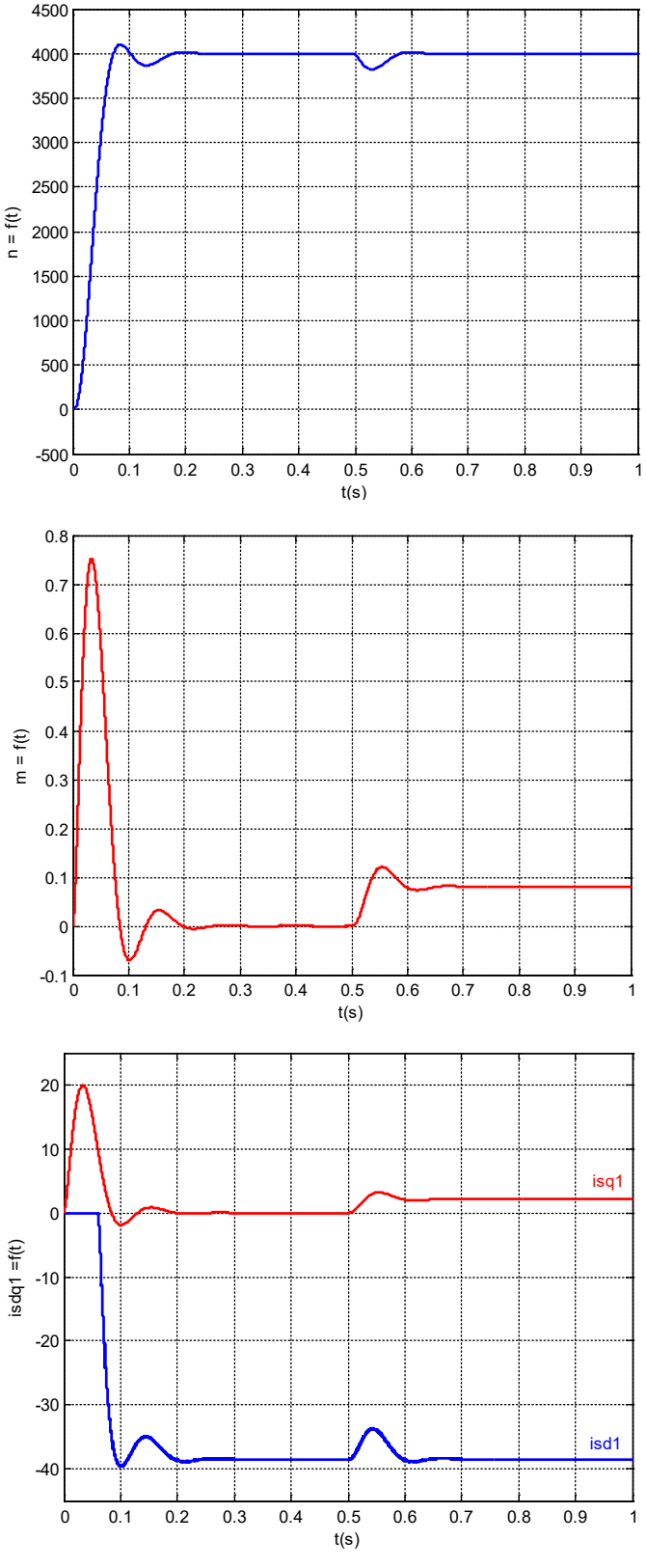


***Figure 13**: Characteristics of AFPM motor when it operates above the nominal speed: n = 4000 rpm, z ≠ 0; the torque m (has the same shape as the current i_{sq}); i_{sd} has large negative value.*

7. Conclusions

The AFPM motor with special structure is used in combination with magnetic bearings to form a system which includes two control loops: the speed control loop and the rotor displacement control loop (with assumtion that the magnetic bearings fullfil their nominal functions). By using rotor flux oriented control method, some results are firstly achieved:

- Speed control above the nominal speed is attained by flux weakening thank to apply a current i_{sd} opposite in direction with the flux ψ_p while sustaining the torque.

- The rotor is always kept at the center of the motor by the rotor axial displacement control loop.

- Further research need to be implemented in order to improve the control quality and sustain the optimal torque when the flux is weakened.

- Experimental study on the AFPM motor.

References

[1] Nguyen Phung Quang and Jörg-Andreas Dittrich, "Vector Control of Three-Phase AC Machines", springer

[2] Akira Chiba, adashi Fukao, Osamu Ichikawa, Masahide Oshima, asatsugu Takemoto and David G. Dorrell, "Magnetic Bearings and Bearingless Drives", Newnes, 2005.

[3] Quang Dich Nguyen and Satoshi Ueno, "Analysis and Control of Non-Salient Permanent Magnet Axial-Gap Self-Bearing Motor", IEEE Transactions on Industrial Electronics, Vol. PP, No. 99, pp. 1-8, 2010 (early access).

[4] Nguyen Van Lien, Nguyen Manh Tien, Doan Quang Vinh, "Control of AC motor with magnetic from semiconductor inverter" NXB KHKT, 2005.

[5] Nguyen Duc Quan,"Modeling, simulation, and control reluctance high speed motor, Journal of Science and Technology – Da Nang University – No 11(96).2015 Book 2.

[6] Nguyen Doan Phuoc: "Nonlinear control theory". NXB KH&KT, 2007.

[7] Trong Duy Nguyen, Gilbert Foo Hock Beng, King-Jet Tseng, Don Mahinda Vilathgamuwa, and Xinan Zhang: "Modeling and Position-Sensorless Control of a Dual-Airgap Axial Flux Permanent Magnet Machine for Flywheel Energy Storage Systems", Journal of Power Electronics, Vol. 12, No. 5, September 2012.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 2, No. 3, 160-170 (2017)
www.astesj.com
Special Issue on Recent Advances in Engineering Systems

ASTES Journal
ISSN: 2415-6698

A synchronizing second order sliding mode control applied to decentralized time delayed multi–agent robotic systems: Stability Proof

Marwa Fathallah*, Fatma Abdelhedi, Nabil Derbel

University of Sfax, Electrical Engineer, Sfax Engineering School, 3000, *Tunisia*

ARTICLEINFO

Article history:
Received: 04 March, 2017
Accepted: 07 April, 2017
On-line: 20 April, 2017

Keywords :
Sliding Mode Control (SMC)
Synchronization
multiple robot manipulators
Cross Coupling
time delay
stability

ABSTRACT

This study investigates the synchronization issue of multiple robot manipulators in the presence of time delay. Since several previous works based on synchronization task neglect the communication delay, in this paper we develop a finite time stability based on a Lyaponov functional for synchronization of a networked robotic system where time delay exists during the communication between robots. To this effect, we consider a second order sliding mode control (SMC) combined with the cross coupling concept in order to ensure the position synchronization of networked robot manipulators. Furthermore, the stability of the proposed controller with communication's delay has been proved. Simulation results illustrate satisfactory performances which prove the efficiency of the proposed approach.

1 Introduction

Over the past few decades, an increasing interest has been noticed on the interconnected systems in several fields of research [1, 2]. In fact, cooperative and coordinated control have attracted several research communities such as: biology [3]; artificial intelligence [4]; wireless sensor networks[5, 6]; control of mobile robot[7, 8]; spacecraft [9, 10]. The synchronized control of robot manipulators has been firstly presented in [11, 12]. Subsequently, further research results on the synchronization of robot manipulators have been published [13, 14]. It is recognized that the existing synchronization works as aforementioned are all for motion control. In this context, motion synchronization of multi agent systems has attracted much attention in various applications including the industrial assembling, automatic control of multi agent systems such as the control of robot manipulators [15, 16].

As an illustration, in [17] position synchronization of multiple motion axes has been studied. Motion synchronization has also been used in more complex mechanical systems especially the nonlinear robot systems. Furthermore, where flexibility and maneuverability are highly recommended [13, 18, 19], with the increasing complexity of evolved and specific applications, manipulability can't be fulfilled by a simple robot. For this reason, the use of cooperative schemes for multiple robots can present a better solution to realize more robust multi agent system controls, where each robots operates cooperatively, and receives feedbacks from each others to achieve a consolidated goal [11, 20].

Furthermore, most of the real systems are known by nonlinearities such as robotic field. For this reason, the formulation for robust control laws is required,in the sense that it is able to ensure the system stability and the robustness via external disturbances and parameters variations [21, 22]. Otherwise, several control methods have been used to synchronize various complex systems such as: adaptative control [23, 24]; sliding mode control(SMC) [25, 26, 27]; neural networks [28] etc.

Among several dynamic behaviors, the synchronized motion control and the stability of complex nonlinear systems are considered between the most important research topics during several years [29, 30]. The mutual synchronization of robotic systems without time delay has been absolutely studied by several communities [11, 13]. Nevertheless, and due to the importance of cooperative research using multiple agents and within its wide range applications, the control algorithm suffers from some disturbing factors that can't be neglected such as time delays communication, communication interruption, packet collisions etc. Time delay communication is commonly

*Marwa Fathallah, Sfax–Tunisia, fathallahmaroua@gmail.com

https://dx.doi.org/10.25046/aj020321

known in biological and physical networks, owing to the finite velocity of communication as well as traffic congestions [31, 32, 33]. In multi–agent robotic systems, the time delay is only considered at the level of the information arriving to a robot and coming from sensors of its neighbors, while information from its own sensors is processed immediately.

The presence of this communication constraint makes the dynamical behaviours more complicated, can lead to undesirable transient response, reduces the performances of the networked systems or even the instability of the system [33, 34, 35]. Referring to the cross-coupling technique, several works have been suggested to improve the performances of the synchronization of multi axis motions [36, 37]. Later on, [38] proposed a control algorithm in order to synchronize networked systems in the presence of time delays [39, 40]. In the light of what was said, we present in this paper a synchronizing SMC algorithm for the control of a time delayed multi agent system in order to prove the effectiveness and the stability of the proposed approach.

1.1 Contribution

The subject of this paper is to realize the motion control of complex networked robotic systems in the presence of time delays.
Since the robot requires an interaction with its environment to achieve its goal, and in the absence of a suitable sensor, the robot remains blind. For this reason, and in order to reproduce human capacities for perception and action in robotic systems, researchers adopt the integration of data from a surveillance camera.
This camera is a great way to provide security to the target location. Nowadays, the surveillance cameras can also be set to be motion activated, recording footage when motion triggers them. Many range of cameras also includes outdoor security cameras and wifi cameras in order to effectively control the monitoring task.
Therefore, a second order sliding mode strategy has been considered and exhibited on a 3 degrees of freedom($3DOF$) surveillance camera system, where we focus on the manipulative arm managing the camera movements.
Then, to guarantee the overall vision of the proposed framework, we combine a muli robot manipulators, where each robot must synchronize its movement with other teammates using the cross-coupling approach .
The main goal of this work is to realize a common and performant motion control task of multi-agent robot manipulators based on the cross coupled second order sliding mode approach design, by reducing the chattering impact, and achieving robust communication between agents which make the system stronger against disturbances, uncertainties, breakdowns and also able to compensate the existence of communication delay.

2 Second Order Sliding mode controller

2.1 Preliminaries

Sliding mode control is a robust nonlinear strategy [41, 46, 47]. Such Variable Structure Control (VSC) is considered as a discontinuous feedback approach where its design is divided into two parts: the reaching phase (system trajectories are forced to reach a specific surface in the state space then to remain on it) and the sliding phase (Figure 1).
The Sliding Mode approach is developed using the Lyaponov Theory in order to ensure the convergence to the sliding surface ($s(x) = 0$):

$$S^T \dot{S} < 0 \tag{1}$$

where the sliding surface S is chosen as:

$$S(x) = Y(x - x_d) \tag{2}$$

in which: x_d is the desired trajectory and Y is a matrix chosen such away x acheives x_d.
The structure of the proposed controller is composed of two terms:

$$u = u_{eq} + \Delta u \tag{3}$$

where u_{eq} is the equivalent control which ensure the "reaching phase" and Δu is the corrective term used to avoid all deviations from the sliding surface.
The expression of the equivalent term can be deduced from the following equation:

$$\dot{S} = F(x) + G(x)u = 0 \quad \mapsto \quad u_{eq} = -[G(x)]^{-1}F(x) \tag{4}$$

where $F(x)$ and $G(x)$ are defined by the affine state equation of a nonlinear system ($\dot{x} = f(x) + g(x)u$).
Moreover, the corrective term can be described as follows:

$$\Delta u = -[G(x)]^{-1} W \text{sign}(S) \tag{5}$$

where W is a definite positive matrix.

2.2 Second Order Sliding Mode Control

As it is mentioned above, this robust approach suffers from the undesirable chattering phenomenon [14, 34] induced by the corrective term Δu, whose impact is manifested by the existence of perturbing high switching frequencies in the control inputs [21, 42]. More precisely, this problem involves fast and sudden changing control signals which lead to low accuracy and even damage the mechanical parts. Therefore, several methods have been developed in order to overcome this annoying phenomenon [22].
In this context, the second order sliding mode approach presents an enhancement of the classical SMC by the introduction of a filtering action in the controller.
Such an action greatly reduces the major drawback of

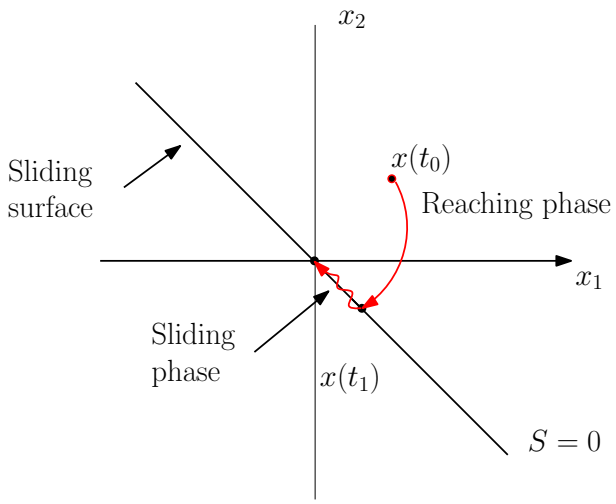


Figure 1: Phases of Sliding Mode Control

the simple SMC which is the chattering phenomenon. Hence, the second order SMC has been considered as the most useful among the high order SMC thanks to its relative simplicity of application, compared to the higher order controls [43, 44, 45].
Then, the sliding surface derivative of the second order SMC approach has been modified as follows :

$$\dot{s_i} = \sigma_i \quad \mapsto \quad \Delta u_i = [Wg(x)]^{-1}\sigma_i \tag{6}$$

The new description of the dynamic control behavior can be written as follows:

$$\begin{cases} \dot{s_i} &= \sigma_i \\ \dot{\sigma_i} &= -a_0\, s_i - a_1\, \sigma_i + v_i \end{cases} \tag{7}$$

where a_0, a_1 are positive scalars and v_i is a variable control of SMC. This study presents multi-input-multi-output systems. Furthermore, the expression of $\dot{\sigma_i}$ can be deduced from the equality $A(p) = 0$ of the following Hurwitz polynomial (which its roots have negative real parts) :

$$A(p) = (p + \mu)^2 \tag{8}$$

where μ is a positive scalar.
In order to ensure the stability, the representation (7) can be reformulated as the condensed form:

$$\dot{Z}_i = \phi Z_i + \Gamma v_i \tag{9}$$

in which:

$$\underbrace{\begin{pmatrix} \dot{s_i} \\ \dot{\sigma_i} \end{pmatrix}}_{\dot{Z}_i} = \underbrace{\begin{pmatrix} 0 & I \\ -\mu^2 I & -2\mu I \end{pmatrix}}_{\phi} \begin{pmatrix} S_i \\ \sigma_i \end{pmatrix} + \underbrace{\begin{pmatrix} 0 \\ 1 \end{pmatrix}}_{\Gamma} v_i \tag{10}$$

where 0 is the null matrix, I is the identity matrix and the discontinuous term v_i is given by .

$$v_i = -Q \text{ sign } (\Gamma^T L S_i) \tag{11}$$

where $Q = [q_1, q_2, ..., q_n]$ and L are positive definite matrix.

3 Mutual SMC synchronization algorithm

3.1 Mathematical model

A robotic manipulator arm designed to be equipped with a surveillance camera system presents the adopted dynamic model of this study (Figure 2).

Using the Lagrangian formulation, the motion equation of a manipulator robot "i" can be written as: [46]

$$M_i(q_i)\ddot{q_i} + C_i(q_i, \dot{q_i})\dot{q_i} + G_i(q_i) = \tau_i \tag{12}$$

where:

- $q_i(t) \in \mathbb{R}^n$ is the measured articulation vector of the manipulator (joint position),
- $\dot{q_i} \in \mathbb{R}^n$ is the velocity vector,
- $\ddot{q_i} \in \mathbb{R}^n$ is the joint acceleration vector,
- $M_i(q_i) \in \mathbb{R}^{n \times n}$ is the symmetric uniformly bounded and positive definite inertia matrix,
- $C_i(q_i, \dot{q_i})\dot{q_i} \in \mathbb{R}^n$ represents the vector expressing Coriolis and centrifugal forces,
- $G_i(q_i) \in \mathbb{R}^n$ is the vector of gravitational torques,
- $u = \tau_i \in \mathbb{R}^n$ denotes the control torque.

3.2 Cross Coupling technique

In this work, we take into consideration the synchronization of multiple robot manipulators. In this context, we propose decentralized control laws for n robots manipulators for which each robot synchronizes its position with the other neighbor agents and track the same desired trajectory. Specifically, using the synchronization approach, manipulators are controlled in a synchronous manner so that the tracking errors and the synchronization errors converge

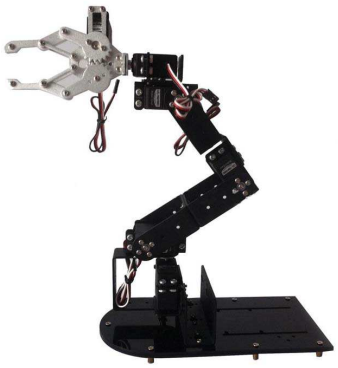


Figure 2: Example of 3*DOF* robot manipulator

to zero where the synchronization error is considered to be a differential position between coordinated agents. In order to achieve a coordinated control motion, an effective synchronization method namely the Cross Coupling concept is carried out where the whole multi-robot manipulators is used as a unique generalized system [47, 48, 49].
The cross coupling design was firstly introduced by [2], then its concept has been used mainly for machine tools [50, 51]. Later on, it has been applied in the robotic fields by [52]. The main idea of such a procedure is to create a global error of the model [36, 53].
In the light of what was said, the tracking error can be written as follows:

$$\eta_i(t) = q_i(t) - q_d(t) \tag{13}$$

where $q_d(t) \in \mathbb{R}^n$ denotes the desired position.
The vector η_i will give insight on the joint positions convergence to the desired trajectory. The proposed cross coupling concept suggests a suitable synchronization error which is defined as follows:

$$\begin{aligned}\xi_i(t) &= \textstyle\sum_{j\neq i}^{p} \Lambda_{ij}[q_i(t) - q_j(t-\tau)] \\ &= \textstyle\sum \Lambda_{ij}(q_i(t) - q_d(t)) - \sum \Lambda_{ij}(q_j(t-\tau) - q_d(t-\tau)) \\ &\quad + \textstyle\sum \Lambda_{ij}(q_d(t) - q_d(t-\tau)) \\ &= \textstyle\sum \Lambda_{ij}\eta_i - \sum \Lambda_{ij}\eta_j(t-\tau) + \sum \Lambda_{ij}[q_d(t) - q_d(t-\tau)]\end{aligned} \tag{14}$$

where Λ_{ij} is a symetric positive definite matrix which reveals an idea about the communication quality between the ith and j^{th} agents.
Thus, in order to attain a robust controller for multi robot systems, and to ensure a synchronous trajectory tracking in the presence of communication data delay, we define the following global error expression (for robot i):

$$\varepsilon_i = \eta_i + \int_{t_0}^{t} \xi_i(\alpha) d\alpha \tag{15}$$

where we note that this error expression includes both synchronization error and trajectory tracking error defined above.
Then, its derivative can be considered as follows:

$$\dot{\varepsilon}_i = \dot{\eta}_i + \xi_i \tag{16}$$

Let's define the sliding mode surface :

$$s_i = \dot{\varepsilon}_i + \lambda_i \varepsilon_i \tag{17}$$

where $\lambda_i > 0$. Then, we obtain:

$$\dot{\eta}_i = -\xi_i - \lambda_i \varepsilon_i + s_i \tag{18}$$

Having in mind the expression of (14), we obtain the following expression:

$$\begin{aligned}\dot{\eta}_i = &-\textstyle\sum \Lambda_{ij}\eta_i + \sum \Lambda_{ij}\eta_i(t-\tau) \\ &-\textstyle\sum \Lambda_{ij}[q_d(t) - q_d(t-\tau)] - \lambda_i \varepsilon_i + s_i\end{aligned} \tag{19}$$

In order to simplify the previous expression, we define:
$d_i = \sum \Lambda_{ij}[q_d(t) - q_d(t-\tau)]$.
Thus:

$$\dot{\eta} = A\eta + B\eta(t-\tau) + d - \Lambda\varepsilon + S \tag{20}$$

where:

$$A = \begin{pmatrix} -\sum \Lambda_{1j} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & -\sum \Lambda_{nj} \end{pmatrix}, \quad \eta = \begin{bmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_n \end{bmatrix}$$

$$B = \begin{pmatrix} 0 & \sum \Lambda_{12} \cdots & \sum \Lambda_{1n} \\ \sum \Lambda_{21} & \sum \Lambda_{23} \cdots & \sum \Lambda_{2n} \\ \vdots & \ddots & \vdots \\ \sum \Lambda_{n1} & \cdots & 0 \end{pmatrix}, \; S = \begin{pmatrix} s_1 \\ s_2 \\ \vdots \\ s_n \end{pmatrix}$$

$$d = \begin{pmatrix} d_1 \\ d_2 \\ \vdots \\ d_n \end{pmatrix}, \; \Lambda = \begin{pmatrix} \lambda_1 I & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n I \end{pmatrix}$$

The equivalent control expression can be deduced from the equality $\dot{s}_i = 0$ then, we obtain:

$$u_{eq_i} = M_i[\ddot{q_d} - \dot{\xi}_i + \lambda_i(\xi_i + \dot{\eta}_i)] + C_i(q_i, \dot{q}_i)\dot{q}_i + g_i(q_i) \tag{21}$$

3.3 Theorem

The control law:

$$u = u_{eq_i} + \Delta u \tag{22}$$

where: u_{eq_i} is expressed in equation (21) and Δu is given by:

$$\Delta u = NR^{-1}[N\varepsilon - P\eta - \Lambda S - KsignS] \tag{23}$$

with N is a diagonal, symmetric definite positive matrix which stabilities the proposed system composed by n manipulators.
Remembering that for the Second Order Sliding Mode Control, the sliding surface becomes $\dot{s}_i = \sigma_i$ and taken into consideration equation (17), we obtain:

$$\ddot{q}_i - \ddot{q}_d + \sum \Lambda_{ij}(\dot{q}_i - \dot{q}_j) + \lambda_i[\dot{q}_i - \dot{q}_d + \sum \Lambda_{ij}(q_i - q_j)] = \sigma_i \tag{24}$$

Retaking the mathematical model equation (12) and substituting it in the previous equation gives:

$$\begin{gathered}\sigma_i = M_i^{-1}[\tau_i - C_i(q_i,\dot{q}_i)\dot{q}_i - G_i(q_i)] \\ +\sum \Lambda_{ij}(\dot{q}_i - \dot{q}_j) + \lambda_i[\dot{q}_i - \dot{q}_d + \sum \Lambda_{ij}(q_i - q_j)] - \ddot{q}_d\end{gathered}$$

And consequently,:

$$\begin{gathered}\tau_i = C_i(q_i,\dot{q}_i)\dot{q}_i + G_i(q_i) + M_i[\ddot{q}_d - \sum \Lambda_{ij}(\dot{q}_i - \dot{q}_j) \\ -\lambda_i[(\dot{q}_i - \dot{q}_d) + \sum \Lambda_{ij}(q_i - q_j)] + \sigma_i]\end{gathered} \tag{25}$$

4 Proof of the Stability Analysis

In order to prove the stability of the considered multi agent system, a first Lyaponov function is chosen as follows:

$$V_1 = \eta^T P \eta \tag{26}$$

Its derivative yields:

$$\dot{V}_1 = \eta^T(PA + A^T P)\eta + 2\eta^T P(B\eta(t-\tau) + d - \Lambda\varepsilon + S) \tag{27}$$

The second function is expressed as:

$$V_2 = \int_{t-\tau}^{t} \eta^T H \eta \, d\alpha \tag{28}$$

where:

$$\dot{V}_2 = \eta^T H \eta - \eta(t-\tau)^T H \eta(t-\tau) \tag{29}$$

Then and in order to simplify the expression of $\dot{V}_1 + \dot{V}_2$, we firstly compute the sub −equation:
$2\eta^T PB\eta(t-\tau) - \eta(t-\tau)^T H\eta(t-\tau)$.

So, we denote that $\zeta_1 = \rho_1\eta(t-\tau) - \frac{1}{\rho_1}D^{-1}B^T P\eta(t)$

Consequently:

$$\begin{gathered}\dot{V}_1 + \dot{V}_2 = \eta^T(PA + A^T P)\eta + 2\eta^T P(d - \Lambda\varepsilon + S) \\ +\eta^T H\eta - \zeta_1 D\zeta_1 + \rho_1^2 \eta^T PBD^{-1}B^T P\eta \\ +\frac{1}{\rho_1^2}\eta(t-\tau)^T D\eta(t-\tau) - \eta(t-\tau)^T H\eta(t-\tau)\end{gathered} \tag{30}$$

After that, we regroup all terms of the previous equation:

$$\begin{gathered}\dot{V}_1 + \dot{V}_2 = \eta^T(PA + A^T P + H + \rho_1^2 PBD^{-1}B^T P)\eta \\ -\zeta_1^T D\zeta_1 - \eta(t-\tau)^T(H - \frac{1}{\rho_1^2}D)\end{gathered}$$

where H and D are positive definite matrix while P, N are symetric positive matrix.
It's obvious from the previous equation that there are three terms relative to d, ε and S which should been developed. Consequently:

$$2\eta^T Pd = -\zeta_2^T\zeta_2 + \rho_2^2\eta^T\eta + \frac{1}{\rho_2^2}d^T P^2 d$$

in which:

$$\begin{cases}\zeta_2 = \rho_2\eta - \frac{1}{\rho_2}Pd \\ -2\eta^T\Lambda\varepsilon = -\zeta_3^T\zeta_3 + \rho_3^2\eta^T\eta\frac{1}{\rho_3^2}\varepsilon^T\Lambda^T\Lambda\varepsilon\end{cases}$$

where $\zeta_3 = \rho_3\eta - \frac{1}{\rho_3}\Lambda\varepsilon$.

In the sequel, we introduce new terms V_3 and V_4 to complete the stability verification of the proposed synchronized control schemes such that:

$$\begin{cases}V_3 = \varepsilon^T N\varepsilon \\ V_4 = S^T RS\end{cases} \tag{31}$$

The differentiation with respect of time gives:

$$\begin{cases}\dot{V}_3 = 2\varepsilon^T N(s - \Lambda\varepsilon) = -\varepsilon^T(N\Lambda + \Lambda^T N)\varepsilon \\ -2\varepsilon^T Ns \\ \dot{V}_4 = 2\dot{S}^T RS\end{cases} \tag{32}$$

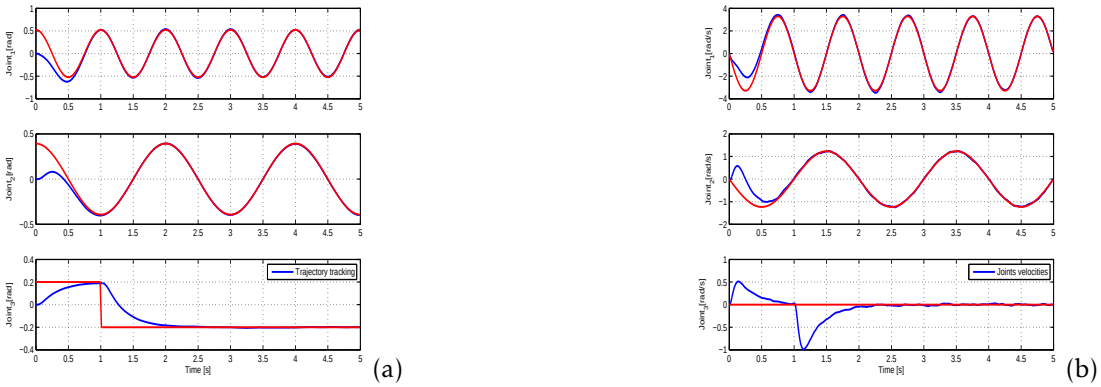
Therefore, we develop the following terms:

$$\begin{cases}2\eta^T PS = -\zeta_4^T\zeta_4 + \rho_4^2\eta^T\eta + \frac{1}{\rho_4^2}S^T P^2 S \\ -2\varepsilon^T NS = -\zeta_5^T\zeta_5 + \rho_5^2\varepsilon^T\varepsilon + \frac{1}{\rho_5^2}S^T N^2 S\end{cases} \tag{33}$$

in which $\zeta_4 = \rho_4\eta - \frac{1}{\rho_4}PS$ and $\zeta_5 = \rho_5\varepsilon + \frac{1}{\rho_5}NS$.

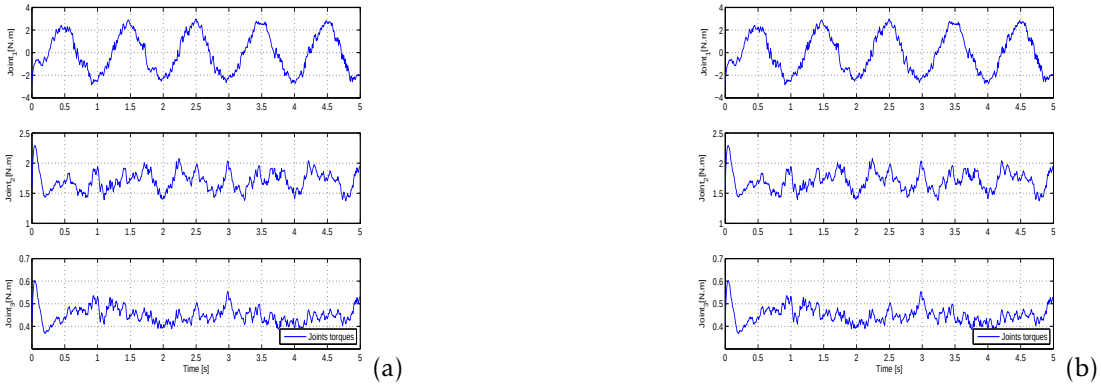
Finally, the derivative of the global Lyaponov function
$V = V_1 + V_2 + V_3 + V_4$ yields:

$$\begin{gathered}\dot{V} = \eta^T(PA + A^T P + H + \rho_1^2 PBD^{-1}B^T P + \rho_2^2 I + \rho_3^2 I \\ +\rho_4^2 I)\eta - \zeta_1^T D\zeta_1 - \zeta_2^T\zeta_2 - \zeta_3^T\zeta_3 - \zeta_4^T\zeta_4 - \zeta_5^T\zeta_5 \\ -\eta(t-\tau)^T(H - \frac{1}{\rho_1^2}D)\eta(t-\tau) - \varepsilon^T(N\Lambda + \Lambda^T N \\ +\rho_5^2 I - \frac{1}{\rho_3^2}\Lambda^T\Lambda)\varepsilon + 2S^T(R\dot{S} - N\varepsilon + P\eta) + \frac{1}{\rho_2^2}d^T P^2 d\end{gathered}$$



(a) (b)

Figure 3: Simulation in the presence of external disturbances: (a) Positions evolutions in the presence of low measurement noises, (b) Velocities evolutions in the presence of low measurement noises



(a) (b)

Figure 4: Simulation in the presence of external disturbances: (a) Torques evolutions of the SOSMC, (b) Measured torques evolutions of the SOSMC

Assume that $\dot{S}$ is defined as follows:

$$\dot{S} = R^{-1}[-N\varepsilon - P\eta - \Lambda S - KsignS] \tag{34}$$

Hence, remember that N, P, Λ, K are positive definite matrices, and assume that they verify the following conditions:

$$\begin{cases} H > \frac{1}{\rho^2} D \gg I \\ PA + A^T P + H + \rho_1^2 PBD^{-1} B^T P + (\rho_2^2 + \rho_3^2 + \rho_4^2) I \\ N\Lambda + \Lambda^T N > \frac{1}{\rho_3^2} \Lambda^T \Lambda \\ 2S^T (R\dot{S} - N\varepsilon + P\eta) + \frac{1}{\rho_2^2} d^T P^2 d = -2S^T \Lambda S < 0 \\ -2K \mid S \mid + \frac{1}{\rho_2^2} d^T P^2 d < 0 \end{cases}$$

The stability is confirmed if:
$\rho_1^2 PBD^{-1} B^T P \ll PA + A^T P$ and the term $\frac{1}{\rho_2^2} d^T P^2 d$ is considered as small bounded so that it can be neglected.
Consequently, we obtain:

$$\dot{V}_i \leq 0$$

Finally, this confirms the stability of the overall system.

5 robustness via measurement noises effect

In order to test the robustness of the proposed controller via uncertainties, an additive measurement errors have been introduced.
Then the measured state can be expressed as follows:

$$x_m(t) = x(t)(1 + b_1(t)) = x(t) + \Delta x(t)$$

where b_1 is an additive bounded measure noise such that:
$\|b_1(t)\| \leq d_1$, and d_1 is a positive constant.
The level of the error effect has been varied from the lower impact to be gradually more intense, aiming to verify the controller's capacity to withstand such disturbances. At the beginning, low noise has been yield (we fluctuate the perturbation from 5 percent to 20 percent), we notice that the positions and the velocities evolutions still remain on the sliding surface which prove the robustness and the insensibility of the second order sliding mode control via disturbances (Figure 3). Then, the level has been increased (almost 30 percent), and in this case, the noise start to affect the tracking evolution of the system(Figure 5b and Figure 6b).
It is obvious from Figure 4 that there is a similarity between the real torques and the measured ones, this

Table 1: Joints parameters

Articulation	Mass	Length	Initial position
q_1	$2.7132(kg)$	$0.2(m)$	$\pi/6(rad)$
q_2	$1.1446(kg)$	$0.15(m)$	$\pi/4(rad)$
q_3	$0.3392(kg)$	$0.1(m)$	$0.2(rad)$

Table 2: Control parameters

Control Parameters	Values
Λ_{ij}	0.8
ω	20
$timedelay$	0.2

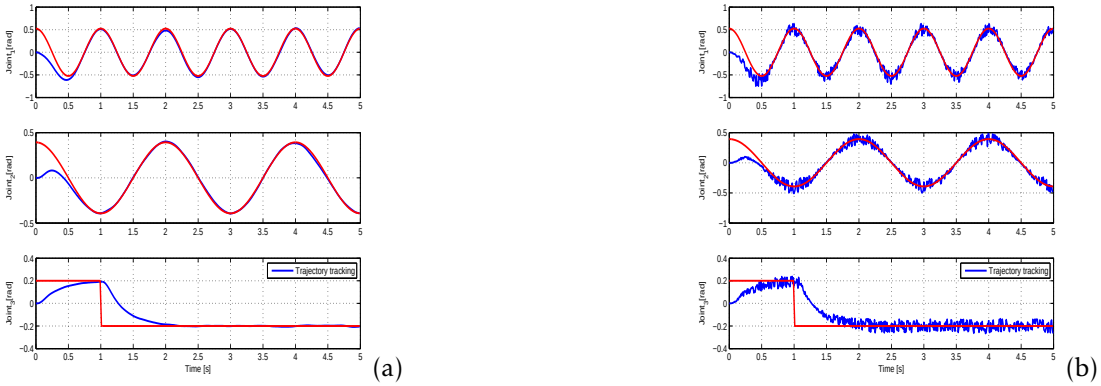


Figure 5: Simulations in the presence of external disturbances: (a) Positions evolutions in the presence of high measurement noises, (b) Measured Position evolutions in the presence of high measurement noises

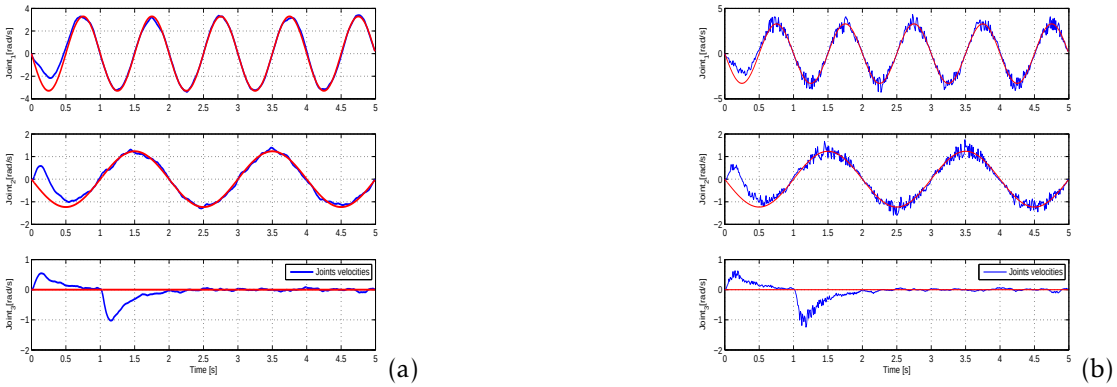


Figure 6: Simulations in the presence of external disturbances: (a) Velocities evolutions in the presence of high measurement noises, (b) Measured Velocities evolutions in the presence of high measurement noises

means that although the system suffers from external disturbances, the evolution of the applied torques do not record any notable increase, subsequently, no additional power consumption needed.

The aim of the comparison between real and measured simulations (Figure 5 and Figure 6) is to prove that the proposed controller exerts a satisfactory compensation action affecting the trajectory tracking and the velocity. This compensation becomes lower while increasing the disturbance effect, and it becomes unable to manage high imposed disruptions, i.e. starting from the presence of 30 present of errors applied to the system. Nevertheless, at this level the proposed controller has still resist to perturbations, and seen that the level of this disruption is relatively high, it can be then considered as a sufficient control, and proves its robustness via external disturbances.

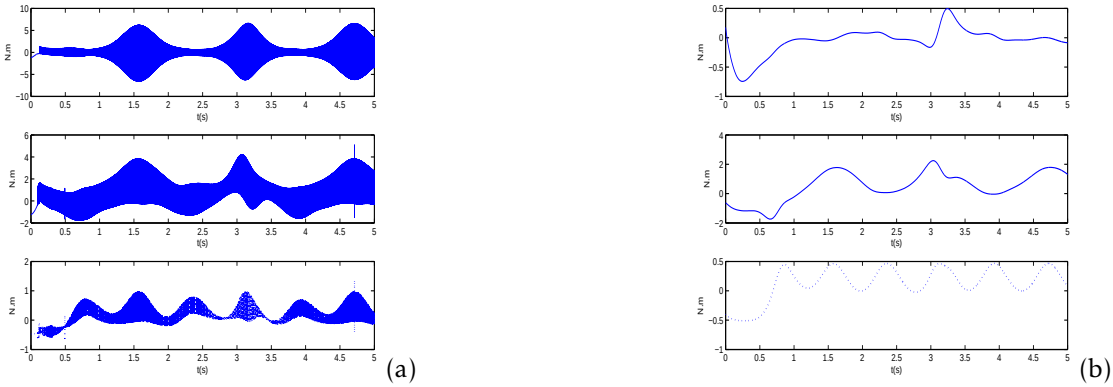


Figure 7: (a) The classical SMC torques evolutions, (b) The second order SMC torques evolutions

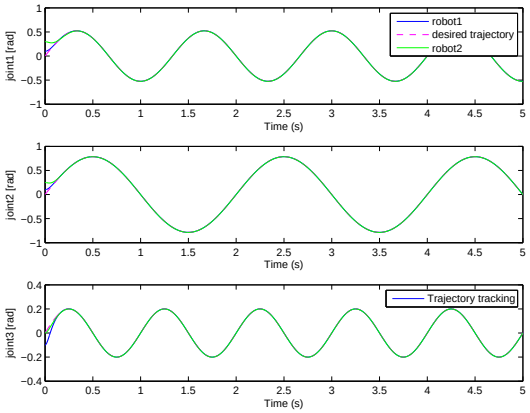


Figure 8: Motion Control and position synchronization

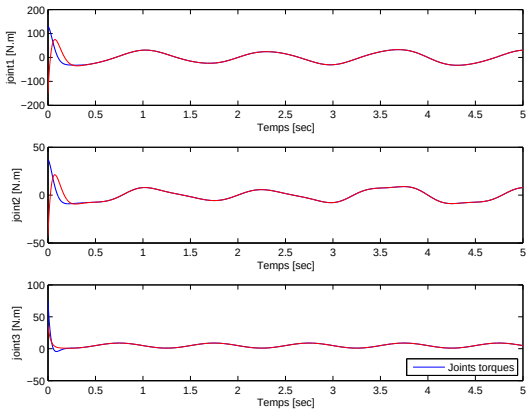


Figure 9: Evolution of the applied torques

6 Simulation Results

In this work, a second order SMC is applied to a networked multi–agent robotic system for a trajectory tracking control task. Indeed, the use of several cameras managed by manipulator robots and controlled by the proposed decentralized control law, allows the interaction of each robot with other agents in the networked system in order to make an overall vision about its environment. Thanks to the cross coupling concept, each agent is able to communicate and exchange information with its neighbors. The desired trajectory is expressed by:

$$q_d(t) = \begin{bmatrix} q_{d1}(t) \\ q_{d2}(t) \\ q_{d3}(t) \end{bmatrix} = \begin{bmatrix} \frac{\pi}{6}\sin(1.5\pi t) \\ \frac{\pi}{4}\sin \pi t \\ \frac{\pi}{2}\sin 4\pi t \end{bmatrix}$$

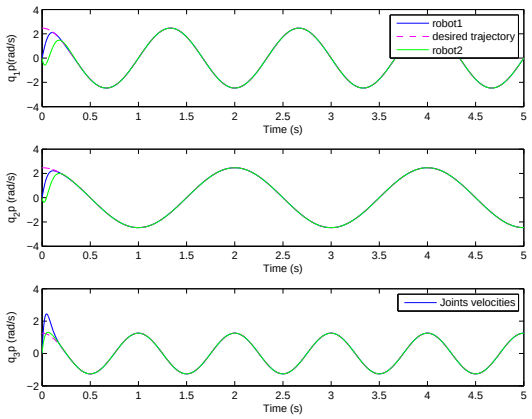


Figure 10: Velocity evolution of the synchronized robots

Parameters of the proposed system model used in simulation are illsutrated in Table 1 and control parameters values are given in Table 2.

Concerning the parameters μ_i and λ_i, they are defined as follows:

$$\lambda_i = 2\begin{pmatrix} 1.125 & & 0 \\ & 1.45 & \\ 0 & & 1.12 \end{pmatrix}$$

$$\mu_i = \begin{pmatrix} 3 & & 0 \\ & 4 & \\ 0 & & 5 \end{pmatrix}$$

Simulation results show the robust synchronization and the smooth evolution in the trajectory tracking based on the second order SMC. The original SMC torques evolutions and the second order SMC ones are shown in Figure 7*a* and Figure 7*b* respectively.

It is obvious from the Figure 7*a* that the classical sliding mode control suffers from the chattering phenomenon whose impact is reflected by the appearance of disrupting high switching frequencies (oscillations). Specifically, the problem consists of rapid and sudden changing control signals which lead to a low control accuracy. On the other side, Figure 7*b* demonstrates that the second order SMC seems to be smooth and able to reduce the chattering phenomenon.

Besides, the presence of time delay between the cooperative robots is slightly reflected during the simulation as shown in Figure 8 and Figure 10 respectively. It may be said otherwise that the communication time delay between robots is clearly compensated and the position synchronization based on the cross coupling concept is obvious, predominately in the (Figure 8). The presence of undesirable phenomenon namely chattering in control torques of each agent is avoided in Figure 9.

7 Conclusion

In this paper, the stability analysis of multi robot manipulator systems with a constant communication delay has been demonstrated. The main goal of this study is to compensate the delayed communication impact and to realize the synchronization between different robots of the system. So the proposed controlled system has succeeded to achieve a performing motion control task in the presence of loss of information during robot's communication, perturbations and also in the presence of delayed communication data. Simulation results show that the multi robots system can achieve the desired motion control task even with presence of a constant time delay.

8 References

1. Sarachik, P., Ragazzini, J. R., "A 2–dimensional feedback control system." Transactions of the American Institute of Electrical Engineers, Part II: Applications and Industry, **76**(2), 55–61, 1957.

2. Koren, Y., "Cross-coupled biaxial computer control for manufacturing systems." Journal of Dynamic Systems, Measurement, and Control, **102**(4), 265–272, 1980.

3. Beni, G., "The concept of cellular robotic system" Intelligent Control, Proceedings., IEEE International Symposium on, **102**(4), 265–272, 1988.

4. Wei, H., Yong, W., Xuanqin, M., Yan, W., "A cooperative fuzzy control method for traffic lights." In Intelligent Transportation Systems, 2001. Proceedings. 2001 IEEE, 185–188, 2001.

5. Lucarelli, D., Wang, I. J. , "Decentralized synchronization protocols with nearest neighbor communication." In Proceedings of the 2nd international conference on Embedded networked sensor systems. ACM, 62–68, 2004.

6. Oh, M. K., Ma, X., Giannakis, G. B., Park, D. J., "Cooperative synchronization and channel estimation in wireless sensor networks." Journal of

Communications and Networks, **7**(3), 284–293, 2005.

7. Liu, S., Sun, D., Zhu, C., "Coordinated motion planning for multiple mobile robots along designed paths with formation requirement." IEEE/ASME transactions on mechatronics, **16**(6), 1021–1031, 2011.

8. Shirkhodaie, A., "Supervised control of cooperative multi-agent robotic vehicles. In System Theory" Proceedings of the Thirty-Fourth Southeastern Symposium on. IEEE, 386–390, 2002.

9. Du, H., Li, S., Qian, C. , "Finite-time attitude tracking control of spacecraft with application to attitude synchronization." IEEE Transactions on Automatic Control, **56**(11), 2711–2717, 2011.

10. Chung, S. J., Ahsun, U., Slotine, J. J. E., Miller, D. W., " Application of synchronization to cooperative control and formation flight of spacecraft" In proceedings of the AIAA Guidance, Navigation and Control Conference and Exhibit. Hilton Head, South Carolina, 2007.

11. Rodriguez-Angeles, A., Nijmeijer, H. , " Mutual synchronization of robots via estimated state feedback: a cooperative approach." IEEE Transactions on Control Systems Technology, **12**(4), 542–554, 2004.

12. Sun, D., Mills, J. K., "Adaptive synchronized control for coordination of two robot manipulators" In Robotics and Automation, 2002. Proceedings. ICRA'02. IEEE International Conference, 976–981, 2002.

13. Liu, Y. C., Chopra, N. , Park, D. J., "Controlled synchronization of heterogeneous robotic manipulators in the task space." IEEE Transactions on Robotics, **28**(1), 286–275, 2012.

14. Zhao, D., Li, S., Gao, F., Zhu, Q., "Robust adaptive terminal sliding mode-based synchronised position control for multiple motion axes systems." IET Control Theory and Applications, **3**(1), 136–150, 2009.

15. Lu, J., Yu, X., Chen, G., Cheng, D, "Characterizing the synchronizability of small-world dynamical networks." IEEE Transactions on Circuits and Systems I: Regular Papers , **51**(1), 787–796, 2004.

16. Tan, Yanxia, and Zhenkun Huang, "Synchronization of drive-response networks with delays on time scales." IEEE/CAA Journal of Automatica Sinica, 2016.

17. Sun, D., "Position synchronization of multiple motion axes with adaptive coupling control." Automatica, **39**(6), 997–1005, 2003.

18. Li, Z., Ge, S. S., Adams, M., Wijesoma, W. S., "Robust adaptive control of uncertain force motion constrained nonholonomic mobile manipulators." Automatica, **44**(3), 776–784, 2008.

19. Yang, C., Ganesh, G., Haddadin, S., Parusel, S., Albu-Schaeffer, A., Burdet, E., "Human-like adaptation of force and impedance in stable and unstable interactions." IEEE transactions on robotics, **27**(5), 918–930, 2011.

20. Widyotriatmo, A., Hong, K. S., "Navigation function-based control of multiple wheeled vehicles." IEEE Transactions on Industrial Electronics, **58**(5), 1896–1906, 2011.

21. Vaidyanathan, S., Azar, A. T., Hybrid synchronization of identical chaotic systems using sliding mode control and an application to Vaidyanathan chaotic systems, In Advances and applications in sliding mode control systems: Springer International Publishing, 549–569, 2015.

22. Shtessel, Y., Edwards, C., Fridman, L., and Levant, A., Introduction: Intuitive theory of sliding mode control. In Sliding Mode Control and Observation, Springer New York, 1–42, 2014.

23. Liao, Teh-Lu, Sheng-Hung Lin., "Adaptive control and synchronization of Lorenz systems." Journal of the Franklin Institute, **336**(6), 925–937, 1999.

24. Zhou, Qi, Chengwei Wu, Peng Shi., "Observer-based adaptive fuzzy tracking control of nonlinear systems with time delay and input saturation." Fuzzy Sets and Systems, **316**, 49–68, 2016.

25. Yang, Shyi-Kae, Chieh-Li Chen, Her-Terng Yau, "Control of chaos in Lorenz system." Chaos, Solitons and Fractals, **13**(4), 767–780, 2002.

26. Wang, J., Gao, Y., Qiu, J., and Ahn, C. K. , "Sliding mode control for non linear systems by Takagi Sugeno fuzzy model and delta operator approaches." IET Control Theory and Applications , 2016.

27. Yu, S., Chau, T. K., Fernando, T., Savkin, A. V., and Iu, H. H. C., "Novel Quasi-Decentralized SMC-Based Frequency and Voltage Stability Enhancement Strategies using Valve Position Control and FACTS Device." in IEEE Access, 2016.

28. Kazemy, Ali., "Complex network synchronization analysis with neural network nodes and time-delays." in Electrical Engineering (ICEE), 24th Iranian Conference on. IEEE, 2016.

29. Wu, Min, Yong He, and Jin-Hua She, Stability analysis and robust control of time-delay systems, Springer: Berlin, 2010.

30. Lakshmanan, S., Prakash, M., Lim, C. P., Rakkiyappan, R., Balasubramaniam, P., and Nahavandi, S., "PSynchronization of an Inertial Neural Network With Time-Varying Delays and Its Application to Secure Communication" in IEEE Transactions on Neural Networks and Learning Systems, 2016.

31. Yao, J., Wang, H. N., Guan, Z. H., and Wang, H. O., "Synchronization of leader follower networks with coupling delays via variable structure control." Fuzzy Sets and Systems, **11**(4), 407–410, 2009.

32. Yeganefar, N., "Definitions et analyse de stabilites pour les systemes retard non lineaires," Ph.D Thesis, University of Science and Technology of Lille-Lille I, 2006.

33. Wu, M., He, Y., and She, J. H., Stability analysis and robust control of time-delay systems, Berlin: Springer, 2010.

34. Jribi, A., Abdelhedi, F., Bouteraa, Y., and Derbel, N., Kinematics and a Comparison Between Two SM Control Strategies for a 5DOF Serial Robot for Tele-Echography, In Applications of Sliding Mode Control, 157–174, 2009.

35. Abdelhedi F., Derbel N., "Adaptive Second order sliding mode control based cross coupling concept for synchronization of robotic systems", Advances in Systems, Signals and Devices, Issues on Systems, Analysis and Automatic Control, 2016.

36. Srinivasan, K., Kulkarni, P. K., "Cross-coupled control of biaxial feed drive servomechanisms." Journal of dynamic systems measurement and control, **112**(2), 225–232, 1990.

37. Koren, Y., Lo, C. C., "Advanced controllers for feed drives." CIRP Annals-Manufacturing Technology, **41**(2), 689–698, 1992.

38. Liu, Yen-Chen, and Nikhil Chopra, "Synchronization of networked mechanical systems with communication delays and human input." Journal of Dynamic Systems, Measurement, and Control, **135**(4), 041004, 2013.

39. Sun, Dong., Synchronization and control of multiagent systems, CRC Press, 2016.

40. Wu, Z. G., Park, J. H., Su, H., Chu, J., "Discontinuous Lyapunov functional approach to synchronization of time-delay neural networks using sampled-data." Nonlinear Dynamics, **69**(4), 2021–2030, 2012.

41. Shafiei, S. E., Sliding mode control of robot manipulators via intelligent approaches, INTECH Open Access Publisher, 2010.

42. Levant, A., "Chattering analysis." IEEE Transactions on Automatic Control, **55**(6), 1380–1389, 2010.

43. Bouaziz, F. A., Bouteraa, Y., and Medhaffar, H., "Distributed sliding mode control of cooperative robotic manipulators" In 10th International Multi-Conference on Systems, Signals and Devices, 2013.

44. Ferrara, A., Magnani, L., "Motion control of rigid robot manipulators via first and second order sliding modes." Journal of Intelligent and Robotic Systems, **48**(1), 23–36, 2007.

45. Zhao, Dongya, and Quanmin Zhu., "Position synchronised control of multiple robotic manipulators based on integral sliding mode." International Journal of Systems Science, **45**(3), 556–570, 2014.

46. Fathallah, M., Abdelhedi, F., Derbel, N., " A surveillance camera algorithm based on the sliding mode approach with measurement's disturbance" In 13th International Multi-Conference on Systems, Signals and Devices, 2016.

47. Abdelhedi, F., Jribi, A., and Derbel, N. , " Robust SMC for synchronization of networked time delay robotic systems" In 28th International Conference on Modelling, Identification and Control (ICMIC), 2016.

48. Nouri, A. S., Bouazi, F. A., and Derbel, N., On the Sliding Control, In Applications of Sliding Mode Control Springer Singapore, 2017.

49. Bouaziz, F. A., Bouteraa, Y., and Derbel, N., " Control energy comparison between 1st and 2nd order sliding mode approach with application to a SCARA robot" In 13th International Multi-Conference on Systems, Signals and Devices, 2016.

50. Koren, Y., "Cross-coupled biaxial computer control for manufacturing systems." Journal of Dynamic Systems, Measurement, and Control, **102**(4), 265–272, 1980.

51. Srinivasan, K., Kulkarni, P. K., "Cross-coupled control of biaxial feed drive servomechanisms." Journal of dynamic systems, measurement, and control, **112**(2), 225–232, 1990.

52. Feng, L., Koren, Y., Borenstein, J., "Cross-coupling motion controller for mobile robots." IEEE control systems, **13**(6), 35–43, 1993.

53. Abdelhedi, F., Bouteraa, Y. and Derbel N., Second Order Sliding Mode Based Synchronization Control for Cooperative Robot Manipulators, Advances and Applications in Nonlinear Control Systems Springer International Publishing, 2016.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 2, No. 3, 171-181 (2017)
www.astesj.com
Special Issue on Recent Advances in Engineering Systems

ASTES Journal
ISSN: 2415-6698

Fault Diagnosis and Tolerant Control Using Observer Banks Applied to Continuous Stirred Tank Reactor

Martin F. Pico*, Eduardo J. Adam

Universidad Nacional del Litoral, Facultad de Ingeniería Química, S3000, Argentina

ARTICLEINFO

Article history:
Received: 19 December, 2016
Accepted: 28 March, 2017
Online: 20 April, 2017

Keywords:
Fault tolerant control (FTC)
Fault detection and diagnosis (FDD)
Observer bank

ABSTRACT

This paper focuses on studying the problem of fault tolerant control (FTC), including a detailed fault detection and diagnosis (FDD) module using observer banks which consists of output and unknown input observers applied to a continuous stirred tank reactor (CSTR). The main objective of this paper is to use a FDD module here proposed to estimate the fault in order to apply this result in a FTC system (FTCS), to prevent a lost of of the control system performance. The benefits of the observer bank and fault adaptation here studied are illustrated by numerical simulations which assumes faults in manipulated and measuring elements of the CSTR.

1 Introduction

The performance of a closed-loop system can be altered by a failure in one of its components, and in some cases can lead to an instability of the control system, causing major damage to it. The goal of fault tolerant control (FTC) is to prevent deterioration of the system, by means of a controller with the ability to compensate faults, by correcting the control action.

An active fault tolerant control systems (AFTCS) have basically two subsystems: a module for fault detection and diagnosis (FDD) and another one for reconfigurable control (RC). The detection and isolation of faults is an important research area in process control due to the improvements that it can be reached in terms of safety and reliability of the plant. For this purpose, several model-based methods for generation of residue, mainly based on observers, are proposed for different authors [5, 15, 8, among others]. However, there are another model-based methods, such as parity equations or stable matrix factorization [6] that were explored by several authors [17, 16, among others].

Firstly, the paper focuses on studying the problem of fault detection and diagnosis using observer banks that they include output and unknown input observers [5, 4]. The main objective is the fault estimation by means of diagnostic module, and then a fault-tolerant control system is developed, to prevent a bad performance [13, 2, 18, among others].

This article focuses on the design issues of FDD module and FTC system applied to a CSTR where an exothermic chemical reaction take place, using a mathematical model presented by several authors as [21, 9, 1, among others] and whose parameters were adopted according to [11]. The simulations show how output and unknown input observers can quickly detect faults in sensors and actuators, estimate the magnitude of the occurred faults and thus quickly correct the control action, for preserving stability and performance of the control system.

This paper is organized as detailed bellow. Section 2 presents the basic theoretical concepts associated with the observer banks here proposed. Then, in section 3 the general scheme of an active FTC and the adopted strategy for recovering performance is presented. Subsequently, section 4 shows the

improvement achieved by using proposed observer banks for fault detection and isolation, by numerical simulations. Finally, in section 5 the conclusions are presented.

2 Fault detection by means of observer banks

2.1 Problem statement

Consider a linear time invariant (LTI) system under the state space representation, with actuators and sensors additive faults as follows:

*Corresponding Author: Martin F. Pico, Universidad Nacional del Litoral, Facultad de Ingeniería Química, S3000, Argentina, martinpico@gmail.com

https://dx.doi.org/10.25046/aj020322

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t) + F_u f_u(t) \\ y(t) &= Cx(t) + F_y f_y(t) \end{aligned} \tag{1}$$

where $A \in \Re^{n\times n}$, $B \in \Re^{n\times q}$ y $C \in \Re^{m\times n}$ are the matrices of state, control and output respectively, $x \in \Re^n$ is the true value of the state vector, $u \in \Re^q$ is the input vector, $y \in \Re^m$ represents the outputs of the system, F_u and F_y are inputs and outputs faults distribution matrices, $f_u(t)$ and $f_y(t)$ are the inputs and outputs additive faults, respectively.

Under a unified framework, the system (1) can be expressed as,

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t) + Lf(t) \\ y(t) &= Cx(t) + Mf(t) \end{aligned} \tag{2}$$

where$f(t) \in \Re^v$ with $v = m + q$, is a actuators and sensors faults representation such that,

$$f(t) = \begin{bmatrix} f_u \\ f_y \end{bmatrix} \tag{3}$$

$L \in \Re^{n\times v}$ y $M \in \Re^{m\times v}$ are actuators and sensors faults matrices, which may written as $L = [B\ 0_{n\times m}]$ and $M = \begin{bmatrix} 0_{q\times m} & I_{m\times m} \end{bmatrix}$.

The goal is to estimate f_u and f_y as soon as possible such that, the monitored data are quickly corrected and at the same time the closed loop control system is able to correct the fault. Consequently, it is possible to prevent that the system is positioned outside the desired operating point or worse, become unstable.

Output observers

For fault detection and isolation (FDI) purposes, the estimation of all states is not required, it is sufficient to estimate of the output variables. For this reason, an output observer[1] is appropriate for FDI [8].

The dynamic of this observer (Fig. 1) is,

$$\dot{z}(t) = Fz(t) + Ju(t) + Ky(t) \tag{4}$$

where $z \in \Re^s$, s denote the observer order which may be the same, greater or lesser than the order of the system (n).

As it shown in Fig. 1, the observer output is the residue

$$r(t) = Wz(t) - Vy(t) \tag{5}$$

where the following conditions must be satisfied [14],

$$\begin{aligned} \Re(\lambda(F)) < 0; \\ KC = TA - FT \\ J = TB \\ WT + VC = 0 \end{aligned} \tag{6}$$

where $\Re(\lambda(F))$ denote the eigenvalue real parts of the F matrix.

According with [14], if these design conditions are satisfied, the dynamic errors and residues are,

$$\dot{e}(t) = Fe(t) - TLf(t) + KMf(t) \tag{7}$$

$$r(t) = We(t) - VMf(t) \tag{8}$$

with

$$e(t) = z(t) - Tx(t) \tag{9}$$

where the residuals depend on faults.

In order to isolate faults, a set of structural or directional residues is designed using this approach. Thus, the design of a structural residue set for sensor faults is straightforward. For example, if the output vector $y = (y_1, ..., y_m)$ is rewritten as $(y_1, ., y_{i-1}, y_{i+1}, .., y_m)$, the residue will be insensitive to the fault of the $i - th$ sensor.

Contrary to the above, to isolate faults in the actuators through structural residues is not straightforward, but can be solved by means of *unknown input observers* (UIO).

2.2 Unknown input observers

This subsection is based on the work of Hou and Müller [7] who proposed to design an observer whose dimension is $(n - q)$ for the purpose of detecting additive actuator faults. To do this, consider the LTI system (1) where A, B, C, F_u are constant matrices with appropriated dimensions. Furthermore, assume that $m \geq q$ and, without loss of generality, $rank(F_u) = q$ and $rank(C) = m$.

Under this assumption, it is possible to choose a nonsingular matrix T as

$$T = [N\ F_u], \tag{10}$$

with $N \in \Re^{nx(n-q)}$.

$$\begin{aligned} \dot{\bar{x}}(t) &= \bar{A}\bar{x}(t) + \bar{B}u(t) + \bar{L}f_u(t) \\ y(t) &= \bar{C}x(t) \end{aligned} \tag{11}$$

where,

$$x = T\bar{x} = T\begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \end{bmatrix}, \tag{12}$$

$$\bar{A} = T^{-1}AT = \begin{bmatrix} \bar{A}_{11} & \bar{A}_{12} \\ \bar{A}_{21} & \bar{A}_{22} \end{bmatrix}, \tag{13}$$

$$\bar{B} = T^{-1}B = \begin{bmatrix} \bar{B}_1 \\ \bar{B}_2 \end{bmatrix} \tag{14}$$

$$\bar{F}_u = T^{-1}F_u = \begin{bmatrix} 0 \\ I_q \end{bmatrix},\ \bar{C} = CT = [CN\ CF_u] \tag{15}$$

con $\bar{x}_1 \in \Re^{n-q}$ and $\bar{x}_2 \in \Re^q$.

Notice that, in Eq. (11), $\bar{x}_2$ is directly affected by unknown fault f_u, while $\bar{x}_1$ it is independent of f_u. Then, is possible to rewrite a system (11) without the unknown input as

$$\begin{bmatrix} I_{n-q} & 0 \end{bmatrix}\dot{\bar{x}} = \begin{bmatrix} \bar{A}_{11} & \bar{A}_{12} \end{bmatrix}\bar{x} + \bar{B}_1 u \tag{16}$$

$$y = CT = [CN\ CF_u]\bar{x} \tag{17}$$

[1] Also designated as diagnostic observer, functional observer or generalized Luenberger observer.

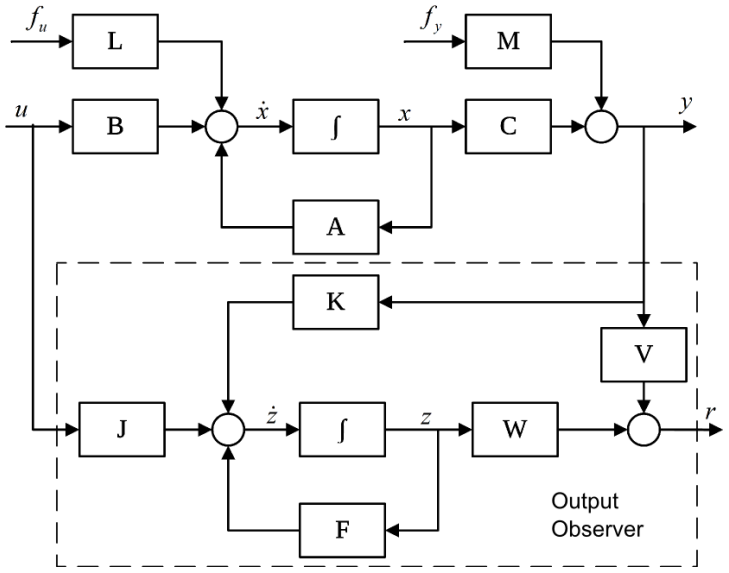


Figure 1: Process and output observer.

Assuming that $\bar{x}_2$ can be obtained from y, the Eq. (16) can be written as a conventional linear system. Indeed, if the matrix $[CF_u]$ is of full column rank, there exist a non-singular matrix

$$U = [CF_u \; Q] \tag{18}$$

Being,

$$U^{-1} = \begin{bmatrix} U_1 \\ U_2 \end{bmatrix} \tag{19}$$

with $U_1 \in \Re^{q\times m}$ y $U_2 \in \Re^{(m-q)\times q}$ and pre-multiplying both sides of Eq. (17) por U^{-1},

$$U_1 y = U_1 CN\bar{x}_1 + \bar{x}_2 x \tag{20}$$
$$U_2 y = U_2 CN\bar{x}_1 \tag{21}$$

and replacing Eq. (20) in Eq. (16) and combining the last equation with Eq. (21) is arrived to

$$\dot{\bar{x}}_1(t) = \bar{A}_1\bar{x}_1 + \bar{B}u(t) + \bar{E}_1 y \tag{22}$$
$$\bar{y} = \bar{C}\bar{x}_1 \tag{23}$$

where

$$\bar{A}_1 = \bar{A}_{11} - \bar{A}_{12}U_1 CN \qquad E_1 = \bar{A}_{12}U_1 \tag{24}$$
$$\bar{C}_1 = U_2 CN \qquad \bar{y} = U_2 y$$

Notice that, it is possible to verify that the $rank(\bar{C}_1) = m - q$.

At this point the pair $\bar{A}_1, \bar{C}_1$ is observable, and following the design procedure of a Luenberger observer, it is possible to design a reduced order observer for the system free of unknown inputs given by Eq. (22) as

$$\dot{w}(t) = \left[\bar{A}_1 - K_e\bar{C}_1\right]w(t) + \bar{B}_1 u(t) - K_e^* y(t),\; w \in \Re^{(n-q)} \tag{25}$$

where $K_e \in \Re^{(n-1)\times(m-q)}$ y $K_e^* = K_e * U_2 + E_1$ and $w \to \bar{\bar{x}}_1$ when $t \to \infty$, and

$$\bar{x} = T\bar{\bar{x}} = T\begin{bmatrix} w \\ U_1 y - U_1 CN w \end{bmatrix} \tag{26}$$

where $\bar{x} \to x$, when $t \to \infty$. Then, the unknown input or disturbance d can be estimated as

$$\bar{f}_u = U_1\dot{y} + G_e w + G_4 y + G_5 u \tag{27}$$

where

$$\begin{aligned} G_3 &= U_1 CNK_e U_2 CN + U_1 CN\bar{A}_{12}U_1 CN - \\ &\quad U1CN\bar{A}_{11} - \bar{A}_{21} + \bar{A}_{22}U_1 CN \\ G_4 &= -U1_C NK_e U_2 - U_1 CN\bar{A}_{12}U_1 - \bar{A}_{22}U_1 \\ G_5 &= -U1CN\bar{B}_1 - \bar{B}_2 \end{aligned} \tag{28}$$

2.3 Observer banks

Figure 2 shows how to implement output and unknown input observer banks to estimate the residues according to the description in the previous two subsections.

3 Fault-tolerant control

The sensor and actuator faults do not impact in a system in the same way. Consequently, the solution for these two types of faults is different.

Fiqure 3 shows the FTC strategy based on [22] and [19] here implemented. Notice that, the detection, isolation and estimation consist in analyzing the results of each residue generator to decide if sensor or actuator is failing.

3.1 Design of FTC system for sensor faults

When a particular sensor fault takes place, a free faults estimation of this particular sensor from observer bank is used, and this signal is feedback. That is, the estimated output variable is considered as good

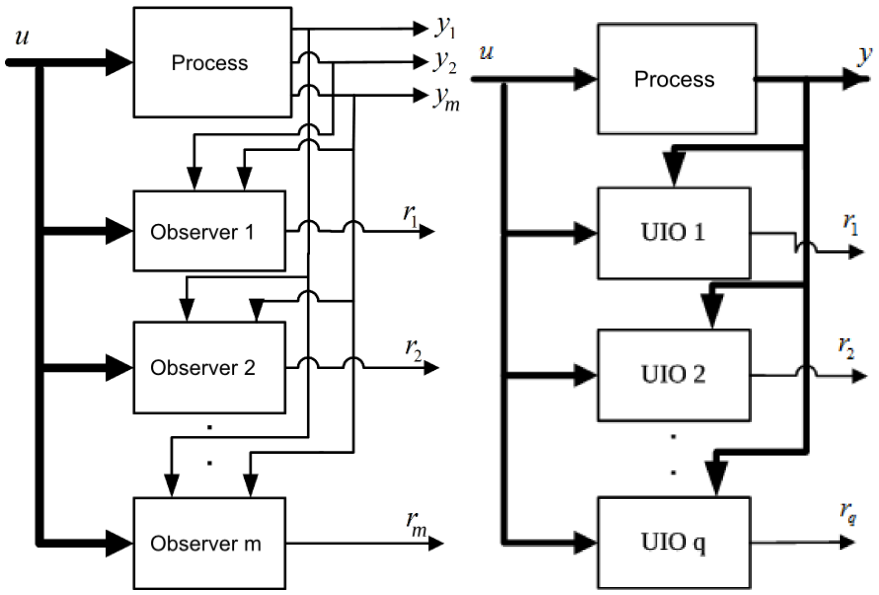


Figure 2: Observer Banks. (a) Output observer (OO) bank for all inputs and outputs except one.(b) Unknown input observer (UIO) bank.

measurement and, the control loop is automatically reconfigured.

The estimation of j-th damaged sensor $\hat{y}_j$ is obtained from an observer which is designed to be insensitive to fault of this j-th sensor. Consequently, the output vector used to implement the control law is

$$y_c(t) = \begin{pmatrix} \check{y}_1(t) \\ \vdots \\ \check{y}_j(t) \\ \vdots \\ \check{y}_m(t) \end{pmatrix} \tag{29}$$

where

$$\check{y}_j(t) = \Big(1 - I(f_j)\Big) y_j + I(f_j)\hat{y}_j(t) \tag{30}$$

with

$$I(f_j) = \begin{cases} 1 \text{ fault in j-th sensor} \\ 0 \text{ sensor without fault} \end{cases} \tag{31}$$

3.2 Design of FTC system for actuator faults

When a fault in an actuator is detected, the control action is compensated by adding to manipulated variable whose value is proportional to the fault in order to minimize the impact.

In this work, a compensation to the control law proposed by [12] is added. Thus, under fault representation presented in Eq. (2), it is propose to add u^{acc} to the nominal control action.

Therefore, when in a particular instant (t_{find}) an actuator fault is detected, the control signal is corrected as follows:

$$u^{FTC}(t) = u(t) + u^{acc} \tag{32}$$

According to this new control law (Eq. (32)), the system described in state space given by Eq. (2) becomes,

$$\dot{x}(t) = Ax(t) + Bu(t) + Bu^{acc} + Lf^u(t) \tag{33}$$

$$y(t) = C_r x(t) \tag{34}$$

where the additional control action must be quickly computed such that the system is recovered as soon as possible. Then, the following condition is requested:

$$Bu^{acc} + Lf^u(t) = 0 \tag{35}$$

The fault estimation ($\widehat{f_u}$) can be obtained by mean of Eq. (35), resulting

$$u^{acc}(t) = -B^{+}L\widehat{f^u}(t) \tag{36}$$

where B^+ is the pseudo-inverse matrix of the matrix B. This compensation control law is represented in Fig. (3) within the module denoted *acomodation and estimation module*.

4 Detecting faults in a CSTR

In order to analyze the system performance with the proposed observer banks above, it is proposed to simulate faults in a CSTR control system. The behavior of the system under actuator and sensor faults was simulated according to the classic non-linear model ([21, 9, 1, among others]) with the parameters suggested by [11].

4.1 CSTR non-linear model

Consider a CSTR represented in Fig. 4, where an irreversible exothermic chemical reaction $A \rightarrow B$ occurs. This reaction takes place in a cylindrical stirred tank

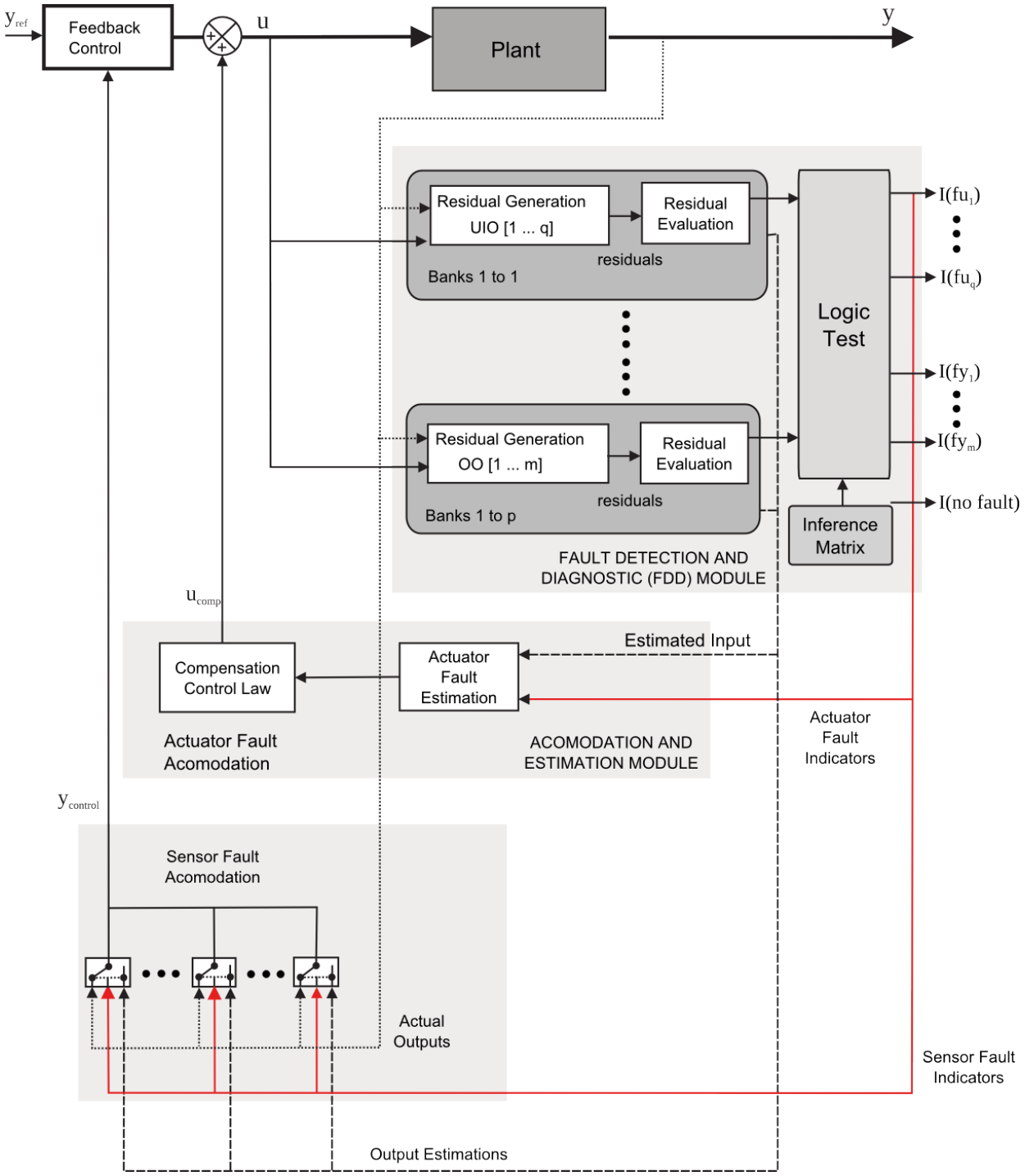


Figure 3: Scheme of a fault tolerant control system including detection and diagnostic module.

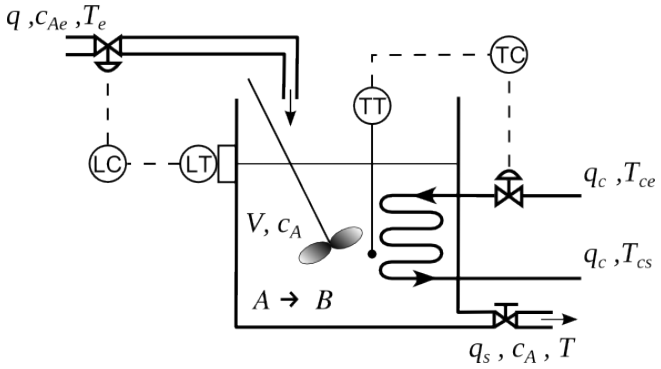


Figure 4: CSTR control schemes.

with a total capacity of $V_{\max} = 120$ l and a transversal section $A = 0.2\text{m}^2$. The principal parameter values are summarized in Table 1.

The complete model (37) can be written according to a non-linear differential and algebraic equation set as follows:

$$\begin{aligned}
\frac{d\,V(t)C_A(t)}{dt} &= q(t)\,(C_{Ae} - C_A(t)) - k_0 Ca(t) e^{-\frac{E}{RT(t)}} \\
\frac{d\,V(t)T(t)}{dt} &= q(t)(T_e - T(t)) - k_1 C_A(t) e^{-\frac{E}{RT(t)}} \\
&\quad + k_2 q_c(t)\left(1 - e^{-\frac{k_3}{q_c(t)}}\right)(T_{ce} - T(t)) \\
\frac{d\,V(t)}{dt} &= q(t) - k_4\sqrt{V(t)} \qquad (37)
\end{aligned}$$

where it is defined the constants,

$$k_1 := \frac{\Delta H k_0}{\rho c_p},\ k_2 := \frac{\rho_c c_{p_c}}{\rho c_p},\ k_3 := \frac{hA}{\rho_c c_{p_c}}, \qquad (38)$$

and the outlet flow rate,

$$q_s(t) = k_v\sqrt{V(t)}. \qquad (39)$$

Using Taylor series expansion around the operation point (C_A^*, T^*, V^*) and, neglecting terms of order two and higher, a matrix-vector equation under state space representation framework can be written. Thus, through this procedure is obtained

$$A = \begin{pmatrix} -\frac{q^*}{V^*} - k_0 e^{-\frac{E}{RT^*}} & A_{12} & -\frac{q^*}{V^{*2}}(C_{Ae} - C_A^*) \\ k_1 e^{-\frac{E}{RT^*}} & A_{22} & A_{23} \\ 0 & 0 & -\frac{k_4}{2\sqrt{V^*}} \end{pmatrix} \qquad (40)$$

where

$$A_{12} = -k_0 {C_A}^* \frac{E}{RT^{*2}} e^{-\frac{E}{RT^*}}$$

$$A_{22} = -\frac{q^*}{V^*} + k_1 {C_A}^* \frac{E}{RT^{*2}} e^{-\frac{E}{RT^*}} - k_2 \frac{q_c^*}{V^*}\left(1 - e^{-\frac{k_3}{q_c^*}}\right)$$

$$A_{23} = -\frac{q^*}{V^{*2}}(T_e - T^*) - k_2 \frac{q_c^*}{V^*}\left(1 - e^{-\frac{k_3}{q_c^*}}\right)(T_{ce} - T^*)$$

By means of parameter values summarized in Table 1 and assuming the operation point suggested by [11],

$$A = \begin{pmatrix} -9.9979 & -0.046787 & -0.009 \\ 1799.6 & 7.3245 & 1.7999 \\ 0 & 0 & -0.5 \end{pmatrix} \qquad (41)$$

$$B = \begin{pmatrix} 0.009 & 0 \\ -0.8854 & -0.8775 \\ 1 & 0 \end{pmatrix} \quad C = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \qquad (42)$$

Finally, according to Fig. 4, the control system has two PI controllers, where the control variables are the temperature T and volume V, which will be controlled by the cooling flow q_c and process flow rate q respectively.

4.2 Design of the observer banks

For this example, two observer bank were developed. One of them consist in three output observer to diagnose faults in the sensors of the volume, temperature and concentration, and the other one include two UIO to diagnose faults in the process and coolant flow rates.

4.3 Output observer design

According to section 2.1, by means of Eqs. (6) the following vectors can be written:

$$\begin{aligned}
J_1 &= (0.1517\ 0.1414)\ F_1 = (-300) \\
K_1 &= (-49.5715\ -0.2975) \\
J_2 &= (38.9958\ -0.8775)\ F_2 = (-200) \\
K_2 &= (8.4374\ -0.0004) \qquad (43) \\
J_3 &= (50.2176\ 48.7760)\ F_3 = (-100) \\
K_3 &= (-1.0005\ -0.0041)
\end{aligned}$$

where the numerical subscript indicates the corresponding OO.

To interpret and isolate faults in the OO bank the isolation patterns are used. For these particular example, is applied a pattern as shown in Table 2.

For example, according to Table 2, if a fault increases the monitored temperature, a decrement of residue related to the reactant concentration (r_{cA}) take place. Furthermore, an increment of temperature residue (r_T) and an invariant residue related to volume (r_v) are observed. As a result of these behaviors, a temperature sensor fault can be inferred.

4.4 Unknown input observer design

According to section 2.2, the following matrices can be written:

$$\begin{aligned}
L_1 &= \begin{pmatrix} 0.009 \\ -0.8854 \\ 1 \end{pmatrix} L_2 = \begin{pmatrix} 0 \\ -0.8775 \\ 0 \end{pmatrix} \\
N_1 &= N_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix} \\
K_{e1} &= \begin{pmatrix} 0.0409 \\ 9.7011 \end{pmatrix} K_{e2} = \begin{pmatrix} 0.0385 \\ 8.7476 \end{pmatrix} \qquad (44)
\end{aligned}$$

where L_1 and L_2 are the actuator fault vectors, N_1 and N_2 are proposals to satisfy Eq. (10), and K_{e1} and K_{e2} come from solving the equation of the observer (25).

4.5 Numerical simulation

The performance of FDI procedure here proposed by means of observer banks applied to additive faults are studied by means of numerical simulations.

Table 1: Nominal CSTR parameter values

Parameter	Description	Value
C_A	Meassured product concentration	0.1 mol l^{-1}
T	Reactor Temperature	438.54 K
q_c	Coolant flow rate	103.14 L min^{-1}
q	Process flow rate	100 L min^{-1}
C_{A0}	Feed concentration	1 mol l^{-1}
T_0	Feed temperature	350 K
T_{c0}	Inlet coolant temperature	350 K
V	CSTR volume	100 L
hA	Heat transfer term	7×10^5 cal min^{-1} K^{-1}
k_0	Reaction rate constant	7.2×10^{10} min^{-1}
E/R	Activation energy term	1×10^4 K
ΔH	Heat of reaction	-2×10^5 cal mol^{-1}
ρ, ρ_c	Liquid densities	1×10^3 g l^{-1}
c_p, c_{p_c}	Specific heats	1 cal g^{-1} K^{-1}
k_v	Valve constant	10 $l^{1/2}$ min^{-1}

Table 2: Signs of residues for different types of additive faults in the CSTR.

	r_{C_A}	r_T	r_V
$\pm f_{C_A}$	$\pm$	0	$\pm$
$\pm f_T$	$\mp$	$\pm$	0
$\pm f_V$	0	0	$\pm$

4.6 Temperature sensor fault

Firstly, set-point changes around the reactor operating point are proposed in order to show that the PI controllers allow tracking changes. In particular, the proposed changes are: the volume $V = 110$ L at $t = 5$ min and the temperature $T = 435$ °K at $t = 15$ min.

Figure 5 shows the reactor dynamic response when a 5 % degradation in the temperature sensor at $t = 30$ min takes place. According with Fig.6, without fault tolerant control, the control loop reduces the coolant flow q_c to adjust the temperature of the reactor to the desired value according to an measured incorrect value, assuming that the reactor needs to accumulate more heat. Consequently, the actual temperature of the reactor is much higher at the end of the transient due to this unnecessary adjustment by the controller. However, this behavior does not take place if the FTC is implemented. Notice that, the same figure shows that when the sensor fault adaptation module (Fig. 3) is activated, the right signal is taken, therefore FTC system correction is involved and the reactor temperature value is maintained in $T = 435$ °K when steady state is reached.

Figure 7 shows how the residues (r_T, r_{c_A} and r_V) indicate the presence of a fault in temperature sensor according to Table 2.

4.7 Volumen sensor fault

In this simulation is shown a change in the volume set-point at $t = 5$ min from $V = 100$ L to $V = 110$ L, and then a 5% degradation in the volume sensor at $t = 35$ min.

Figure **??** shows the system dynamic response without fault-tolerant control. The feedback signal from the volume sensor fails abruptly into −5%, monitoring a wrong value ($V = 104.5$ L). As a result of this fault, the control loop corrects the measurement error in order to maintain the level at 110 L. Thus, according to the fault assumed the monitored volume by the sensor suggests to the controller that is correcting properly.

Furthermore, in this figure can also be seen how the FTC reacts an instant later in which it follows that a fault has occurred, the sensor fault adaptation module switches the signal from the observer bank according to Fig. 3. Then, the controlled variable retrieves the set-point value while the measured signal from the sensor continues to the wrong value.

Notice that according to the introduced changes in temperature set-point in $t = 15$ min does not influence the dynamics of the reactor volume, this being consistent with the physics of the problem.

Figure 8 shows how the reactor temperature evolves for both volume set-point change and with volume sensor fault. Notice that when a volume set-point change is introduced, a change in the reactor temperature is observed, forcing the temperature control loop to correct it ($t = 5$ min). Then, in the instant $t = 35$ min, the temperature corrections are made by the control loop. Figure 8 shows this corrections with and without FTC.

Figure 9 shows the dynamic response of temperature and the manipulated coolant flow for the problem with and without FTC, for the time interval between $t = 30$ min and $t = 50$ min. Based on the dynamic response without FTC it is possible to conclude

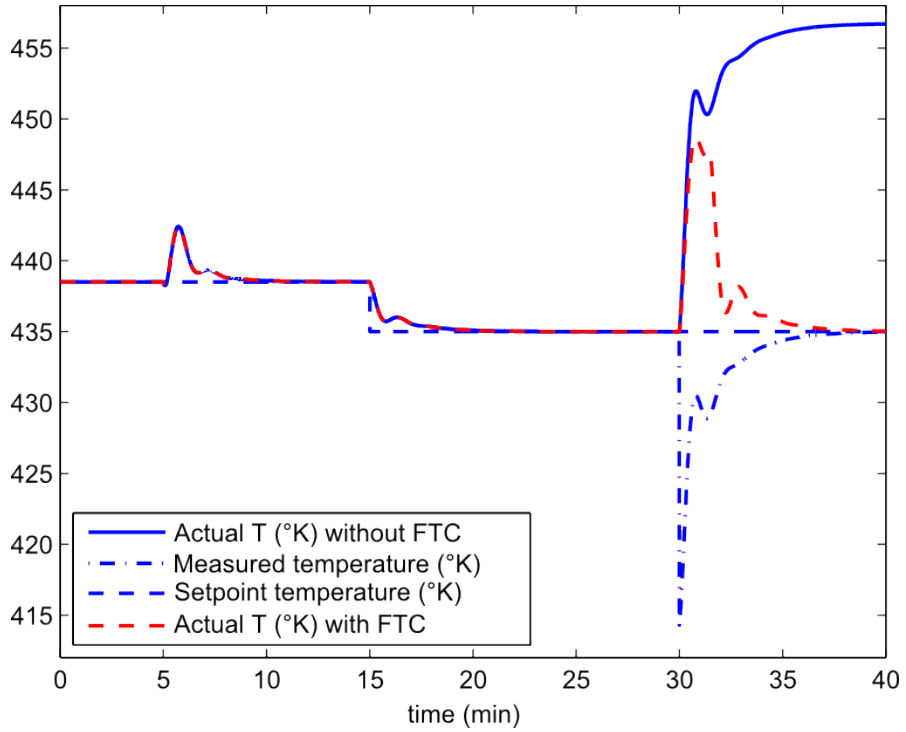


Figure 5: Actual CSTR temperature without FTC, measured temperature in the CSTR, set-point and temperature with FTC. Notice that the following changes was implemented: (1) a set-point change in the reactor volumen at $t = 5$ min, (2) a set-point change in the CSTR temperature at $t = 15$ min and, (3) a 5 % degradation in the temperature measuring element at $t = 30$ min.

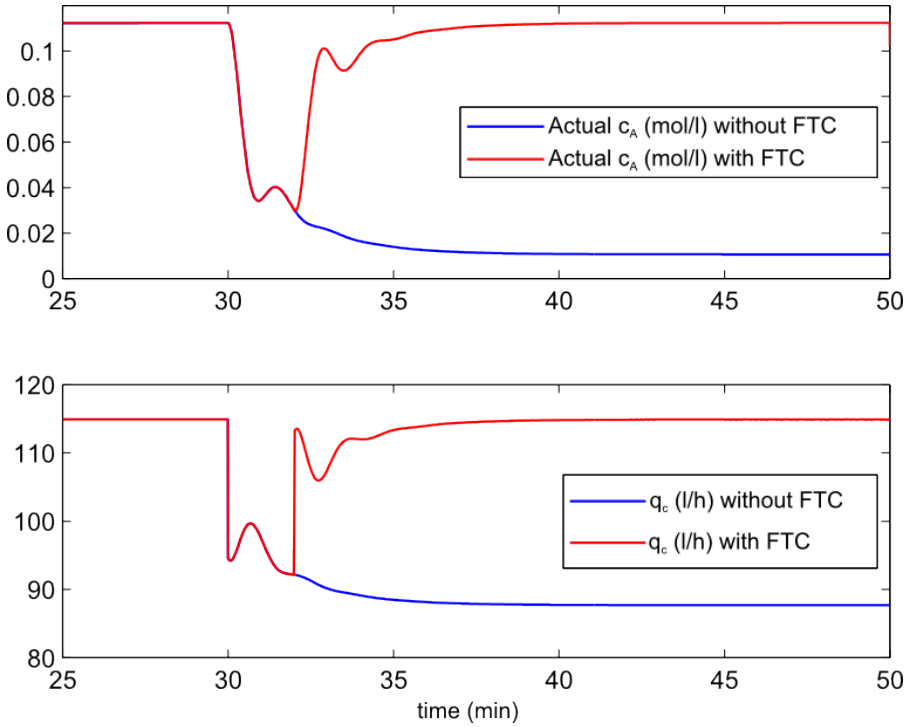


Figure 6: (a) Dynamic response of reactant concentration (c_A) when a degradation of temperature sensor take place with and without FTC. (b) Dynamic response of manipulated variable (q_c) with and without FTC.

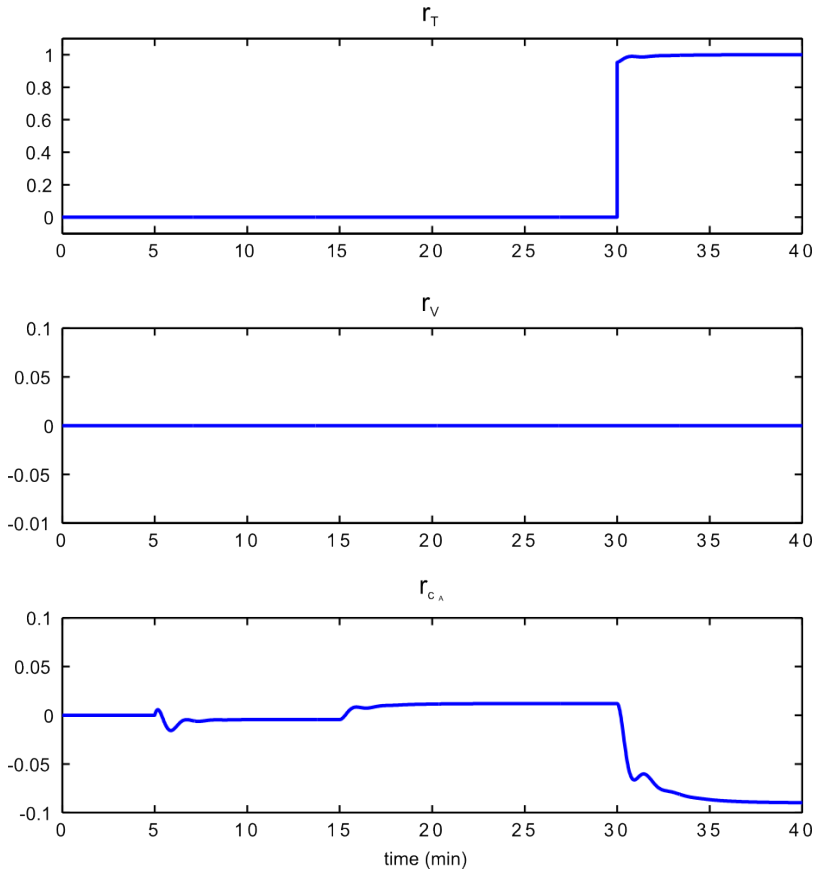


Figure 7: Residues of r_T, r_V and r_{c_A} to a fault in the temperature sensor.

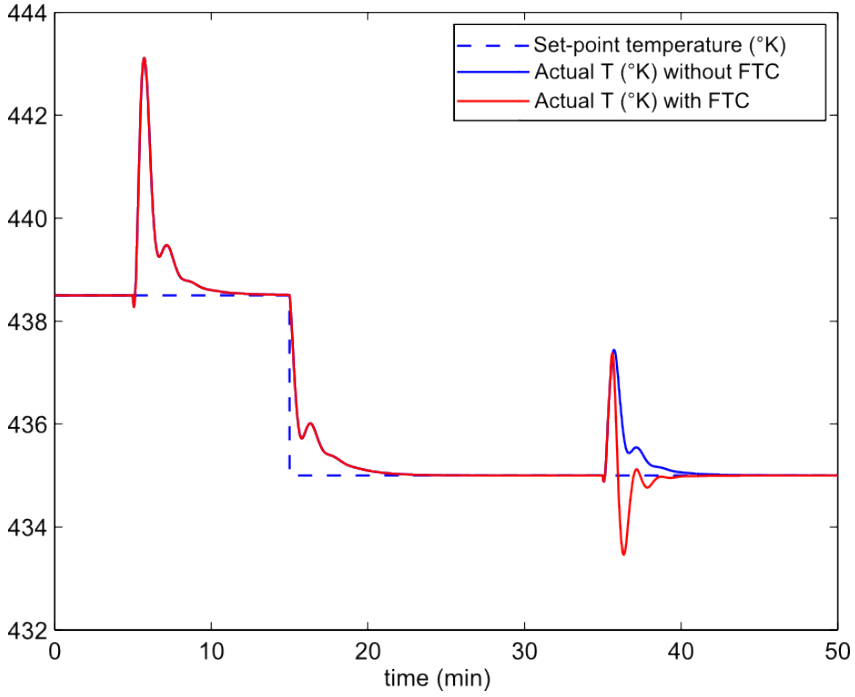


Figure 8: Actual and measured temperature and CSTR set-point. Notice that the following changes: (1) a set-point change in the volume at $t = 5$ min, (2) a set-point change in the CSTR temperature at $t = 15$ min and, (3) a 5 % degradation in the volume measuring element at $t = 35$ min.

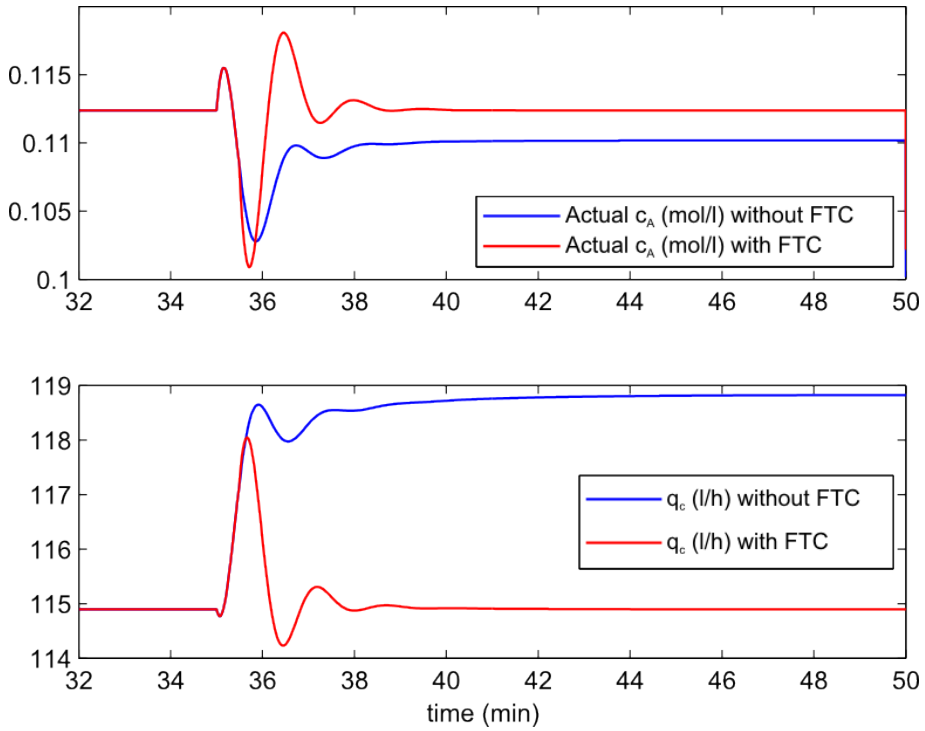


Figure 9: (a) Dynamic response of reactant concentration change (c_A) when a fault in the volume sensor take place with and without FTC. (b) Dynamic response of manipulated variable (q_c) with and without FTC.

that due to the actual volume of the reactor is not in the set-point value, it is necessary to increase the coolant flow to extract the heat required to bring the temperature back to the set-point.

5 Conclusions

In this paper, the design of an active fault tolerant control was presented, showing an adaptation to a fault in both an actuator and a sensor. To achieve this, two schemes that allow to compute residues through analytical redundancy was presented. The main proposal of this work is the use of observer banks to diagnostic sensor and actuator faults. In this context, output observers were used to detect sensor fault while an unknown input observer was used to identify manipulated variable fault.

This fault diagnosis strategy is applied to a CSTR, showing that the methodologies presented lead to satisfactory results. Furthermore, it is emphasized that the chosen example has specially interest in the chemical industry and, although there is many examples in the literature with non-linear systems using this methodology ([13, 8, among others]), but there is few references in FDI applied to CSTR [20, 3].

Finally, this work allows to understand even better the behavior of the CSTR when different faults taken placed.

Conflict of Interest The authors declare no conflict of interest.

Acknowledgment The authors would like to thank the financial support received by the Universidad Nacional del Litoral by means of its CAI+D program.

References

1. E. J. Adam, Instrumentación y Control de Procesos. Notas de Clase, Ediciones UNL, 2nd edition, 2014.

2. M. Blanke, M. Kinnaert, J. Lunze, and M. Staroswiecki, Diagnosis and Fault-Tolerant Control Systems, Springer, 2006.

3. F. Caccavale, M. Iamarino, F. Pierri, and V. Tufano, Control and Monitoring of Chemical Batch Reactors, Springer, 2011.

4. J. Chen, R. J. Patton, and H. Y. Zhang, "Design of unknown input observers and robust fault detection filters" Int. Journal of Control, **63**(1), 85–105, 1996.

5. P. M. Frank, "Fault diagnosis in dynamic system using analytical and knowledge-based redundancy – a survey and some new results" Automatica, **26**(3), 459–474, 1990.

6. J. Gertler, Fault detection and diagnosis in engineering systems, Marcel Dekker, Inc., 1998.

7. M. Hou and P. C. Müller, "Design of observers for linear systems with unknown inputs" IEEE Trans. On Automatic Control, **37**(6), 871–875, 1992.

8. R. Isermann, Fault-Diagnosis System, Springer, 2006.

9. W. Luyben, Process Modeling, Simulation and Control for Chemical Engineers, Mc Graw Hill, 1990.

10. M. Mahmoud, J. Jiang, J. Lunze, and Y. Zhang, Active Fault Tolerant Control, Springer, 2003.

11. J. D. Morningred, B. E. Paden, D. E. Seborg, and D. A. Mellichamp, "An adaptive nonlinear predictive controller" Proc. Amer. Contr.Conf., 2, 1614–1619, 1990.

12. H. Noura, D. Sauter, F. Hamelin, and D. Theilliol, "Fault-tolerant control in dinamic systems: Application to a winding" IEEE Control Systems Magazine machine, **20**(1), 33–49, 2000.

13. H. Noura, D. Theilliol, J.C. Ponsart, and A. Chamseddine, Fault-Tolerant Control Systems, Springer, 2009.

14. J. O'Reilly, Observers for Linear Systems, Academic Press, 1983.

15. R .J. Patton and J. Chen, "Observer-based fault detectiona and isolation: Robustness and applications", Control Eng. Practice, **5**(5), 671–682, 1997.

16. M. F. Picó and E. J. Adam, Advantages in residual methods for fault detection and isolation problem: application in laboratory experimental dc-motor in XIV Reunión de trabajo en Procesamiento de la Información y Control, RPIC 2011, Oro Verde, Argentina, 2011.

17. M. F. Picó and E. J. Adam, Comparing fault detection and isolation methods applied to a laboratory experimental servomotor, CAIP 2011. X Cong. Interam. de Computacin Aplicada a la Industria de Procesos, Girona, Espaa, 2011.

18. M. F. Picó and E. J. Adam, Control tolerante a fallas en un sistema de tanques aplicando redundancia analtica en el diagnstico de fallas, CIETA 2012. IX Cong. Int. Electrnica y Tecnologas de Avanzada, Pamplona, Colombia, 2012.

19. M. F. Picó, "Diseño de Controladores Tolerante a Fallas Aplicados a Procesos de la Industria Química", Mag. Thesis, Univeridad Nacional del Litoral, 2015.

20. O.A.Z. Sotomayor and D. Odloak, "Observer-based fault diagnosis un chemical plants", Chemical Engineering Journal, 112:93–108, 2005.

21. G. Stephanopoulos, Chemical Process Control, Prentice Hall, 1984.

22. D. Theilliol, H. Noura and J.C. Ponsart, "Fault diagnosis and accommodation of a three-tank system based on analytical redundancy", **41**(3), 365–382, 2002.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 182-185 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Development and Validation of a Heat Pump System Model Using Artificial Neural Network

Nabil Nassif[1*], Jordan Gooden[2]

[1]*Associate Professor, Department of Civil and Architectural Engineering, North Carolina A&T State University, 27411, USA*

[2]*Graduate Student, Department of Civil and Architectural Engineering, North Carolina A&T State University, 27411, USA*

ARTICLE INFO

Article history:
Received: 11 March, 2017
Accepted: 09 April, 2017
Online:24 April, 2017

Keywords :
Heat Pump
Artificial Neural Network
Air Conditioning

ABSTRACT

Modeling of direct expansion (DX) air conditioning and heat pump systems can be necessary in developing energy saving methods required to reduce energy consumption in buildings. The artificial neural networks (ANN) can be simple and reliable as compared to traditional methods. A properly trained artificial neural network can provide accurate results, while being relatively straightforward and easy in development. This paper discusses the implementation and validation of an artificial neural network modeling technique to predict the performance of a DX air conditioning and heat pump. The model predicts the compressor power as a function of airflow rate, humidity ratio, ambient and mixed air temperatures. Three different leaning algorithms were compared and validated versus the actual data using statistical indexes to determine the most accurate model structures. Experiments were conducted on a 3-ton DX split-system heat pump fully implemented. The heat pump ran over the course of several months to obtain a wide range of measurements. The results showed that artificial neural network can provide very accurate predictions and this ANN model technique can effectively be used for many energy-efficiency heat pump applications.

1. Introduction

As energy concerns continue to grow, the need for developing more efficient building energy systems has increased. The global contribution from buildings toward energy consumption, both residential and commercial, has steadily increased over the years, with estimates between 20% and 40% in developed countries [1]. Unitary air conditioning and air-source heat pump systems are one of the most common energy systems for both commercial and residential HVAC applications in USA. Those systems can provide the desired thermal comfort for various building types and sizes. Significant advancements have been made in the design of air conditioning and heat pump systems, including advanced cycle components with support for new refrigerants, improved cycle designs, and expanding their range of applications. One study found the potential to reduce energy consumption by up to 80% through the development of better compressor technology [2]. Many of these systems in use today are designed improperly, and with newer systems becoming increasingly complex; more effective modeling methods are required [3].

Modeling of DX air conditioning system that can be integrated into many energy solution tools is necessary to reduce energy operation cost of those systems [4].

This paper proposes the use of data-based model using artificial neural networks to accurately predict the performance of heat pump systems. The artificial neural network capabilities were extended from linear to nonlinear domains in 1974, for developing a method as general as linear regression with nonlinear capability [5]. Artificial neural network modeling has become increasingly popular for a variety of disciplines, including signal processing and character recognition, stock market prediction, logistics, among others. The main advantages of ANN modeling compared to other methods such as expert systems are its simplicity, speed, and ability to handle a multivariable problem to solve complex relationships between variables. Over the last two decades, the use of ANN systems in HVAC system field has been steadily increasing to solve these complicated problems [6-8]. Bachtler and

*Corresponding Author: Nabil Nassif, Civil and Architectural Engineering, North Carolina A&T State University, Greensboro, USA. Email: nnassif@ncat.edu

others used a generalized radial basis function neural network for predicting the steady-state performance of a vapor-compression liquid heat pump [8]. Nassif predicted and optimized various HVAC systems using artificial neural networks and genetic algorithm [4,9].

The proposed modelling technique was tested by performing experiments on a 3-ton DX split-system heat pump. The experiment was prepared utilizing a specially integrated unit in a controlled setting. The heat pump ran under various outside conditions and in heating and cooling modes under. Various ANN model structures and training algorithms were tested and evaluated.

2. Artificial Neural Network

Artificial Neural Networks (ANNs) are non-linear mapping structures that are based on the principle functions of the human brain. They are extremely powerful tools for modeling, especially when the mathematical relationship of a given data is unknown, or not easily discerned. Over the years, they have become the focus of attention, largely due to their wide range of applicability and the ease in which they can work with complicated problems. ANNs are capable of identifying and learning correlated patterns between the input data sets and corresponding target values. After training, ANNs can be used to predict the outcome of new independent input data. They have been used for a large number of applications where statistical methods are traditionally used. Neural Network models are now being employed to solve problems, where in the past was normally solved through logistic regression, multiple regressions, Bayes Analysis and other classical methods. They consist of simple computational units called perceptrons, which are also known as neurons, which are highly interconnected.

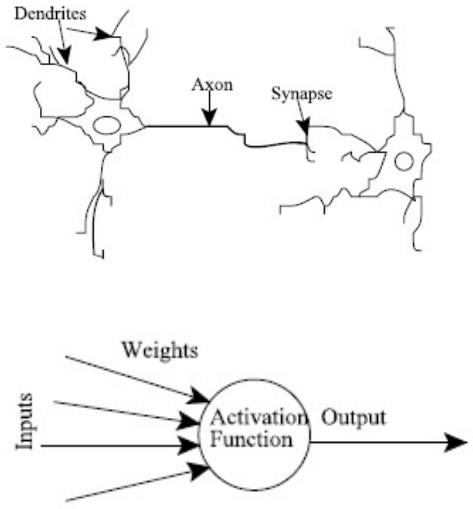


Fig. 1. An artist conception of a natural neuron (top) and an artificial neuron (bottom)

While in actual neurons the dendrite receives electrical signals from the axons of other neurons, with artificial neurons these electrical signals are represented as numerical values. At the synapses between the dendrite and axons, electrical signals are modulated by various quantities. This is also modeled in artificial neurons by multiplying each input by a value called the weight. The higher a weight of an artificial neuron is, the stronger the input which is multiplied by it will be. The calculated weighted sum of the inputs represents the total strength of the input signals. This is then applied to a step function (also known as a Sigmoidal Function) to the sum to determine its output. The weights in most neural networks can be either positive or negative, which have activation or inhibitory influence to each input. Fig. 1 shows an artist conception of a natural neuron (top) and an artificial neuron (bottom). Artificial neural networks need appropriate learning algorithms to be trained. Three learning algorithms were tested: (1) Levenberg Marquardt, (2) Bayesian Regulation, and (3) Conjugate Gradient algorithm. The MATLAB ANN toolbox was used in this study.

3. Experimental Setup

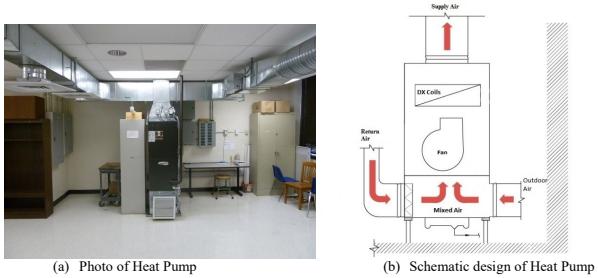
The experimental tests were conducted on a typical heat pump fully instrumented and installed in the HVAC laboratory located at North Carolina A&T State University. This laboratory is equipped with a fully integrated 3-ton DX split-system air conditioning AC as shown in Fig. 2 and controlled by Building Automation and Control network (BACnet) building automation system (BAS). The BAS collects measurement data of nearly all components, including the required measurements such as the outdoor air conditions, temperature and humidity ratio entering and leaving the DX coil, airflow rates, damper positions, and compressor power. Airflow can be changed by varying the position of the return and outside dampers. The airflow through the variable air volume (VAV) box is controlled by two DDC PID cascade control loops; the first control loop determines the set point of airflow rate, and the second controls the VAV damper position. The data collected from BAS can be accessed online, or from a device connected directly to the server's network. The heat pump ran at various intervals covering a wide range of operation conditions. The data collected were the airflow rate, outside, supply, and mixed dry-bulb temperatures, humidity ratio, and compressor power. The ANN developed uses the airflow rate, outside mixed air temperatures and humidity ratio as model inputs. The ANN model outputs were the supply air temperature and compressor power, which were consequently designated as the target variables. In the heating mode, the same input variables applied, except the humidity ratio, as it was not a factor in effecting compressor power or supply air temperature. The artificial neural networks ran multiple times using three different training algorithms, and by varying the number of neurons. These results were then compared against the actual data to determine the most accurate neural networks and model parameters.

The heat pump was first tested to check if its components are all working appropriately. It was found that the mixed air temperature sensor gave readings higher than expected one on the system monitor. This is a common mixed air temperature reading problem in many DX systems. A study conducted by Avery [10] found that mixed air temperature sensors are often unreliable, due to air stratification and the accompanying design documents which rarely detail proper sensor installation. To resolve this problem, the mixed air temperature T_m and humidity ratio W_m are calculated by applying energy conservation equation as function of outdoor air humidity W_0 and temperature T_0 and return air temperature T_r and humidity ratio W_r:

$$W_m = \frac{m_o}{m_a} \cdot W_o + \frac{m_a - m_o}{m_a} \cdot W_r \quad (1)$$

$$T_m \approx \frac{m_o}{m_a} \cdot T_o + \frac{m_a - m_o}{m_a} \cdot T_r \qquad (2)$$

Where the m_a is mass flow rate of the supply air and m_o is mass flow rate of outdoor air. The equations 1 and 2 are applied only when the difference between the return air temperature T_r and outdoor air temperature T_o is relatively large (higher than 5 ºF).



(a) Photo of Heat Pump (b) Schematic design of Heat Pump

Fig 2. Split System Heat Pump for experimental Studies

To test neural network model, a wide range of operating performance data is needed. Therefore, the system was run at different outside conditions and various times of the year to have more extreme outdoor conditions. At different conditions, the heat pump ran for a few hours, this repeats throughout the months of April through September. The building automation system was set to record the measurements every five-minute interval, and more than 891 performance data points were collected. Approximately 618 data points are used to train the ANN. Roughly 70% of each dataset were designated specifically for training, and the remaining 30% used solely for testing. While there are several different model parameters within an artificial neural network which affects its performance, the most significant factors are number of hidden layer neurons and how the training is performed (training algorithm). In this paper, the neural networks as shown in Fig 3 were configured to run under every permutation between the three different training algorithms and different number of neurons, ranging from 5 to 45. Furthermore, these networks were tested using multiple and single variable outputs.

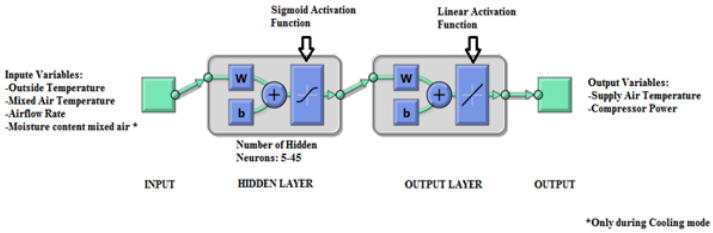


Fig. 3. ANN Model with inputs and outputs

4. Results

As mentioned before, the 891 performance data points were collected and divided into two sets: (1) training data set (618 data points) and (2) testing data points (273). Three learning algorithms Levenberg Marquardt, Bayesian Regulation, and Conjugate Gradient were used to train the neural network model to predict the DX heat pump power. After completing the ANN training, the model was then tested and evaluated. The evaluations were done by comparing the resulted coefficient of variance (COV) for each model options. The COV is calculated by divided the square mean square error to the mean data. The training and testing are done for various number of neurons to find the best or optimal model. The optimal model is determined based on the lowest COV obtained from the testing data set. Fig.4-6 show the COV results for the training and testing data sets and with various training algorithms.

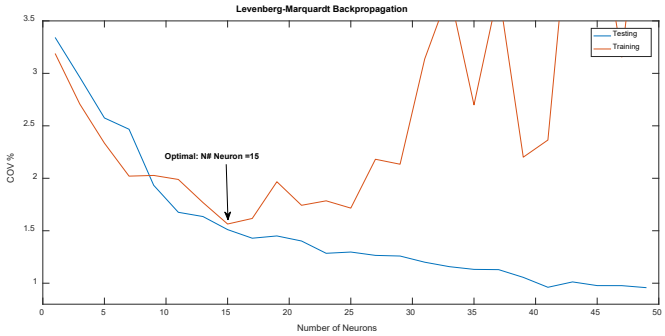


Fig 4. COV results as a function of number of neurons for Levenberg-Marquardt training algorithm

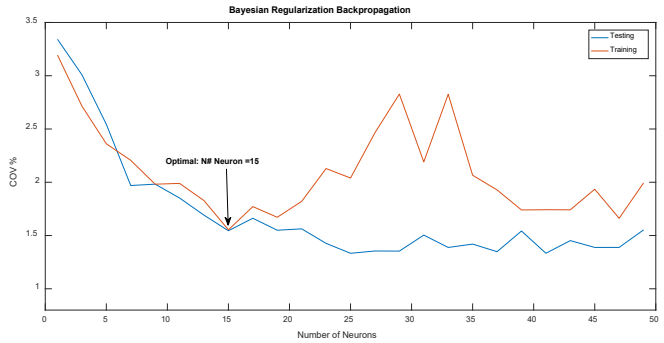


Fig 5. COV results as a function of number of neurons for Bayesian Regularization training algorithm

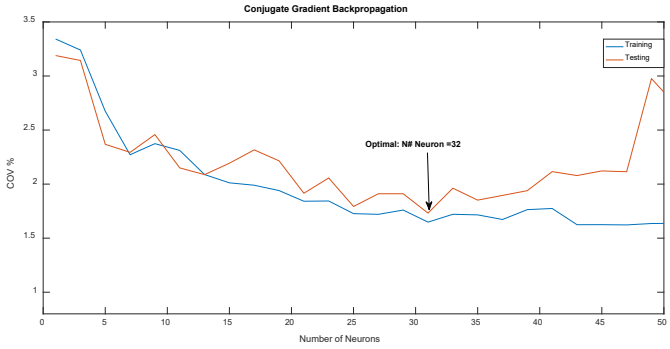


Fig 6. COV results as a function of number of neurons for Conjugate Gradient training algorithm

By increasing the number of neurons, the COV is decreased for training data set. However, for the testing data set, the COV is decreased and then increased. There is an optimal value of number of neurons with the lowest resulted COV that makes the model predict accurately the power for training data set. The optimal number of neurons are 15, 15, and 32, and the lowest COVs are 1.54%, 1.56%, 1.73% for training algorithms Levenberg-Marquardt, Bayesian Regularization, and Conjugate Gradient. Fig. 7 shows the comparison between the simulated and measured heat pump power. The simulation is done by the ANN model with the optimal number of neurons (#15) and training algorithm Levenberg-Marquardt. These results show that the ANN model

can predict accurately heat pump power with COV of 1.54%. The ANN model technique can be used for diffident energy efficiency applications.

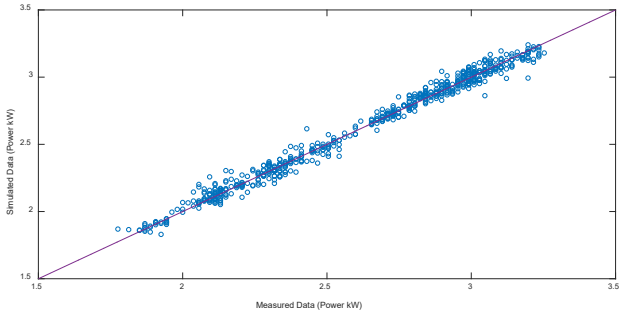


Fig 7. Comparisons between the measured data and simulated data calculated by ANN

5. Conclusion

Various artificial neural network model structure and leaning algorithms were developed and tested. The model predicts the compressor power as a function of airflow rate, humidity ratio, ambient and mixed air temperatures. Experiments were conducted on a 3-ton DX split-system heat pump to collect a wide range of performance measurements. The measurements were divided into training and testing sets. The models were evaluated by comparing the resulted coefficient of variance COV for each model structure and learning algorithm options. The optimal number of neurons are 15, 15, and 32, and the lowest COVs are 1.54%, 1.56%, 1.73% for training algorithms Levenberg-Marquardt, Bayesian Regularization, and Conjugate Gradient. The validation results showed that artificial neural network can provide very accurate predictions and this ANN model technique can effectively be used for many energy-efficiency heat pump applications

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This work was performed under the following financial assistant award 70NANB14D1319 from the U.S. Department of Commerce, National Institute of Standards and Technology.

References

[1] Pérez-Lombard, L., J. Ortiz, and C. Pout, "A review on buildings energy consumption information" Energy and buildings **40**(3), 394-398, 2008

[2] Chua, K. J., S. K. Chou, and W. M. Yang, "Advances in heat pump systems: A review" Applied Energy **87**(12), 3611-3624, 2010

[3] Nassif, N., Rice, K., and Shen, B, "Development of a Matlab-Based Integrated Model for Optimal Design and Operation of Heat Pumps", CD 2015 ASHRAE Annual Conference, Atlanta, GA

[4] Nassif, N, "Modeling and Optimization of HVAC systems using Artificial Neural Network and Genetic Algorithm", International Journal of building Simulation **7**(3), 237-245, 2014.

[5] Werbos PJ, "Beyond regression: new tools for prediction and analysis in the behavioral sciences", Ph.D. dissertation. Cambridge, (MA): Harvard University, 1974.

[6] Mohanraj M, Jayaraj S, Muraleedharan C, "Energy assessment of direct expansion solar assisted heat pump working with R22 and R407C/LPG mixture", International Journal of Green Energy (7): 65–83, 2010

[7] Dauphin, Yann and Bengio, Yoshua, "Big Neural Networks Waste Capacity" University of Montreal, QC Canada. March 2013.

[8] Bacthler H, Browne MW, Bansal PK, Kecman V. "Neural networks—A new approach to model vapour compression heat pumps", International Journal of Energy Research (25), 591–9, 2001.

[9] Nassif, N., M. Arida, and R. Talib, "Development and Testing of Building Energy Model Using Non-Linear Auto Regression Neural Networks", 2016 ASHRAE Meeting, St Louis, MO.

[10] Avery, Gill, "Do Averaging Sensors Average?" ASHRAE Journal (12), 42-43, 2002

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 186-190 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Assessment of the usefulness and appeal of stigma-stop by psychology students: a serious game designed to reduce the stigma of mental illness.

Adolfo J. Cangas[1,*], Noelia Navarro[1], Juan J. Ojeda[2], Diego Cangas[2], Jose A. Piedra[2], José Gallego[4]

[1]*University of Almería, Department of Psychology, 04120, Almería, Spain*

[2] *University of Almería, Department of Computer Science, 04120, Almería, Spain*

[3] *University of Almería, Department of Education, 04120, Almería, Spain*

ARTICLE INFO

Article history:
Received: 05 March, 2017
Accepted: 11 April, 2017
Online: 24 April, 2017

Keywords :
Stigma
Mental health
Video games

ABSTRACT

The present work describes the first serious game designed to reduce the stigma among students towards mental health problems. The game is called Stigma-Stop, and it features characters who suffer from various mental disorders. Players are provided with information about different mental illnesses, and they are presented with several options on how to act when they encounter characters with these problems. In addition, the game questions the participants as to whether they have ever felt like the individuals portrayed in the game, with the goal of fostering empathy for those that suffer from these types of disorders. Stigma-Stop was applied to a sample of second-year university Psychology students to evaluate the game's usefulness and appeal. The results demonstrate the importance of this game and that these students consider it to be highly useful. The most notable characteristics are described in depth in the present paper.

1. Introduction

The present work is an extension of the study presented at the 8th International Conference on Games and Virtual Worlds for Serious Applications (VS-GAMES) called "Stigma-Stop a Serious Game against the Stigma in Mental Disorders" [1]which focused on developing the first serious game designed for the purpose of reducing the stigma towards mental illnesses.

The game features characters with various mental disorders (schizophrenia, agoraphobia, depression, and bipolar disorder). The objective is for the player to convince all the characters to work towards a common goal, which is to form a group to participate in a contest. Over the course of the game, the player is given different options on how to behave when interacting with the characters; the game also provides them with information on which is the most appropriate choice. It gives further details about each of the mental disorders and asks the players questions which seek to foster empathy towards those who suffer from these problems. The video game also features four mini-games, whose ultimate goal is to provide knowledge and dispel traditional myths about mental health [2-5].

In a previous study, the game was successfully used with a sample of 168 secondary school students. In this case, stigma towards mental health disorders was significantly reduced among the students who played the video game, compared to the control group which played other video games and did not display any changes [6].

In this way, Stigma-Stop seeks to familiarize young people with mental disorders, particularly at this age because most of these problems manifest themselves during adolescence. The game intends to increase their knowledge on this subject and strengthen their awareness, while emphasizing the importance of biographical and/or contextual experiences related to these illnesses and that anyone can experience similar problems over the course of their life. The game teaches how to interact with those who suffer from these disorders, fostering interest in helping these individuals. This aspect is of vital importance considering young people are

*Corresponding Author: A.J. Cangas, Departament of Psychology, University of Almeria. Crta Sacramento, s/n CP04120, La Cañada de San Urbano, Almería, Spain. E-mail: ajcangas@ual.es

acquiring and consolidating values they will act on as future citizens.

The present study intends to analyze Psychology students' opinion of Stigma-Stop. Their point of view is quite relevant considering, firstly, they will be future working professionals in the field of mental health, and, secondly, they are at an age (the sample was of second-year Psychology students) when they are very familiar with the use of video games.

The game is optimized to be used in schools and can be applied on an individual basis or in groups. This flexibility makes it possible to initiate discussions in class about these topics, where they can be addressed "naturally" and any misconceptions can be identified and clarified.

2. Method

2.1. Sample

The sample consisted of 26 students in the second year of their Psychology degree at the University of Almería (Spain). Participants were between the ages of 18 and 28 (X=20.12; S.D.=2.85). In this group, 76.92% were female, and the remaining 23.08% were male.

2.2. Tools

Stigma-Stop. This is a serious game designed in a non-immersive virtual reality environment. Stigma-Stop was developed with Unity3D software for three platforms (PC, website and Smartphone application) and it is aimed primarily at young people between the ages of 14 and 21. The player can choose the sex of the avatar they are going to play with. During the game, different characters appear who suffer from various mental disorders (depression, schizophrenia, bipolar disorder and agoraphobia). Figure 1 shows some of the scenes from the program.

Figure 1. Scenes from "Stigma-Stop".

The objective of the player is to interact with each one of the characters and convince them to contribute their knowledge to a project and work towards a common goal: design a video game for a contest. The player does not follow a predefined plot in the game. The player's freedom to choose the character who they visit at any given point in the game was incorporated using a machine of finite states. This machine contains each one of the situations that the user could experience and keeps track of the situations that the user has already completed so each character is visited only once. The user interacts with the characters through dialogs which reveal some of the symptoms of the various disorders. Once the objective of the dialog has finished, the player is presented with three options for solving different dilemmas that arise, only one of which is correct. If one of the less appropriate choices is selected, information is provided explaining why a different option should be chosen and the player is given another opportunity to select a new one. Similarly, a short form will appear with questions about general knowledge concerning each disorder and the user is able access additional information about the current condition of the characters and what circumstances led them to develop these types of problems. Finally, after meeting each character, the player is asked if they think the character is emotionally well, if they themselves have ever felt the same, and if they think they would be able help that person (specifying how). User answers are stored in a database located in an external server. Communication between said database and the serious game is carried out using RESTful Web Services equipped with PHP at the server end. The information transmitted by these Web Services will be in JSON format. When the user chooses the correct option, they will be redirected to the "mini-game" state, where they can play with one of four mini-games included in the program. Each one focuses on a mental illness and provides the user with more relevant information, apart from that already shown throughout the game. The four mini-games are (figure 2):

• Memory: In this mini-game players must match faces of famous people who suffered from different mental disorders to make pairs. These famous individuals include Leonardo da Vinci, Edgar Allan Poe, Vincent van Gogh, John Forbes Nash, Charles Darwin and Robert Schuman. The objective is to complete the game in as few turns as possible. In this way, players will learn that various famous individuals throughout history suffered from some type of disorder. Once they have matched the two cards of the same character, information is displayed about that person (name, profession, mental illness, etc.).

• Trivial: In this mini-game the user will have to answer questions related to mental disorders, specifically whether depression is something that only happens to "weak" people; If we are ever depressed, it means that we will be depressed forever; Whether a depressed person can be helped; If people with mental disorders, especially those diagnosed with schizophrenia, are more dangerous, aggressive and unpredictable than those who do not have a mental disorder; and whether people with mental disorders can work. The questions have been designed in the form of affirmation so that the user answers 'True' or 'False'. In case the person is wrong the program tells then why it is incorrect and they will have to respond again.

• Running with Dali: In this mini-game the user must move through a scene in as little time as possible, while avoiding obstacles and collecting pieces of paintings. If the player hits an obstacle, their time will increase by a specific increment, which depends on the obstacle itself. When they collect one piece of a

painting, the time on the clock will decrease. Once the game finishes, the pieces of the painting collected will be displayed along with a brief biography of the painter Salvador Dali, who also suffered from a mental illness.

• Stigma Shooter: This mini-game contains concepts related to mental illness. The job of the player is to "trap" positive concepts (such as overcoming, improving, independence, acceptance, able to work) and "destroy" negative concepts (unpredictable, dangerous, vague or aggressive) by shooting them.

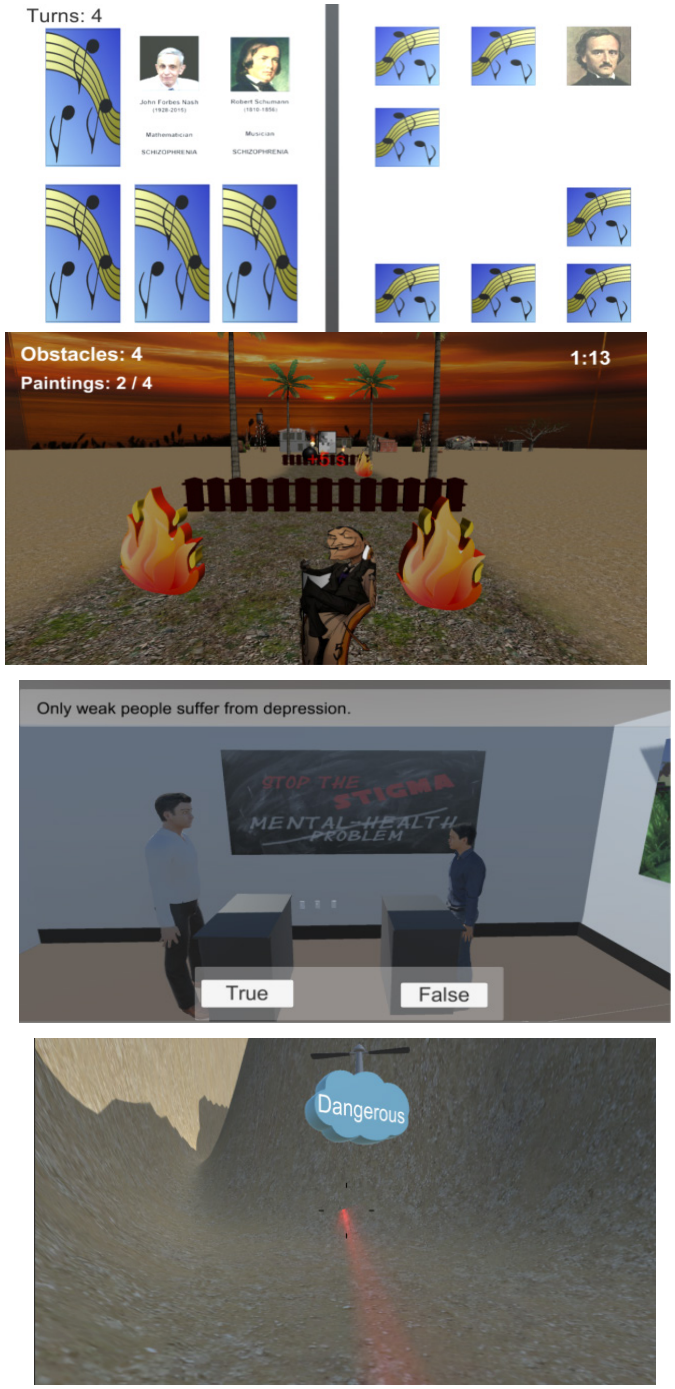


Figure 2. Four mini-games

At the end of each mini-game, the player's results are displayed, and users can see details about their performance. These results will also be stored in the database previously mentioned with the objective of ranking participants and motivating players.

In this serious game artificial intelligence resources have been included to make the game seem more realistic. Routes have been created to simulate the movement of pedestrians and cars through the city, thus giving the sensation of activity. These paths are formed by a set of waypoints [7], which contain the position of the three-dimensional space and are assigned to the position of the 3D model. In addition, each route has a certain speed associated with the movement of the 3D model to ensure that the behavior of pedestrians and cars do not have the same pattern, which would give a feeling that their behavior is predictable and artificial.

The cars, apart from using the waypoints, have an algorithm implemented so that when they detect the presence of a character on the road they do not pass over it and stop until the character is at a safe distance, at which time the vehicle continues its path. The scenes of the city where the user interacts with their character are designed so that the avatar can only cross at the pedestrian crossings and when it is detected that the user wants to pass the road, an invisible wall is created to prevent passage and force them to cross at the crosswalk.

Decision trees [8] is another strategy that has been used in this video game regarding the animation of the characters. As mentioned before the protagonist has to visit his friends to achieve the goal. In each of the visits to his friends, the protagonist will engage in a conversation with his friend. These dialogues are composed of a series of animations that give the 3D models movement. Decision trees determine the animation of the different characters; depending on who the protagonist is talking to, certain actions or others will be shown. The options that the user chooses in response to the situation that arises will be another factor that influences the animations chosen during the dialogue.

Stigma-Stop assessment questionnaire. It evaluates on a scale of 0-10 the degree of attractiveness of the video game and also the interest it arouse. They should also describe what they have learned from the program and whether or not they would recommend its use to a friend. Likewise, for each of the characters that appear in the video game, participants are asked to indicate whether they believe they are psychologically well, if they think they could help the character and if they have ever felt like them.

3. Procedure

The experimental application of the video game took place at the University of Almería in quiet room equipped with tables, chairs and a desktop computer connected to a projector. The activity was conducted in group format in which four participants interacted directly with the video game. Each member of the group visited one character while the others observed on the projector screen. After finishing the program all students completed the Stigma-Stop assessment questionnaire.

4. Results

In terms of the quantitative data, on a scale of 1-10, the participants gave the program an average score of 8.08 (SD=1.35) for its usefulness and 6.58 (SD=1.21) for its entertainment value. In addition, the 26 participants (100%) indicated that they would recommend the game to a friend. As for the question: What do you think you have learned from Stigma Stop? responses were

classified as follows (bearing in mind that some responses may be included in more than only category):

1. Answers that emphasized the educational value of the video game in terms of the information it provides regarding the symptoms of the various disorders in the game, making them easier to recognize. (e.g. answers: "more easily diagnose a mental illness since the symptoms are clearly shown"; "try to diagnose the characters and help them").

2. Answers that specifically highlighted the acquisition of intervention strategies for dealing with the disorders, promoting prosocial and helpful behavior. (e.g. answers: "teaches ways in which you can help people with these disorders"; "how to treat these individuals and how to behave"; "how we should help to improve the situation of the person").

3. Answers that mainly emphasize the usefulness of the game in raising awareness against stigma, dispelling myths, and normalizing the disorders by considering them responses to specific adverse biographical and/or contextual circumstances. (e.g. answers: "end specific stigmas associated with the label of mental disorder"; "the experiences of people are very important for understanding their unusual behavior").

The students' answers were distributed quite evenly among the three categories. 56% of the responses were classified under the second option (teaching intervention strategies) and 52% were categorized under the first and third options (usefulness in providing information on symptoms and capacity to raise awareness against social stigma, respectively).

Table 1. Participant responses regarding characters in percentages

	Panic disorder with agoraphobia (%)		Schizophrenia (%)		Bipolar Disorder (%)		Depression (%)	
	YES	NO	YES	NO	YES	NO	YES	NO
Do you think the character is emotionally well?	4	96	4	96	27	73	0	100
Do you think you could help this person?	100	0	88	12	79	21	92	8
Have you ever felt like this person?	11	89	0	100	31	69	42	58

With respect to the questions where participants indicated whether they thought the character was emotionally well, nearly 100% responded 'no' in every case except for bipolar disorder, which registered at 73 %. These percentages remained similar regarding whether they thought they could help the characters after discovering their problems; virtually all responded 'yes' except in the case of bipolar disorder (79 %). With respect to question as to whether participants had themselves ever felt like one of the characters, 42 % responded 'yes' in the case of depression, less for bipolar disorder (31 %) and considerably less for panic disorder with agoraphobia (11 %). None of the participants identified themselves as having ever experienced schizophrenia (Table 1).

5. Discussion

Serious Games have shown to be highly beneficial in various fields of education [9], problem solving [10], psychological interventions [11] and even job-oriented development of university students [12]. Nevertheless, "Stigma Stop" is the first Serious Game to date that focuses on providing information about mental illnesses and raising awareness among young people about this topic, thereby promoting sensitivity and empathy towards people who suffer from mental health problems.

In the case of the Psychology students in this study, their assessment of the game was highly favorable. This is very significant considering we are dealing with future healthcare professionals, most of whom will work with individuals suffering from mental health disorders. Their evaluations were especially positive in terms of the game's usefulness, yet they were somewhat lower with regard to its entertainment value. In addition, the students highlighted three main benefits of the serious game, namely the fact that it provides intervention strategies, information on symptoms, and raises awareness against social stigma. In regard to this last aspect, although it may have received a lower score than expected, we must bear in mind that many healthcare professionals also display high stigma towards mental disorders and do not recognize them as problems [13-15]. The students similarly scored the other characteristics which they were learning about in their course studies, such as the identification of disorders and intervention and also raising awareness about stigma.

In terms of other issues, most of the students recognized that the characters in the serious game had some psychological problem. However, the percentage for bipolar disorder was much lower (73 %) than for the remaining three (panic disorder with agoraphobia, schizophrenia and depression), which were above 90%. We believe that this aspect is important and reflects the video game's accurate portrayal of the different psychological disorders. Furthermore, it is noteworthy that the psychology students stated they had felt, at some time, similar to the characters with depression and bipolar disorder (42 % and 31 %), something which did not occur in the case of panic disorder with agoraphobia (11 %) or schizophrenia (0%). Finally, as expected, the majority felt capable of helping the people with different disorders.

Ultimately, "Stigma Stop" fulfills three functions: inform on disorder symptoms, provide intervention strategies, and dispel misconceptions and myths; thereby normalizing disorders. Thus, it educates young people about psychological disorders, while always emphasizing the importance of biographical and/or contextual experiences related to these problems. Moreover, it teaches users that we are all susceptible, given specific situations, to experiencing similar problems during our lives. Finally, the game provides players with basic intervention strategies that make them more prone to help individuals with such disorders.

6. Conclusion

Finally, the interest shown by the Psychology students who participated in this study and their favorable assessment of the game's usefulness constitute strong testaments to its effectiveness. It is also especially noteworthy that they unanimously stated that they would recommend using the program to others. As regards the limitations of the present study, we can cite its small sample size, Being necessary to replicate it with a broader sample, as well as to carry out a controlled experimental study, not only descriptive like in this case. Also, the present study did not evaluate the mental

state of students (whether or not they have a mental disorder), an aspect that could also influence the results.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This work was funded by the Regional Government of Andalusia (Excellence Research Projects SEJ-7204).

7. References

[1] J.A. Piedra-Fernandez, A.J. Cangas, J.J. Ojeda-Castelo, D. Cangas, A.J. Fernandez-Garcia, "Stigma-Stop a Serious Game against the Stigma in Mental Disorders", *2016 8th International Conference on Games and Virtual Worlds for Serious Applications (VS-Games)*, vol. 00, no. , pp. 1-4, 2016, doi:10.1109/VS-GAMES.2016.7590367

[2] N. Rusch, M.C. Angermeyer, P.W. Corrigan, "Mental illness stigma: Concepts, consequences, and initiatives to reduce stigma"Eur Psychi.,20, 529-539, 2005.

[3] M. López, M. Laviana, A. López, A, "La lucha contra el estigma y la discriminación en salud mental. Una estrategia compleja basada en la información disponible"Rev. Asoc. Esp. Neuropsiq., **28**, 43-83, 2008.

[4] M. Muñoz, E. Pérez, M. Crespo, A.I. Guillén, Estigma y enfermedad mental. Análisis del rechazo social que sufren las personas con enfermedad mental, Madrid-Editorial Complutense, 2009.

[5] G. Thornicroft, Shunned: Discrimination against People with Mental Illness, Oxford- Oxford University Press, 2006.

[6] N. Navarro, A. J. Cangas, J.M. Aguilar, J.J. Ojeda, J.A. Piedra, D. Cangas, M. Gregg," Stigma-Stop: A serious Game against stigma in mental health" (submitted).

[7] R. Graham, H. McCabe, & S. Sheridan, (2015). Path finding in computer games. *The ITB Journal*, ***4***(2), 6, 2015.

[8] S. Cheema, M. Hoffman, & J. J. La Viola. 3D gesture classification with linear acceleration and angular velocity sensing devices for video games. *Entertainment Computing*, ***4***(1), 11-24, 2013.

[9] M. Varga, "Innovation in Education through Partnerships: A Case Study of the Johnny Money Online Game Partnership", Dissertation available from ProQuest.

[10] H., Spires, J. Rowe, B. Mott, J. Lester, "Problem Solving and Game-Based Learning: Effects of Middle Grade Students' Hypothesis Testing Strategies on Learning Outcomes."J. Educ. Comput. Res.,**44,** 453-472, 2011

[11] J. A. Carmona, A.J. Cangas, A. Langer, "Applications of 3D simulation in Mental Health: Utilities and new developments", in Mental Illness. Evaluation, Treatment and Implications, L. Labate Ed.Rijeka, Intech, 2012, pp.37-56.

[12] M. Guenaga, S. Arranz, I. Rubio, E. Aguilar, A. Ortiz de Guinea, A. Rayón, M. Bezanilla, I. Menchaca, "Serious Games para el Desarrollo de Competencias Orientadas al Empleo" VAEP-RITA., **1**, 35-41, 2013.

[13] B. Angell, A. Cooke., K. Kovac, "First person accounts of stigma", in on the stigma of mental illness: practical strategies for research and social change, P. W. Corrigan, Ed. Washington D.C: American Psychological Association., 2005, pp. 69-98.

[14] D. Kingdon, T. Sharma, D. Hart, "The Schizophrenia Subgroup of the Royal College of Psychiatrists' Changing Minds Campaign What attitudes do psychiatrists hold towards people with mental illness?" BJPsych Bulletin.,**28,** 401–406, 2004.

[15] N. Aydin, A. Yigit, T. Inandi, I. Kirpinar, "Attitudes of hospital staff toward mentally ill patients in a teaching hospital, Turkey" Int. J. Soc. Psychiatry.,**49**,17–26, 2003.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 191-197 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Kinect-Based Moving Human Tracking System with Obstacle Avoidance

Abdel Mehsen Ahmad*, Zouhair Bazzal, Hiba Al Youssef

Department of Computer and Communication Engineering, Lebanese International University, Lebanon

Emails: abdelmehsen.ahmad@liu.edu.lb, zouhair.bazzal@liu.edu.lb, 20830324@students.liu.edu.lb

ARTICLE INFO

Article history:
Received: 04 March, 2017
Accepted: 14 April, 2017
Online: 24 April, 2017

Keywords:
Kinect, Arduino, Tracking, Depth Image, Processing.

ABSTRACT

This paper is an extension of work originally presented and published in IEEE International Multidisciplinary Conference on Engineering Technology (IMCET). This work presents a design and implementation of a moving human tracking system with obstacle avoidance. The system scans the environment by using Kinect, a 3D sensor, and tracks the center of mass of a specific user by using Processing, an open source computer programming language. An Arduino microcontroller is used to drive motors enabling it to move towards the tracked user and avoid obstacles hampering the trajectory. The implemented system is tested under different lighting conditions and the performance is analyzed using several generated depth images.

1. Introduction

For the past years, the development of robot technology has significantly increased due to industrial and military applications. Human tracking robot is one of the applications that could be implemented under robot technology.

Because of its human following capability, human tracking robot can work as solutions for many problems or as assistants for humans in various situations. One of them would be to fight wars in place of humans as tracking and following the enemy to unknown places in order to reduce human casualties. It can also be used to help people with physicals problems to carry their objects [1] , [2].

To achieve this, it is required to use a sensor that measures variables within an environment where these variables can be processed and analyzed to detect a person.

There are several ways to implement computer vision. The first one is to use the 2D scanning like RGB cameras or stereo cameras that have 2 lenses to simulate human binocular Vision [3].

*Corresponding Author: Abdel Mehsen Ahmad, Department of Computer and Communication Engineering, Lebanese International University, Lebanon
Email: abdelmehsen.ahmad@liu.edu.lb

The second way is the 3D scanning like the time of flight that uses both the distance and time to calculate the point in 3D and the structured light that uses a pattern of projected light to calculate the depth image [4].

For example J. Hoang and J. Zhou designed an algorithm based on a color light sensor to model a person after a color histogram for person segmentation [5].

Another example was when a group of researchers created a person following robot using a light sensor capable to detect a person based on the color within a certain region of their clothing [3].

Some researchers have utilized an omnidirectional camera in conjunction with a laser based distance sensor to track a person. Those sensors are capable to detect light outside the visible spectrum [6].

This paper has involved the integration of Kinect sensor that beats any other camera by its ability to capture a direct depth image with high accuracy rate. This small, portable device provided developers with the foundation needed to create and deploy interactive applications that respond to people's natural movement, gestures and voice commands [7].

One of the most amazing capabilities of the Kinect used is the skeleton tracking. This concept is used in this paper to map depth image to make a robot able to track human movements [8].

Another important part of this paper is the obstacle avoidance. The robot may face obstacles while tracking human, these obstacles must be observed and detected by the Kinect so that it would be able to move to the right direction without collide with them.

Robots have proven to be more useful for a wide range of industrial and commercial applications, principally in indoor environments since the outdoor environment brings more challenges and complications. In an indoor environment, the small size of rooms restricts the visual perception where tables, chairs and even walls are encountered as occlusions but it still more beneficial and robustness compared to the outdoor environment where the occlusions there: plants, bushes, etc. are more shaped complexly. In addition, when operating in an outdoor environment, lighting intensity is more dynamic, reducing the robustness of visual perception. So the third part of the paper is to study the flexibility of the Kinect while operating under various lighting conditions [9].

This paper is an extension of work originally presented and published in IMCET conference [1]. In this extended work, we first present the proposed system design, and then we show the evaluation of the Kinect performance in different lighting conditions.

2. Proposed System

2.1. System Design

The project deals with making a robot that will move towards a specific user and avoid obstacles encountered in its path.

The proposed system as shown in Figure 1consists of three main parts: Kinect sensor, Arduino and Processing.

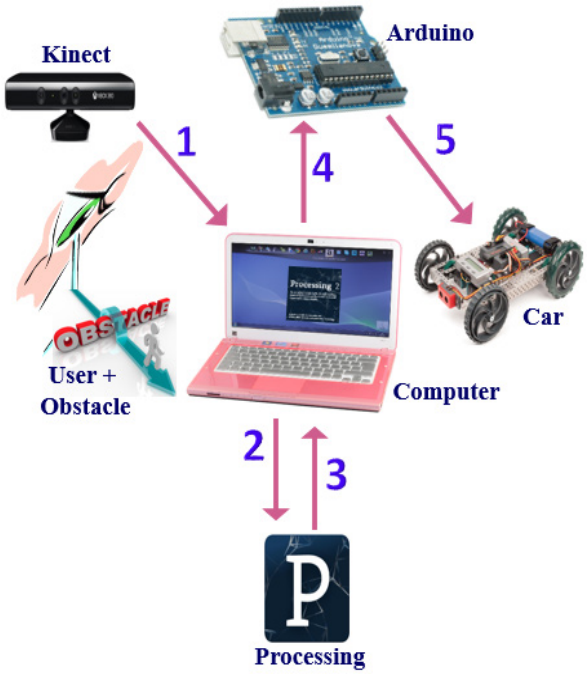


Figure 1: System Design

The Kinect is a motion sensing input device by Microsoft that consists of object detection, object tracking and reorganization. It enables users to control and interact with the applications using full skeleton tracking, gesture recognition and facial recognition and it can also be used to detect and distinguish between different kinds of objects. It has three eyes; 2 cameras and an infrared projector. The IR projector shines a grid of infrared dots and based on the triangulation between the observed and the reference patterns, the depth will be calculated [10].

Thus, the Kinect is used as an effective tool for discovering the surrounding area, calculating the depth image, detecting and tracking a specific human and locate their individual joints and body parts as illustrated in Figure 2.

The Arduino, which is an open source physical platform or microcontroller board, can take input from a variety of sensors and control a variety of motors, lights and others physical outputs [11]. In this project and in order to achieve the full control over the robot, the Arduino will be connected to the car motors.

The processing is an open source programming language used to access the depth data coming from the Kinect, analyze it and display it on the screen [8].

The robot was decided to be a car since it can serve part of the system's demands in regards to the power, rotation and control. It consists of two motors; one is used for the motion and the other for controlling the motion direction of the car. Concerning the Kinect it is necessary to be located on the front of the car to avoid seeing the car parts as obstacle [1].

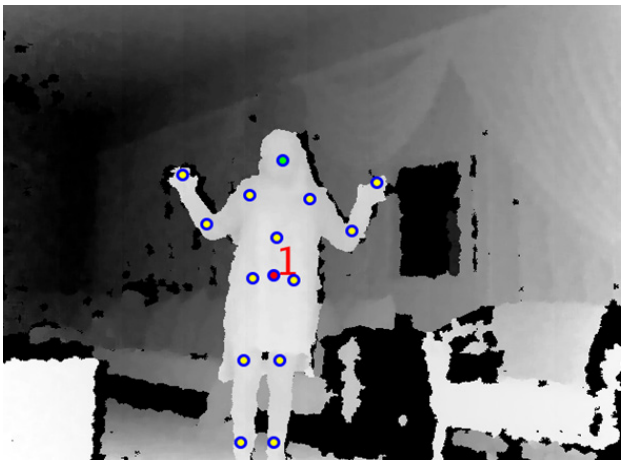


Figure 2: Skeleton Tracking

When turning the system on, the Kinect sensor, which is the main part in the system, starts capturing the depth image of the whole scene in front of it. The information captured by the Kinect must then be transferred to the computer where the processing is running. The processing will analyze the depth image and once it detects a user, it will determine his/ her center of mass which is the main point in tracking him/her and it has also to detect the obstacles between that user and the robot. Based on the analyzed data, the processing will send commands (like: left, right, forward, stop, etc…) to the Arduino serially via the computer, these commands will be already defined between processing and Arduino. Finally, the Arduino which will be connected to the car motors will receive the commands coming from computer, analyze them and order the robot to move in the right way.

Now, the robot can start tracking the user and avoid obstacles according to the information obtained from its "brain", the Arduino, and build on scenes coming from its "eyes", the Kinect [1].

2.2. Sequence Diagram

The sequence diagram in Figure 3shows the interactions between different parts of the system in sequential order.

1- Start with the Kinect that captures depth image at 30 Hz [8].
2- Pass by the processing that will analyze the depth image, detect the center of mass and send commands to the Arduino.
3- Move to the Arduino, serially via the computer, which will receive the commands coming from processing, analyze them and order the Robot to move in the correct direction.
4- End with a robot able to move towards a user and avoid obstacles detected in its path.

2.3. State Diagram

The robot will not move until a user is detected in front of the Kinect. Once this happens, the processing will track his/ her skeleton specifically according to his/her center of mass. If it failed to track the center of mass for a certain period of time, the processing will return back and wait for a new detection. Once the center of mass is determined, the robot should move towards the detected user and processing should maintain the distance between the robot and the user to check if any obstacle exists there. Now, the Robot should avoid obstacles detected and continue moving in the same way the user does. These steps are represented in the state diagram of Figure 4**.**

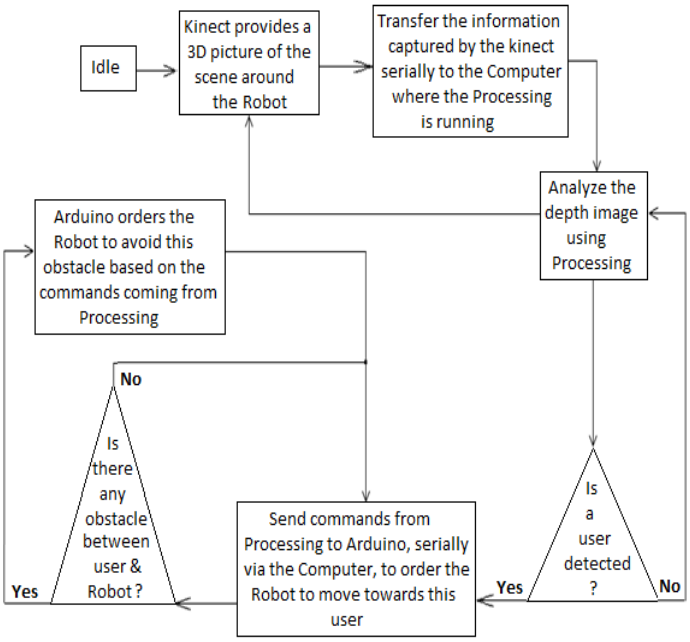


Figure 3: Sequence Diagram.

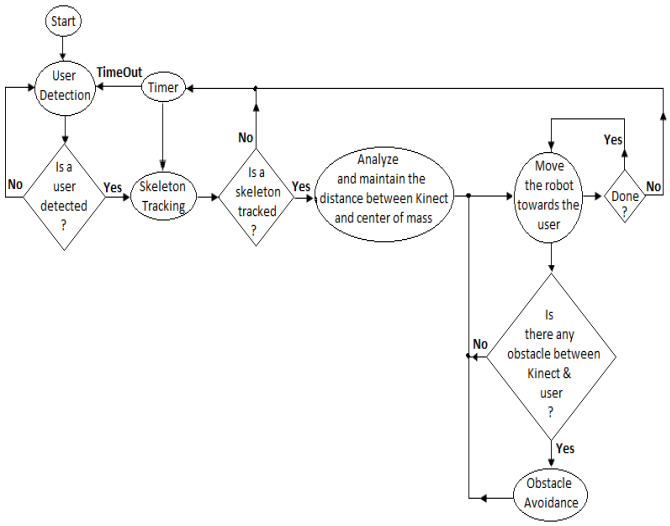


Figure 4: State Diagram.

3. Implementation

The project needs two kinds of software to be implemented, the first one which is the Processing is responsible for analyzing the surrounding information, detecting obstacles, recognizing the user and tracking him/her. The second software is the Arduino Software IDE which is responsible for controlling the robot movement based on commands coming from the Processing.

3.1. Processing

The processing code imports the necessary steps for performing user tracking and obstacle avoidance from the Simple Open NI library.

The first of these steps is to tell the Simple Open NI to turn on user tracking. That happens within the *setup ()* by using *enable User()* function and to access depth data the *enable Depth()* function is used.

The *Depth Map ()*function is used to get the depth array and to save the coordinates of the closest value, each row in the depth image must be checked. In addition, each pixel the row must be observed and the corresponding value must then be pulled out from the depth array.

In skeletal tracking, the body parts of human such as head, neck, hands, legs, shoulders and arms are represented by a number of joints. Each joint is represented by its 3D coordinates [2].Once a user has been detected and successfully calibrated, it uses the joint position data to have circles track the user's joints where the main joint to track him/her is the center of mass (CoM). The real coordinates saved, shall be then translated into projective coordinates to match the 2 dimensional depth images.

As illustrated in Figure 5, the image is divided into three parts in width: left, middle and right where x is the width coordinate. If the closest value saved within 1 meter and the x value of the center of mass are in the same part of the image, then this closest value is an obstacle detected between the user and the robot [1].

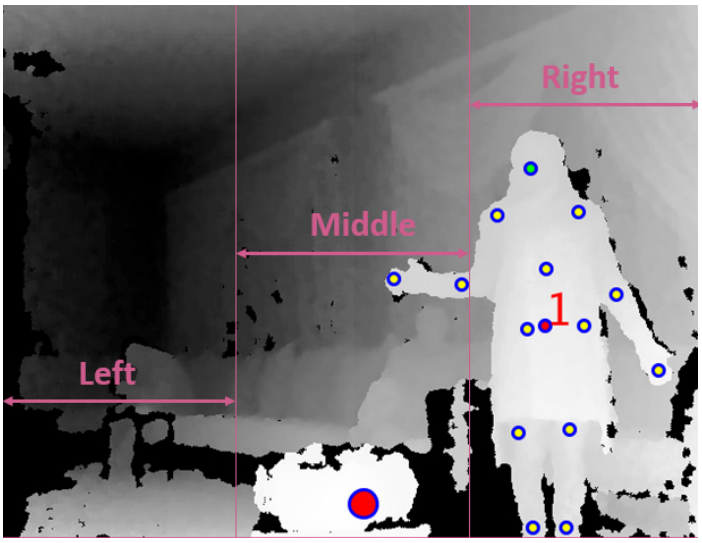


Figure 5: The three parts of the image

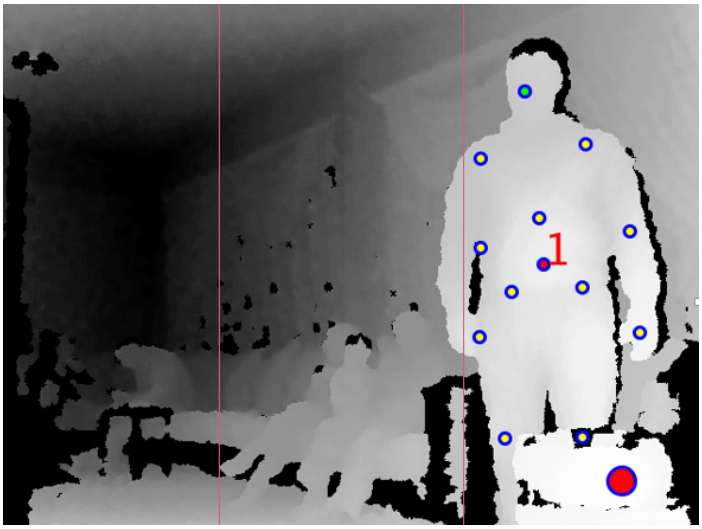


Figure 6: User and Obstacle in the same range (left side)

Depending on the Kinect data, there is an output send to the Arduino to be able to control the robot's movements. For instance, to move: left, forward, right and stop.

The whole process begins when the user enters the view of the Kinect. At this point the *Simple Open NI* detects that the user is present and is a candidate for tracking within the *On New User*() function.

Now, in the case in which the user is located at a distance less than or equal to 1 meter the robot should stop immediately. Otherwise, it checks if the closest value and the center of mass of the user are in the same range; if so, then there is an obstacle that hampering the trajectory of the robot while tracking the user detected as represented in Figure 6. While if no obstacle is detected the robot will continue in its way toward the user as Figure 7 illustrates. This process ends when the user is no longer in the Kinect range by calling the *On Lost User ()* function [8].

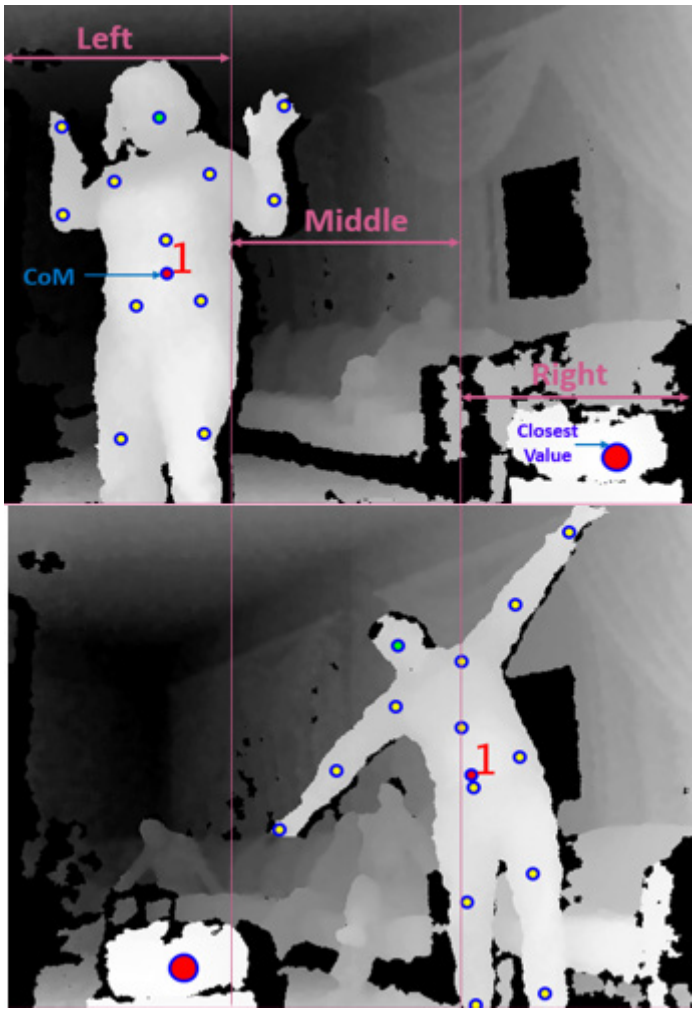


Figure 7: The user and the obstacle are in different range

The final task within processing was to send data from processing to Arduino serially via the computer, by importing the serial library, declaring a serial port and after matching the port on which the Arduino is connected a predefined value is sent on that port to enable the robot to move in the correct direction by using the *write* method. For example, in the case in which the user is in the left range and there is no obstacle in that range, the value '3' is written on the port.

3.2. Arduino Software

The brain of the robot will be an Arduino microcontroller. So, the Arduino Mega 2560 board will be used for its large memory and available ports that make the installation more flexible [12].

To order the robot to move in the correct direction based on the data captured by the Kinect, the processing will communicate with the Arduino serially via the computer. Serial is a simple protocol that allows Processing to send information to the Arduino and vice versa [8].

The basic thing to do is to program the input and output ports, and give the order to apply a specific function at the output once a specific input is received on a specific input port [1].

The Arduino will establish a serial connection to the computer by using the *serial.begin ()* function and then listen for data available on the serial port by using the *serial.available ()* function. This function will return true if any data is available to read and false otherwise. To read the available data the *serial.read ()* function is used [11].

If the received output from the processing is '3', the Arduino should give its command to the system to move left, and thus digital Write function is used to set the pins to the required values. All other directions are defined in the same; by setting the output pins as desired once receiving a specific input.

3.3. Lighting Conditions

The purpose of this part is to analyze the influence of different lighting conditions in the range of Kinect performance.

Six different scenarios, as shown in Table 1, were made under various light conditions producing a set of depth images from the Kinect. For example, scenario 3 is the case in which the Kinect was placed under sunlight while the user was under environmental light [13].

Scenario	Kinect	User	Light Intensity levels
1	Sunlight	Sunlight	High
2	Environmental light	Sunlight	High
3	Sunlight	Environmental light	High
4	Environmental light	Part of the user under sunlight	Medium
5	Environmental light	Environmental light	Medium
6	Artificial light	Artificial light	Low

Table 1: Lighting Conditions Scenarios

4. Simulation Results

The first stage within the processing code was to detect a user, draw his/ her skeleton, detect his/ her center of mass and give him/her a specific ID. This stage was successfully done where the result in Figure 8 shows the skeleton of three different users with their specific IDs.

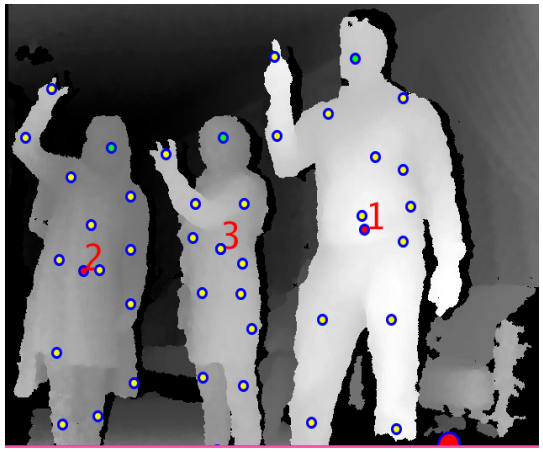


Figure 8: The skeletons of three different users

The next task was to keep track of the first user entering the Kinect range so that the output on the console indicates the movement and position of that user. Figure 9 and Figure 10illustrate the expected results.

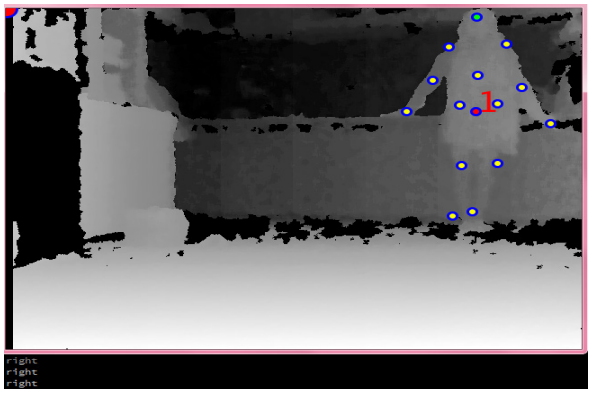


Figure 9: The output is "right"

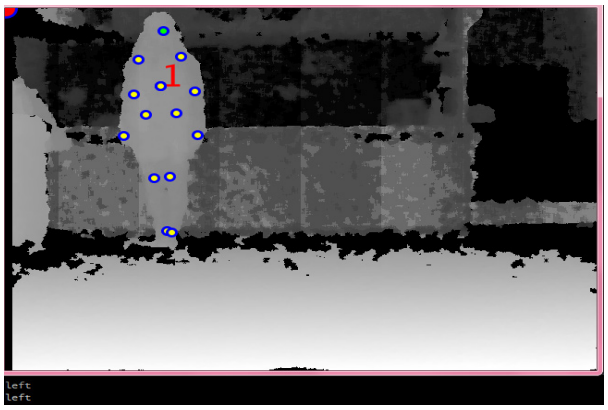


Figure 10: The output is "left".

As displayed in Figure 11, Figure 12 and Figure 13the obstacle detection and avoidance stage was successfully done so that the output on the console indicates where the robot should move to avoid any detected obstacle.

Figure 11: Incoming obstacle, go to the left.

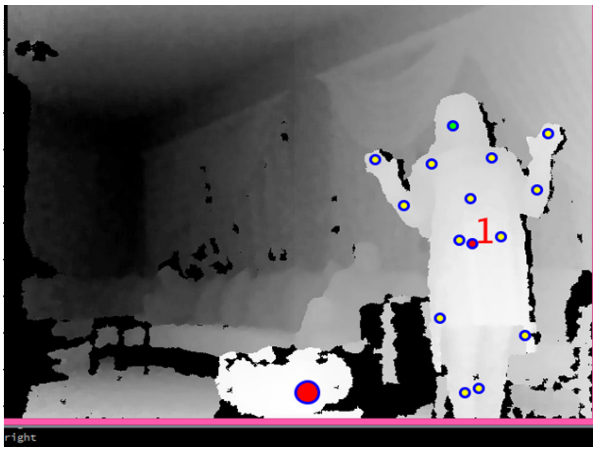


Figure 12: Go right towards the user.

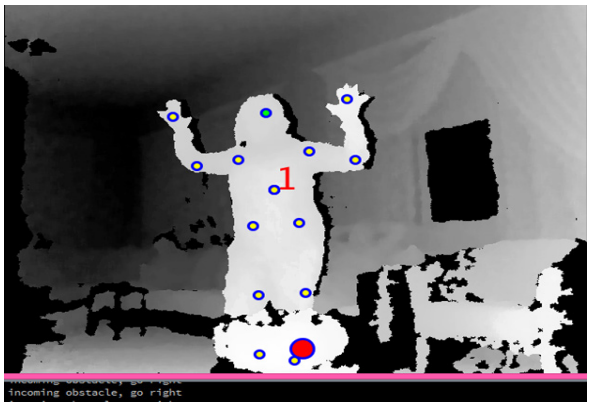


Figure 13: Incoming obstacle, go to the right.

Figure 14: Final prototype of 3D sensor-based Moving Human Tracking Robot with Obstacle Avoidance.

The Arduino was connected to the laptop and the Processing code was run where it started to print the results: Forward, right, stop…etc., according to the state of the user and the allocated obstacles in front of it. The Arduino output is then measured using a multimeter where the measurement ensures the successive work in connecting the Arduino to the Kinect serially.

Finally, all parts were assembled together to get the final system illustrated in Figure 14.

Figure 15 presents the expected results such that tracking the human movement and avoiding the obstacles coming in its way became feasible.

Figure 15: A User followed by the Robot.

5. Lighting Conditions Results

Concerning the lighting conditions, the results of the six scenarios studied, shown in Figure 16, were analyzed.

In scenario 1, in the presence of sunlight over both the Kinect and the user, and in scenario 2, where the Kinect was placed under environmental light and the user still under sunlight the Kinect was not able to obtain the depth information of the user.

Scenario 3 shows a higher intensity over the user with no skeleton joints detected since the Kinect was placed under the sunlight.

While in scenario 4, where only a part of the user was under sunlight, the Kinect was able to capture the depth information of the second part of the user.

For scenarios 5 and 6, where environmental and artificial lights were present, the Kinect performance is acceptable.

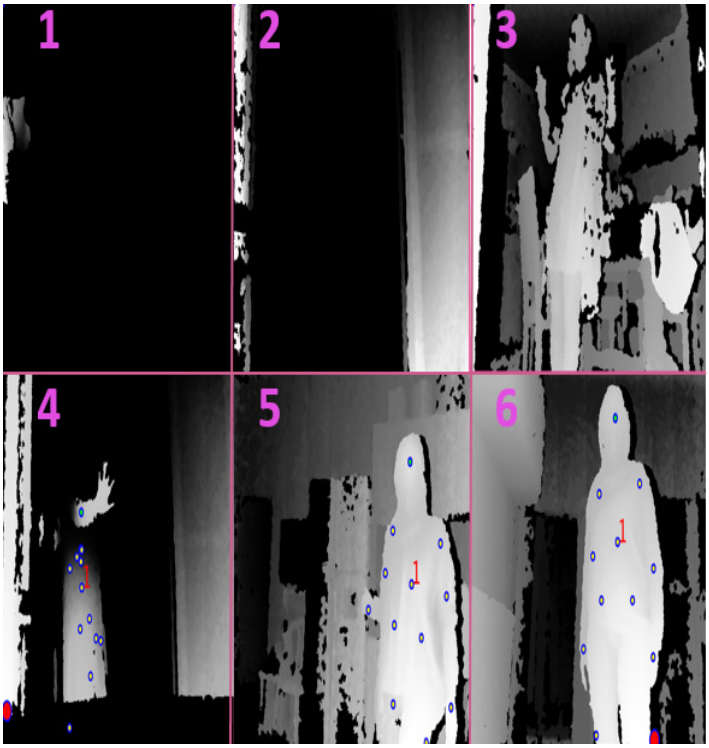


Figure 16: The depth images of the Kinect under various lighting conditions.

It was found that lighting conditions affect Kinect's performance and that the sunlight is the main enemy for the Kinect. So try to avoid having sunlight on the user and don't let the Kinect directly faced it.

6. Conclusion

The paper described the specifications of the Kinect sensor and how it can be integrated as the vision system of the robot. It also presented an experimental study of its performance under various light conditions.

The information of the Kinect were accurate enough to integrate a human tracking robot able to track a specific user in an indoor environment and avoid any obstacle detected in its path.

The project used the Kinect skeleton data to find the location of the user's center of mass. This value is then sent to the Arduino so that it can match the movement of the user with the robot's motor.

Our testing for this platform shows that the Kinect can be used as a powerful and useful device in many fields and that the combination between the Kinect and Arduino is as an advanced step in 3D Kinematics and interactive projects. Many system improvements will come in further works.

References

[1] Abdel Mehsen Ahmad, Hiba Al Youssef, "3D- Sensor based Moving Human Tracking Robot with Obstacle Avoidance," in IEEE International Multidisciplinary Conference on Engineering Technology (IMCET), Lebanon, 2016.

[2] Satish Prabhu, Jaly Kumar, AmanKumar Dabhi, Pratik Shatty, "Real Time Skeleton Tracking Based Human Recognition System using Kinect and Arduino," in National Conferene on Role of Engineers in Nation Building (NCRENB-15).

[3] T. Yoshimi, M. Nishiyama, T. Sonoura, H. Nakamoto, S. Tokoura, H. Sato, F. Ozaki, N. Matsuhira, and H. Mizoguchi, "Development of a person following robot with vision based detection," in Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE, Beijing, China, 2006.

[4] K. Jungling and M. Arens, "Feature based person detection beyond the visible spectrum," in Proceedings of the IEEE International Computer Society Conference on Computer Vision and Pattern Recognition Workshop, IEEE, Miami, Florida, 2009.

[5] J. Zhou and J. Hoang, "Real time robust human detection and tracking system," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, San Diego, California, 2005.

[6] M. Kobilarov, G. Sukhatme, J. Hyams , and P. Batavia, "People tracking and following with mobile robot using omnidirectional camera and a laser," in Proceedings of the IEEE International Conference on Robotics and Automation, IEEE, Orlando, Florida, 2006.

[7] Jose Henrique, "Using Kinect fo Smart Costume Design," August 2012. [Online]. Available: http://www.macs.hw.ac.uk/cs/msc_dissertations/2012/3.pdf.

[8] Borenstein, Greg, Making Things See, O'REILLY, Ed., 04 - 01 -2012.

[9] R. Seggers, "People Tracking in Outdoor Environments," Amesterdam, 2015.

[10] Martin Kvalbein, "The use of a 3D sensor (Kinect™) for robot motion compensation.," 12 May 2012. [Online]. Available: https://www.duo.uio.no/handle/10852/9069.

[11] "Arduino Official Website," [Online]. Available: http://www.arduino.cc/en/Main/Software.

[12] "Arduino Board Mega," [Online]. Available: http://www.arduino.cc/en/Main/arduinoBoardMega.

[13] J.R. Ruiz Sarmiento, C. Galindo and J.Gonzalez Jimenez, "Experimental Study of the Performance of the Kinect Range Camera for Mobile Robotics," Universidad de Malaga, Andalucia Tech, Departamento de Ingeniera de Sistemas., Malaga, Espana.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 198-204 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

A security approach based on honeypots: Protecting Online Social network from malicious profiles

Fatna Elmendili, Nisrine Maqran, Younes El Bouzekri El Idrissi, Habiba Chaoui

Computer Sciences Systems Engineering Laboratory, National School of Applied Sciences, Ibn Tofail University, Kenitra, Morocco

ARTICLE INFO

Article history:
Received: 15March, 2017
Accepted: 16 April, 2017
Online: 24 April, 2017

Keywords :
Social network
Social honeypots
Feature based strategy
Honeypot feature based strategy
Profile
Security
Spam attacks
WEKA
Malicious Profiles

ABSTRACT

In the recent years, the fast development and the exponential utilization of social networks have prompted an expansion of social Computing. In social networks users are interconnected by edges or links, where Facebook, twitter, LinkedIn are most popular social networks websites. Due to the growing popularity of these sites they serve as a target for cyber criminality and attacks. It is mostly based on how users are using these sites like Twitter and others. Attackers can easily access and gather personal and sensitive user's information. Users are less aware and least concerned about the security setting. And they easily become victim of identity breach. To detect malicious users or fake profiles different techniques have been proposed like our approach which is based on the use of social honeypots to discover malicious profiles in it. Inspired by security researchers who used honeypots to observe and analyze malicious activity in the networks, this method uses social honeypots to trap malicious users. The two key elements of the approach are: (1) The deployment of social honeypots for harvesting information of malicious profiles. (2) Analysis of the characteristics of these malicious profiles and those of deployed honeypots for creating classifiers that allow to filter the existing profiles and monitor the new profiles.

1. Introduction

Social networks are now part of our daily life, we have certainly several accounts said "social", which are related to our daily lives (Facebook, Twitter), our professional life (Viadeo or LinkedIn), our sporting life or associative (we can cite jogg.in) and why not on sites of meeting (Adopteunmec, Meetic, Lovoo tinder or).

We can divide social networks on those which do not promote the anonymity (Facebook, Viadeo, LinkedIn) or the others which are promoting the anonymity (of general way all the sites of meeting, but also of services of Visio-conference as Skype), we have certainly received requests for connections from persons completely unknown. These applications are, in most of the cases, issued by malicious profiles.

That they emanate from the robots or that they are created to spoof the identity of a user, the malicious profiles are in constant increase on the Internet. On social networks, the malicious profiles can be generated by machines or be the result of identity theft and their motivations are various: Spy a PERSON, Increase the number of fans of a page (Facebook, Twitter…), spammer friends in impunity, fit all types of scams (very often of blackmail), harm the reputation of a person or a company, etc.

In particular, social spammers are increasingly targeting those systems as part of phishing attacks. To disseminate malware, and commercial spam messages, and to promote affiliate websites [1]. In only the past year, more than 80% of social networking users have "received unwanted friend requests, messages, or postings on their social or professional network account". Unlike traditional email based spam, social spam often contains contextual information that can increase the impact of the spam (e.g., by eliciting a user to click on a phishing link sent from a "friend") [1].

Successfully defending against these social spammers is important to improve the quality of experience for community members, to lessen the system load of dealing with unwanted and

*Corresponding Author: Elmendili Fatna, Computer Sciences Systems Engineering Laboratory, National School of Applied Sciences, Ibn Tofail University, Kenitra, Morocco, Email:f.elmendili@gmail.com.

sometimes dangerous content, and to positively impact the overall value of the social system going forward. However, few information is known about these social spammers, their level of sophistication, or their strategies and tactics. Filling this need is challenging, especially in social networks consisting of 100s of millions of user profiles (like Facebook, Myspace, Twitter, YouTube, etc.). Traditional techniques for discovering evidence of spam users often rely on costly human-in the-loop inspection of training data for building spam classifiers; since spammers constantly adapt their strategies and tactics, the learned spam signatures can go stale quickly. An alternative spam discovery technique relies on community-contributed spam referrals (e.g., Users A, B, and C report that User X is a spam user); of course, these kinds of referral systems can be manipulated themselves to yield spam labels on legitimate users, thereby obscuring the labeling effectiveness, and neither spam discovery approach can effectively handle zero-day social spam attacks for which there is no existing signature or wide evidence [1]. With these challenges in mind, we propose and evaluate a novel Honeypot-based approach for uncovering social spammers in online social systems. Concretely, the proposed approach is designed to (1) the deployment of social honeypots for harvesting information of malicious profiles [8, 9]. (2) Analysis of the characteristics of these malicious profiles and those of deployed honeypots for creating classifiers that allow to filter the existing profiles and monitor the new profiles [2]. Drawing inspiration from security researchers who have used honeypots to observe and analyze malicious activity (e.g., for characterizing malicious hacker activity ,generating intrusion detection signatures,),In This extended paper we deploy and maintain social honeypots for trapping evidence of spam profile behavior, so that users who are detected by the honeypot have a high likelihood of being a spammer [1].This paper is an extension of work originally presented in conference 2016 4th IEEE International Colloquium on Information Science and Technology (CIST) in Tangier Morocco, we describe the processes of the proposed approach, starting with the deployment of social honeypots, the use of both feature based strategy and honeypot feature based strategy methods for collecting data, and finally we give the results and the test of this approach by using a dataset of profiles in machine learning based classifiers for identifying malicious profiles [2].These results are quite promising and suggest that our analysis techniques may be used to automatically identify the malicious profiles in social network.

2.1. Overall Framework

Malicious profiles are increasingly targeting Web-based social systems (like Facebook, Myspace, YouTube, etc.) as part of phishing attacks, to disseminate malware and commercial spam messages, and to promote affiliate websites. Successfully defending against these social spammers is important to improve the quality of experience for community members, to lessen the system load of dealing with unwanted and sometimes dangerous content, and to positively impact the overall value of the social system going forward. However, little is known about these social spammers, their level of sophistication, or their strategies and tactics [3].

In our ongoing research, we are developing approach for uncovering and investigating malicious user. Concretely, the Approach for detecting malicious profiles based Social Honeypot to (1) the deployment of social honeypots for harvesting information of malicious profiles. (2) Analysis of the characteristics of these malicious profiles and those of deployed honeypots for creating classifiers that allow to filter the existing profiles and monitor the new profiles [3]. Drawing inspiration from security researchers who have used honeypots to observe and analyze malicious activity. The Approach for detecting malicious profiles based Social Honeypot deploys and maintains social honeypots for trapping evidence of malicious profile behavior. In practice, we deploy a social honeypot consisting of a legitimate profile and an associated bot to detect social spam behavior. If the social honeypot detects suspicious user activity (e.g., the honeypot's profile receiving an unsolicited friend request) then the social honeypot's bot collects evidence of the spam candidate (e.g., by crawling the profile of the user sending the unsolicited friend request plus hyperlinks from the profile to pages on the Web-at-large). What entails suspicious user behavior can be optimized for the particular community and updated based on new observations of spammer activity [3].

While social honeypots alone are a potentially valuable tool for gathering evidence of social spam attacks and supporting a greater understanding of spam strategies, the goal of the Approach for detecting malicious profiles based Social support ongoing and active automatic detection of new Honeypot is to and emerging spammers. As the social honeypots collect spam evidence, we extract observable features from the collected candidate spam profiles (e.g., number of friends, text on the profile, age, etc.). Coupled with a set of known legitimate (non-spam) profiles which are more populous and easy to extract from social networking communities, these spam and legitimate profiles become part of the initial training set of a spam classifier [3].

As the social honeypots collect evidence on the Malicious Behaviors, the characteristics of profiles are extracted from the data of malicious profiles (for example: number of friends, Text, on the Profile, the age, etc.). Coupled to a set of legitimate profiles which are easy to extract the communities of social networks .This is called type of strategy by "Feature based strategy."[2].

A new method used in our approach to improve our classification and increase the ability to detect an attacker on the social networks that is "honeypot feature based strategy", this strategy uses the whole of char-Characteristics of honeypots that interact with users to refine our ranking [2].

The whole of the data collected are becoming an integral part for the training of a classifier of malicious profiles. By iterative refinement of selected characteristics using a set of algorithms for automatic classification which are implement on "Weka Machine Learning Toolkit" we can explore the more wide space of malicious profiles[14]. Fig. 1 present the approach for detecting malicious users.

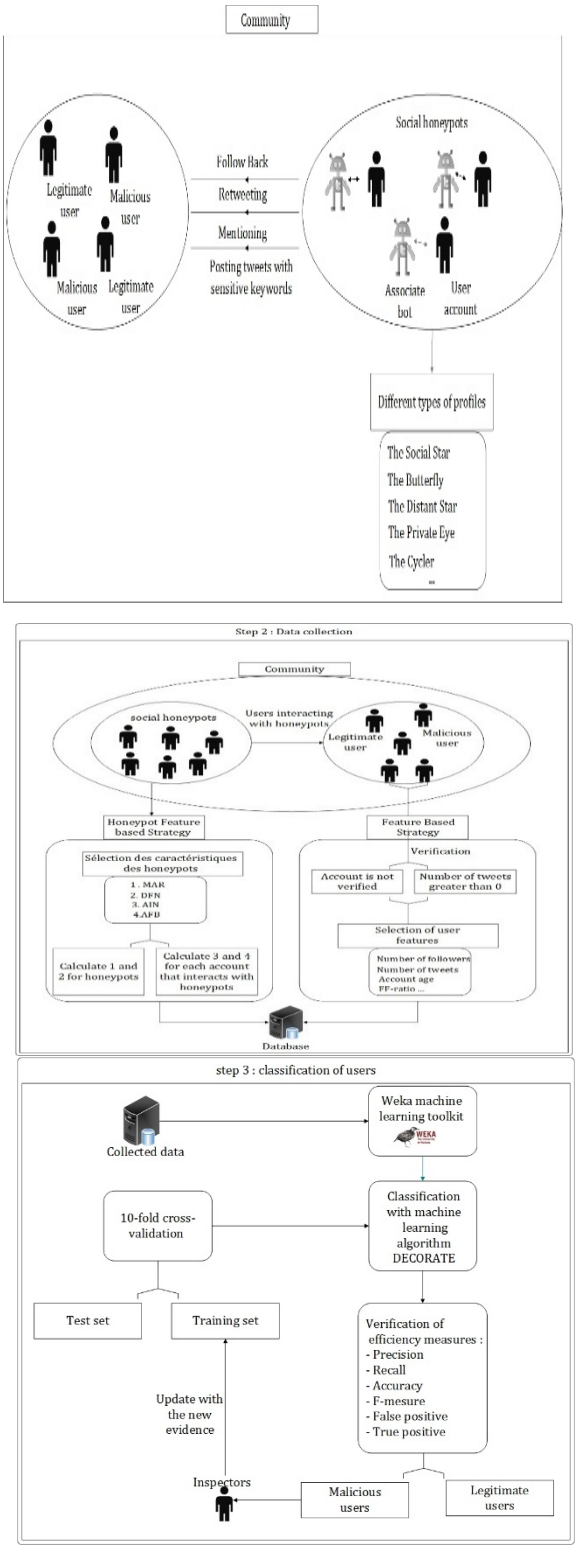


Fig.1 Overall Framework [2].

2. Malicious Profiles Detection Results:

In this section we present the results of our experiment. Creating of social honeypots: To develop our approach, we created 20 honeypots on twitter to trap malicious users and we analyze all the characteristics of users and deployed honeypots using weka that provides a platform algorithms of artificial intelligence and machine learning including Decorate algorithm that we used in our approach. We investigated various characteristics of the 90 friend requests that we collected with our social honeypots.

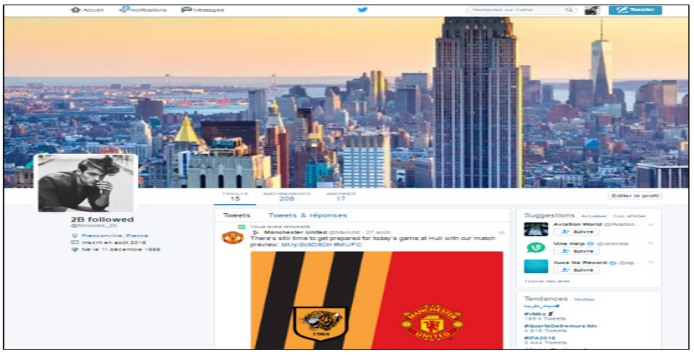


Fig. 2: Example of created honeypots

2.1. Development Tools:

- ***Twitter***

Twitter has become in the space of a few months a media phenomenon on the Internet. Everyone is put to talk in good or bad but without ever remain indifferent to them. Pushed by a few influential on the web, this small service extends more and faster its community.

Twitter has become in the space of a few months a media phenomenon on Internet. Everyone is put to talk in good or bad but without ever remain indifferent to them. Pushed by a few influential on the web, this small service extends more and faster its community.

Twitter is a tool managed by the enterprise Twitter Inc. It allows a user to send free short messages, known as tweets, on the Internet, by instant messaging or by SMS. Those messages are limited to 140 characters. The concept was launched in March 2006 by the company obvious based in San Francisco. The service rapidly became popular, up to bring together more than 500 million users in the world at the end of February 2012. At the 6 May 2016, twitter account 320 million active users per month with 500 million tweets sent by day and is available in more than 35 languages [4].

The first objective of Twitter and the reason of its deployment is to be able to provide a simple answer to the question: what am I doing? The use is very simple and free: We have 140 characters to disseminate our messages to whoever wants to receive if we specify our account in public or to our network only if it is in private. At the same time, we choose the members of Twitter which we want to follow the publications and these Members can we also add in return in their own network.

To publish our messages, several means are available: via the Twitter website, via our mobile phone by SMS, via Instant messaging type Google Talk or via software/third party Web sites based on the api free to twitter [5].

- ***Weka machine learning toolkit:***

Weka (for Waikato environment for knowledge Analysis) is a tool for data search open-source (GNU license) developed in Java. It was created at the University of Waikato New Zealand, by a group

of researchers from the automatic learning, recognition of forms and the search of data. The software allows you to deal with different sources of data: files of various formats, including the format Attribute-Relation File Format (RTTW), developed for Weka; URLs; SQL databases. The analysis can be performed using most of the techniques of existing search.

The bibliographic reference attached to the Software is the book: data mining, practical machine learning tools and techniques with Java implementations, Ian H. Witten & Eibe Frank[6].

2.2. Classification results:

- ***Collection of data:***

After implementing our honeypots and interact with different types of users, we selected 90 profiles among 300 profile strapped by our honeypots and for each profile we selected traditional characteristics «traditional features" such as (Follower number, FF-ratio, Account age, Tweet Number, Mention ratio, ratio URL ...) and features based on honeypots "honeypot based features" such as (the number of honeypots with whom one interacts account, the daily average of new followers fora honeypot ...). The size of the database is 90 profiles.

TABLE 1: Dataset of users

Name	creation date	account age	Followers	followings	Tweets	profile image	average tweets per day	FF ratio followers / following	url ratio (RU)	Mention ratio (MR)	Retweet count (RT)	Tweet similarity (TS)	Number of honeypots a user interact with	class
#Boussi Global Art	07/11/2014	668	94457	108915	82798	1	66.9	0.9	0.37	1.65	90.5%	30,00%	15	mal
Corrioland Rubben	27/08/2016	9	3	40	2	1	0.1	0.1	0.50	0.50	0.0%	0,00%	10	mal
Jérémy A	11/03/2013	1274	10153	937	712	1	1.6	10.8	0.00	0.68	54.5%	25,00%	0	leg
MAXIME CORRY	08/03/2013	1277	21005	9914	2566	1	0.6	2.0	0.78	0.67	0.0%	36.8%	3	leg
Ndrouot	27/01/2012	1683	52855	422	5098	1	0.7	125.0	0.20	0.40	80.0%	59.8%	10	leg
Theresa Longo	29/09/2009	2533	70884	39692	7026	1	4.6	1.8	0.71	1.18	35.4%	33.3%	4	leg
Élysabeth	19/10/2015	322	12902	9869	100	1	0.3	1.3	0.00	0.00	0.0%	94.8%	20	mal
Mokded	14/08/2016	22	30	142	10	1	0.7	0.2	0.00	1.00	0.0%	1,00%	0	leg
CITATIONS D'amour	05/09/2015	566	62826	19711	1246	1	2.4	3.2	0.03	0.06	2.9%	87.9%	5	mal
Stay Curious Music	15/08/2014	754	4508	4043	344	1	1.4	1.1	1.11	0.53	47.4%	71.9%	8	mal
matt	06/12/2015	274	866	1113	827	1	5.4	0.8	0.05	0.92	93.3%	74.8%	10	leg
Capitaine Tweets	03/11/2012	1402	127440	109208	4851	1	4.2	1.2	0.17	0.02	0.0%	96.2%	9	mal
Zack Saldandra	18/08/2016	18	52	73	5	1	0.4	0.7	0.40	0.60	0.0%	30.5%	16	mal
Gutembergh Freitas	02/12/2015	278	34	39	2	1	0.1	0.9	0.00	0.50	0.0%	9,00%	1	leg
Travis Moet	29/03/2009	2717	17078	13481	8704	1	20.7	1.2	0.85	1.19	61.0%	44.5%	2	leg
GRABVITY	09/12/2015	271	198	859	407	1	2.4	0.2	0.58	1.45	84.8%	48.4%	4	leg

Further, the data is converted to ARFF (Attribute Relation File Format) format to process in WEKA. An ARFF file is an ASCII text file that describes a list of instances sharing a set of attributes. User Classification with Weka: After preparing our data file, we will use the classification algorithms implemented in Weka for test methods, solve problems, focusing us on the use and the results provided by these implementations without having to rewrite every time algorithms.

The framework within which we will work, and algorithms that we will study and use are based on the following functional diagram:
-There is a set of examples, each example being defined by Description: This is a set of values defining this example. The class that he was associate (with the help of a human expert) [7].

-This set of examples is provided to a program that will generate a classifier. A classifier is a program that, when provided him an example, try to guess its class. In other words, the program tries to guess the class of an instance from its description. After processing the ARFF file in WEKA the list of all attributes, statistics and other parameters can be utilized as shown in Fig. 3.

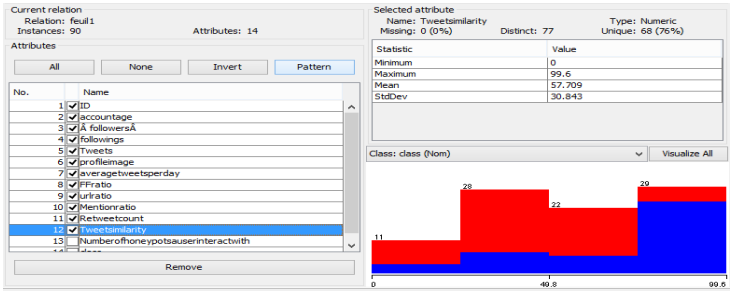


Fig. 3: Processed ARFF file in WEKA

In the above shown file, there are 90 profiles data is processed with different attributes like followers, FF ratio, Retweet-count, Mention-ratio, etc. [10,11,12].

The processed data in Weka can be analyzed using different data mining techniques like, Classification, Clustering, Association rule mining, Visualization, etc. The Fig. 4 shows the few processed attributes which are visualized into a 2 dimensional graphical Representation.

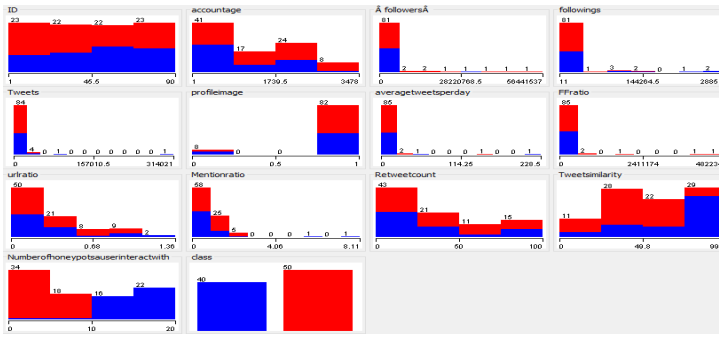


Fig. 4: Graphical visualization of processed attributes

Now that we have loaded our dataset, we used DECORATE machine learning algorithm to model the problem and make predictions and we choose Cross-validation, which lets WEKA build a model based on subsets of the supplied data and then average them out to create a final model[15].

In previous work, generally the métas Classifiers (Decorate, Lo-gitBoost, etc.) Product of Best performance that the classifiers to trees (BFTree and FT) and classifiers based on functions (SimpleLogistic and libsvm) [7]. For our approach we chose decorate as classifier. For different reasons:
-The speed and the execution time;
-Best Performance Compared to Other;
-Of Results approximately correct.

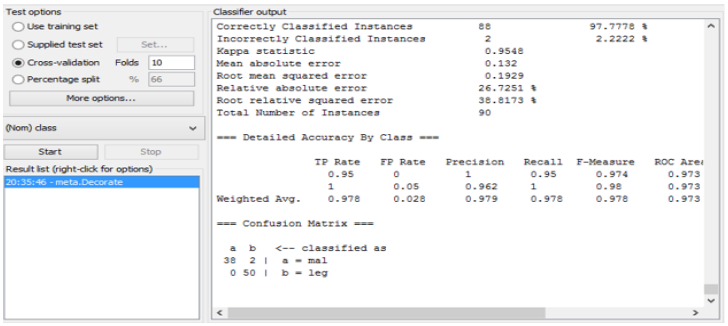


Fig. 5: Classifier output

Fig.5 shows estimates of the trees predictive performance, generated by WEKAs evaluation module [16]. It outputs the list of statistics summarizing how accurately the classifier was able to predict the true class of the instances under the chosen test module. The set of measurements is derived from the training data. In this case 97.7778% of 90 training instances have been classified correctly. This indicates that the results obtained from the training data are optimistic compared with what might be obtained from the independent test set from the same source. In addition to classification error, the evaluation output measurements derived from the class probabilities assigned by the tree. More specifically, it outputs mean output error (0.132) of the probability estimates, the root mean squared error (0.1929) is the square root of the quadratic loss. Theme an absolute error calculated in a similar way by using the absolute instead of squared difference. The reason that the errors are not 1 or 0 is because not all training instances are classified correctly. Kappa statistic is a chance-corrected measure of agreement between the classifications and the true classes. It's calculated by taking the agreement expected by chance away from the observed agreement and dividing by the maximum possible agreement. The Kappa coefficient is calculated as follows:

$$K = \frac{\text{Po}-\text{Pe}}{1-\text{Pe}} \qquad (1)$$

$With\ P_0$: is the proportion of the sample on which the two judges are of the agreement.
(i.e. the main diagonal of the matrix of confusion).
And $p_e = \frac{\sum_i p_i\, p.i}{n^2}$ (2)

Where:

p_i is the sum of the elements of the line i;
p.i is the sum of the elements of the COLUMN I;
n: in a sample size.

The Kappa coefficient takes values between -1 and 1:
- It is maximal when both judgments are the same:

All examples are on the diagonal, and P0 = 1.
- It is 0 when both judgments are independent (P0=Pe).

- It is -1 when the judges disagree. Detailed Accuracy by Class demonstrates a more detailed per class break down of the classifiers prediction accuracy.

- The True Positive (TP) rate is the proportion of examples which were classified as class x, among all examples which truly have class x, i.e. how much part of the class was captured. It is equivalent to Recall. In the confusion matrix, this is the diagonal element divided by the sum over the relevant row, **i.e.** 38/ (38+2) =0.95 for class malicious and 50/ (50+1) =1 for class legitimate in our example.

- The Precision is the proportion of the examples which truly have class x among all those which were classified as class x. In the matrix, this is the diagonal element divided by the sum over the relevant column, **i.e.** 38/ (38+0) =1 for class malicious and 50/ (2+50) =0.962 for class legitimate. From the Confusion matrix in Fig. 6 we can see that two instances of a class "legitimate" have been assigned to a class" malicious", and zero of class "malicious" are assigned to class "legitimate".

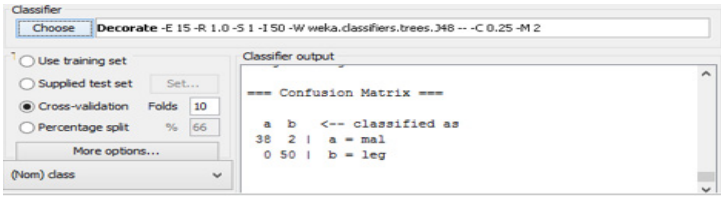


Fig. 6: Confusion matrix

Fig. 7 present the threshold curve for the prediction. This shows a 97.28% predictive accuracy on the malicious class.

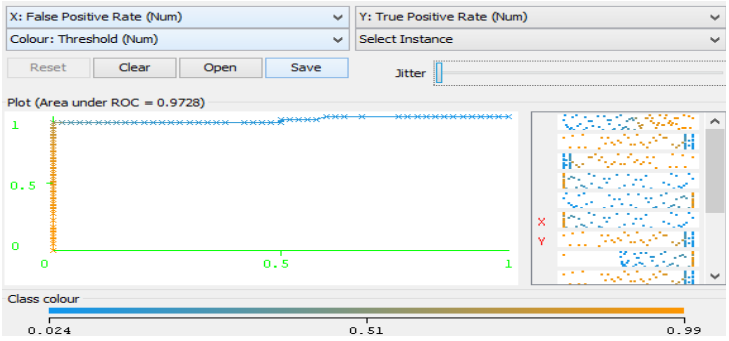


Fig. 7: Threshold curve

In order to find the optimal value of the threshold, we perform the cost/benefit analysis. Consider attentively the window for the Cost/Benefit Analysis. It consists of several panels. The left part of the window contains the Plot: Three should Curve frame with the Threshold Curve (called also the Lift curve).However, the axis X in the Threshold curve corresponds to the part of selected instances (the Sample Size). In other words, the Threshold curve depicts the dependence of the part of active compounds retrieved in the course of virtual screen in gup on the part of compounds selected from the whole data set used for screening. The confusion matrix for the current value of the threshold is shown in the Confusion Matrix frame at the left bottom corner of the window. We observed that the confusion matrix for the current value of the threshold sharply differs from the previously obtained one. In particular, the classification accuracy 74.4444% is considerably less than the previous value 97.7787%, the number of false positives has greatly increased from 0 to 21.

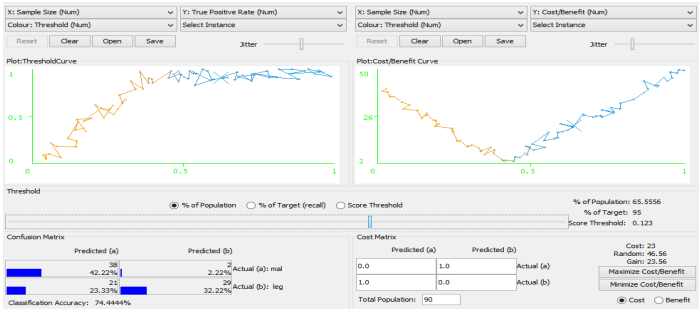


Fig. 8: cost/benefits analysis

In Fig.9, we generated a plot illustrating the prediction margin; the margin is defined as the difference between the probabilities predicted for the actual class and the highest Probability predicted for the other classes.

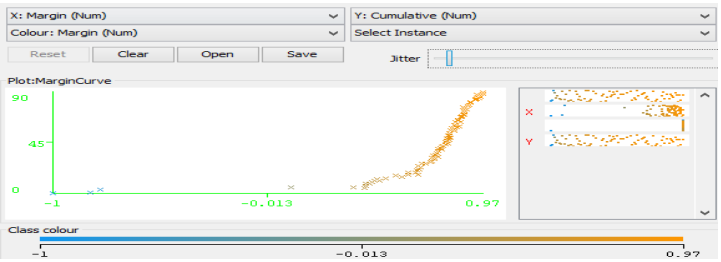


Fig. 9: Margin curve

In Fig.10 we present the performance of our malicious users detector trained with traditional features, honeypot based features and two sets of features together, respectively.

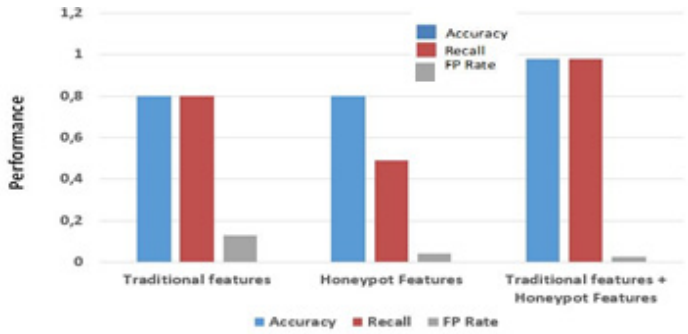


Fig. 10: Malicious user's detector performance with different feature set

We can find that after we combine traditional feature set with honeypot based feature set, we can achieve an accuracy of 0.979, a recall of 0.978 and a false positive rate of 0.028. The accuracy and recall are much better than simply using the other two feature sets independently. Though the FP rate is higher than simply using honeypot based feature set, we can modify the threshold to make a trade-off between FP rates and recall [2].

From the results obtained we can point out that with the hybrid method used (traditional feature set with honeypot based feature), we can detect more wide space of malicious users on social networks and why not apply the same approach to other communities. This hybrid approach gives results relevant to other methods.

3. Conclusion

In this paper, we have presented the results of a novel social honeypot-based approach to detect malicious profiles in social networking communities already published [2]. Our overall research goal is to investigate techniques and develop effective tools for automatically detecting and filtering malicious users who target social systems. Specifically, our approach deploys social honeypot profiles in order to attract malicious accounts. By focusing on twitter community, we use a set of user's characteristics and honeypots deployed characteristics to create a malicious profiles classifier based on machine learning algorithm Decorate for identifying malicious accounts with high precision and allow rate of false positives. In our ongoing work, we are using our analysis results to automatically identify malicious users in twitter social network. We have tried to apply our proposed approach using Weka to classify a set of users caught by our honeypots. From the results obtained we can point out that with the method of classification used, we can detect the more wide space of malicious users on social networks and why not apply the same approach to other communities. We hope this will give us a more macroscopic picture of the privacy awareness of general OSN (Online social network) users [13], and we want to use this to raise the awareness of privacy not only from the user sides but also for the OSN designers. Together with our research on detection malicious profiles in social network, we hope being able to contribute on making OSNs a safer place for the ordinary users [17, 18].

References.

[1] Kyumin Lee, James Caverlee, Steve Webb, Uncovering Social Spammers: Social Honeypots +Machine Learning, Department of Computer Science and Engineering Texas A&M University,2010.

[2] EL Bouzekri El idrissi Younes, El mendili Fatna, Maqrane Nisrine, A security approach for social networksbased on honeypots, 2016 4th IEEE International Colloquium on Information Science and Technology (CIST),Janvier, 2017.

[3] Kyumin Lee, James Caverlee and Steve WebbThe Social HoneypotProject: Protecting Online Communities from Spammers .2010.

[4] https://fr.wikipedia.org/wiki/Twitter

[5] http://blog.cozic.fr/twitter-presentation-du-concept-de-ses-differents-usages-et-de-quelques-applications-connexes-188

[6] Intelligence Artificielle Weka : Présentation février 2004

[7] Prem Melville : Creating Diverse Ensemble Classifiers : The University of Texas at Austin, 2003

[8] D. H. Joshua Goodman and R. Rounthwaite , A Stopping spam .Scientific American, 292(4):4288, April 2005.

[9] A. Nazir, S. Raza, and C.-N. Chuah , Unveiling facebook: a measurement study of social network based applications .In SIGCOMM, 2008.

[10] S. Ghosh, B. Viswanath, F. Kooti, N. K. Sharma, G. Korlam, F.Benevenuto, N. Ganguly, and K. P. Gummadi , Understanding and combating link farming in the twitter social network .In proceedings of the 21st international conference on World Wide Web, pp. 6170, ACM, 2012.

[11] C. Yang, R. Harkreader, J. Zhang, S. Shin, and G. Gu , Analyzing spammers social networks for fun and profit: a case study of cyber criminal ecosystem on twitter .In proceedings of the 21st international conference on World Wide Web, pp. 7180, ACM, 2012.

[12] V. Sridharan, V. Shankar, and M. Gupta , Twitter games: How successful spammers pick targets .In proceedings of the 28th Annual Computer Security Applications Conference, pp. 389398, ACM, 2012.

[13] G. Stringhini, C. Kruegel, and G. Vigna , Detecting spammers on social networks .In proceedings of the 26th Annual Computer Security Applications Conference, pp. 19, ACM, 2010.

[14] Hidemasa Naruoka*, Masafumi Matsuta, Wataru Machii, Tomomi Aoyama, Masahito Koike, Ichiro Koshijima and Yoshihiro Hashimoto, ICS Honeypot System (CamouflageNet) based on attacker's human factors. In 6th International Conference on Applied Human Factors and Ergonomics (AHFE 2015).

[15] Dr. Sudhir B. Jagtap, Dr. Kodge B. G., Census Data Mining and Data Analysis using WEKA. In International Conference in Emerging Trends in Science, Technology and Management-2013, Singapore (ICETSTM2013).

[16] Saini Jacob Soman and Murugappan, A study of spam detection algorithm on social media network, Department of Computer Science and Engineering, Sathyabama University, India, Journal of Computer Science 10 (10): 2135-2140, 2014

[17] Zhiyong Zhang a, b,∗, Brij B. Gupta c; Social media security and trustworthiness: Overview and new direction, a Information Engineering College, Henan University of Science and Technology, Luoyang 471023, People's Republic of China b Department of Computer Science, Iowa State University, Ames 50010, USA c National Institute of Technology, Kurukshetra, India, 2016 Published by Elsevier B.V.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 205-209 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Pulse Generator for Ultrasonic Piezoelectric Transducer Arrays Based on a Programmable System-on-Chip (PSoC)

Pedro Acevedo*, Martín Fuentes, Joel Durán, Mónica Vázquez, Carlos Díaz

Universidad Nacional Autónoma de México, Departamento de Ingeniería en Sistemas Computacionales y Automatización, Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, 04510, México.

ARTICLE INFO

Article history:
Submitted: 03 March, 2017
Accepted: 16 April, 2017
Online: 24 April, 2017

Keywords:
Ultrasonic Pulse generator
PZT and PVDF transducer arrays
Programmable System-on-Chip (PSoC) module

ABSTRACT

This paper describes the design of a pulse generator to excite PZT and PVDF ultrasonic transducer arrays, based on the Programmable System-on-Chip (PSoC) module. In this module, using programmable logic different pulses were implemented; these pulses are required in ultrasonic applications for multiple channels to excite PZT and PVDF transducer arrays. To excite multiple elements, bursts are required which can be generated simultaneously or out of phase, generating dynamic wave fronts. For medical applications where bidirectional blood flow is detected burst and quadrature pulses are used. These pulses can be generated independently or in combinations, as simultaneous pulses, shift pulses or burst. This module can operate with programmable frequencies from 3-74 MHz; its programming may be versatile covering a wide range of ultrasonic applications.

1. Introduction

In ultrasonic applications where it is required to focus the ultrasonic radiation fields, piezoelectric transducer arrays are designed, in order to obtain specific radiation fields. To get a specific response it should be consider the type of piezoelectric material, geometry of the array and the type of excitation for each element [1].

The focusing of the radiation fields depends largely on the form of excitation of the array elements; it is for this reason that the pulse generator for driving ultrasonic transducers arrays plays an important role in ultrasonic systems. The excitation may be in continuous or pulsed mode. The parameters associated with the pulse generator are: the operating frequency of the transducer, the repetition frequency of the pulse burst, the voltage amplitude and the excitation current. The pulsed mode allows determining the distance between the transmitter and receiver transducers, determining the time of flight of the ultrasonic signal, which is affected by the propagation medium. The ultrasonic transducers can be manufactured with PZT (Lead Zirconate Titanate) or PVDF (Polyvinylidene Fluoride) which are piezoelectric materials [1]. The type of transducer depends on the required specific application; PZT transducers can support an excitation of hundreds of volts, whereas PVDF can be driven with maximum voltages of 30 Volts.

Figure 1 shows different types of pulses most commonly used in ultrasonic applications.

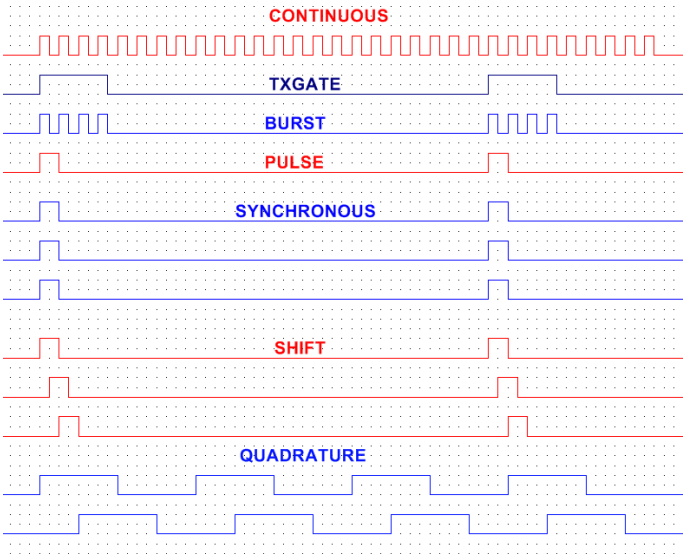


Figure 1 Excitation pulses for an ultrasonic system, a) continuous fix frequency, b) pulse gate, c) burst, d) single pulse, e) synchronous and simultaneous channels, f) shift pulses, g) quadrature pulses.

*Corresponding Author: pedro.acevedo@iimas.unam.mx

Figure 1 shows; a) an excitation signal with fixed frequency and continuous mode, b) gate pulse, c) an excitation signal in pulse burst mode in which the repetition frequency and the number of cycles may be programmable for systems in pulsed mode, d) single pulse, e) several channels with simultaneous excitation pulses for excitation of transducer arrays, f) multiple channels with shift pulse excitation for arrays with focusing applications, g) 2 channels with excitation in quadrature. In medical applications where bidirectional blood flow is detected, signals in quadrature are used for the demodulation of Doppler signals [2-3]. These pulses can be generated independently or in combinations, such as simultaneous pulses and shift or burst.

In applications where it is required to focus dynamically the ultrasonic beam, transducer arrays are used with multiple piezoelectric elements, such as that shown in Figure 2. For excitation of multiple elements pulse bursts are required which can be generated simultaneously or out of phase (with time delays between them) in order to generate dynamic wave fronts.

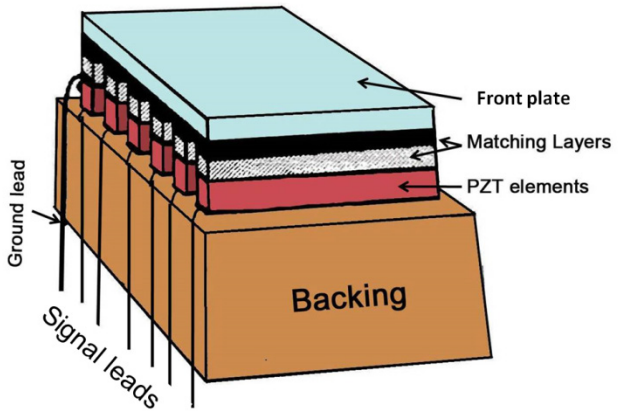


Figure 2. Linear transducer array

The focal depth is determined by the time delay between the electrical pulses. This can be changed electronically to focus pulses to give a good image detail at various depths within the body rather than just one depth with fixed focus transducer. One approach is to create an image by using a sequence of pulses, each one focused to a different depth or zone within the body. The acoustic parameters of the resulting beam, such as incidence angle, focal length and focal point size can be manipulated by the generated front waves in the array, controlling the excitation pulses via software [2].

A pulse generator to excite PZT and PVDF transducer arrays is presented. This pulse generator is based on a PSoC-5LP (Programmable System on Chip) programmable module, which is a USB port programmable development kit, from Cypress Semiconductor PSoC. [2-4]

The PSoC-5LP module has a microcontroller and a set of analog and digital programmable peripherals, all in one integrated circuit along with a 32-bit CPU (ARM Cortex family), The PSoC in the development kit belongs to the 5LP family; this family has an ARM Cortex M3 CPU with a clock rate that can reach up to 80 MHz [4].

The main difference between the PSoC microcontroller and conventional microcontrollers is that in the latter the set of peripherals are defined by the manufacturer, while the PSoC due to its programmable logic composed by Universal Digital Blocks (UDBs) and a set of analog and digital peripherals can be programmed (customize) in order to perform the required functions [5].

The implementation of the design of the programs for PSoC is made using its own IDE (Integrated Design Environment) PSoC Creator 3.2 Service Pack 2[6]. The programming process requires of a schematic diagram design, then an assignment (input/output port), and finally if required to write a C program, this process is integrated in programmable analog logic and digital functions.

2. Methodology

In this work, a programmable CY8CKIT-059, PSoC 5LP module was used.

The PZT transmitter transducer used is a 64 elements linear array (Krautkramer 116-001-264 model) [9], this PZT array has an operating frequency of 3.5 MHz (Figure 3a).

A PVDF Piezoelectric Transducer Sensor (SDT1-28K model) from Shielded Measurement Specialties [7-8] (Figure 3b) and an Ultrasonic Pulser Receiver, Olympus 5072PR (30 dBs amplifier) were used for the reception of the ultrasonic pulses

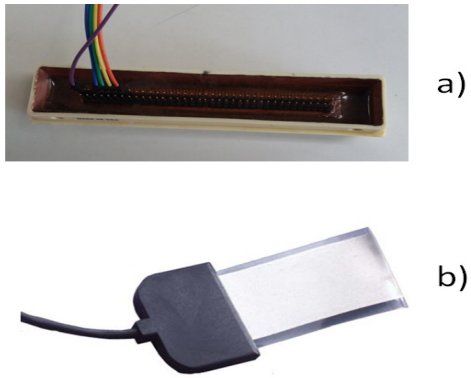


Figura 3: a) Top view of the PZT transmitter array; b) PVDF Piezo transducer Sensor, SDT1-28K model, (16 mm x 30 mm and thickness 75 µm).

To excite the PZT transducer array, it was required to generate 3.5 MHz pulse bursts corresponding to the operating frequency of the array, with a 10 kHz repeating frequency. These pulse bursts are shown in Figure 4.

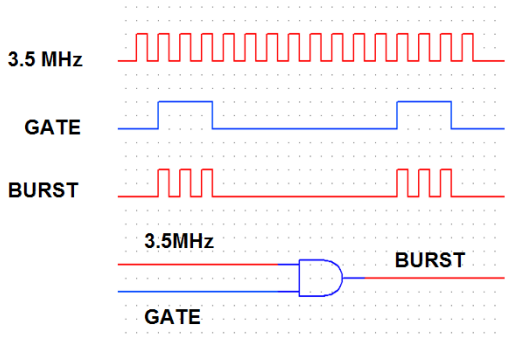


Figure 4. Time diagram of the excitation pulses of the PZT array.

In order to generate the operating frequency of the PZT transducer, the internal PSoC clock was programmed at 40 MHz. An UDB in PWM mode (PWM_35M) was also implemented to obtain a 3.5 MHz frequency, as illustrated in Figures 5 and 6.

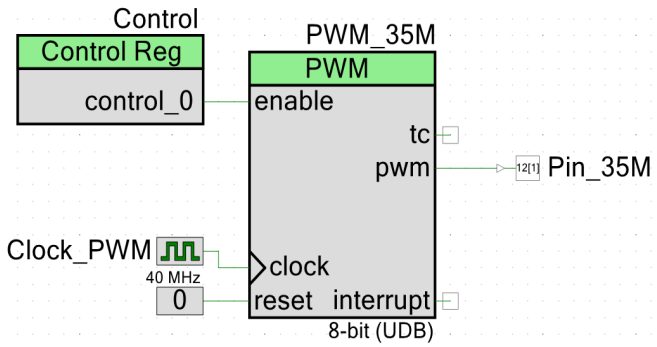


Figure 5: Implementation of the 3.5 MHz clock.

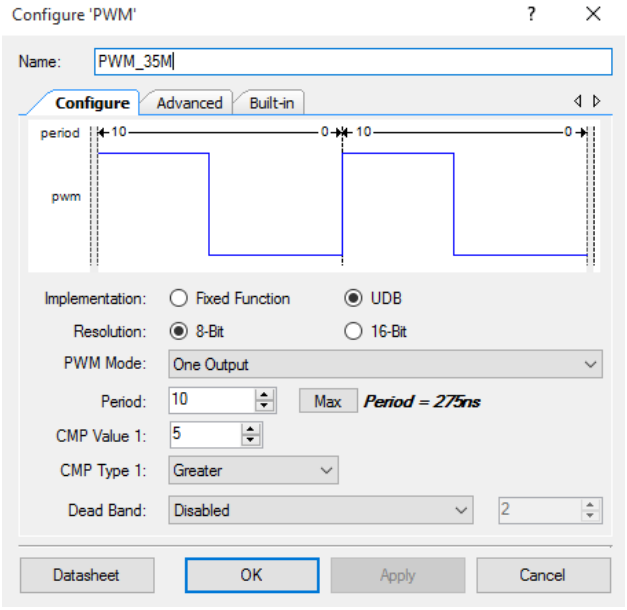


Figure 6. UDB PWM dialog window

In order to generate the 10 kHz repetition frequency a second UDB in PWM mode (PWM-Gate) was used in conjunction with an AND gate to generate the burst pulse, as shown in Figure 7.

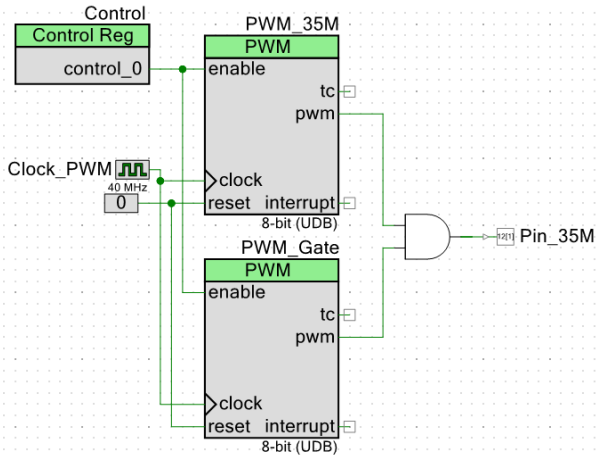


Figure 7 . Implementation of the excitation pulse bursts.

To test the excitation pulse generator, an experimental platform like the one shown in Figures 8 and 9 was implemented. A PSoC CY8CKIT-059-5LP module was used to generate the pulses and excite the PZT transducer elements. A PVDF transducer was used as a receiver, then the signal is amplified 30 dB and acquired with a digital oscilloscope, finally this signal was processed on a personal computer.

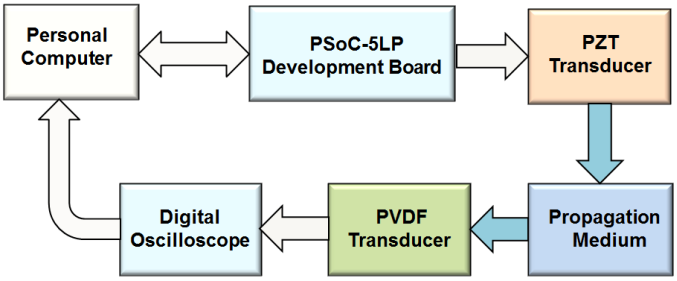


Figure 8: Block diagram of the experimental setup.

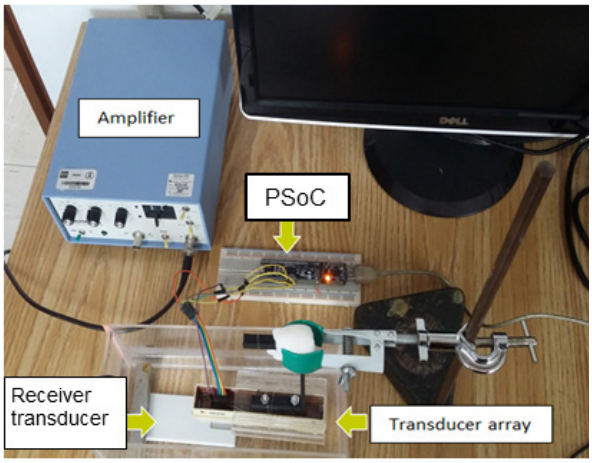


Figure 9: Experimental setup.

3. Results

To test the pulse generator, pulse bursts were programmed in four output ports, which can deliver pulse bursts simultaneously or out of phase as shown in Figures 10 and 11.

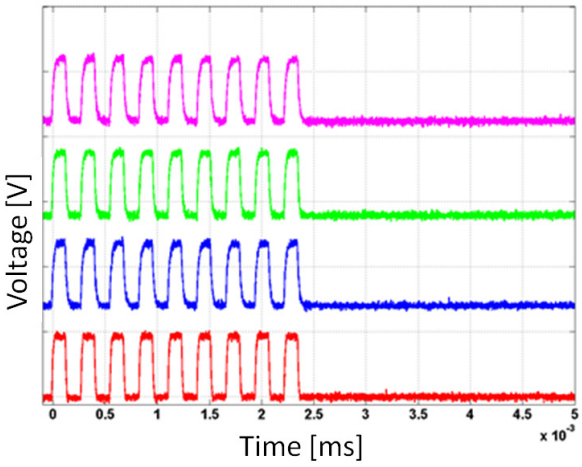


Figure 10. Simultaneous pulse bursts.

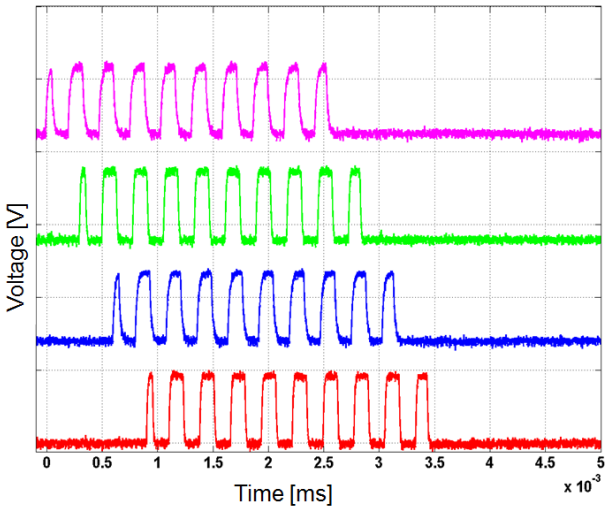


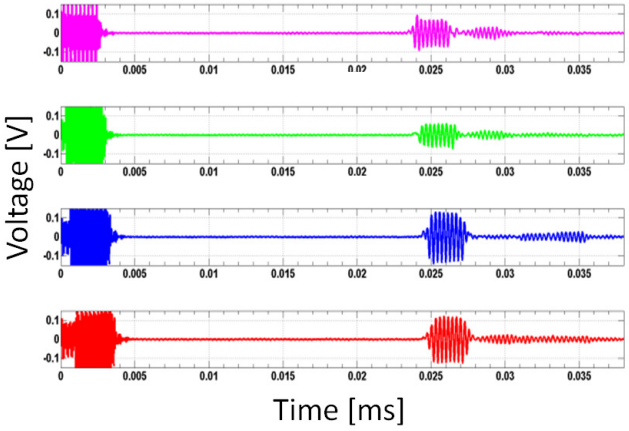
Figure 11. Out of phase pulse bursts.

Pulse generation of the four output ports was obtained by repeating the implementation shown in Figure 7 for simultaneous outputs and programming a 25 ns time delay between them for the out of phase outputs.

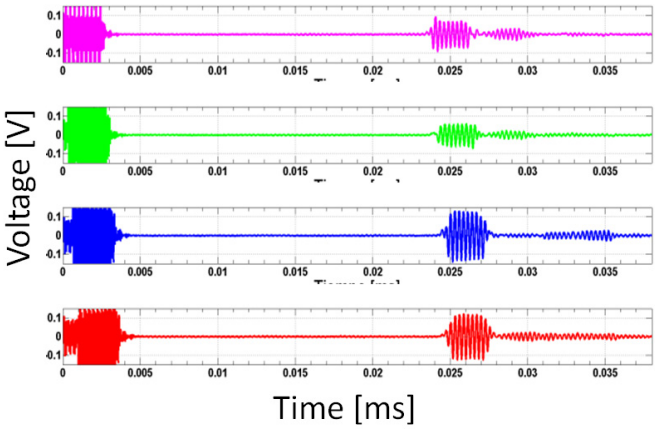
Using an experimental platform like the one shown in Figure 8, two experiments were conducted. The first experiment consisted in generating pulse bursts in four output ports to simultaneously excite four piezoelectric elements of the PZT 64 elements array (Figure 2), then more pulse bursts were generated in four ports to exciting the PZT array but this time with a 25 ns delay between each port. The results of this experiment are shown in Figure 12. Figure 12a shows the acquired signals with simultaneous excitation and Figure 12b shows the acquired signals with phase shifted excitation. The signals were acquired using a linear PVDF transducer array.

The second experiment was similar to the first one, the only difference consisted that in the acquisition of the signals, a single PVDF transducer was used. Figure 12a shows the acquired signals with simultaneous excitation and Figure 12b shows the acquired signals with phase shifted excitation.

The wave fronts generated by the four elements of the linear PZT array are shown in Figure 12, and correspond to a distance of 30 mm between the PZT and PVDF transducers. The signals have a 30 dB gain with a 500 ohms input impedance.

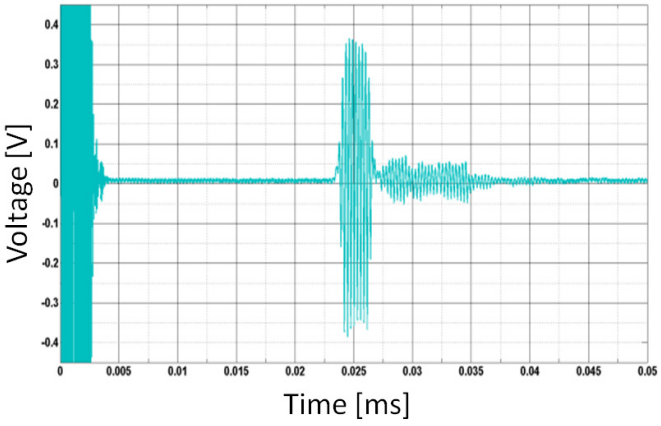


(a)

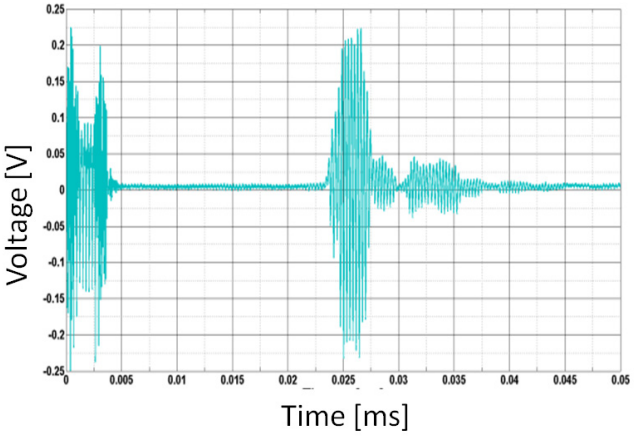


(b)

Figure 12. PVDF transducer response as a receiver, with one element at a time, with amplification of 30 dB (a) Simultaneous and (b) out of phase.



(a)



(b)

Figure 13. PVDF transducer response as a receiver, receiving the signal from four elements of the array at the same time, with amplification of 30 dB (a) Simultaneous and (b) out of phase.

4. Discussion

The precedent of this research work was the implementation of a pulse generator using the Arduino Mega platform [3], in this implementation it was not possible to obtain synchronous pulses in multiple output ports with frequencies above 120 kHz, since its programming does not allow it. Instead the PSoC modules have a

programmable hardware that allows programming the operation clock signal and thus programming output ports with a wide frequency range. It is also possible to program simultaneous output ports and phase shifts between ports, since such delays between signals are implemented with a delay by software of one cycle of the master clock. There is also a previous article in which this subject is presented with less extension [10].

A pulse generator based on a PSoC-5LP to excite PZT and PVDF transducer arrays is presented, it is possible to excite PZT or PVDF elements with 5 volts of amplitude, for applications requiring more power additional voltage power supply and current drivers are required. Figure 14 shows two PZT driver circuits, a) single pulse and b) burst.

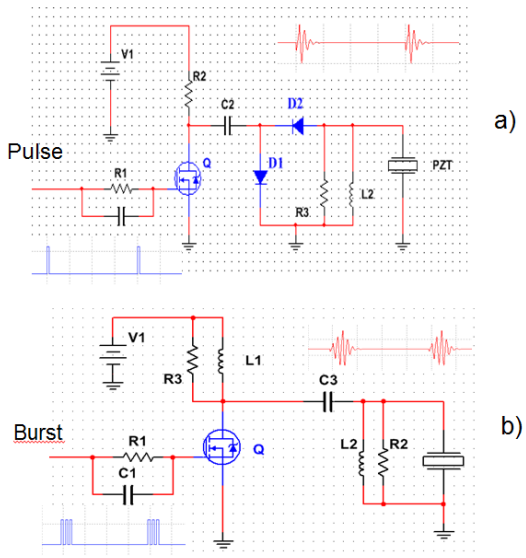


Figure 14. PZT driver circuit, a) driver circuit for single pulse, b) driver circuit for burst pulse.

5. Conclusions

The pulse generation with PSoC-5LP modules is a good alternative because as demonstrated in this work, it is possible to generate different pulses to excite transducers, either in one or multiple output ports, continuous or pulse burst, simultaneously or out of phase.

The advantage of the pulse generator implemented using the PSoC-5LP module with respect to commercial generators is its versatility. It can operate programmable frequencies from 3-74 MHz, which covers a wide range of ultrasonic applications

With the pulse generator implemented in the PSoC-5LP module, it is possible to excite PZT or PVDF elements with 5 volts of amplitude, resulting in an excellent alternative for low power ultrasound applications.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

Authors thank DGAPA for its support, projects PAPIIT IN106016 and PAPIIT IT101316.

References

[1] Antonio Arnaud Vives. Piezoelectric Transducers and Applications. Second Edition. Berlin Heidelberg, Springer, 2008.

[2] Pedro Acevedo, Mónica Vázquez, Joel Durán, Rodolfo Petrearce. “A pulse generator based on an Arduino Platform for ultrasonic applications,” 2015 International Congress on Ultrasonics, 2015 ICU Metz Physics Procedia 70, 1096-1099, Elsevier 2015

[3] Pedro Acevedo, Carlos Díaz, Mónica Vázquez, Joel Durán. Design of a Pulse Generator Based on a Programmable System-on-Chip (PSoC) for Ultrasonic Applications. International Journal of Mechanical, Aerospace, Industrial, Mechatronic and Manufacturing Engineering Vol:10, No:3, 453-456, 2016. World Academy of Science, Engineering and Technology. USA.

[4] PSoC 5LP. Architecture Technical Reference Manual. January 2016. Cypress Semiconductor Corporation. Web site: http://www.cypress.com/file/123561/download.

[5] PSoC 5LP Prototyping Kit Guide. January 2016. Cypress Semiconductor Corporation. Web site: http://www.cypress.com/file/157971/download

[6] PSoC Creator TM Release Notes. January 2016. Cypress Semiconductor Corporation. Web site: http://www.cypress.com/file/213966/download

[7] Piezo Film Sensors. Technical Manual. December 2015. Measurement Specialties. Web site: http://www.meas-spec.com/piezo-film-sensors.aspx.

[8] SDT Shielded Piezo Sensors. December 2015. Measurement Specialties. Web site: http://www.meas-spec.com/downloads/SDT_Series.pdf

[9] Krautkramer Transducer Certification. Model 116-001-264. Krautkramer Branson Inc. 1999.

[10] Carlos Díaz, Mónica Vázquez, Martin Fuentes, Joel Durán, Pedro Acevedo. “Pulse Generator for Ultrasonic PZT and PVDF Transducer Arrays Based on a Programmable System-on-Chip (PSoC)”, 2016 13th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE), Mexico, City. México. Septiembre 26-30, 2016. 978-1-5090-3511-3/16 ©2016 IEEE.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 210-216 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Enabling Toy Vehicles Interaction With Visible Light Communication (VLC)

M. A. Ilyas[1*], M. B. Othman[1], S. M. Shah[2], Mas Fawzi[3]

[1]*Optical Communication and Network Group (OpCon), Faculty of Electrics and Electronics Engineering, Universiti Tun Hussein Onn Malaysia, Parit Raja, Batu Pahat, 86400 Johor,Malaysia*

[2]*Research Centre for Applied Electromagnetic (Emcenter), Universiti Tun Hussein Onn Malaysia,86400 Parit Raja, Batu Pahat, Johor, Malaysia*

[3]*Centre for Energy and Industrial Environment Studies (CEIES), Universiti Tun Hussein Onn Malaysia, Parit Raja, Batu Pahat, 86400 Johor, Malaysia*

ARTICLE INFO

Article history:
Received: 05 March, 2017
Accepted: 12 April, 2017
Online: 28 April, 2017

Keywords:
Visible light communication
Vehicle to vehicle communication
Wireless optic

ABSTRACT

An effective communication among the road users is crucial for safety precautions. Nowadays, a light emitting diode (LED) are commonly installed on a motor vehicle as the lighting system. The LED lights are more reliable and flexible in terms of the design and power adjustments. To further utilize the LED applications in a vehicular system, a simple data transmission system based on LED lights for short distance is developed. Toy vehicles are used to demonstrate the functionality of the system. This paper explains the components used and the communication protocols of A and B that have been developed and tested. The implementations of a software-based modulation and communication protocols by using a microcontroller can be considered as the initial approach that can be further enhanced for future development. In this paper, the prototype of the system as well as the experimental results are presented. The prototypes are able to send short messages by using the LED and photodetector (PD) which could reach up to 2-meter distance. In addition, it is also believed that this project can be implemented on the real motor vehicles subject to a few improvements.

1. Introduction

This paper is an extension of the work originally presented in 2016 IEEE Student Conference on Research and Development (SCOReD) [1]. The previous work focused on the development of communication protocol and data transmission performance of the toy vehicle VLC system. On the other hand, this work presents the application and prototype of the toy vehicle interaction protocol that have been mentioned in the previous paper.

This work is inspired from one situation that involves a multiple road users. The situation occurs in the middle of the night, on a single lane road, where the driver of a car needs to overtake a slow truck. However, due to the size of the truck, the car's driver is not able to see the traffic from the other side of the road without slightly moving to the opposite lane which is dangerous and might be fatal. Therefore, in order to have a safer driving experience, standard vehicles such as cars and trucks must have several signalling devices such as indicators, brake light, hazard warning lights, headlights and reverse lights [2]. However, the current technology can only provide a limited number of signalling signs such as turning left or right, brake signal, reverse signal, and hazard warning signal.

In the year 2012, the car manufacturer, i.e.: Volvo, started to build a system that enables the communication of a car-to-car and car-to-medium. These types of communication ensures safer driving experience. The technology relies on the GPS system and a dedicated wireless network, in which the car-to-car communication has already acquired a separate range of radio frequency for that purpose. Vehicles within a given radius would be automatically linked together and able to exchange information such as the position, speed, and direction [2,3].

*Corresponding Author: M. A. Ilyas, Universiti Tun Hussein Onn Malaysia, Malaysia | Email: arif.b.ilyas@gmail.com

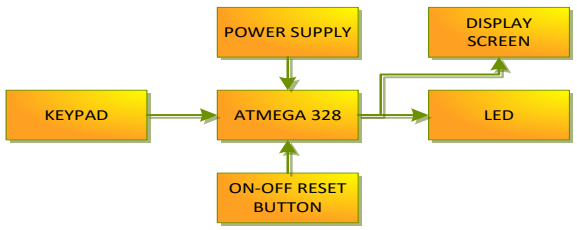
Nowadays, the radio frequency spectrum is highly congested due to the demand for wireless communication applications [3]. As a natural source of energy, a visible light can be thought as an alternative to the RF communication. The visible light communication (VLC) possesses a few key advantages over the wireless RF transmitters. It has a nearly infinite bandwidth, free and unregulated channels, poses no health hazards and concerns, and consumes much less power per given transmission [4,5]. VLC is also a potential low-power and low-cost alternative to the traditional short-range wireless RF communications [6-8].

In this work, a data transmission system for two toy vehicles has been developed by using a light emitting diode (LED) which is highly reliable for short distance communication. It has also been demonstrated that the current LED signalling technology can be further utilized by using the VLC. The LED provides a visual indicator as well as a communication channel for both vehicles to exchange information. Additionally, it provides a medium for the front toy vehicle to convey any information efficiently to the rear toy vehicle.

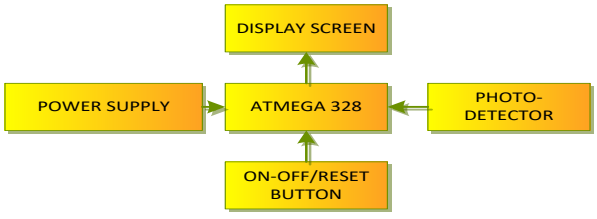
This paper is organized as follows: Section 2 introduces the design method for the VLC system which is followed by Section 3 that discusses the results and discussion. Finally, the paper is concluded in Section 4.

2. Visible Light Communication (VLC) Design

This section focusses on the VLC circuit development starting from the LED selection until the successful implementation of communication between the LED and photodetector (PD). As shown in Fig.1, the transmitter circuit consists of a keypad, LED, and display screen. On the other hand, the receiver circuit only has a PD and display screen. Both circuits have a common microcontroller which is known as Atmega328. It is the brain of the circuit and is responsible for all operations that include the modulation, signal processing and keypad response. When the keypad is pressed, the microcontroller gives an order to the LED to transmit a signal. The data are transmitted using on-off keying modulation (OOK). At the receiver, the PD will detect the signal which will then trigger the microcontroller to demodulate the transmitted signal as well as activating the display screen to show the message.



(a) Transmitter circuit



(b) Receiver circuit

Figure 1: Block diagram of VLC circuit [1]

2.1. Optical hardware selection

In order to establish the VLC communication, an appropriate LED and PD selection must be performed. In this work, the choice of LED must take into account the cost, efficiency, and performance. The LED model chosen in this work is COM-00531. It is classified as a super bright LED and emits white colour. The LED viewing angle is very narrow at 10°, as can be seen from the specification in Table 1 and further illustrates in Fig. 1. The narrow angle will lead to a more focused beam of light that provides better detection. In addition, its emission spectrum is very wide, ranging from 400 nm to 800 nm.

Table 1: Specifications of COM-00531 LED

Material of construction	InGaN/GaN
Emitted colour	White
Lens color	Water clear
Viewing angle	10°
Typical luminous intensity at 20mA	10cd
Typical forward voltage at 20 to 30mA	3.5V
Reverse Current	10 uA

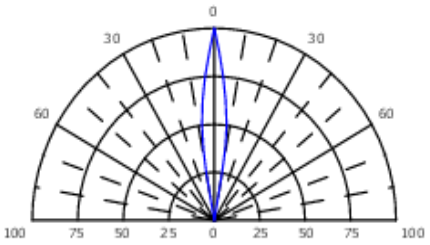


Figure 2: LED viewing angle

As for the PD, a 5-mm PD (flat head) is selected as the light sensor for this work. Table 2 shows the specification of the PD. The justification of this selection is that the PD could detect a light from 400 nm to 1100 nm that covers the visible light range of wavelengths. In addition, it also has a high reverse voltage that could be used to alter the sensitivity of the photo detection for higher detection performance.

Table 2: Specifications of 5-mm PD (flat head)

Reverse voltage	35V
Power dissipation	50mW
The peak wavelength sensor	940nm
Spectral bandwidth	400 - 1100nm
Reverse light current	40uA

A few experimental procedures have been conducted to evaluate the performance of the LED and PD. An oscilloscope is used to observe the behavior and functionality of the LED and PD. The main objective of these experiments is to determine the maximum rate of data transmission for the LED-PD pair. By comparing the signal generated by the PD to the signal provided by the LED, a performance analysis can be examined.

2.2. Electrical Domain

In this section, the VLC design process in an electrical domain is discussed which consists of OOK modulation, communication protocol and system integration.

2.2.1 On-Off Keying (OOK) Modulation

In this work, a non-return-to-zero on-off-keying (NRZ-OOK) with direct detection (DD) is used for modulation as shown in Fig. 3. From the figure, Signal 1 is represented by 5 V and Signal 0 by 0 V, respectively. The PD is continuously detecting the data by using a time delay method that samples the data continuously for every period of time. The modulation is implemented in a software level on the microcontroller as to ease the design process.

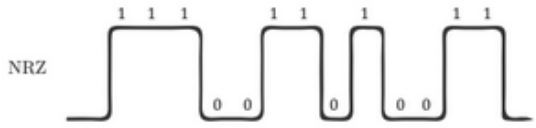


Figure 3: Bit sequence modulation

In this experiment, the possibility of sending signals between the LED and PD is verified. Furthermore, the maximum possible speed of operation for a certain communication distance is investigated. The programming code for this experiment has the following roles:

- To receive the binary input signal from the user's computer (convert the ASCII characters to binary data) – in the transmitter.
- To send data by using an OOK modulation.
- To detect these signals by using a PD.
- To demodulate the message and recover the original binary string.

Once the programming code has been developed and uploaded to each circuit, the experiment is conducted by placing the transmitter and the receiver at a separation distance of 1 cm and 5 cm. Initially, to test the system, a time delay of 1000 milliseconds (ms) is setup for the delay function. Depending of the successful achievement of the experiment, the delay time is then shortened accordingly. For a start, a random binary string is chosen: 01010101 (ASCII: 'U'). If the serial monitor's output produced the same binary string, the experiment can be considered as successful. The experiment setup with the programming code consists of a LED as the transmitter and a PD in the receiver. A binary data sent by the transmitter should be displayed on the receiver.

2.2.2 LED-PD Communication: The Setup of Communication Protocol

This section discusses the setup of communication protocol in the communication system namely, Protocol A and B.

In the previous experiment, the binary string of 01010101 is not the only signal received as the receiver cannot distinguish whether the signal is an actual data or a noise. The serial monitor's output can be viewed Fig. 4. In order to solve this problem, a communication protocol must be established.

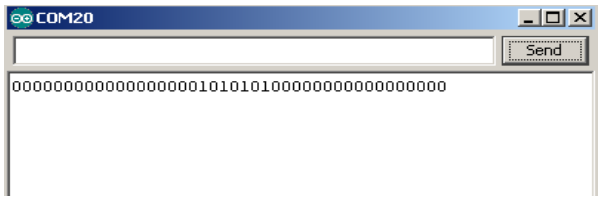


Figure 4: Receiver serial monitor output

Since the objective is to develop a working interactive communication between the two toy vehicles by using a LED signalling, the needs to convert the received binary data to a real and readable information is compulsory. Hence, another programming code is written that would have all the existing features from the previous experiment and also the ability to send an actual text.

In order to get the new code working, two additional features need to be added which are:

- The conversion from text to binary – in the transmitter.
- The conversion from binary to text – in the receiver.

2.2.2.1 The Setup of Communication Protocol A

The best way to address the problem is to add another LED for the transmission to provide a synchronous data transfer for both devices. The communication system would now work with the following procedures:

i. Once the message has been input on the transmitter's serial monitor, the LED begins sending the message through OOK.
ii. If the receiver secondary PD receives any signals from the transmitter secondary LED, it tells the receiver primary PD to start receiving the data sent from the transmitter primary LED.
iii. Once the transmitter primary LED finishes sending the message, the transmitter secondary LED turns off. Thus, the receiver secondary PD will stop receiving any signals and will tell the receiver primary PD to stop receiving data.
iv. After conversion is done, it will then be displayed on the monitor.

The hardware experimental setup consists of two LEDs as the transmitter and two PDs in the receiver circuit. The primary LED is used for data transmission while the secondary LED and PD are used to indicate the start and end of a transmission period.

2.2.2.2 The Setup of Communication Protocol B

Now that the transmitter needs two LEDs to set up a successful communication to the receiver, it is not practical to be implemented in real life situation because the LEDs need to be placed directly at the opposite of the receiver. Therefore, a new protocol is developed so that only one LED is needed for every data transmission. Fig. 5 shows the protocol that has been implemented in the transmitter and receiver system. Every data transmission begins with a start bit, followed by the information on the total bit number and the actual data packet.

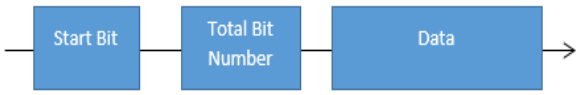


Figure 5: Communication protocol B

Thus, this experimental setup only comprises of one LED as the transmitter and one PD in the receiver circuit. A string data is sent by the transmitter by using a new protocol and will then be displayed on the receiver.

2.3. VLC Complete Circuit Development

This section focuses on the complete VLC circuit development which consists of a 4 × 4 keypad, 5110 LCD, PD, and LED.

2.3.1 Transmitter Circuit with 4 × 4 Keypad and LCD

In this section, a programming code is written to combine the transmitter circuit with a 4 × 4 keypad. Each button has been assigned to a specific string of characters and it is used as a source of information to be transmitted. When a button is pressed, the string will be displayed on the 5110 LCD before being converted to the binary string by using the ASCII codes. The binary data will then be used for OOK modulation. Table 3 shows the characters that has been assigned to each button.

Table 3: String of characters assigned to each keypad button

Button	Proposed LED Technology (Visual and Data Transmission)
1	Turn left
2	U-turn
3	Turn right
4	Thanks
5	Overtake
6	Welcome
7	Emergency
8	Too bright
9	Open light
B	Fire truck
C	Police
D	Ambulance

Fig. 6 shows the hardware setup of the experiment which consists of a LED, 5110 LCD, 4 × 4 keypad and Arduino microcontroller as the programming platform.

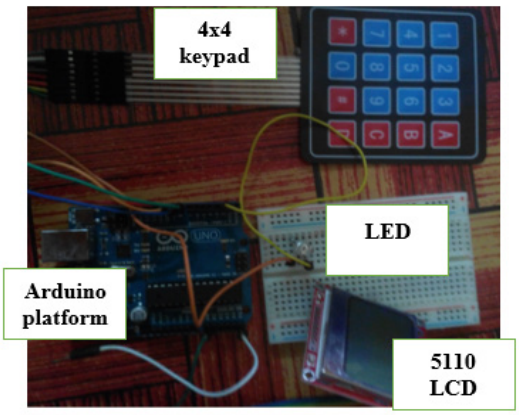


Figure 6: Transmitter circuit with 4 × 4 keypad and LCD

2.3.2 Receiver circuit with LCD

In this section, the receiver is used to display the received characters on a LCD screen. As the receiver received the characters, it is sent to the LCD to be displayed. There is no predetermined character on the LCD. All the received characters are totally based on the received signals.

Fig. 7 shows the hardware setup of the experiments which encompasses of a LED, 5110 LCD and Arduino microcontroller as the programming platform. The performance of the combined receiver circuit is evaluated with the combined transmitter circuit.

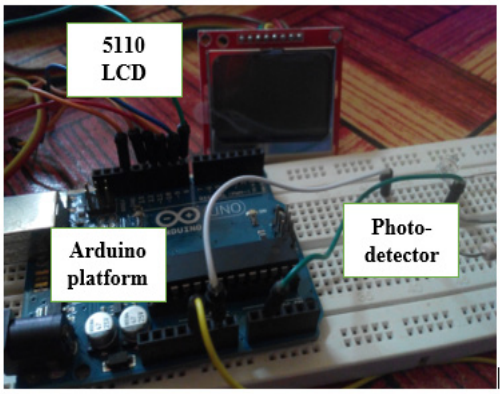


Figure 7: Receiver circuit with LCD

3. Results and Analysis

In this section, the results in optical domain, electrical domain, and circuit integration of each experiment are presented and analysed.

Fig. 8 is the oscilloscope screen that displays the result of the experiment to test the functionality and performance of a PD. Signal 1 is the square signal that is injected into the LED with a function generator. On the other hand, Signal 2 is the behavior of PD in the presence of light from the LED. It is observed that when the LED is having a high voltage that represents bit 1, there is a

voltage rise on the PD. Conversely, when there is no voltage on the LED, the voltage at the receiver decreases. However, the rate of voltage falling on the PD is significantly slower as compared to the rate of voltage rising. This indicates the maximum data rate that the photodiode could handle. In the experiment, the frequency used for injecting the signal to the LED is approximately 300 Hz. In order to have a reliable system, the maximum data rate should be lower than 300 Hz so that the fall time of PD is not an issue during data processing.

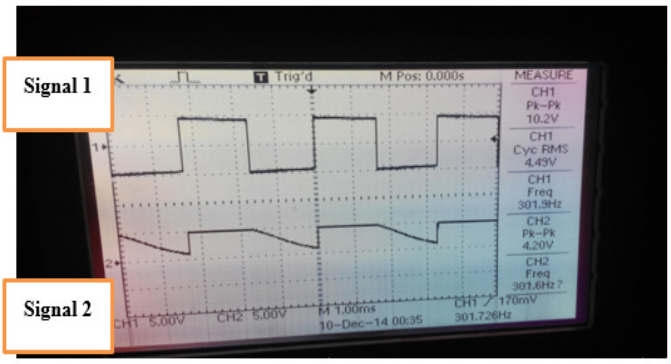


Figure 8: Oscilloscope display

3.1. Electrical Domain

In this section, the results for electrical domain experiments that cover data modulation and communication protocol are presented and analysed.

3.2.1 OOK Modulation Performance

Table 4 and 5 show the experiment results in Section 2.2.1 that compare the delay time and the status of the received string as the distance is increased from 1 cm to 5 cm. The experiment is conducted to find the limit of data transmission rate that could be handled by the system. The results indicate that the maximum time delay that the transmission could handle is 5 milliseconds. This work does not require a fast data transmission rate, but rather a reliable system. However, the distance covered is not long enough due to the limitation of using only one LED. A LED produces a noncoherent and diverging light output pattern. At greater distances, the intensity of the light are becoming much lower and restricts the receiver LED to receive the data.

Table 4: Results for 1 cm distance

Delay time (ms)	*Status of the received string*
1000	Correct
100	Correct
10	Correct
5	Correct
1	Wrong

Table 5: Results for 5 cm distance

Delay time (ms)	*Status of the received string*
1000	Correct
100	Correct
10	Correct
5	Correct
1	Wrong

3.2.2 Communication Protocol A and Protocol B

Fig. 9 and Table 6 show the results of the experiments in Section 2.2.2.1 and 2.2.2.2. The experiments are conducted to determine the performance of the communication protocols when implemented on the actual system. Both protocols are developed to enable the receiver to distinguish between the actual data and unwanted data. Therefore, the receiver will only process the transmitted data and ignore the unwanted data. It also helps in synchronizing the transmitter and receiver transmission timing so that an accurate and reliable data transfer could be established. However, only communication Protocol B is used in the final design since it offers more flexibility as compared to Protocol A. Further explanation can be read in [1].

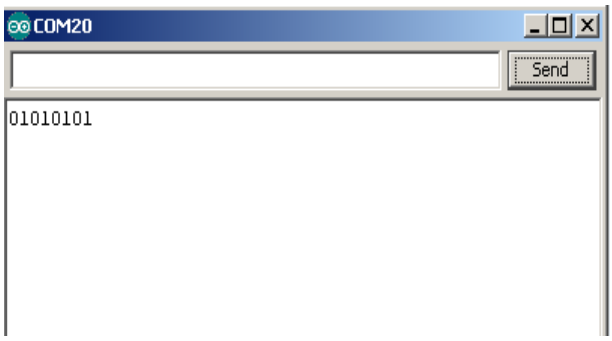


Figure 9: Receiver serial monitor output

Table 6: Results for communication Protocol A and B

Character transmitted	*Status*
Turn left	Successfully Received
U-turn	Successfully Received
Turn right	Successfully Received
Thanks	Successfully Received
Overtake	Successfully Received
Welcome	Successfully Received
Emergency	Successfully Received
Too bright	Successfully Received
Open light	Successfully Received
Fire truck	Successfully Received
Police	Successfully Received
Ambulance	Successfully Received

3.2. VLC Complete Circuit Integration

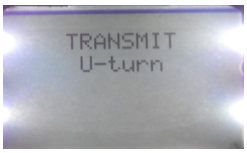
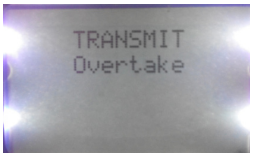
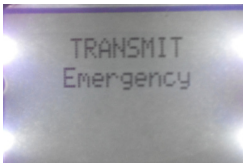
In this section, a complete circuit of VLC system is presented. The transmitter and receiver circuits have been successfully developed.

3.3.1 Complete Transmitter Circuit

The transmitter circuit has been successfully combined with a 4 × 4 keypad and 5110 LCD. The keypad is used to initiate the data transmission and the characters are displayed on the 5110 LCD. A minor adjustment is made where the user has to press button A to initiate a transmit mode. The changes are made so that the transmitter is compatible with the receiver circuit when they

are combined. The program uses 10,962 bytes (33%) of the program storage space that have a total of 32,256 bytes. On top of that, the global variables use 1,204 bytes (58%) of the dynamic memory, leaving only 844 bytes for local variables. Table 7 shows the results of the transmitter circuit that has been integrated with a 4 × 4 keypad and 5110 LCD.

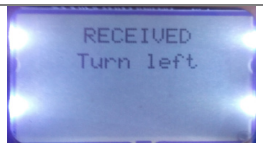
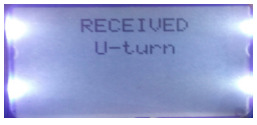
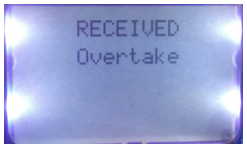
Table 7: Results for transmitter circuit

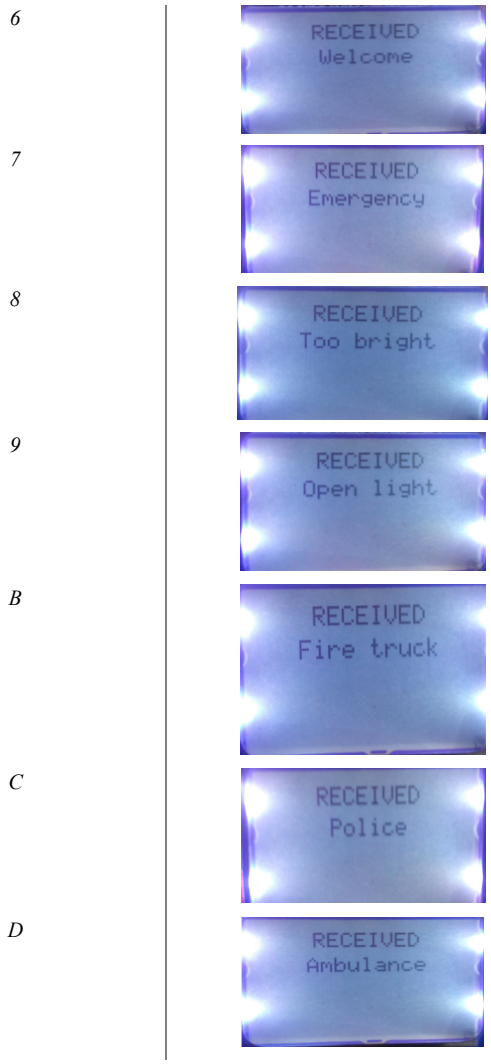
Button Pressed	5110 LCD
1	
2	
3	
4	
5	
6	
7	
8	
9	
B	
C	
D	

3.3.2 Complete Receiver Circuit

A combination of the receiver circuit and 5110 LCD has been successfully performed. The receiver circuit will continuously check the condition of PD. If any transmission is detected, the received characters will be displayed on the 5110 LCD. The program uses 10,466 bytes (32%) of the program storage space while the global variables use 1,299 bytes (63%) of the dynamic memory, leaving only 749 bytes for local variables. Notice that the global variable consumes so much dynamic memory which can make the program to be slightly unstable. Some of the original data have to be simplified so that a minimum number of global variables is used. Table 8 shows the results of an integration of the receiver circuit with 5110 LCD.

Table 8: Results of receiver circuit

Transmitter Pressed Button	Received signal displayed in 5110 LCD
1	
2	
3	
4	
5	



4. Conclusion and Future Work

In this work, a visible light data communication that can function as a transmitter and receiver has been developed. A LED and PD are used as the light source and light sensor, respectively. In addition, an interactive communication between the two prototype toy vehicles by using a LED signaling has also been constructed in which the transmitter could transmit any characters or information to the receiver by using a single source of light. Currently, the characters that are being assigned to the transmitter are turn left, turn right, U-turn, police indicator, ambulance indicator, and fire truck indicator.

There are, however, a few limitations of this work, which can be attributed from the circuit connections, the written programming code and the platform that is being used. The proposed suggestions are listed below to overcome the limitations which can be considered for future work:

- Faster Processing Platform
  A faster processing platform such as a FPGA can be used to enable data to be processed at a faster speed. In addition, it also helps with the synchronization process between data transmission and enables a large number of bits to be sent correctly.
- Bluetooth Device Enhancement.
  The visible light data communication system can be equipped with other systems such as a Bluetooth to add to the versatility. It will enable the user to change the characters manually for each button by using a smartphone. In other word, it will give the user more freedom to customize the system.
- Brighter LED
  Brighter light source implementation will make the system to be a good candidate to be used in a real motor vehicle. Not only it can increase the distance of transmission, but a brighter LED also provides higher flux density within a certain area. However, a larger power source needs to be used so that it can support the brighter LED

Conflict of Interest

The authors declare no conflicts of interest.

Acknowledgment

This authors would like to acknowledge the Ministry of Higher Education Malaysia (MOHE) and Universiti Tun Hussein Onn Malaysia (UTHM) for the financial support under the Fundamental Research Grant Scheme (FRGS) (Vot No: 1416).

References

[1] M. A. Ilyas, M. B. Othman and M. F. M. Ali, "Two toys vehicles interactions using communication protocol for visible light communication," 2016 IEEE Student Conference on Research and Development (SCOReD), Kuala Lumpur, 2016, pp. 1-6. doi: 10.1109/SCORED.2016.7810083

[2] Burkard Woerdenweber, Jörg. Wallaschek, Peter Boyce. “Automotive Lighting and Human Vision., Springer. pp. 95–96. ISBN 3540366970, 2010.

[3] “Visible Light Communication”, www.supremearchitecture.com .

[4] H. Elgala, R. Mesleh, and H. Haas, “Indoor optical wireless communication: Potential and state-of-the-art,” IEEE Communications Magazine, Optical Communications Series, vol. 49, pp. 56–62, September 2011.

[5] “Visible Light Communication”, lbsitbytes2010.wordpress.com

[6] Spiros Louvros and David Fuschelberger, “VLC Technology for Indoor LTE Planning”, System-Level Design Methodologies for Telecommunication, DOI 10.1007/978-3-319-00663-5_2, Springer International Publishing Switzerland 2014.

[7] Cecie Starr. “Biology: Concepts and Applications”, Thomson Brooks/Cole. ISBN 0-534-46226-X, 2005

[8] Tanaka, Y.; Haruyama, S.; Nakagawa, M.; "Wireless optical transmissions with white colored LED for wireless home links," Personal, Indoor and Mobile Radio Communications, 2000. PIMRC 2000. The 11th IEEE International Symposium on , vol.2, no., pp.1325-1329 vol.2, 2000.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 217-220 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Analysis of Fractional-Order 2xn RLC Networks by Transmission Matrices

Mahmut Ün*, Manolya Ün

Biomedical Engineering Department, İstanbul Yeni Yüzyıl University, 34010 İstanbul , Turkey

ARTICLE INFO

Article history:
Received: 28 March, 2017
Accepted: 17 April, 2017
Online: 28 April, 2017

Keywords:
2xn fractional-order network
Transfer function analysis
Transmission matrices
Fractional- order elements

ABSTRACT

In this study, a new method is devised and presented for the dynamic analysis of a 2xn RLC circuit network modeled with the fractional-order circuit elements. The analysis method is based on the principles of dynamic analysis with transfer function approximation. Firstly, the fractional-order 2xn RLC circuit network of interest which is divided into n equal cells connected in cascaded form, then related transmission matrices are individually calculated for each cell correspondingly, Secondly. the transmission matrix of the whole circuit network is calculated based on the properties of the cascaded connection. By means of this transmission matrix and the two-part connection, diagram circuit functions, such as transfer function and the equivalent input impedance of the whole circuit network, are obtained depending on the number of cells (n) and the fractional-order values. Finally, essential dynamic system analysis, such as frequency, step and pulse responses, and the impedance characteristics of the network, are simulated using necessary MATLAB programs depending on cell number n and the fractional-order values.

1. Introduction

Circuit networks have become more attractive in recent years because they can be employed for the modelling of both electrical and non electrical systems, such as biological and chemical ones [1,2]. A study on resistance networks of graphine which indicates the existence of the planar circuit networks in nature won the 2010 Nobel Prize in Physics [3-5]. Over the past few decades, researchers have published some papers on the integer circuits networks. However, the last efforts have been focused mainly on analyzing the resistors or capacitors of the fractional order single component circuit networks [6,7]. Comparatively little attention has been paid to the impedance of the fractional-order multiple-component circuit networks.

Fractional-order mathematical models developed for inductors and capacitors could describe the electrical characteristics more accurately, which allows for higher flexibility, more freedom, best fit and better optimization techniques. In other words, the actual inductors and capacitors are fractional-order in nature [8]. Fractional-order inductors could be designed based on skin effect. Moreover, the fractional order capacitors have been created with different electrolytes recently [9-11].

Nowadays some researchers have been studying to design and realize the fractional elements [12-16]. Furthermore, some researchers have also concentrated on the study of fractional-order circuit theory [17-20]. However, only few researchers are interested in the electrical characteristics of 2xn circuit networks in fractional-order sense. Therefore, we focus on impedances and transfer functions of the network in fractional domain.

The following research contents could make our research more attractive. Firstly, network transfer function and equivalent input impedance formula are obtained depending on cell numbers used in network and fractional-order values. Cascade connection method with the transmission matrices is developed for this derivation. Secondly, the characteristics of circuit network impedance and the transfer function are investigated with respect to cell numbers and fractional-order values by means of MATLAB simulation programs.

2. Analysis of Fractional-Order 2xn RLC Circuit Network

Figure 1 represents the diagram of the fractional-order 2xn RLC circuit network, where z_0 refers to the resistors, z refers to the fractional order capacitors, and z_1 refers to the fractional-order

*Corresponding Author: Mahmut Ün, Biomedical Engineering Department, C İstanbul Yeni Yüzyıl University, 34010 İstanbul, Turkey
Email: mahmut.un@yeniyuzyıl.edu.tr

inductors [17]. The represented network model is planar, and contains only passive circuit elements.

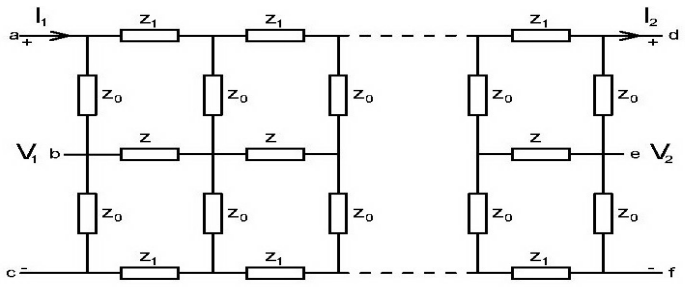


Figure 1: Fractional-order 2xn RLC circuit network model

In order to obtain equivalent impedance and transfer function formula of the fractional-order circuit network, the basic circuit network theory, the two-port transmission matrix, and the matrix tranformation methods are employed. It is clearly seen that the circuit network model shown in Figure 1 includes the network submodel shown in Figure 2a, and the network submodel shown in Figure 2b, which are connected in cascaded form. One can easily verify that the whole circuit network is formed by cascading these submodels. With the help of the calculated transmission matrices for each submodel equivalent transmission matrix, the whole network model can be computed. From the equivalent transmission matrix, the transfer function and the equivalent impedance of the complete network can easily be found.

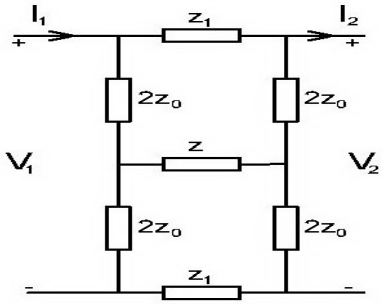


Figure 2a: Two-port subnetwork model for basic unit cell

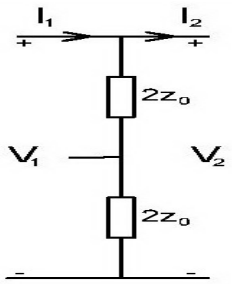


Figure 2b: Two-port subnetwork model used at input and output

Transmission matrice for the two-port network shown in Figure 2a is calculated by

$$T_1 = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \quad (1)$$

Where parameter values of the transmission matrice given by (1) are defined respectively

$$a = \frac{16z_0^2+12zz_0+5zz_1+zz_0}{2z_0\,(4z_0+z+z_1)} \qquad b = 2z_1 \quad (2a)$$

$$c = \frac{16z_0^2+8zz_0+8z_0z_1+2zz_1+zz_0}{2z_0\,(4z_0+z+z_1)} \qquad d = \frac{z_1}{2z_0} \quad (2b)$$

Transmission matrice for the two-port network shown in Figure 2b is computed by the following equation.

$$T_2 = \begin{bmatrix} 1 & 0 \\ \frac{1}{4z_0} & 1 \end{bmatrix} \quad (3)$$

Transmission matrix for the whole network connected in cascaded form becomes

$$T_= T_2 T_1^n T_2 \quad (4)$$

Where ;

$$T_1^n = \begin{bmatrix} a_n & b_n \\ c_n & d_n \end{bmatrix} \quad (5)$$

T_1^n is defined as nth power of T_1 matrice. From Cayley-Hamilton theorem, the following equation can be written as

$$T_1^n = \alpha_0 I + \alpha_1 T_1 \quad (6)$$

Where T_1(2x2) is given by (1) and I(2x2) is unit matrix [21]. Where α_0 and α_1 coefficients are calculated by

$$\begin{bmatrix} \alpha_0 \\ \alpha_1 \end{bmatrix} = \frac{1}{\lambda_2-\lambda_1}\begin{bmatrix} \lambda_2\lambda_1^n - \lambda_1\lambda_2^n \\ \lambda_2^n - \lambda_1^n \end{bmatrix} \quad (7)$$

Where λ_1 ve λ_2 coefficients are non repeated eigen-values of T_1 matrix. If the equations (3) and (5) are substituted in the equation (4), one can easily obtain the following equation.

$$T_= \begin{bmatrix} a_n + \frac{b_n}{4z_0} & b_n \\ \frac{b_n}{16z_0^2} + \frac{a_n+b_n}{4z_0} + c_n & \frac{b_n}{4z_0} + d_n \end{bmatrix} \quad (8)$$

Equivalent input impedance and transfer function can easily be obtained by means of T transmission matrix given by (8). Equivalent input impedance and transfer function are given below, respectively.

$$Z_{in} = \frac{V_1}{I_1} = \frac{4z_0a_n+b_n}{\frac{b_n}{4z_0}+a_n+b_n+4z_0c_n} \quad (9)$$

and

$$H(s) = \frac{V_2}{V_1} = \frac{4z_0}{4z_0a_n+b_n} \quad (10)$$

The derived equivalent input impedance and transfer function equations are employed for the dynamic analysis of fractional-order 2xn circuit network. Finally, the analytical equations for the equivalent impedance and transfer functions are presented in the rational equation form. These equations vary with cell numbers and fractional-order values, thus dynamic characteristics of the network can be investigated with respect to cell numbers and fractional-order values.

3. MATLAB Simulation of Network

The given example of the circuit network shown in Figure 1 with n=10 cell is simulated by using MATLAB program. Firstly, the equivalent input impedance and the transfer function of the circuit network are computed by means of symbolic programming in MATLAB. Then, by using the first order approximations presented in literature for the fractional elements, the transfer function and the equivalent input impedance are correspondingly obtained in the rational function form. Finally, the characteristics of the transfer function and the equivalent input impedance of the 2xn circuit network, such as frequency response, step and impulse responses are investigated with MATLAB simulations.

The parameter values of the proposed network in fractional domain are chosen as $z_0 = 1$, $z_1 = s^{0.5}$, $z = s^{-0.5}$, $n = 10$, $\alpha = \beta = 0.5$. In the simulation, $s^{0.5} = (3s+1)/(s+3)$ the first order approximation are used fort he fractional terms [14]. Firstly, frequency response, step and impulse responses of transfer function given by (10) are obtained by using MATLAB simulations. Graphics of these responses are shown in Figure 3a, 3b, 3c, respectively.

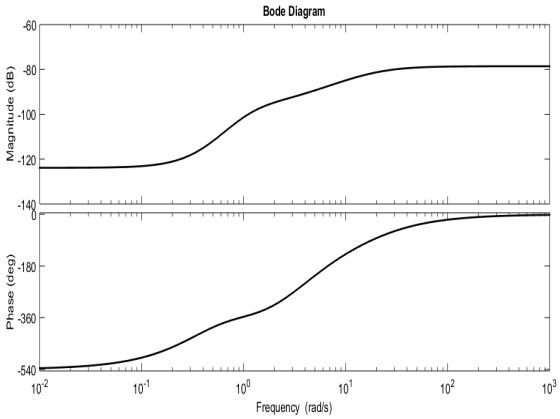


Figure 3a: Frequency responses of the transfer function

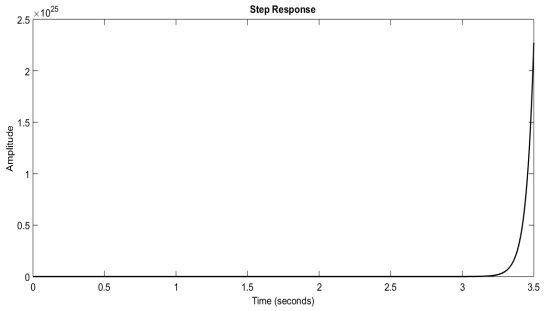


Figure 3b: Step response of the transfer function

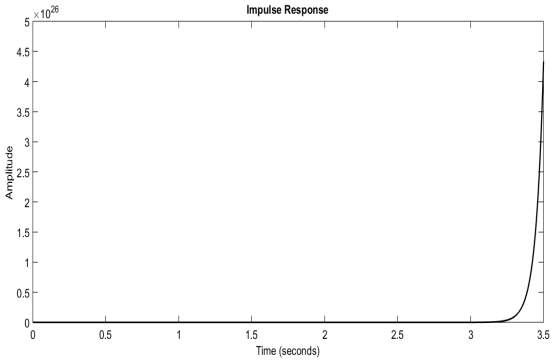


Figure 3c: Impulse response of the transfer function

From the simulation results, we can conclude the following interpretations about transfer function of the network. It is clearly seen that the characteristics of the transfer function has high pass filter characteristic, and the transfer function is stable.

Later, by using the same parameter values and the same approximations for the equivalent input impedance defined in (9), the MATLAB simulations are repeated in fractional domain. Frequency response, step and impulse responses of equivalent input impedance corresponding results are shown in Figure 4a, 4b, 4c., respectively.

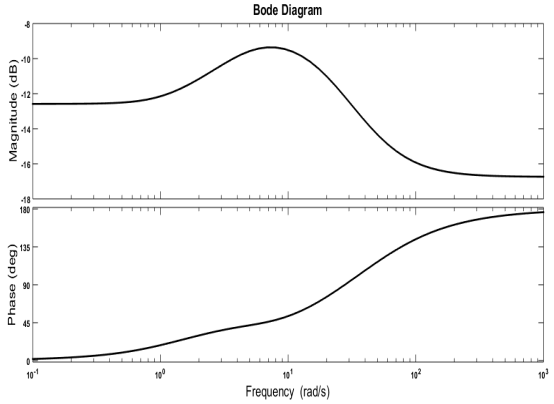


Figure 4a: Frequency responses of the equivalent input impedance

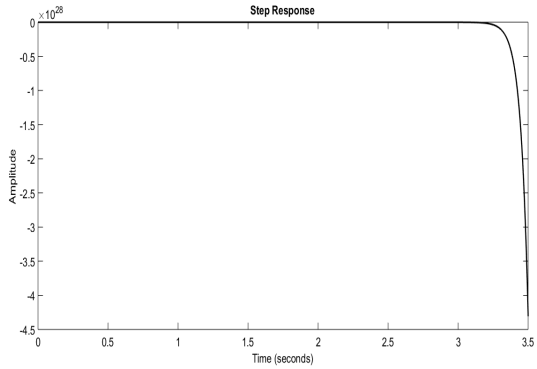


Figure 4b: Step response of the equivalent input impedance

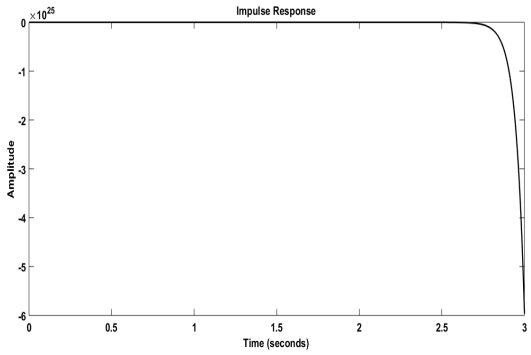


Figure 4c: Impulse response of the equivalent input impedance

From the simulation results, the following interpretions about impedance function can be concluded. It is clearly seen that the characteristics of the equivalent input impedance function has low pass filter characteristic, and the impedance of the network is stable

4. Results and Conclusion

Mathematical model of the fractional- order 2xn RLC circuit network are developed by means of two-port transmission matrices in fractional domain. Mathematical model is in the form of fractional order rational function. By using rational function approximation for the each fractional-order term, the transfer function and the equivalent input impedance of the 2xn circuit network are presented in the form of rational transfer function. Although the fractional- order circuit network contains nonlinear elements, such as fractional- order inductors and capacitors, the transfer function approximation can be employed by using the Laplace transform properties. We have proposed a new method to model and to analyze the fractional- order circuit network in fractional domain.

The characteristics of the equivalent impedance and transfer functions of fractional-order circuit network are investigated by using the MATLAB simulation programming. Network characteristics, such as the frequency, step and impulse responses are obtained graphically for both equivalent impedance ant transfer functions. The dynamic stability and the characteristics of the fractional- order 2xn circuit network are also investigated with respect to the model parameters, such as cell numbers and fractional –order values.

References

[1] Von Busse V, Swartz S. M., and C. C. Voigt C. C., "Flight metabolism in relation to speed in Chiroptera: Testing the U-shape paradigm in the short-tailed fruit bat Carollia perspicillata," *J. Exp. Biol.*,. 216(11),. 2073–2080, 2013.

[2] Simpson R., Jaques A., Nunez, C H.. Ramirez, and Almonacid A.,"Fractional calculus as a mathematical tool to improve the modeling of mass transfer phenomena in food," *Food Eng. Rev.*, 5(1), 45–55, 2013.

[3] Bolotin K. I., Ghahari F., Shulman M. D., Stormer H. L., and Kim P.,"Observation of the fractional quantum Hall effect in graphene," *Nature*, 475(7), 7354-7358,. 2011.

[4] Feldman B. E., Krauss B., Smet J. H., and Yacoby A., "Unconventional sequence of fractional quantum hall states in suspended graphene," *Science*, 337(6099), 1196–1199,. 2012.

[5] Owaidat M. Q., Hijjawi R.S. and Khalifeh J.M.," Network with two extra interstitial resistor", *Int. J. Theor. Phys*., Vol: 10, 3152 – 3159, 2012.

[6] Yao N. Y., Gorshkov A. V., Laumann C. R., Lauchli , J. Ye, and. Lukin M. D, "Realizing fractional Chern insulators in dipolar spin systems," *Phys. Rev. Lett.*, 110(18), 2013.

[7] Hijjawi R.S., Assad J.H and Sakaji H.J, "Perturbation of an infinite network of identical capacitors", *Int. J. Mod. Phys*., Vol: B2, 199 – 209, 2007.

[8] Tripathy M. C. and Behara S."Modelling andanalysis of fractional capacitors", IJAS, Vol.2(10), 29-32, 2015.

[9] Freeborn T. J.,"A survey of fractional order circuit models for Biology and Biomedicine", IEEEJ in Circuits and Systems", Vol.3(3), 2013

[10] Westerlund S., "Dead Matter has Memory, Causal Consulting Kalmar", Sweden, 2002.

[11] Tenreiro M.J.A. and Galhano A.M.S.F.," Fractional order inductive phenomena based on the skin effect", *Nonlinear Dyn*., Vol: 1-2, 107 – 115, 2012.

[12] Radwan A. G. and Salama K. N., "Passive and active elements using fractional L_β C_α circuit", *IEEE T. Circuits-I,* Vol: 10, 2388 – 2397, 2011.

[13] Das S., Sivaramakrishna M.,. Biswas K and Goswami B., "Performance study of a constant phase angle based impedance sensor to detect milk adulteration, sensors and actuators", *A-Physical,* Vol*:* 2, 273-278, 2011.

[14] Krishna M.S., Das S., Biswas K and Goswami, B. "Fabrication of a fractional order capacitor with desired specifications; A study on process identification and characterization", *IEEE Trans. Electron Dev*., Vol: 11, 4067 – 4073, 2011.

[15] Haba T., Ablart G., Camps T. and Olivie F.," Influence of the electrical parameters on the input impedance of a fractal structure realised on silicon", *Chaos Solitons Fract.,* Vol: 2, 479 – 490, 2005.

[16] Elshurafa A.M., Almadhoun M.N., Salama K.N. and Alshareet H.N., "Microscale electrostatic fractional capacitors using reduced graphene oxide percolated polymer composites", *Appl. Phys. Lett.,* Vol: 23, 1234-1237, 2013.

[17] RadwanA.G. and Salama K.N.," Fractional – order RC and RL circuit", *Circuits Syst. Signal Process,* Vol:6, 1901 – 1915, 2012.

[18] Jia H.Y., Chen Z.Q. and.Qi G.Y, "Topological horsesshoe analysis and circuit realization for a fractional-order Lu system", *Nonlinear Dynam*., Vol: 1 – 2, 203 – 212, 2013.

[19] Chen P. and He S. B., "Analysis of the fractional-order parallel tank circuit", *J. Circuits sys. Comput*., Vol: 6, 452-458, 2013.

[20] Maundy B., Elwakil A.S. and Freeborn T.J," On the practical realization of higher order filters with fractional stepping", *Signal Process,* Vol: 3, 484 - 491, 2011.

[21] Kuttler K, Linear Algebra , Ventus Publusing ApS, 2012

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 221-226 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ
ISSN: 2415-6698

Fire extinguishing system in large underground garages

Ivan Antonov[1], Rositsa Velichkova[1*], Svetlin Antonov[1], Kamen Grozdanov[1], Milka Uzunova[2], Ikram El Abbassi[2]

[1]*Hydroaerodynamis and hydraulic machines, Technical university of Sofia 1000, Bulgaria*

[2] *Graduate School of Engineering, ECAM-EPMI, France*

ARTICLE INFO

Article history:
Received: 02 March, 2017
Accepted: 17 April, 2017
Online: 28 April, 2017

Keywords:
Water curtains
Numerical investigation
Fire in garages

ABSTRACT

In the work is considered an acceptable constructive scheme from a practical point of view at fire extinguishing in underground garages. The garage space is divided into quadrants which covering, for example, 2 cars. In case of ignition on one of them, a sprinkler nozzle system is triggered by the effect of the vertical convective jet. A protective curtain preventing the spread of fire to adjacent vehicles is realized. The solution is based on an integrated method which allows the calculation from hydrodynamic point of view on extinguishing time of the fire extinguishing system.

1. Introduction

Fire safety in garages for cars is essential and requires taking into account such actions to ensure it. The complexity of the problem requires a development of methods for numerical simulation of heat conduction and mass processes which are solved with the modern methods of CFD.

In the present work a simple method (from a technologically point of view) is offered for solution of the complex problem. It is suggested to isolate the parked in the garage cars in pairs by operating a thick curtain of water at arisen burning. The necessary insulation for solid non-combustible barriers are replaced at this way.

2. Principal of action

Referring to Fig. 1 cars are placed to ensure the possibility between the pairs to have enough distance for the implementation of water curtains. In case of burning over the car is formed upward convective flow, due to differences of the density of the products of combustion and the environment. This stream is proportional to the lift force:

$$dF_A = -\left[\int_f (\rho - \rho_{ok})\, g df\right] dx \ , \quad (1)$$

where f is the area of fire ignition, dx elementary stretch at vertical direction.

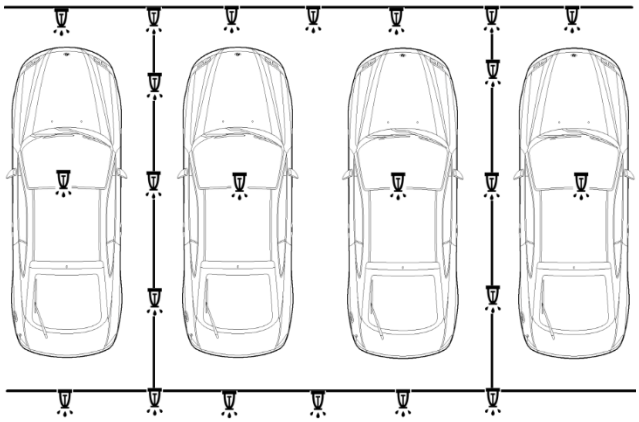


Fig.1 – Distribution of cars in the garages

The power of convective updraft is determined by the number of Archimedes

*Corresponding Author: Rositsa Velichkova, Technical university of Sofia, Bulgaria | Email: rositsavelichkova@abv.bg

$$Ar = \left(\frac{\rho_{ok}}{\rho} - 1 \right) \frac{g d_h}{u_0^2}, \qquad (2)$$

where d_h is the hydraulic diameter of the outbreak of fire; u_0 - initial value of the velocity of the upward flow. The velocity is determined according to Drayzdeyl [6]:

$$u_0^2 = 1{,}9 Q^{1/5}, \qquad (3)$$

where Q, kW is the power of the fire.

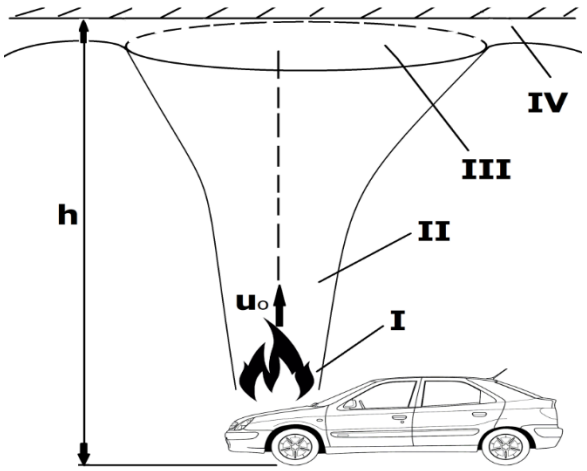


Fig.2 – Distribution of the fire

The formed convective flow is shown in Fig. 2. The conditional arisen stream can be divided into the following areas:

In zone I is formed a convective flow (Fig. 3). In the zone of fire enters from all directions ambient air which is heated and reverses direction in a vertical direction. The second zone is a free convective flow, which continues until it reach the ceiling of the room where the flow changes its character (zone III). In this area the jet is transformed into radial semi-confined stream and spreads along the ceiling of the garage (zone IV).

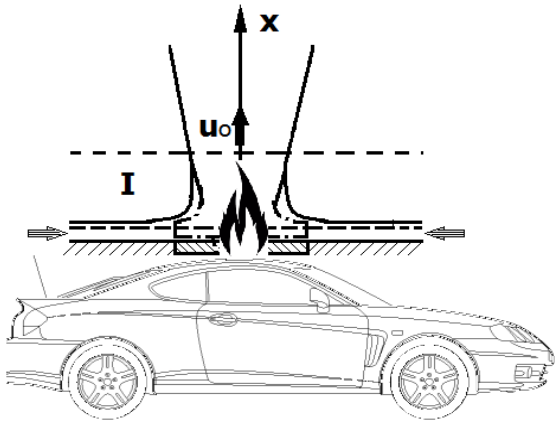


Fig.3 – Sketch of the convective flow

The system includes fire sprinklers- fast response and standard response sprinklers. Reaching the ceiling of the garage under the influence of its temperature convective flow turns on the sprinkler for rapid response and the burning car is flushed with water spray. This is how the process of fire fighting starts in the initial stages Further, spreading out as a radial semi-confined jet reaches the "standard response sprinklers which turns on the water curtain.

This stage is defined as isolating the burning cars from the surrounding area and no affected other pairs.

3. Mathematical model of convective non-isothermal jet

For the purpose of solving the task is used an integral method according to [1] [4]. The used equations are as:

-for amount of movement

$$\frac{d}{dx} \int_0^b \rho u^2 \left(\pi y^j \right) dy = -g \int_0^b \left(\rho - \rho_{ok} \right) \left(\pi y \right)^j dy \qquad (4)$$

-to preserve enthalpy flow

$$\frac{d}{dx} \int_0^b \rho \Delta h u y^j dy = 0 \qquad (5)$$

-for export vertical upward mass flow

$$\frac{d}{dx} \int_0^b u \left(\rho - \rho_{ok} \right) \left(\pi y \right)^j dy = 0 \qquad (6)$$

A simple decision can be made as (5) of enthalpy is replaced by a linear dependence on the widening of the jet

$$b = cx \qquad (7)$$

In (4) on the right side is written the Archimedes buoyancy. The included symbols are as follows: u - jet velocity; y - transverse coordinate, ρ - current density, ρ_{ok} - density of the environment, Δh - enthalpy of the stream. The exponent j marks: at j = 0 -flat stream; j = 1 - axis jet. The coordinate x is directed vertically upwards.

There is a correlation between density and temperature:

$$\rho = \frac{p}{RT}, \qquad (8)$$

where p - pressure of the environment, R - gas constant, T - absolute temperature. Similarity to transverse distribution of the velocity and the density (temperature) are initiated [4] [13], where solving (4) and (6) leads to the parameters of the upward convective stream:

- the velocity of the upward stream

$$u_m = B_u^{"} D_0^{1/3} \Delta T_{пл}^{4/9} x^{-1/3} \qquad (9)$$

- the temperature difference

$$\Delta T_m = T_m - T_{env} = B_{\Delta T}^{"} D_0^{1/3} \Delta T^{8/9} x^{-5/3}, \qquad (10)$$

where D_0 - initial diameter of the heat source of fire (the burning car); $\Delta T_1 = T_{fl} - T_{env}$; $\bar{x} = \frac{x}{D_0}$; the constants B_u^* and $B_{\Delta T}^*$ have values $B_u^* = 0,222\left[m^3 K^{9/4}\right]$, $B_{\Delta T}^* = 0,71\left[m^{1/3} K^{9/8}\right]$.

These values correspond to the case at $\bar{x} = \frac{x}{D_0} \geq 3 \div 3,5$. When adopted fire size created from a burning car $D_0 = 0,5m$ and height $H = 3 \div 4,5m$, $\bar{x}$ of the garage will be always greater than the above values.

At the relatively short distance to the ceiling, the high power of fire (the accepted conditions are $Q = 1500W$ and $T = 600K$), the velocity and the temperature and the rising convective stream doesn't change significantly.

The initial velocity calculated by (3) is $u_0 = 8,2m/s$ and the time when the convective stream will reach the ceiling at different heights of the garage is given in Table 1.

Table 1

h, m	3	3,5	4	4,5
Δt	0,36	0,43	0,49	0,55

This means that less than $1s$ sprinklers over the burning car will be activated and extinguishing stream will flow over the burning car.

The expansion (increasing of thickness) of the jet in height can be determined by the expression:

$$\frac{db}{dx} = 0,22\frac{\rho_0 + \rho_{ок}}{2\rho_{ок}} \quad , \qquad (11)$$

respectively

$$b = \left[0,22\frac{\rho_0 + \rho_{ок}}{2\rho_{ок}}\right] \quad . \qquad (12)$$

The density of the jet in the opening section is defined by (8): at $R = 287J/kgK$, $T_0 = 600K$, $p = 10^5 Pa$ in which for ρ_0 is recived $\rho = 0,58kg/m^3$. The density of the environment is $\rho_{env} = 11,2kg/m^3$ at the same pressure and temperature $T_{env} = 293K$. At this density the widening of the jet in the present case is:

$$b = 0,163x \qquad (13)$$

For the different height in the garage the parameter b is given in Table 2.

Table 2

h, m	3	3,5	4	4,5
b_1, m	0,501	0,507	0,652	0,73
$b_{изт}$, m	0,651	0,77	7,74	1,98

The last row in Table 2 is given the extension of the isothermal jet $(T_0 \approx T_{env})$. Obviously a slight extension of non-isothermal convective flow comparing with the isothermal.

Reaching the ceiling vertical convective stream is transformed into radial jet (Fig. 4).

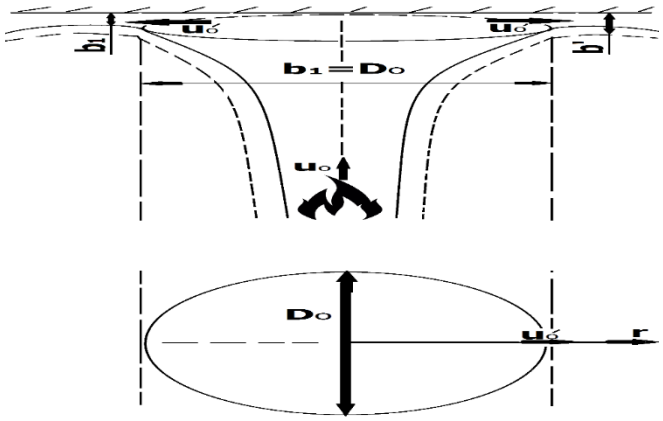


Fig. 4 – Sketch of radial jet

Due to the weak widening of the jet and the short distance to the ceiling the mass flow is not increased significantly because the temperature, density and relatively low mileage ceiling have not changed. The jet has retained its temperature, density and velocity accordig to Abramovich [1] may be determinated:

$$u_{\max} = 0,88u_0 \quad , \qquad (14)$$

where for $u_0 = 8,2m/s$ results as $u_{\max} = 7,2m/s$.

It is assumed that the starting size of the radial jet is equal to the obtained in Table 2- b_1, i.e. $D_0 = b_1$.

Width of the radial jet b_0 is determined by the flow rate Q at the intersection of the reverse flow. This flow rate is amount of initial flow rates Q_0 and increase its height due to suction of air from the environment. Ejecting fluid is considered proportional to the square of the relative increase in the width of the jet $\left(\frac{b_1 - b_0}{b_0}\right)^2$ and the distance x divided by the duration of the process Δt.

$$Q_{ej} \approx \left(\frac{b_1 - b_0}{b_0}\right)^2 \frac{x}{\Delta t}, m^3/s \qquad (15)$$

In this total flow rate is obtained as the sum of normal and ejecting flow rate:

$$Q = Q_0 + Q_{ej}, m^3/s \quad , \qquad (16)$$

where $Q_0 = u_0\frac{\pi d_n^2}{4}$, when $d_n = 0,5m$ and $u_0 = 8,m/s$ we have:

$$Q = Q_0 + Q_{ej} = u_0\frac{\pi d_n^2}{4} + \left(\frac{b_1 - b_0}{b_0}\right)^2 \frac{x}{\Delta t} \qquad (17)$$

The flow rate of the respective heights $x = 3;3,5;4;4,5m$ of the garage is shown in Fig. 5 where is defined by the relationship (12), respectively in case of a leak (13).

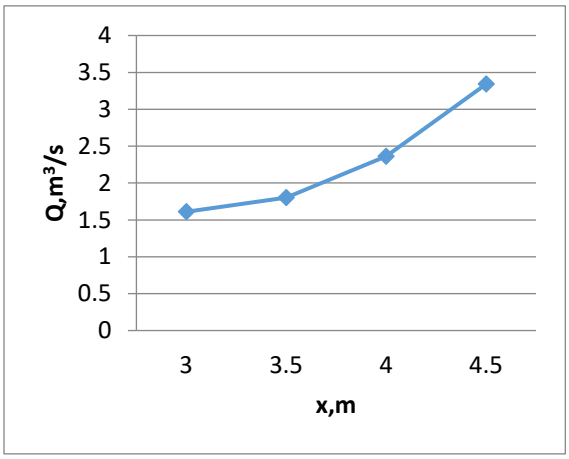


Fig. 5 – Change of the flow rate at different height of the garage

According to Figure 4 is assumed $D_0 = b_1$ that in already known flow rate (16) original width of the radial jet $b_0^{'}$ is calculated:

$$b_0^{'} = \frac{Q}{\pi D_0} \quad . \qquad (18)$$

The relationship $b_0(x) = f(x)$ is given in Figure 6.

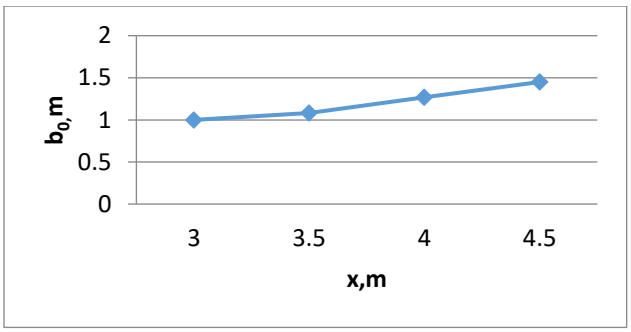


Fig. 6 – Change of the initial weight at different height of the garage

The cross-section of the radial jet, as a function of r is determined by the expression:

$$S = 2\pi r b^{'} \quad , \qquad (19)$$

where r is the current radius, $b^{'}$ is the width of the jet to the corresponding r.

Since the resulting stream is parietal and has parietal boundary layer which thickness is approximately $0,1b^{'}$, than (19) can be recast in the form:

$$S = 2,2rb^{'} \qquad (20)$$

The width $b^{'}$ is calculated by (12) and for the case (13) by replacing x with r then we have:

$$b^{'} = \left[0,22\frac{\rho_0 + \rho_{ok}}{2\rho_{ok}}\right] r = c^{'} r; c^{'} = 0,22\frac{\rho_o + \rho_{ok}}{2\rho_{ok}}, \qquad (21)$$

respectively $b^{'} = 0,163r$.

Substituted in (19) follows:

$$S = 2,2\pi c^{'} r^2 \quad . \qquad (22)$$

The average velocity of the ceiling of the room depending on r is obtained by:

$$u_m = \frac{Q}{S} \quad , \qquad (23)$$

respectively:

$$u_m = \frac{Q}{2,2\pi c^{'} r^2}, m/s \qquad (24)$$

Parking average velocity depending on r about the four heights is shown at Fig. 7. Fig. 8 shows the time to reach the appropriate distance.

$$\Delta t = \frac{r}{u_m}, s \qquad (25)$$

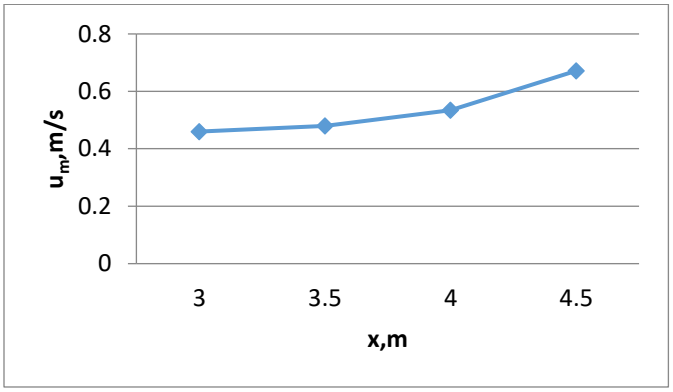


Fig. 7 - Average velocity at different height of garage

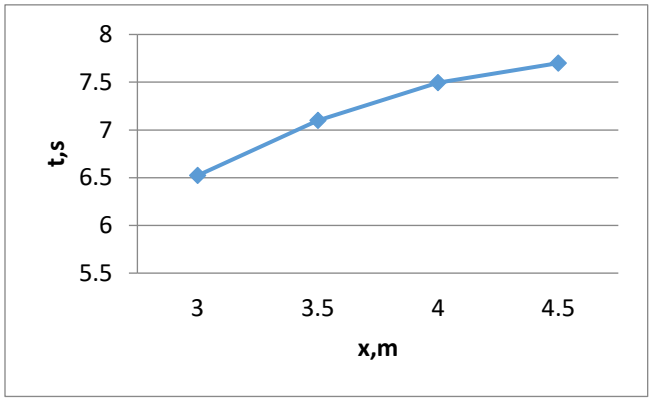


Fig. 8 Time to reach the sprinklers at different height of garage

This means that in the first two seconds all sprinklers at distance of 2 m away from the burning car will be triggered. For longer distances the remote sprinklers will act at a condition if the temperature of the burning car doesn't decrease too quickly. For maximum calculated time of 7,7 seconds could not be expected too much decrease of the temperature, which leads to the conclusion that the ceiling temperature will be much greater than the starting temperature of "fast" sprinklers so that at $t_p = 57^0 C$

or T = 330K will always remain less than the temperature of the wall jet which initial temperature is 600 K.

With the removal from the water curtain it is possible to turn on the other ceiling sprinklers which are in the range.

In the vicinity of the burning car to sprinkler curtain $l \le 2m$ will trigger 3 (to 5) fast sprinklers. At a longer distance it will trigger maximum of three quick sprinklers of water curtain plus the main ones over the burning car and eventually these which are lying in the range of $l = 4m$ ceiling sprinklers so that the number of activated sprinklers will increase [5].

To create a smokeless zone under a layer of smoke floating [6] it is designed and installed air exhaust systems for smoke and hot gases. An exhaust ventilation system for smoke and hot gases is a scheme of safety equipment designed to perform a positive role in spin fire. The smoke is drawn in the direction of the non carrier partition EI from of 2 m / s to 5 m / s velocity. Quoted above standard allowed velocity of 5 m / s, but it should be taken into consideration that this velocity would affect negatively and lead to the merging of streams of pure air.

From [6] the density of the thermal load in the premises for the storage of combustible materials according to their purpose, is determined the heat capacity of the prevailing materials. Ventilation systems to remove smoke and heat (VSRSH) have to reach its designed performance level within 60 seconds of receiving the command signal. Each VSRSH systems has to ensure receipt of sufficient fresh air that enters the room for the expense of the flue products.

4. Thermal impact

Heat transfer by convection and radiation is defined according to [5]. Thermal effects are expressed by the intensity of the heat flow $h_{nbt}, W/m^2$ to the surface of the element is determined taking into account the heat transfer by convection and radiation, such as:

$$h_{nbt} = h_{nbt,c} + h_{nbt,r}, W/m^2 \quad , \qquad (26)$$

where: heat transfer by convection $h_{nbt,c}$ is given by the relationship

$$h_{nbt,c} = \alpha_c \left(\theta_g - \theta_m\right), W/m^2 \qquad (27)$$

radiation heat transfer $h_{nbt,r}$ is given by the dependence:

$$h_{nbt,r} = \Phi \varepsilon_m \varepsilon_f \sigma \left[\left(\theta_1 r + 273\right)^2 - \left(\theta_m + 273\right)^4 \right], W/m^2 \quad (28)$$

Convection component of the intensity of the heat flow is determined by:

$$h_{nbt,t} = \alpha_c \left(\theta_g - \theta_m\right), W/m^2 \qquad (29)$$

where: α_c is the heat transfer coefficient by convection $\left[\frac{W}{m^2}K\right]$; θ_g is the gas temperature near the exposed fire element [^{0}C], θ_m is the surface temperature of the element [^{0}C]

The coefficient of heat transfer by convection α_c is determined by the nominal curves corresponding to "temperature-time". On indirectly heated surface elements the intensity of heat flow h_{nbt} is determined by the relationship (16) where $\alpha_c = 4\left[\frac{W}{m^2}K\right]$. The coefficient of heat transfer by convection has value $\alpha_c = 9\left[\frac{W}{m^2}K\right]$, considering that the effects of heat transfer by radiation are included. Radiating components of net heat flux per unit surface area is defined as:

$$h_{nbt,r} = \Phi \varepsilon_m \varepsilon_f \sigma \left[\left(\theta_1 r + 273\right)^2 - \left(\theta_m + 273\right)^4 \right], W/m^2 \quad (17)$$

where: Φ factor of configuration, ε_m emitting surface element, ε_f transmission of fire, $\sigma = 5,67.10^8\left[\frac{WK^4}{m^2}\right]$ constant of Stefan - Boltzmann, θ_r is the effective temperature of the radiation environment [^{0}C], θ_m is the surface temperature of the element [^{0}C]. Transmission of fire is taking equals $\varepsilon_f = 1$

5. Determination of the intensity of water curtain

Because of the difficulties associated with the construction of fire walls, experiments are conducted so that these areas to be reduced to such proportions that the primarily split up do not disturb of the process. In many cases, such as in buildings of first degree of fire resistance, as already noted, firewalls did not provide the detriment of fire safety. In connection with this arises a need of using such fire barriers that could effectively limit the spread of fire and at the same time would give some freedom for internal layout of buildings with different functions which is the case of the water curtain.

When calculating the water curtains, the assumption of the simultaneous satisfacties the following conditions:

Structural parts of the building to withstand the effects of fire on one side and the passage of flames or hot gases to be prevented by the transfer of heat to the unexposed side. The ability of the structural parts of the building to withstand the effects of fire on one side and prevent the transfer of heat from the exposed to the unexposed side. The transfer is limited so that it does not ignite neither unexposed surface, nor any other material in the immediate vicinity. The structural element is designed to serve as a barrier against the heat and to ensure the protection of people who are close to it.

The effectiveness of water curtains is assessed according to the amount absorbed heat.

It is known that the dependence of the growth temperature of the source of radiation of maximum energy is moving to the side of the shorter waves. This follows from the law of Vin:

$$\lambda_{\max,T} T = 0,29 = const \qquad (30)$$

where: λ is the wavelength in m, T - temperature at the surface of water curtain, ^{0}K.

Good enough inter-phase and heat absorbing surface has a water drops at size 200.10^{-6}. It is considered that in the best case, sprinklers spray water less then $1000\mu m$.

6. Required flow rate for air curtain

The current aim is decide on the following characteristics - density of the radiation heat flux is 1500 W/m^2; density of the irradiation protected material is 900 W/m^2; height of the hole -4 m; length of the hole - 6 m; pressure of water in sprinkler - 0.6 MPa (6 atm) and the radius of the water drops -0,0006 m ($600\mu m$).

Opacity density of the curtain:

$$\delta = \frac{2,303 \log q_{изл}}{q_{кр}} = 0,51 \quad (31)$$

Thickness of the curtain:

$$R = \frac{\delta}{c} = \frac{0,51}{2,8} = 0,182m \quad (32)$$

Flow rate of the water curtain for 1 m^2 of lateral surface is defined by:

$$Q = 0,666\mu\rho\frac{rR}{H}\sqrt{2gh} = 0,467\frac{l}{s}m^2 \quad (33)$$

For the whole surface of the water curtain:

$$Q^H = 11,2l/s \quad (34)$$

Water curtains are constructed so that the entire hole is irrigated with finely dispersed water. For this purpose sprinklers are placed over the hole and next to it. When they are placed at the top of the hole it is possible to remain unprotected areas through which is possible a penetration of hot gases to occur.

Sprinkler heads which are used to spray jets are spaced 0.5 m in protecting small holes and 1.25 - 1.5 m in protecting large holes. For sprinkler heads water curtains which are situated at a distance greater than 3 m, it is required head pressure of the water 4-6 mH_2O.

7. Conclusion

It is made an approximate fast and easy to implement analysis and calculation of water curtain, which replacing the fire walls with water curtains**.**

It is advisable that nozzles for spraying the water to be directed vertically upward to cover the entire space of the garage ceiling to the floor, thus help to keep heat away from the curtain.

These minimum values of pressure in their heads provided sufficiently stable curtain of water flows under the pressure of combustion products.

Water extinguishing systems can completely extinguish a fire or to locate it only on condition that they provide the necessary delivery rate extinguishing substances over the area of the fire.

The proposed firefighting system has some limitations in technical field: secure enough water for the sprinkler system; the entering into the garage cars must be with no Liquefied Petroleum Gas fuel (this condition is not easy to control); the fire extinguishing and prevention of the thermal radiation diffusion system ensure state of safety in the first 10 – 15 minutes from the occurrence beginning- after that time the firefighting department must act in this zone.

8. Additional contributions

It is developed a new fire extinguishing system for configuration and the fire spread to neighboring cars in the garage.

On the basis of an integral method it is made a mathematical model for calculating the extinguishing system and it is given also the respective fire extinguishing times.

It is developed a system for fire extinguished in one of the largest underground garages in Sofia using this conseptual design. The fire extinguishing system described in this article is implemented in fire extinguishing practice after its adoption by specialized scientific councils by authorized institutions (institutions).

References

[1] Abramovich, Theory turbulentnyiy Strings, Moscow 1984

[2] Annica Nordmark, Fire and life safety for underground facilities: present status of fire and life safety principles related to underground facilities: ITA working group 4, “subsurface planning”,Tunnelling and Underground Space Technology, Volume 13, Issue 3, July–September 1998,pp 217-269

[3] Annex № 9 of art. 122, para. 3 NSTPNOBP

[4] Antonov IC, Applied Fluid Mechanics, Sofia 2010

[5] A.W. Marshall, M. di Marzo, Modeling Aspects of Sprinkler Spray Dynamics in Fires, Process Safety and Environmental Protection,vol 82, Issue 2, March 2004, pp 97-104

[6] Drysdale D. An introduction to fire dynamics, John Wiley sons, 2011

[7] Eurocode 1 - Impact on stoitelni structures, Part 1-2

[8] EN-12845, Installation and Maintenance of fixed sprinkler systems in buildings and industrial plants

[9] Ivan Antonov, Rositsa Velichkova, Kamen Grozdanov, Milka Uzunova, The possibility of replacing solid walls with water curtain applicable to a large underground garage, EFEA’2016, Belgrade, Serbia, http://ieeexplore.ieee.org/document/774880/,**DOI:** 10.1109/EFEA.2016.7748805, **Electronic ISBN:** 978-1-5090-0749-3

[10] G.D. Lougheed, J.K. Richardson, Sprinklers in Combustible Concealed Spaces, Journal of Fire Protection Engineering, May 1989; vol. 1.2 pp.49-61.

[11] Jian-yong Liu, De-ming Zhu, Zhe Zhao, Dong Liang ,Experimental Study on Ultra-Fine Water Mist Extinguishing Heptane Cup Fire in Confined Space, Procedia Engineering,vol 52, 2013,pp 225-229

[12] Noureddine Benichou, Ahmed H. Kashef, Irene Reid, George. V. Hadjisophocleous, David A. Torvi, Gaetan Morinville, FIERAsystem: A Fire Risk Assessment Tool to Evaluate Fire Safety inIndustrial Buildings and Large Spaces,Journal of Fire Protection Engineering, August 2005;vol.15,pp.145-172.

[13] Rositsa Velichkova, Ivan Antonov, Kamen Nikolov, Kamen Grozdanov, Milka Uzunova, Modeling of the occurrence of fire in closed cars garages, EFEA’2016,http://ieeexplore.ieee.org/document/7748807/,**DOI:** 10.1109/EFEA.2016.7748807, **Electronic ISBN:** 978-1-5090-0749-3

[14] V. Stoyanov, A. Terziev , M. Uzunova, "Numerical study of distribution of smoke and hazards in underground parking areas considering the operation of ventilation,"Environmental Friendly Energies and Applications (EFEA), 2014 3rd International Symposium on, St. Ouen, 2014, pp. 1-6

[15] Vijay Kumar Cheeda, Amit Kumar, K. Ramamurthi, Influence of height of confined space on explosion and fire safety Fire Safety Journal,vol 76,August 2015,pp 31-38

[16] Xin Ye, Jian Ma, Yi-xin Shen, Long-yuan Lin, Suppression Effect of Sprinkler System on Fire Spread in Large Commercial Buildings , Procedia Engineering,vol 135,2016,pp 454-461

[17] Xiangyang Zhou , Characterization of interactions between hot air plumes and water sprays for sprinkler protection, Proceedings of the Combustion Institute,vol 35, Issue 3,2015, Pages 2723-2729

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 227-232 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTES Journal
ISSN: 2415-6698

Directional Antenna Modulation Technique using A Two-Element Frequency Diverse Array

Shaddrack Yaw Nusenu*

[1] *Koforidua Technical University, Electrical / Electronic Department, Faculty of Engineering, Ghana.*

ARTICLEINFO

Article history:
Received: 04 April, 2017
Accepted: 26 April, 2017
Online: 01 May, 2017

Keywords :
Frequency diverse array
Directional modulation
Physical-layer security
Range-angle dependent

ABSTRACT

Direction dependent antenna modulation (DAM), as a potential physical-layer security technique, is able to enhance the security of wireless communication systems by virtue of the capability of its direction-dependent signal modulation format transmission. In this paper, DAM using frequency diverse array (FDA) to produce the modulation is presented. By phase shifting each element, the desired amplitude and phase of each symbol can be produced during a particular transmission in a given range and angle position determined by the switching speed of the phase shifters. Accordingly, the proposed method offers a robust security for wireless transmission, as the transmitted signal can be purposely distorted in undesired positions, while preserving a clear constellation along pre-specified position(s). Numerical examples are given to illustrate the effectiveness of the proposed method.

1 Introduction

Directional antenna modulation (DAM) transmitter has been recently proposed [1, 2, 3] which has the capability of projecting deliberately distorted signals in all directions in free space other than the pre-specified secure communication directions. In doing so, only intended receivers located along this the secure communication direction are able to capture the signal leading to successful demodulation, while distorted signal emitted along undesired directions make signal demodulation by an eavesdropper a difficult task. Unlike classical cryptographic technique adopted at higher protocol layers to provide encryption, DAM technology does not share any keys, and secures data transmission directly at physical-layer which enables information theoretical security [4].

The basic concept of DAM is that it transmits a signal with a known modulation format to the desired receiver. The transmitted signal can be successfully recovered and demodulated by a receiver with a-prior modulation scheme [3]. For any directions distinct from the desired direction(s), the modulation format imposed on the transmitted signal will change. Thus making it difficult for the eavesdroppers to demodulate. In [2] a similar concept directional modulation (DM) was proposed which employs phased-array antenna. Other researches has proposed different DM transmitters such [5, 6, 7, 8]. Nevertheless, up to now, all published DAM transmitters employs phased-array. Although the DAM techniques using phased-array can steer the beam to an arbitrary direction, but the steering is independent of the range. This is because phased-array produces fixed steering direction at an angle for all the ranges [9].

In fact, the ability to control range dependent transmit energy distribution becomes an important requirement in many applications, for instance, range-dependent interference suppression [10, 11, 12] and directional communications [2]. To address the current needs,frequency diverse array (FDA) provides new opportunities to realize DAM. The FDA concepts was initially introduced in 2006 [13, 14]. They reported that the range-angle-time dependent far-field array pattern formation are the core attributes of the concept. FDA has been defined by [15] as an antenna aperture where the frequency of each antenna element is controlled independently to steer the main beam, depending on the time domain of each frequency component in the far-field. The most significant feature of FDA is that, it employs a small frequency increment across its antenna array elements. This frequency increment makes the array beam-pattern to change as a function of the range, angle and time [9, 13, 14].

In this contribution FDA elements is configured using DAM technique to achieve physical-layer security in range-angle dependent. The range and angle dependent DAM is achieved with 2-bit phase controls across the array elements for secure point-to-point communications. The main contributions can be summarized as: 2-bit phase shifters combi-

*Shaddrack Yaw Nusenu, Koforidua Technical University, Ghana, Email: nusenu2012gh@yahoo.com

https://dx.doi.org/10.25046/aj020331

nations is formulated for the FDA-based DAM to achieve directional range-angle dependent M-ary signal constellations. Compared with conventional phased-array based method, our proposed system achieves narrower demodulation beamwidth for the desired receiver.

In Section II of this paper, FDA DAM for secure point-to-point communications is presented. Numerical results are provided in Section III. Finally, this paper is concluded in Section IV.

2 FDA DAM for Secure Point-To-Point Communications

The FDA elements can be excited by either the same waveform or different waveforms. Here, the waveform radiated from each array element is assumed to be identical with a frequency offset of Δf Hz. Consider a uniform linear array (ULA) FDA composed of identical transmit elements with element spacing d. The radiated frequency of a standard FDA from the mth element is given as

$$f_m = f_0 + m\Delta f,\ m = 0, 1, \ldots, M-1 \tag{1}$$

where f_0 and Δf are the carrier frequency and frequency offset, respectively, and M is the number of array elements. The transmit beampattern can be represented by (2), [16] with the common phase Φ_0 given in (3), where c_0 denotes the speed of light and $r_m \approx r - md\sin\theta$, is the target slant range for the mth element and $d = \frac{c_0}{2f_0}$ denotes the element spacing. Equation (2) implies that FDA has range, angle and time dependent transmit beampattern.

To describe directional antenna modulation for M-ary transmission, a two-element FDA antenna with 2-bit phase control which provides four state units $\alpha_A = \alpha_B = [\alpha_1, \alpha_2, \alpha_3, \alpha_4]$ is configured for each FDA array element as shown in Figure 2. Each FDA element is controlled by the RF switch embedded to specifically select one of four possible phase state units. The four phase state units $[\alpha_1, \alpha_2, \alpha_3, \alpha_4]$ herein represents $\left[0^0, 90^0, 180^0, 270^0\right]$ phase shifters, respectively. The signals are modulated individually by controlling the element phase shifters with the RF switchers. In doing so, the array radiation pattern will change at each symbol in the digital modulation scheme. A desired constellation pattern can be produced along pre-specified range-angle positions. At the same time, constellation patterns generated for other positions will be distorted.

The array factor for this FDA based DAM transmitter configuration can be expressed as

$$AF(\theta, r) = \sum_{m=1}^{M} e^{j\alpha_{AB}(t)} \times e^{-j\left[(m-1)\frac{2\pi}{\lambda}d\sin\theta - \frac{2\pi\cdot\Delta f\cdot r}{c_0}\right]} \tag{4}$$

where $e^{j\alpha_{AB}(t)}$ is a time dependent phase shifters. From Figure 2, the directional modulated range and angle dependent signals of two-element configuration can be written as

$$AF(\theta, r) = e^{j\alpha_A(t)} + e^{j\left(\alpha_B(t) + \phi - \pi\sin\theta - \frac{2\pi\cdot\Delta f\cdot r}{c_0}\right)} \tag{5}$$

where e^{jA} and e^{jB}, denotes the phase shifters for each element A and B, respectively. The frequency offset Δf and phase offset ϕ are used here as a fixed value in a particular signal transmission. The Δf and ϕ can be adjusted when considering different signal transmissions. The phase shifter produce by the ϕ is assumed to be chosen between $-80°$ and $80°$. It is important to note that, the phase shifters provided for FDA elements are based on assumption and their corresponding values are shown in Table I.

It is worth noting that, because the FDA DAM technique depends on switching the antenna elements at each symbol rate, a series of high frequency single-pole multi-throw (SPMT) or single-pole-4-throw (SP4T) switches with nanosecond or sub-nanosecond speeds is required [17, 18].

3 Simulation Results and Analysis

3.1 Directional Modulated 16-ary Transmission Analysis

Analogous to [3], sixteen data points was calculated for the desired receiver at the range-angle position $(R, \theta) = (90\ km, 60°)$. The proposed transmitter system can modulates up to 16-ary unique signal constellation at the far field. When 16-ary transmission is considered, Figure 3 shows the 16-ary unique signal constellation at the intended receiver position $(R, \theta) = (90\ km, 60°)$. The modulated signal can be successfully decoded easily by the receiver. Note that in Figure 3, Δf = 70Hz and $\phi = 45°$ was adopted. These arbitrary values of both frequency offset Δf and phase offset ϕ was chosen to generate 16 unique points at the desired position.

Next, the effect of changing the arbitrary values of both frequency offset Δf and phase offset ϕ are examined. As shown in Figure 4, it can be noticed that for distinct values of frequency offset Δf and phase offset ϕ, such as Δf = 70Hz, $\phi = 45°$; Δf = 120Hz, $\phi = 65°$; Δf = 300Hz, $\phi = -75°$; respectively, generate different distributions of constellation points. It is important to note that Δf and ϕ may be considered as transmission signatures and can provide a series of different modulation format.

Figure 5 demonstrate how the proposed system provides additional level of security. As shown in Figure 5, the 16 unique data point generated corresponds to Δf = 70Hz, $\phi = 45°$ and range R = 90 km for distinct observation angles of $\theta = 60°$ (desired angle direction), $\theta = 40°$ and $\theta = 80°$, respectively. It can be observed that at angles distinct from the desired angle direction, new distributions of modulation format are achieved. This implies that apart from the desired angle direction, all other angle directions received signal constellation changes locations making demodulation by an eavesdropper very difficult.

Similarly, Figure 6 illustrate 16 unique constellations for distinct ranges such as R = 90 km (desired range position), R = 110 km and R = 150 km which corresponds to Δf = 70Hz, $\phi = 45°$ and $\theta = 60°$. It is noticed that at other range positions apart from the desired range position, the signal constellation have different distributions making it difficult for the eavesdroppers to intercept information.

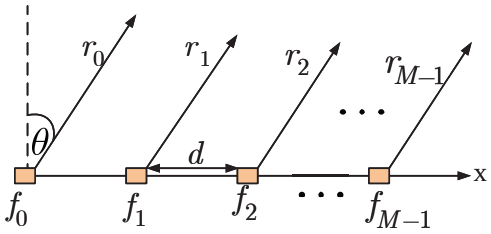


Figure 1: ULA FDA elements with frequency offset Δf.

$$AF(t;r,\theta)=\sum_{m=0}^{M-1}\exp\left\{-j2\pi f_m\left(t-\frac{r_m}{c_0}\right)\right\}\approx e^{-j\Phi_0\frac{\sin\left[M\pi\left(\Delta f t-\frac{\Delta f r}{c_0}+\frac{d f_0\sin\theta}{c_0}+\frac{\Delta f d\sin\theta}{c_0}\right)\right]}{\sin\left[\pi\left(\Delta f t-\frac{\Delta f r}{c_0}+\frac{d f_0\sin\theta}{c_0}+\frac{\Delta f d\sin\theta}{c_0}\right)\right]}} \tag{2}$$

$$\Phi_0=2\pi f_0(t-r/c_0)-\pi(M-1)\Delta f r/c_0+\pi(M-1)f_0 d\sin\theta/c_0+\pi(M-1)\Delta f d\sin\theta/c_0 \tag{3}$$

Table 1: Possible phase shifters combinations for FDA DAM using two-element configuration.

Number	FDA element $[\alpha_A \quad \alpha_B+\phi]$	Number	FDA element $[\alpha_A \quad \alpha_B+\phi]$
1	$[0^0 \quad 0^0]$	9	$[180^0 \quad 0^0]$
2	$[0^0 \quad 90^0]$	10	$[180^0 \quad 90^0]$
3	$[0^0 \quad 180^0]$	11	$[180^0 \quad 180^0]$
4	$[0^0 \quad 270^0]$	12	$[180^0 \quad 270^0]$
5	$[90^0 \quad 0^0]$	13	$[270^0 \quad 0^0]$
6	$[90^0 \quad 90^0]$	14	$[270^0 \quad 90^0]$
7	$[90^0 \quad 180^0]$	15	$[270^0 \quad 180^0]$
8	$[90^0 \quad 270^0]$	16	$[270^0 \quad 270^0]$

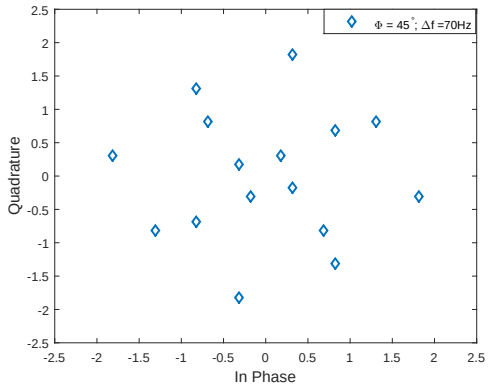


Figure 3: Illustration of 16-ary unique points when Δf = 70Hz and ϕ = 45° is adopted.

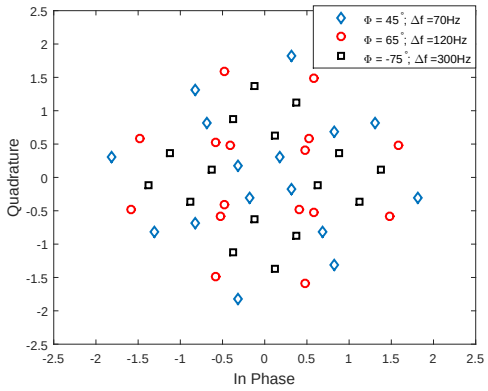


Figure 4: Comparisons of 16 unique constellations for distinct Δf and ϕ.

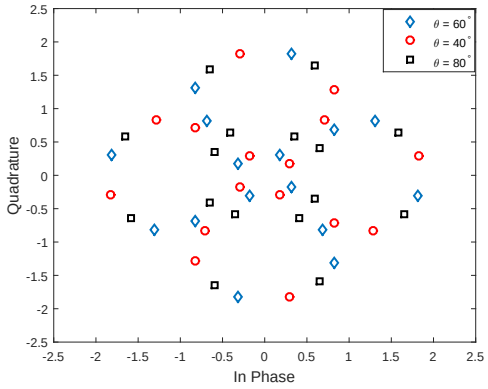


Figure 5: Comparisons of 16 unique constellations for distinct angles.

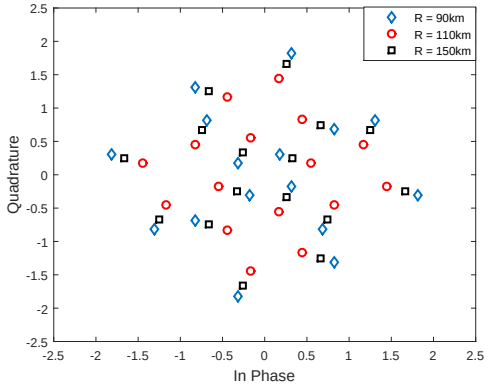


Figure 6: Comparisons of 16 unique constellations for distinct ranges.

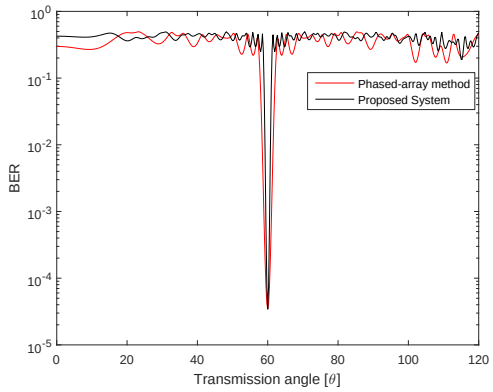


Figure 7: BER performance in angle dimension.

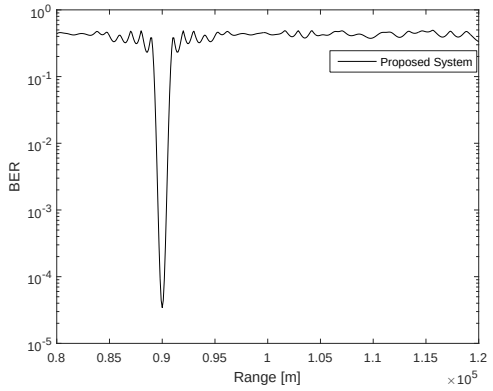


Figure 8: BER performance in range dimension.

3.2 Bit error rate (BER) Analysis

To demonstrate how the proposed system can be applied to a secure communication system, the simulated bit error rate (BER) of the system is calculated in angle and range dimensions using an algorithm based on minimum distance decoding [2].

The essence of Directional modulation technique is to achieve low BER in a narrow beamwidth toward the intended receiver and still enforce a high BER in other undesired directions. The resulting BERs are shown in Figure 7 and Figure 8 . It is fair to ensure that, all BERs are normalized until they are the same at the desired position. In this fashion, the security in undesired directions can be compared while they have the same performance in the intended direction.

In angle dimension, the proposed system achieves a lower BER in a narrow beamwidth with a relatively constant high BER in other directions than phased-array based method. In range dimension as shown in Figure 8, the phased-array-based method has no range resolution capability. In contrast, the proposed system still has good BER performance in the range dimension due to the frequency offsets Δf. These results confirm that, the proposed method achieves good security performance in both range and angle dimensions, not just in the angle dimension as the phased-array-based method and thus yields the worst security performance.

4 Conclusion

This paper has presented a directional antenna modulation scheme using a two-element FDA antenna. The directional manipulation of the constellation points is possible because the signals are modulated at antenna-level with phase shifters changing in each symbol transmission. The proposed system achieves both range and angle dependent DAM for secure 16-ary transmitted symbols for wireless communications. The effectiveness of the proposed system is validated with the aid of constellation diagrams and BER performances. Compared with the phased-array-based DAM method, which scrambles the signal constellations only in undesired angles, the proposed scheme scrambles the signal constellation in both range and angle dimensions. Future work will includes the design of genetic algorithm for phase shifters to improve the security performance.

Conflict of Interest The author declare no conflict of interest.

Acknowledgment The author would like to thank Respected Professor Wen-Qin Wang (Senior Member - IEEE) of UESTC for his great contributions, support and encouragement.

References

[1] A. Babakhani, D. B. Rutledge, and A. Hajimiri. "Transmitter architectures based on near-field direct antenna modulation," IEEE Journal of Solid-state Circuits, vol. 43, no. 12, pp. 2674–2692, December 2008.

[2] M. P. Daly and J. T. Bernhard. "Directional modulation technique for phased arrays," IEEE Transactions on Antennas and Propagation, vol. 57, no. 9, pp. 2633–2640, September 2009.

[3] H. Z. Shi. "Physical layer communication using direction antenna modulation," PhD. dissertation, University Sheffield, London, 2014.

[4] M. Bloch and J. Barros. "Physical-Layer Security: From Information Theory to Security Engineering." Cambridge University Press, 2011.

[5] Y. Ding, Y. Zhang and V. Fusco. "Fourier rotman lens enabled directional modulation transmitter," International Journal of Antennas and Propagation, vol. 2015, Article ID 285986, pp. 1–13, 2015.

[6] T. Hong, M.-Z. Song, and Y. Liu. "Dual-beam directional modulation technique for physical-layer secure communication," IEEE Antennas Wireless Propag. Lett., vol. 10, pp. 1417-1420.

[7] N. Valliappan, A. Lozano, and R. W. Heath. "Antenna subset modulation for secure millimeter-wave wireless communication," IEEE Transactions Commun., vol. 61, no. 8, pp. 3231-3245, 2013.

[8] Nafel N. Alotaibi and Khairi Ashour Hamdi. "Switched Phased-Array Transmission Architecture for Secure Millimeter-Wave Wireless Communication," IEEE Transactions on Commun., vol. 64, no. 3, pp. 1303–1312, March 2016.

[9] W.-Q. Wang. "Frequency diverse array antenna: New opportunites,"IEEE Antennas and Propagation Magazine, vol. 57, no. 2, pp. 145-152, April 2015.

[10] C. Wen, T. Wang, and J.-X. Wu. "Range-dependent clutter suppression approach for non-sidelooking airborne radar based on orthogonal waveforms,"IET Radar, Sonar and Navigation, vol. 9, no. 2, pp. 210-220, February 2015.

[11] J. X. Wu, T. Wang, L. F. Zhang, and Z. Bao. "Range-dependent clutter suppression for airborne sidelooking radar using MIMO technique,"IEEE Transactions on Aerospace and Electronic Systems, vol. 48, no. 4, pp. 36473654, October 2012.

[12] W. S. Yan, X. P. Fang, and J. B. Li. "Formation optimization for AUV localization with range-dependent measurements noise,"IEEE Communications Letters, vol. 18, no. 9, pp. 15791582, September 2014.

[13] P. Antonik, M. C. Wicks, H. D. Griffiths, and C. J. Baker. "Range dependent beamforming using element level waveform diversity," in Proceedings of the International Waveform Diversity and Design Conference, Las Vegas, USA, pp. 1-4, January 2006.

[14] P. Antonik, M. C. Wicks, H. D. Griffiths, and C. J. Baker. "Frequency diverse array radars," in Proceedings of the IEEE Radar Conference, Verona, NY, April 2006, pp. 215-217.

[15] Wen-Qin Wang, H. C. So, and Alfonso Farina. "An Overview on Time/Frequency Modulated Array Processing ," DOI 10.1109/JSTSP.2016.2627182, IEEE Journal of Selected Topics in Signal Processing.

[16] P. F. Sammartino, C. J. Baker, and H. D. Griffiths. "Frequency diverse MIMO techniques for radar,"IEEE Transactions on Aerospace and Electronic Systems, vol. 49, no. 1, pp. 201222, January 2013.

[17] M. Madihian, L. Desclos, K. Maruhashi, K. Onda, and M. Kuzuhara. "A high-speed resonance-type FET transceiver switch for millimeterwave band wireless network," in Proc. Microwave Conference, vol.2, pp. 941 944, 1996.

[18] J. H. Lee, S. Hong, W. K. Kim, J. W. An, and M. Y. Park. "A switched array antenna module for millimeter-wave wireless communications," in Proc. Global Symp. Millimeter Waves, pp. 161163, 2008.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 233-240 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Classifying region of interests from mammograms with breast cancer into BIRADS using Artificial Neural Networks

Estefanía D. Avalos-Rivera[1*], Alberto de J. Pastrana-Palma[2]

[1]*Engineering Faculty, Research and Postgraduate Division, Universidad Autónoma de Querétaro, 76010, Mexico*

[2]*Accounting and Administration Faculty, Research and Postgraduate Division, Universidad Autónoma de Querétaro, 76010, Mexico*

ARTICLE INFO

Article history:
Received: 05 March, 2017
Accepted: 20 April, 2017
Online: 02 May, 2017

Keywords:
Breast cancer
Mammograms
Computer-aided diagnosis system
Artificial Neural Network
High pass Gaussian filter
BIRADS
Morphological descriptors
Principal Component Analysis

ABSTRACT

Breast cancer is one of the most common cancers among female diseases all over the world. Early diagnosis and treatment is particularly important in reducing the mortality rate. This research is focused on the prevention of breast cancer, therefore it is important to detect micro-calcifications (MCs) which are a sign of early stage breast cancer. Micro-calcifications are tiny deposits of calcium which are visible on mammograms as they present as tiny white spots. A computer-aided diagnosis system (CAD) is created with the development of computer technology that way radiologists are aided improving their diagnostics while using CAD as a second reader. We are aiming to classify into BIRADS 2, 3 and 4 which are the stages when the cancer can be prevented and a fourth category called No lesion which are veins and tissue that our high pass Gaussian filter detects. This research focuses on classification using ANN (Artificial Neural Network). Experimenting with the categories to classify into using ANN, the results were the following: into the four mentioned before an overall accuracy of 71% was obtained, then joining categories BIRADS 2 and 3 into one and classifying into 3 categories gave an 80% of accuracy. Joining this two categories was the result of analizing the ROC curve and observation of the ROI images of the MCs as the regions measured are very alike in this two categories and variation is that MCs are more present in BIRADS 3 than in BIRADS 2. Data matrix was reduced using PCA (Principal Component Analysis) but it did not gave better results so it was discarded as the ANN accuracy to classify was reduced to a 69.8%.

1. Introduction

This paper is an extension of work originally presented in *IEEE CACIDI 2016 - IEEE Conference on Computer Sciences* [1]. Breast cancer is one of the most common cancers among female diseases all over the world. Early diagnosis and treatment is particularly important in reducing the mortality rate. Currently, the most effective method for early detection of breast cancer is mammography [2]. This research is focused on the prevention of breast cancer, therefore it is important to detect micro-calcifications which are a sign of early stage breast cancer. Micro-calcifications are tiny deposits of calcium which are visible on mammograms as they present as tiny white spots [3]. As micro-calcifications are barely visible in a mammogram it is frequent radiologists missing them in an evaluating screening [2]. A computer-aided diagnosis system (CAD) is created with the development of computer technology, the advances of digital image processing, pattern recognition and artificial intelligence, radiologists are aided improving their diagnostics and using CAD as a second reader [2][3].

The interpretation of micro-calcifications is very difficult due to their fuzzy nature, low contrast and low distinguishability from their surroundings. They are very small with various sizes, shapes, and distributions. To deal with said problems, it is very important to suppress the noise, to enhance the contrast between the region of interest (ROI) and background in the image [2]. Particularly in this research the image database used is of a good quality and high resolution so the finding of micro-calcification clusters (MCCs) it is not as problematic as in previous works that had worked with for example a low quality image free database as the MIAS (Mammographic Image Analysis Society) [4,5,6,7,8,9].

*Corresponding Author: Estefanía D. Avalos-Rivera, Universidad Autónoma de Querétaro, 76010, Mexico | Email: eavalos15@alumnos.uaq.mx

Image database used in this research is acquired from the Medical Specialized Unit on Detection and Diagnosis of Breast Cancer (UNEME DEDICAM) in Querétaro, Mexico. This dataset includes its diagnosis into the BIRADS system (Breast Imaging Reporting and Data System) which was published by the American College of Radiology in an effort to standardize mammography reports [10]. This classification system aims to have a standard way of communicating the results of a mammogram, because it allows radiologists to use the same words and terms. In this research we classify into categories in which the advance of cancer can be prevented, that is BIRADS 2, 3 and 4, we added a fourth category called *No Lesion*, which includes false positives like veins, tissue, and what is detected by the filter that is not a MCC.

Although the images worked with are of good image resolution there was the need to enhance the MCCs so we could threshold the ROI images with accuracy and use those to measure the pixels with morphological descriptors. The approaches for enhancement of MCCs, including various filtering methods, global and local thresholding methods, histogram equalization, mathematical morphology transformations, statistic methods, wavelet transformations, neural networks, stochastic models, fractal models, high-order statistic methods, fuzzy logic approaches, etc. [2], but as the Autonomous University of Querétaro has been researching on detection of MCCs as in [11, 8], this research was improved thanks to [8] because from that a filter bank was made in which a ROI was given many filters were applied to it, and the script gave many enhanced images of the ROI, those images were analyzed and the proper filter for our image database was found, which is a High Pass Gaussian Filter.

Once images were binarized the MCCs and the lesions in ROIs were a region of pixels that were measured by morphological descriptors which are a set of numbers that describe a given shape. The regions may be described based on the boundaries of an object or be described based on regions properties [12]. The descriptors used in this study to quantify the binarized images obtained from the ROIs were *area, perimeter, centroid, Euler number, major axis length, minor axis length* and *orientation*. A data matrix was constructed of the measurements of the regions of the ROIs selected from the mammograms. This matrix was used to feed an Artificial Neural Network (ANN) for training, testing and validation of ROIs with anomalies.

In this research a feed forward neural network (FFNN) was used as ANN is a machine learning technique that has been widely used in different fields as they are good at recognizing patterns. It has been used in [4,13,14,15] to classify MCCs into benign and malign but not to classify into BIRADS categories.

2. Previous researches

In this section, we review some researches that have been done on CAD in recent years, main focus in the classification stage. Our direct previous research is [4] in 2014, thesis that classify into Le Gal using ANN as classifier. In this the categories were benign or malignant classification according to Le Gal with a sensitivity greater than 93.26%, the disadvantage of having a high sensitivity in this classification has an impact on the specificity. While many researches regarding classification of MCCs have been done, using BIRADS to classify is not very common, most common is being or malignant.

Few times classification into BIRADS was done, most recent research about classifying into BIRADS is [16], using Fuzzy Logic they introduced morphological descriptors as linguistic variables, the images were analyzed by a group of doctors and those evaluations were introduced to the fuzzy algorithm, an accuracy of 76.67% to 83.34%, said accuracy was affected by discrepancies of radiologists in evaluating the MCCs.

In 2005 a study [15] comparing several machine learning methods —support vector machine (SVM), kernel Fisher discriminant (KFD), relevance vector machine (RVM)— for classification was conducted, demonstrated that the kernel based methods (i.e., SVM, KFD, and RVM) yielded the best performance, outperforming that of Feed Forward Neural Network (FFNN), again this time the classification was into Malignant and Benign, and SVM was used as a binary classifier, outperforming FFNN, it is not useful for our objective of classifying into more that those two categories.

Another CAD system achieving a really good 91.4% and 90.1% classification accuracy using SVM as classifier was Görgel, Sertbas, and Uçan [6] classifying again into benign and malignant.

Most recent research [7] from 2015 that used an ANN as classifier where a ROI image is classified as normal or abnormal (benign or malignant) using a Probabilistic neural network (PNN) shows that their proposed model performance is good at achieving high sensitivity of 97.27% and specificity of 94.38%.

[9] classifies detected MCCs into benign and malignant cases, eight features such as fractal dimension variations, entropy and wavelet coefficients were proposed to classify both malignant and benign cancerous zones, those are identified and utilized in radial basis function neural network.

FFNN is able to classify into more than 2 categories and even if fuzzy can classify into many categories it fails in accuracy because of subjective diagnosis. In the option of using SVM we have not found the use of it to classify into more than two categories, it is an area we want to experiment but SVM as multiclass classifier.

3. Methodology

3.1. BIRADS

The BIRADS is a quality control system, its daily use implies an evaluation in numerical categories of a mammogram, assigned by the radiologist after interpreting the mammography consists [17]. This allows for a consistent and concise radiographic report and can be understood by multiple doctors or hospital centers. It consists of 7 different classes according to their staging, category 6 was added in the 4th edition of the mammography atlas [18].

Category 0: Insufficient X-ray, need an additional evaluation with another study, it is not possible to determine some pathology.

Category 1: Negative mammography to malignancy, no lymph nodes or calcifications. 0% chance of cancer.

Category 2: Mammography negative to malignancy, but with benign findings (intramammary ganglia, benign calcifications, etc). 0% chance of cancer.

Category 3: Result with probable benignity, but that requires control to 6 months. It may have circumscribed nodules or a small group of rounded and punctate calcifications. 2.24% chance of cancer.

Category 4: Doubtful result of malignancy. It requires histopathological confirmation. It consists of 3 degrees according to its percentage of malignancy ranging from 3 to 94%

A. Low suspicion of malignancy. 3 to 49%

B. Intermediate suspicion of malignancy. 50 to 89%

C. Moderate suspicion of malignancy. 90 to 94%

Category 5: High suspicion of malignancy. Requires biopsy to confirm diagnosis. > 95% chance of malignancy.

Category 6: Malignancy ascertained by biopsy.

In this research, we focus to classify into categories were cancer can be prevented, so the categories to classify our findings are BIRADS 2, 3 and 4 and an extra category called *No Lesion,* which is of veins and tissue that the high pass Gaussian filter detects of the ROIs.

Figure 1 shows examples of samples ROIs used for each category. ROIs were manually obtained from the mammograms.

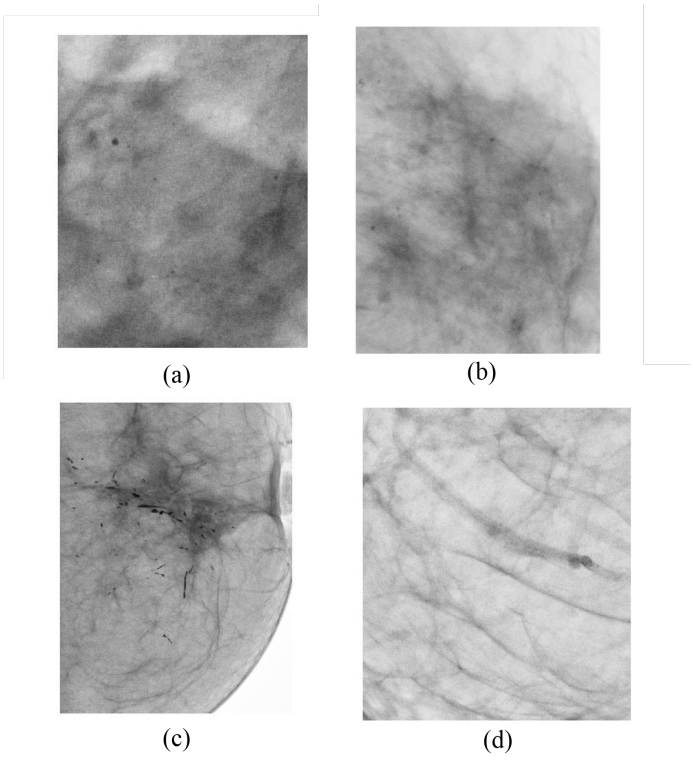


Figure 1. Samples of ROIs of each category taken from the image database from UNEME-DEDICAM: a. BIRADS 2 b. BIRADS 3 c. BIRADS 4 d. No Lesion

3.2. Image Database

Image dataset consists of mammograms of 10 patients for each BIRADS category, 2, 3 and 4. Each patient has 4 mammograms of respective cranial–caudal (CC) view and medio lateral oblique view (MLO) for each breast. For each category 70% of the mammograms were taken for the training stage of the ANN that is mammograms of 7 patients for each category were going to be used to obtain ROIs.

Images format is DICOM (Digital Imaging and Communications in Medicine). DICOM is a standard used worldwide to store, exchange, and transmit medical images. Incorporates standards for imaging modalities such as radiography, ultrasonography, computed tomography, magnetic resonance imaging, and radiation therapy. It also includes protocols for image exchange (e.g., via portable media such as DVDs), image compression, 3-D visualization, image presentation, and results reporting [19].

From the selected mammograms ROIs were manually selected, as seen in Figure 2 meaning areas where MCCs were found in the image according to the respective diagnosis given from UNEME-DEDICAM for each patient study. Display resolution of the images is 3540×4740 pixels.

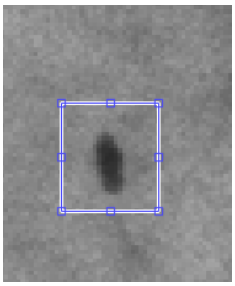


Figure 2. Selecting a micro calcification manually.

3.3. Image Enhancement

First a complement of the grayscale ROI image is applied so that black and white pixels are reversed, and MCCs that were originally black pixels are now white pixels, this done in MATLAB. A filter bank script made using MATLAB environment is used to enhance the ROIs previously extracted, said script is a modified version of what research [8] did to detect MCCs. Script consist of applying three filters to the ROI images, Ideal and Gaussian.

The script uses the High Pass Filters which attenuates low frequencies while keeping high frequencies unchanged. Since the high frequencies correspond in the images to sudden changes of density, this type of filters is used, because among other advantages, it offers improvements in the detection of borders in the space domain, since these contain many of these frequencies. It reinforces the contrasts found in the image. This is important as it is possible to detect by sharpening those areas where MCCs are.

Mathematical description of what the high pass Ideal filter does in the script is described by the following transfer function

$$H(u,v) = \begin{cases} 0 \; if \; D(u,v) \leq D_0 \\ 1 \; if \; D(u,v) > D_0 \end{cases} \quad (1)$$

High pass Gaussian filter transfer function is

$$H(u,v) = e^{-D^2(u,v)/2{D_0}^2} \quad (2)$$

Where D_0 is a specific non negative number which represents the frequency cut-off of the filter and $D(u, v)$ is the distance from point (u, v) to the center of the filter. The script test with different values of D_0.

The script gives ROI images in a loop so it is stopped manually to the point where there are around 1000 images and visually select the best ROI image with the MCCs highlighted the best so when the threshold is to be applied the binary mask obtained reflects accurate the region. In this case the configuration to enhance the

MCCs was a high pass Gaussian filter with a D_0 value of 0.0021 as seen in Figure 3 for each category.

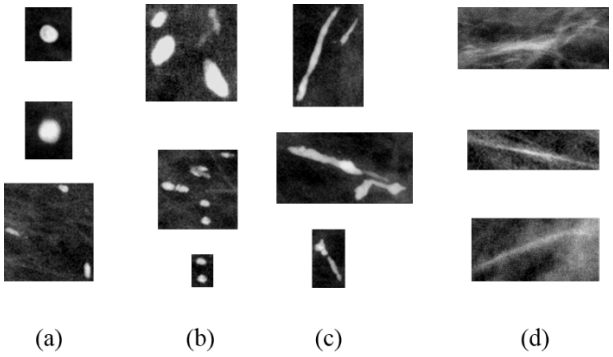


Figure 3. Samples of ROIs with a high pass Gaussian filter with a D0 value of 0.0021 of each category taken from the image database from UNEME-DEDICAM: a. BIRADS 2 b. BIRADS 3 c. BIRADS 4 d. No Lesion

Once the ROI images are enhanced a threshold is applied and a binary mask is obtained with the region of the MCCs ready to be measured. In Figure 4 there a binary mask of a micro calcification (MC).

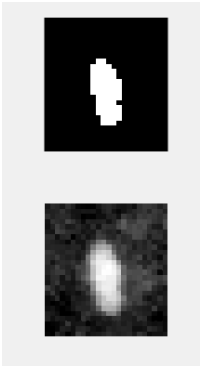


Figure 4. Above, binary image of the ROI. Below ROI with the Gaussian filter applied to the ROI of the BIRADS 4 image.

3.4. Data Matrix

To construct a data matrix it is necessary to measure the ROI images that is to measure the region of white pixels in the binary mask, which will give measurements of the MC region. To measure it, it is used *regionprops* from MATLAB. Region properties selected in this research are:

Area: is the actual scalar number of pixels in the region.

Perimeter: is a scalar that specifies the distance around the boundary of the region.

Centroid: is the center of mass region, in this there are two values, centroid in x and centroid in y.

Equivalent diameter: is the diameter of a circle having the same area with the region.

Euler number: is the number of objects in the region minus the number of holes in those objects.

Major axis length: is the length (in pixels) of the major axis of the ellipse that has the same second moments as the region

Minor axis length: is the length (in pixels) of the minor axis of the ellipse that has the same second moments as the region.

Eccentricity: belongs to the ellipse that has the same second moments as the region, and it is the ratio of the distance between the foci of the ellipse and its major axis length.

Orientation: means the angle (in degrees) between the x-axis and the major axis of the ellipse that has the same second moments as the region [12].

In total 10 features are used to measure MCCs and create a data matrix. Gathering measurements from all the mammograms a 1736 × 10 data matrix. This matrix is standardized, that is take all of the columns of the matrix and standardize / normalize the data so that each data sample exhibits zero mean and unit variance. This means that after this transform, the mean value of any column in this matrix would be 0 and the variance would be 1. This is a very standard method for normalizing values in statistical analysis, machine learning, and computer vision. Following formula describes the calculation of a raw data *x* into a standard data:

$$z = \frac{x - \mu}{\sigma} \tag{3}$$

Where μ is the mean of the population and σ is the standard deviation of the population [20].

Another property used is *Bounding box* that returns the smallest rectangle containing the region and it is used to mark the MCs. Figure 5 shows red boxes marking the MCs it is done using the bounding box property.

Figure 5. Bounding box marking MCs found in the binary mask of a ROI image from a BIRADS 4 mammogram.

3.5. Classification

An ANN is created with the MATLAB toolbox for neural networks to classify the findings in the ROIs in categories BIRADS 2, 3, 4 and *No lesion*, using the pattern recognition tool.

Neural Network Toolbox™ provides algorithms, functions, and apps to create, train, visualize, and simulate neural networks. You can perform classification, regression, clustering, dimensionality reduction, time-series forecasting, and dynamic system modeling and control [12].

In the Neural Pattern Recognition app data to classify is selected, then a network is created. The network created is a two-layer feed-forward with sigmoid hidden and softmax output neurons, can classify vectors arbitrarily well, given enough neurons in its hidden layer. The network will be trained with

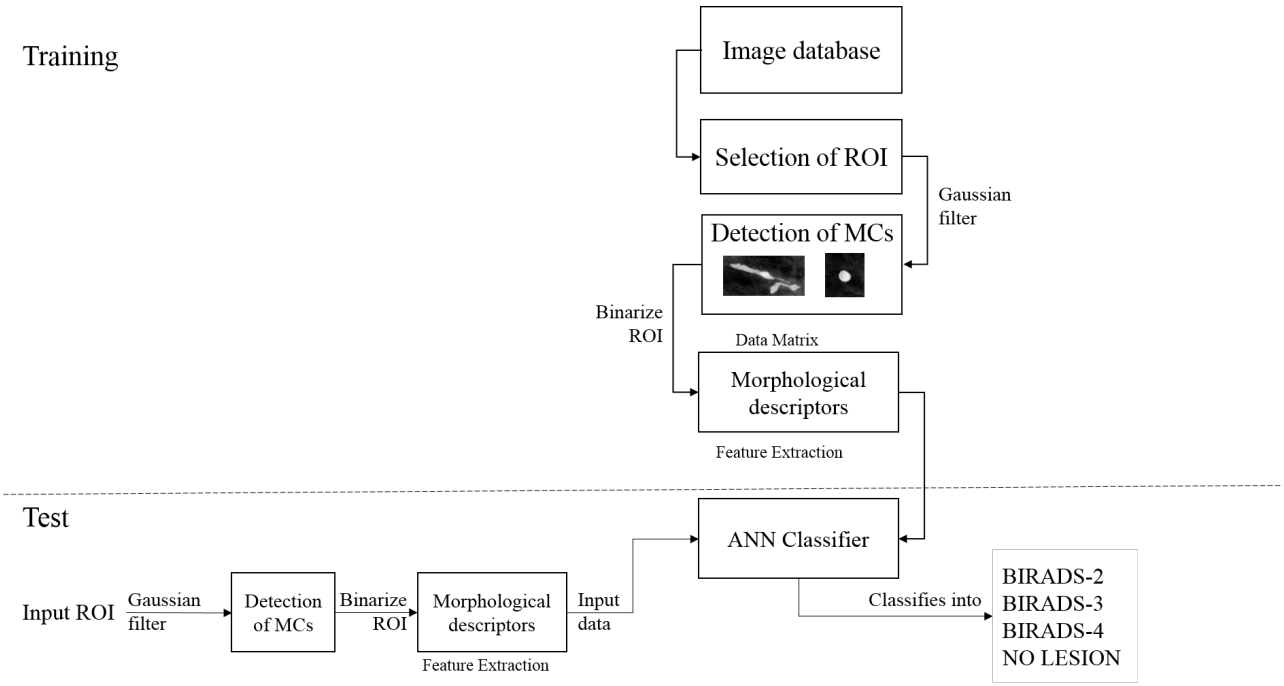


Figure 6. Methodology followed in this research.

scaled conjugate gradient back propagation, in this research it is used ten neurons in its hidden layers.

Next step is to select the data that is the matrix of 1736 × 10 as input data and as output data is a 1736 × 4 matrix with 1's indicating to which category the vector belongs. Each column represents a category, being first column the BIRADS 2 and last column No lesion category. If the data vector belongs to category BIRADS 2 then 1 is in that cell in column 1 and the rest is a zero.

In the next stage the Validation and test data, is where data is distributed into Training, Validation and Testing. In here our 1736 sample are randomly divided. For training 70% of the samples, these are presented to the network during training and the network is adjusted to its error. For validation 15% these samples are used to measure network generalization and to halt training when generalization stops improving. And last 15% for testing, these samples have no effect on training and so provide an independent measure of network performance during and after training. Finally our ANN looks as seen in Figure 7.

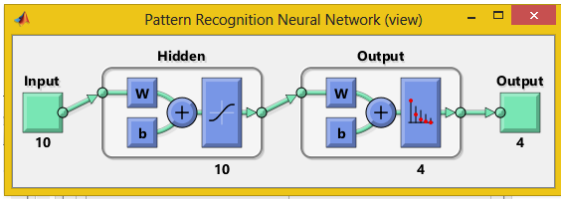


Figure 7. ANN used of 10 input of each morphological descriptor and 4 outputs for each category to classify into. 10 hidden neurons.

Last part is to measure the performance of the ANN. Tools given are the confusion matrices for training, testing and validation and the plot of the Receiver Operating Characteristic (ROC) curve.

The overall methodology used in this research is shown in the diagram in Figure 6. In Figure 8 it is shown the main flow chart of the code used.

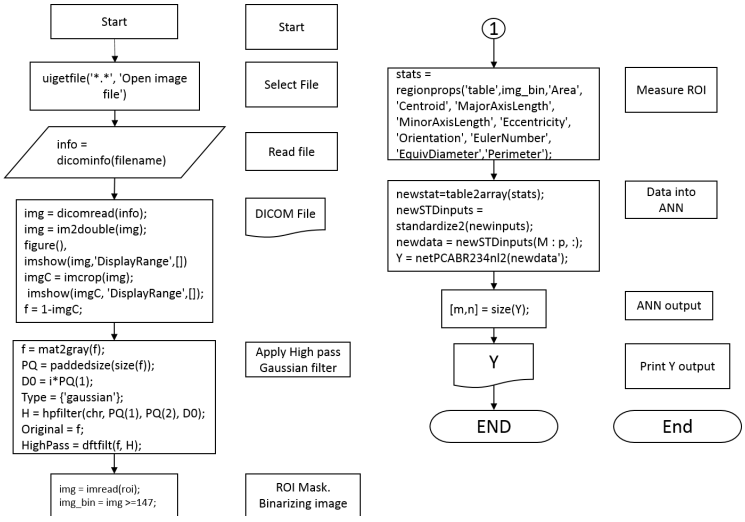


Figure 8. Main flow chart of the code used.

4. Experimental results

In the confusion matrices the green squares are the correct responses and in the red squares are the incorrect responses. The lower right blue squares illustrate the overall accuracies. In this case a 71% of accuracy in general was obtained, but the Test confusion matrix is the most important as the ANN is classifying samples it didn't know and a 75.8% was obtained, see Figure 9.

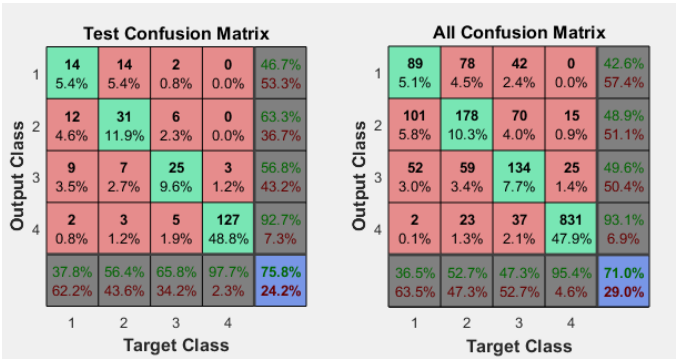


Figure 9. Confusion matrices obtaining a 75.8% in testing and an overall 71%

In the plot of the ROC curve the colored lines in each axis represent the ROC curves. The ROC curve is a plot of the true positive rate (sensitivity) versus the false positive rate (1 - specificity) as the threshold is varied. Sensitivity measures the proportion of positives that are correctly identified as such [21] and the false positive rate is the proportion of all negatives that still yield positive test outcomes, i.e., the conditional probability of a positive test result given an event that was not present. A perfect test would show points in the upper-left corner, with 100% sensitivity and 100% specificity [12]. ROC obtained are seen in Figure 10.

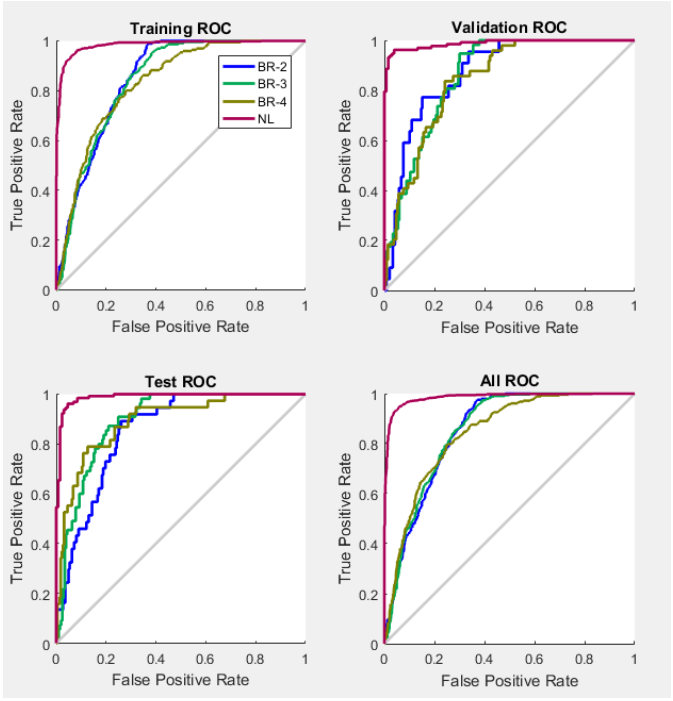


Figure 10. In the ROC plots it is observed that the No lesion category has the best performance, BIRADS 2 and 3 are having trouble to differentiate from each other.

To obtain a better performance of the ANN, it is proposed to apply Principal Component Analysis (PCA) to the data matrix. PCA is a variable reduction technique. It is used when variables are highly correlated. It reduces the number of observed variables to a smaller number of principal components which account for most of the variance of the observed variables. It is a large sample procedure [22]. So this way we expected to discard a region property of the original 10 selected. PCA is applied to the data matrix and outputs 6 descriptors as principal components, those 6 descriptors are area, centroid in x, centroid in y, major axis length, minor axis length and eccentricity. A second network is generated but the overall accuracy gets reduced to a 69.8%, as seen in Figure 11, so using less than the 10 initial descriptors is reducing accuracy so using PCA is discarded.

While experimenting it is notable that the network is having trouble to differentiate the categories BIRADS 2 and 3, as the MCCs in those are very alike in shape, but the difference between them is that the presence of MCCs increases in BIRADS 3. So it is proposed to join this two categories into one and then a new network is generated to train into 3 categories but using the 10 initial descriptors.

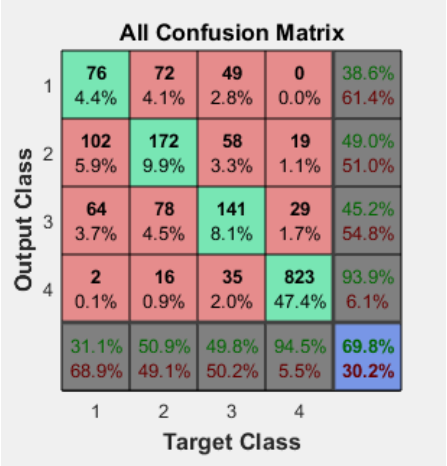


Figure 11. ANN performances gets reduced when only using the first 6 descriptors.

Following the last proposal now the ANN is classifying into 3 categories instead of 4, the overall accuracy is 80% meaning that our previous observation of BIRADS 2 and 3 being alike is correct. In Figure 12 it is observed that the ANN performance improved.

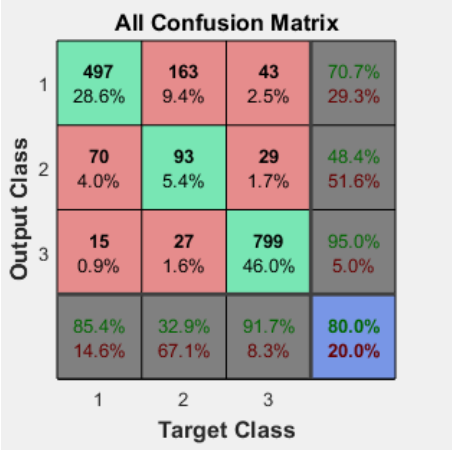


Figure 12. Joining alike categories BIRADS 2 and 3 gave an improvement.

In the experimentation stage the hidden neurons were added up to 15 but it did not improved the performance of the ANN, reducing the hidden neurons reduced the percentage of the overall accuracy of the ANN for a very tiny 1%. So modifying the default 10 hidden neurons is discarded.

5. Comparison with state-of-art

As we discussed in the section *Previous researches,* most of the related researches focus in classifying into two categories, malignant and benignant. As this research focus in the classification method and the system to classify the cancerous findings into BIRADS, our research will be compared against [16] from 2015 and [23] from 2000 which are the more alike as they classify into BIRADS.

First the research conducted by [23] presented an automatic detection and classification of MCCs. A block region growing and K-means clustering-based thresholding is employed to extract the breast region. Then, a blanket method finds and locates the suspicious areas of possible MCCs clusters. The MCCs detection module is developed to automatically extract the MCCs from the ROIs. Among the image processing that are involved in this module are gradient enhancement, contrast enhancement and Gaussian filters. The segmentation of MCCs from the background is done using entropy-based thresholding. Shape cognitron which is based on a neural network-like shape recognition systems is introduced as a classification technique of MCCs. The system achieved as high as 95% classification rate with 93% detection rate [24]. It is important to mention that this particular research has a processing time of 72 seconds as it is completely automatic and it is a more complex methodology, our research is simple, once a ROI is manually selected, goes into a Gaussian filter for denoise, then a binary threshold is applied, regions are measured and finally the data gets classified with a FFNN, all this takes 0.77 seconds of CPU processing time. We obtained an 80% of accuracy, lower against the 95% of [23], it is important to mention they used, 104 cases while we used 120 mammograms. Another difference in methods are that we measure ROIs using morphological descriptors. About the classification technique used they obtained a better accuracy percentage using a neural network-like classifier in our case we are using a more intelligent machine learning technique that is a FFNN that can be tuned every time it is trained and it can learn to be effective. Our research is at disadvantage as it is less accurate, but our methodology is not as complex, and therefore it is faster and barely consume CPU resources.

Next comparing against [16] a computer-aided diagnosis tool for automatic BI-RADS categorization of breast lesions is developed. The user provides parameters such as contour, shape and density and the system gives a suggestion about the BI-RADS classification. Initially, values of malignancy were defined for each image descriptor, according to the BI- RADS standard. When analyzing contour, for example, this method considers the matching of features and linguistic variables. Next, it is created the fuzzy inference system. The generation of membership functions was carried out by the Fuzzy Omega algorithm, which is based on the statistical analysis of the dataset. This algorithm maps the distribution of different classes in a set. Images were analyzed by a group of physicians and the resulting evaluations were submitted to the Fuzzy Omega algorithm. The results were compared, achieving an accuracy of 76.67% for nodules and 83.34%. In this case we are having and ambiguous characterization of the MCCs as it depends of what the user considers is seeing in the mammograms as it is relative to what the user defines as very or little malignancy, as it uses linguistic variables. Contrary to our research where we are actually measuring and obtaining data from the region of the MCC. This data feeds the ANN and helps to obtain an automated classifier.

According to [15] were many classifiers were compared the best is SVM, when classifying binary, next step in our research is explore Multiclass SVM. The advantage of using FFNN is the possibility of classify into more than two categories, therefore achieving to classify into BIRADS. Disadvantage is that our approach is too simple and fails in accuracy, implementing the mesuremnt of MCCs as a cluster could be an improvement for our method as in [23] this approach gave better results than our measuring single MCCs.

6. Conclusions

With the ROC plot it is observed that the best category to classify was the No lesion, from the other 3, meaning that our ANN is accurately discerning from a non-lesion and a lesion or MC. Also the initial 10 descriptors are the better choice and reducing them did not gave a better performance so in future works more descriptors will be added. We are already experimenting with Multiclass SVM as seen in surveys that SVM is better than ANN but as a binary classifier, in our case we'll classify into more categories.

Acknowledgment

We wish to thank Universidad Autónoma de Querétaro (UAQ) for the academic support for this project research and CONACyT, México for the financial support.

References

[1] E. D. Avalos-Rivera and A. D. J. Pastrana-Palma, "Classifying microcalcifications on digital mammography using morphological descriptors and artificial neural network," *IEEE Congr. Argentino Ciencias la Informática y Desarro. Investig.*, 2016.

[2] H. D. Cheng, X. Cai, X. Chen, L. Hu, and X. Lou, "Computer-aided detection and classification of microcalcifications in mammograms: a survey," *Pattern Recognit.*, vol. 36, no. 12, pp. 2967–2991, Dec. 2003.

[3] W. Jian, X. Sun, and S. Luo, "Computer-aided diagnosis of breast microcalcifications based on dual-tree complex wavelet transform.," *Biomed. Eng. Online*, vol. 11, p. 96, 2012.

[4] R. Audelo-Valverde and A. Pastrana-Palma, "Clasificación de Microcalcificaciones utilizando Descriptores Morfológicos," Universidad Autonoma de Queretaro, 2014.

[5] A. Tirtajaya and D. D. Santika, "Classification of Microcalcification Using Dual-Tree Complex Wavelet Transform and Support Vector Machine," in *2010 Second International Conference on Advances in Computing, Control, and Telecommunication Technologies*, 2010, pp. 164–166.

[6] P. Görgel, A. Sertbas, and O. N. Uçan, "Computer-aided classification of breast masses in mammogram images based on spherical wavelet transform and support vector machines," *Expert Syst.*, vol. 32, no. 1, pp. 155–164, 2015.

[7] B. Joseph, B. Ramachandran, and P. Muthukrishnan, "Intelligent Detection and Classification of Microcalcification in Compressed Mammogram Image," *Image Anal. Stereol.*, no. 1997, pp. 183–198, 2015.

[8] L. A. Salazar-Licea, J. C. Pedraza-Ortega, and A. Pastrana-Palma, "Algoritmos de procesamiento digital de imagenes para la deteccion de microcalcificaciones a partir del analisis de mamografias," Universidad Autonoma de Queretaro, 2015.

[9] A. Shirazinodeh, H. A. Noubari, H. Rabbani, and A. M. Dehnavi, "Detection and classification of Breast Cancer in Wavelet Sub - bands of Fractal Segmented Cancerous Zones," vol. 5, no. 3, 2015.

[10] P. Ricci, A. Cruz, M. Rodríguez, H. Sepúlveda, I. Galleguillos, F. Rojas, V. Peña, R. Carvajal, M. Bravo, R. Castillo, and C. Núñez, "Trabajos Originales MICROCALCIFICACIONES BIRADS 4 : EXPERIENCIA DE 12 AÑOS," *Rev Chil Obs. Ginecol*, vol. 71, no. 6, pp. 388–393, 2006.

[11] R.-A. Gibran and A. Pastrana-Palma, "Implementación de un Diagnóstico Asistido por Computadora para la Detección de Microcalcificaciones en Mamografías," Universidad Autonoma de Queretaro, 2014.

[12] *MATLAB R2016a Product Help*. .

[13] C. J. V. Y. Jiang, R. M. Nishikawa, E. E.Wolverton, C. E. Metz, M. L. Giger, R. A. Schmidt, "Malignant and benign clustered micro-calcifications: automated feature analysis and classification," *Radiology*, vol. 198, pp. 671–678, 1996.

[14] D. K. Jiang Y, Nishikawa RM, Schmidt RA, Metz CE, Giger ML, "Improving breast cancer diagnosis with computer-aided diagnosis," *Acad. Radiol.*, vol. 6, pp. 22–33, 1999.

[15] R. M. Nishikawa, "A study on several Machine-learning methods for classification of Malignant and benign clustered microcalcifications," *IEEE Trans. Med. Imaging*, vol. 24, no. 3, pp. 371–380, Mar. 2005.

[16] G. H. B. Miranda and J. C. Felipe, "Computer-aided diagnosis system based on fuzzy logic for breast cancer categorization," *Comput. Biol. Med.*, vol. 64, pp. 334–346, 2015.

[17] F. P. Kestelman, G. A. De Souza, L. C. Thuler, G. Martins, V. Aguilera, R. De Freitas, and E. D. O. Canella, "BREAST IMAGING REPORTING AND DATA SYSTEM – BI-RADS ®: POSITIVE PREDICTIVE VALUE OF CATEGORIES 3 , 4 AND 5 . A SYSTEMATIC LITERATURE REVIEW *," vol. 40, no. 3, pp. 173–177, 2007.

[18] American College of Radiology, *Breast Imaging Reporting and Data System Atlas (BI-RADS Atlas)*. Reston (VA), 2003.

[19] C. E. Kahn, J. A. Carrino, M. J. Flynn, D. J. Peck, and S. C. Horii, "DICOM and Radiology: Past, Present, and Future," *J. Am. Coll. Radiol.*, vol. 4, no. 9, pp. 652–657, Sep. 2007.

[20] E. Kreyszig, *Advanced Engineering Mathematics*, 4th ed. 1979.

[21] "Detector Performance Analysis Using ROC Curves – MATLAB & Simulink Example." [Online]. Available: www.mathworks.com.

[22] P. M. Jain and V. . . Shandliya, "A survey paper on comparative study between Principal Component Analysis (PCA) and Exploratory Factor Analysis (EFA)," *Int. J. Comput. Sci. Appl.*, vol. 6, no. 2, pp. 373–375, 2013.

[23] S. Lee, C. Lo, C. Wang, P. Chung, C. Chang, C. Yang, and P. Hsu, "A computer-aided design mammography screening system for detection and classification of microcalcifications.," *Int. J. Med. Inform.*, vol. 60, no. 1, pp. 29–57, 2000.

[24] N. Yusof, N. Ashidi, M. Isa, H. Amylia, and M. Sakim, "Computer-Aided Detection and Diagnosis for Microcalcifications in Mammogram : A Review," vol. 7, no. 6, pp. 202–208, 2007.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 241-244 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Magnetically Levitated and Guided Systems

Florian Puci[*]*, Miroslav Husak*

Department of Microelectronics, FEE CTU in Prague, 166 27, Czech Republic

ARTICLE INFO

Article history:
Received: 29 March, 2017
Accepted: 25 April, 2017
Online: 02 May, 2017

Keywords:
magnetic levitation technology
planar micro-actuator
stator, translator

ABSTRACT

The paper describes the fundamentals of magnetic levitation technology. A general background of the magnetic levitation is given in this article, including applications of this technology, several comparisons with other types of technologies, the real stage of its development, etc. Further in the paper, the two main types of magnetically levitated systems are compared within their subgroups, on characteristics and specifications basis. A comparison between the AC and DC power supplies for these systems, including the pros and cons of each type, is also provided in the paper.

1. Introduction

The up and coming era of lithography requires a high exactness organize, which is perfect with a high vacuum condition and for this, a magnetic levitation stage with two or more degrees-of-freedom is considered state-of-the-art technology innovation. Nowadays the size of wafer is moving up to 12 inch to enhance the efficiency of a batch process. Thusly, a high-resolution process and a large operating range simultaneously with the application into a super-clean condition are the prerequisites for the related smaller scale actuators. The non-contact characteristic of magnetic levitation technology (MAGLEV) empowers high accuracy positioning as well as no particle generation. The assembling procedure of a current semiconductor IC, forces exceptionally extreme requirements on the preparing exactness as well as the workplace. However, the heat is inevitably generated, while using electromagnetic actuators for levitation, which deforms the structures and degrades accuracy of the stage, and though a gravity compensator is required.

Between the parts with relative motion of the new sort of wafer transporters, spoke to for instance by belt transporters or articulated robots, there are mechanical contacts. For the most part, to limit and diminish the friction and enhance smooth operation, lubricants (oils) are used in the contact regions. In any case, because of the rubbing activity, there are constantly created a lot of tiny solid particles which are mainly absorbed by the lubricant. During the production process, it can happen that some particles can still flee the wafer transporter. Eventually, a portion of the particles will drop back onto the surface of wafers, damaging the fragile and newly manufactured integrated circuit (IC) device.

A fully operational system with magnetic levitation technology is strongly required to have the following characteristics:

- Gravity compensation with no power consumption
- Large and homogenous force density
- Zero stiffness between the translator and stator part
- Position-independent dynamic force
- Enough workspace for operation in all directions

In the nowadays microelectronic industry, the wafer transporter is ranked the number one in the list of particle source extrinsic to the fabrication process. The presence of lubricant in the contact regions has a negative impact to the process cleanness and this results from the volatility of lubricant. The surface properties are strongly affected by the migration and the diffusion of the gaseous elements into the surface of wafers.

An answer can be that the rate of evaporation is enhanced by the application of a vacuum environment, since the undesirable gaseous elements are evaporated to the processing environment.

[*]Corresponding Author: Florian Puci, Department of Microelectronics, FEE CTU in Prague, 166 27, Czech Republic, Contact No: +420734235015 | Email: puciflor@fel.cvut.cz

Obviously, reduction or if possible total elimination of the contact between relatively moving components, is the most ideal approach to eliminate wafer transport contamination. In this way, a particle-free and oil-free environment can be obtained and consequently both particle generation and lubricant evaporation will disappear simultaneously. For instance, in a wafer transporter, mechanical contacts typically happen and are persistently present in the area between the part that conveys the wafer, so the wafer carrier, and the transporter base that supports the carrier. By isolating the wafer carrier from the transporter base, these mechanical contacts can be dispensed with. This can be accomplished by means of levitation, where non-contact physical forces are applied between the carrier and the base, so the carrier is separated from the transporter base. In the known physical world, there exist three principle strategies that the free-floating levitation can be accomplished:

a) air-bearing b) electrostatic levitation c) maglev

To determine the most suitable and effective technique for our application, is required a general physical understanding of each method's working theory including some fundamental information of incorporated circuit manufacture innovation. An air-bearing system essentially requires two very important components: a special filter used to exclude particles generated by the air pump from entering the transporter and an additional circulation system used to enclose the air. In comparison, with the air-bearing system, in the electrostatic levitation one, the particle residues which are generated in the first installation or accidentally generated during wafer loading and unloading process will be pulled by the electrostatic field. Consequently, this can cause possible particle contamination, with the particles being stick to the transporter due to the electrostatic field.

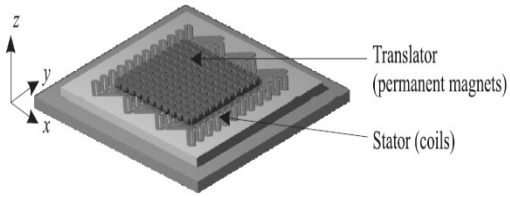
Given the above contentions and to achieve various performance targets, the precision stage using a novel contact-free planar actuator with magnetic levitation is highly recommended and this paper consider it the best solution for IC manufacture. Generally, as MAGLEV eliminates the friction due to a mechanical contact, it has a variety of applications requiring super-cleanness environment, including here the semiconductor wafer transfer. The way of usefulness and functionality of MAGLEV eliminates the problems caused by using the above technologies. It makes a steady state with no mechanical contact when the gravitational force is solely counterbalanced by magnetic forces. This system incorporates levitating tracks, stabilizing tracks, and propulsion coils [2].

2. Magnetically levitated planar actuators

Different types of conventional transportation systems such belt-type conveyors or articulated robots generate dusts and pollution due to the mechanical friction or lubrication. These systems are inadequate and improper to satisfy the environmental demands. The MAGLEV technology has the advantages of being contact free, eliminates the mechanical components e.g., gears, guide etc., reduces the mechanical alignment and maintenance cost, hence it satisfies the environmental demands. In this manner, research on contact-free type transportation system and actuator has been actively performed by worldwide researchers. Present day applications of this technology in equipment like magnetic bearings and magnetically levitated vehicles have given renewed impetus to research efforts in the direction of electromagnetic levitation [3-4].

There exist two main categories of planar actuators, referred as either with moving coils or moving magnets. The actuators from the first type consist of moving air-core coils and stationary magnets. The main advantage of such system is the usage of a small number of coils and their amplifiers, as the stroke force can be easily increased by adding a few more magnets in the magnet array. Additionally, the simple design of these actuators allows torque control on the translator part by using different coil topologies. A major disadvantage for this configuration is the presence of a cable connecting the translator and the stator part, as the coils require power and cooling. The following Figure 1 shows an example of these planar actuators.

a)



b)

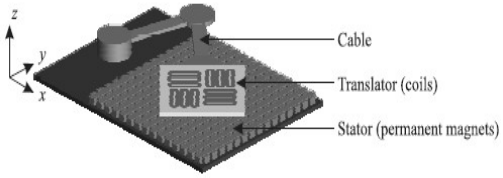


Figure. 1. Planar actuators: a) moving magnets, b) moving coils

The other category is represented by actuators with moving magnets and stationary air-core coils. In opposite with the other class of actuators, it doesn't require a cable for connecting the translator and stator part, which is a big advantage from the design point of view. With the coils located and powered in the stator platform, the amount of disturbances delivered to the levitated translator is then reduced significantly. On the other hand, the torque decoupling as a function of position is more complex than in the moving-coils planar actuators. During manufacturing process, working environment affects the quality of the precision products [3].

3. Designed planar actuator with moving magnets

The moving coil principle in microstructure designs requires powering and cooling of the coils. Additionally, including here the guidance of the current lines on the thin etched silicon beams. The springs which are driven by the Lorentz force, are moving in the magnetic field generated by an external magnet. Strong currents are required to produce large deflections, though this can cause thermally induced bending or buckling of the thin microstructures. Hence, the principle with a moving magnet is more effective, including less complex circuit design.

The carrier consists of several permanent magnets and essentially its stabilization is required. Ideal case is a uniform object with the center of mass coincident with the center of geometry. In numerical analysis, the magnetic flux is firmly related with the levitation distance between the levitated magnet and the cylindrical solenoid underneath. If such distance is very small, then the magnetic flux approaches a constant value. The aim of the stator configuration is the arrangement of the cylindrical solenoids and their floating currents in such a way, so together would form a uniform magnetic field over a planar surface slightly above the coil array. The high-precision positioning performance requires proper setup of carrier control and levitation height. The Figure 2 shows an example of a planar actuator with moving magnets.

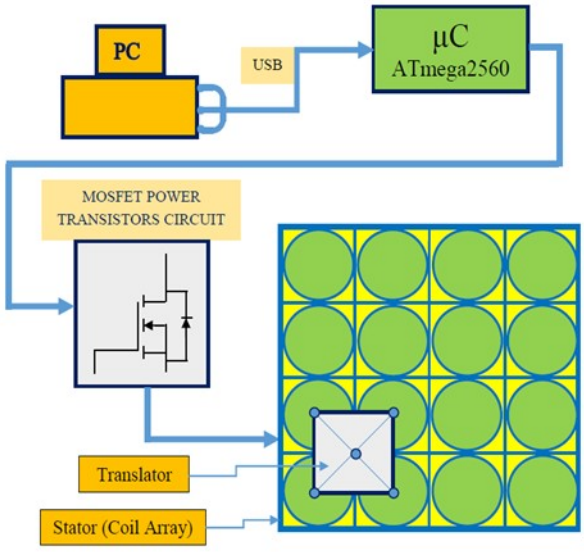


Figure 2. Block diagram of the planar actuator with moving magnets

Referring to the type of planar actuator described here, the translator is represented by a uniform shaped object with specific characteristics. This is embedded with one stabilizing permanent magnet on each corner so it can give sufficient control force against the lateral forces, and with one permanent magnet on the bottom of the carrier to counteract the weight of the carrier. For achieving a high-precision positioning performance, it is required to set up properly the control and the height of the carrier. Coils are supplied by the current flowing through Arduino board pins which are connected to the circuits consisting of MOSFET power transistors. In order to increase the output current from the board pins and strengthen the magnetic field around the coil, 4 such circuits with power MOSFET transistors were designed. For full control of the floating current of the coils, four such circuits are needed. In the stator platform, there are located 16 identical coils, forming together a 4x4 array. The current flowing through a relevant coil is controlled from the output Pulse-Width-Modulation (PWM) signal of one of the board pins of Arduino Mega 2560 [5].

The table below shows the accomplished parameters of the designed microactuator with moving magnets.

Table. Microactuator overall parameters

Parameter	Value			Unit
	min	typical	max	
Supply voltage range	12	-	20	V
Height of magnetically levitated translator	-	5	-	mm
Coil's current	-	0.5	-	A
Electromagnetic field strength	-	62	-	µT
Coil's resistance	-	4.9	-	Ω

4. Supply sources for MAGLEV systems

There exist two main approaches for designing a controlled force magnetic levitation system: the first one uses the attractive forces between an electromagnet and a magnetic flux closure, and the other is based on repulsive forces in interaction between electromagnets and permanent magnets.

Electromagnetic levitation technology uses either an Alternative Current (AC) or Direct Current (DC) source to drive the electromagnets. Despite the fact that there have been manufactured and tested several experimental frameworks which use as the supply an AC source, this method for adjustment is fitting for applications where mass of the levitated object is relatively small. The AC method of stabilization is inappropriate for heavy payloads due to the losses caused by the effect of the eddy currents, the complex circuit design and control of power modulation.

Conversely with the above AC method, the DC method, known also as the electromagnetic levitation system (EMLS), is characterized by simpler circuit configuration and favorable power requirement. The DC EMLS circuits use a switched mode power amplifier to control and utilize the current as well as the attraction force of the electromagnets. An electromagnetic levitation system consists of four main components: (i) Actuator and Rail, (ii) Position Sensor, (iii) Controller, (iv) Power amplifier [6].

The electrodynamic system utilizes the repulsive forces to levitate the carrier, conversely with the other system which levitates the carrier based on the attractive forces. The gap between the magnet pole-face and the ferromagnetic object is sensed and measured by a position transducer. The output signal from the transducer is fed back to a comparator and the process continues with output signal of the comparator being applied to a position controller, giving in this way the reference current for the current loop. The current sensor compares the actual current supplying the coil with the reference current. The power amplifier, a major component in this configuration, produces necessary currents in the actuator coils after receiving a command from the current controller. Preceding this, the current controller completes the current error process task. The output currents from the power amplifier are used to supply the coils. These generate requisite magnetic forces which are the key for levitating the translator.

Applications of electromagnetic levitated and guided systems, can be found for example, in the field of transportation vehicles, frictionless bearings, conveyor systems and lithography scanners [6].

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

Research described in the paper was supervised by Prof. Ing. Miroslav Husak, CSc., FEE CTU in Prague and has been supported by the CTU project No. SGS14/195/OHK3/3T/13 "Development of Smart Devices and Systems in the Field of Microelectronics, Nanoelectronics, Optoelectronics and Micro-nano-optoelectronic structures, elements and systems".

References

[1] Y. M. Choi, M. G. Lee, D. G. Gweon, J. Jeong, "A new magnetic bearing using Halbach magnet arrays for a magnetic levitation stage", Rev. Sci. Instrum., **80** (4), 1-9, 2009.

[2] H. Park, S. K. Lee, J. H. Yi, S. H. Kim, Y. K. Kwak, I. A. Wang, "Contactless magnetically levitated silicon wafer transport system", J. Mech., **6** (5), 591-610, 1999.

[3] C. M. M. van Lierop, "Magnetically levitated planar actuator with moving magnets: Dynamics, commutation and control design", Ph.D Thesis, Eindhoven University of Technology, 2008.

[4] J. W. Jansen, "Magnetically levitated planar actuator with moving magnets: Electromechanical analysis and design", Ph.D Thesis, Eindhoven University of Technology, 2007.

[5] F. Puci, M. Husák, "Translator with Magnetic Levitation" in 10th International Conference on Advanced Semiconductor Devices and Microsystems, Bratislava, 2014.

[6] P. K. Biswasa, S. Banerjeeb, "Design and ANSYS Software Based Simulation of U-I Type Actuator and Rail Used in Electromagnetic Levitation System", Int. J. Appl. S. Eng., **12** (3), 225-239, 2014.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 2, No. 3, XX-YY (2017)
www.astesj.com
Special issue on Recent Advances in Engineering Systems

ASTES Journal
ISSN: 2415-6698

Energy-Efficient Mobile Sensing in Distributed Multi-Agent Sensor Networks

Minh Tuan Nguyen*

Thai Nguyen University of Technology (TNUT), Vietnam

ARTICLEINFO

Article history:
Received: 23 March, 2017
Accepted: 26 April, 2017
Online: 10 May, 2017

Keywords:
Mobile Sensor Network
Compressive Sensing
Data Collection

ABSTRACT

In this paper, we exploit an integration between Compressive sensing (CS) and the random mobility of sensors in distributed mobile sensor networks (MSNs). A small number of distributed mobile sensors are deployed randomly in a sensing area to observe a large number of positions. The distributed mobile sensors sparsely sample the sensing area for data collection. At each sampling time, the sensors collect data at their random positions and exchange their readings to the others through their neighbors within the sensor transmission range to form one CS measurement at each sensor. After a certain number of rounds for moving, sensing and sharing data, each mobile sensor creates a sufficient CS measurements to be able to reconstruct all readings from all positions that need to be observed. Network performance is analyzed considering the number of sensors deployed in the networks, the convergence time and the sensor transmission range. Expressions for transmission power consumption are formulated and optimal low power cases are identified.

1 Introduction

1.1 Motivation

Energy efficiency is the most important issue for mobile sensor networks (MSN) that can be useful for measuring data in many applications including environmental monitoring, event detection, intrusion detection, etc. These networks are constructed from sensors, control algorithms and other dynamic factors which depend on specific purposes or application scenarios [1, 2]. We define the vector $\underline{\mathbf{X}} = [x_1\, x_2 \ldots x_N]^T$ to represent all sensor readings from the locations to be observed. These readings are typically highly correlated and compressible and could be an object for energy saving.

Compressive sensing (CS) [3, 4, 5] offers to sample and to reconstruct sparse or compressible signals using fewer samples than the Nyquist-Shannon theorem would suggest. CS can be applied effectively with wireless sensor networks (WSN) and MSNs to reduce the amount of data and power required by the network. The technique can recover all data from $\underline{\mathbf{X}}$ based on a small number of CS measurements ($\underline{\mathbf{Y}} = [y_1\, y_2 \ldots y_M]^T$) compared to the number of nodes or positions ($M \ll N$) as $\hat{\underline{\mathbf{X}}} = arg\,min \parallel \underline{\mathbf{X}} \parallel_1$, $subject\ to\ \underline{\mathbf{Y}} = \Phi\underline{\mathbf{X}}$, where Φ is the measurement matrix, also called routing matrix in wireless networks. Each CS measurement can be collected from all sensing nodes or from some random nodes.

In this paper we propose a novel frame work that exploits CS sampling by mobile sensors deployed randomly in a sensing area. At each sampling time, each sensor samples its own data as x_i which is shared with others L connected sensor nodes in the MSN. We assume the mobile sensors move into random positions in the area. After each round of moving, data sampling, and sharing, a linear combination of readings is computed as $y_i = \sum_{j=1}^{L} x_j$. To achieve a desired error-target, each mobile sensor needs to move M times and share data to obtain M CS measurements ($\underline{\mathbf{Y}} \in \mathbf{R}^M$). A CS recovery algorithm can be applied to reconstruct all readings from positions in the sensing area ($\hat{\underline{\mathbf{X}}} \in \mathbf{R}^N$).

With a set transmission range, R, mobile sensors may move out of range and become disconnected from the group. With the proposed method the unequal CS measurements do not negatively impact CS performance due to a sparse binary measurement matrix.

*Corresponding Author Email: mtn@ostatemail.okstate.edu

https://dx.doi.org/10.25046/aj020334

1.2 Related work

The integration between CS and data collection methods in WSNs is being exploited effectively [6, 7]. It offers to reconstruct all readings from the networks based on a small number of CS measurements. Each CS measurement is collected from some random sensors. Some data collection methods utilizing CS are proposed as energy efficient algorithms to reduce power consumption for sensors. In [8, 9, 10] random walks with CS provide distributed routing methods for WSNs. Cluster based [11, 12] and tree based [13, 14] data collection methods significantly show the power reduced based on the combination with CS.

Recent research studies have exploited the mobility of sensors and CS. Wang [15] applied CS to monitoring vehicle networks. Mostofi built maps in mobile networks [16] and robot networks [17] while the mobile sensors and robots were deployed outside the sensing areas. Huang [18] reconstructed a scalar field using MSNs and information fusion. Nguyen [19, 20] used flocking control to lead a group of distributed robots to sample data and applied CS to recover data from the field based on a certain number of CS measurements.

The previous work has not focused on the integration between the random mobility of sensors and CS. In this paper, we complete our previous work [21]. The mobile sensors move only a certain number of times and share data among them to generate a desired number of CS measurements. Each sensor can reconstruct all raw data based on the measurements. The important results in this paper include:

1. A proposed new distributed compressive mobile sensing data collection method.
2. Formulations and analysis for estimating network transmission power consumption.
3. Analysis and simulation of factors such as the number of mobile sensors, the convergence time and the sensor transmission range leading to choices of minimizing power consumption.

2 Problem Formulation

2.1 Network Model

We assume an area to be observed with N unknown positions corresponding to N readings for collection. There are L mobile sensors randomly distributed in the sensing area to sense data. They are allowed to move with random direction and random velocities. The connections between sensors are created by radio links having a maximum sensor transmission range denoted as R. An appropriate value of R is chosen for all sensors. As shown in Figure 1, all mobile sensors are connected as an undirected graph $G(L, E)$, where L is the set of vertices representing the mobile sensors and E is the set of edges representing the connections between sensors.

We further assume that the distributed sensors move randomly in the area. Each mobile sensor shares data to the others through its neighbors within the transmission range R.

2.2 Compressive Sensing (CS)

Compressed Sensing techniques [3, 4, 5] bring us amazing work to recover a compressible signal from undersampled random projections. They are also called measurements. A compressible vector signal $\underline{\mathbf{X}} \in R^N$ ($\underline{\mathbf{X}} = [x_1\, x_2 \ldots x_N]^T$) is k-sparse ($\underline{\mathbf{X}}$ has k non zero elements) or dense but sparse in Ψ domain $\underline{\mathbf{X}} = \Psi\underline{\Theta}$ (where $\underline{\Theta}$ is k-sparse vector) will be sampled and then recovered precisely with CS algorithm. The huge gain when we apply CS is the number of measurements are much less than the number of original vector values. Vector $\underline{\mathbf{Y}}$, called the measurement vector, contains data sampled from N sensor readings; $\underline{\mathbf{Y}} \in R^M$ ($\underline{\mathbf{Y}} = [y_1\, y_2 \ldots y_M]^T$), where $M \ll N$.

Signal Sampling: The random measurements are generated by $\underline{\mathbf{Y}} = \Phi\underline{\mathbf{X}}$ where $\Phi \in R^{M\times N}$ is often full-Gaussian matrices or binary matrices, called projection matrices. $\underline{\mathbf{Y}} \in R^M$ is the measurement vector with $y_i = \sum_{j=1}^{n} \varphi_{i,j} x_j$, where $\varphi_{i,j}$ are all entries on the ith row of our projection matrix Φ.

Signal Recovery: The number of CS measurements required to reconstruct the original signal perfectly with high probability is $M = \mathcal{O}(k\, log\, N/k)$ following the l_1 optimization problem given by [3].

$$\hat{\underline{\mathbf{X}}} = arg\,min \|\underline{\mathbf{X}}\|_1, subject\,to\ \underline{\mathbf{Y}} = \Phi\underline{\mathbf{X}}, \qquad (1)$$

or in case we need sparsifying matrix to have $\underline{\mathbf{X}}$ sparse in Ψ domain (Ψ can be Wavelet or DCT depending on our signal properties) as follows.

$$\hat{\underline{\Theta}} = arg\,min \|\underline{\Theta}\|_1, subject\,to\ \underline{\mathbf{Y}} = \Phi\Psi\underline{\Theta}, \qquad (2)$$

where $\|\Theta\|_1 = \sum_{i=1}^{N} |\Theta_i|$. The l_1 optimization problem can be solved with linear programming techniques such as Basis Pursuit (BP) [3].

In reality, we have to consider the noise while sampling and sending the measurements (in our case we collect measurements and send to the base-station) : $\underline{\mathbf{Y}} = \Phi\underline{\mathbf{X}} + \underline{e}$, with $\|e\|_2 < \epsilon$ and recover

$$\hat{\underline{\mathbf{X}}} = arg\,min \|\underline{\mathbf{X}}'\|_1, subject\,to\ \|\underline{\mathbf{Y}} = \phi\underline{\mathbf{X}}'\|_2 < \epsilon. \qquad (3)$$

2.3 Data Collection in Distributed MSNs utilizing CS

The idea of applying CS into MSNs is that the sensing field can be sampled randomly based on the mobility of mobile sensors randomly deployed in the field. In general, each CS measurement is a collection from some random positions as shown in Figure 2. L mobile sensors are deployed randomly in the sensing area that they can collect sensing data randomly. If these sensors exchange their own data to each other, they can all achieve the data sampled randomly as shown

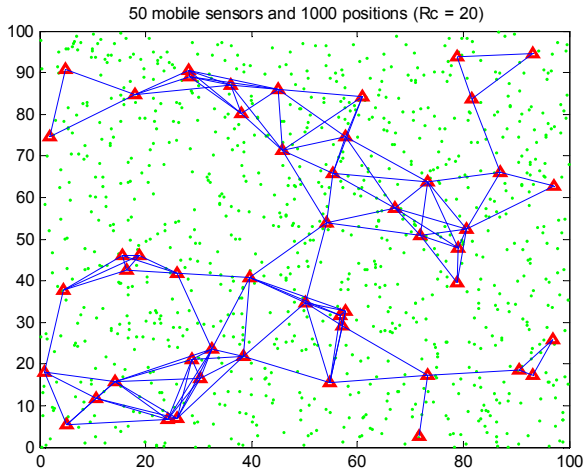


Figure 1: Sparse sampling a square sensing area dimensioned $[0,100]^2$ units: one CS measurement created by a fully connected MSN with 50 distributed mobile sensors.

in Figure 2. At this point, one CS measurement is collected at each distributed mobile sensor. After M moving steps, each sensor has M measurements. This number of measurements defined in CS background in Section 2.2 helps to reconstruct all sensory readings from N positions in the area.

After M periods of time, L distributed mobile sensors visit approximately N unknown positions. The readings are contained in vector $\underline{\mathbf{X}}$ which is compressible or sparse in proper domains as

$$\underline{\mathbf{X}} = \Psi \underline{\mathbf{\Theta}}, \tag{4}$$

where $\underline{\mathbf{\Theta}} = [\theta_1\, \theta_2 \ldots \theta_N]^T$ has k non-zero elements and

$$|\theta_1| \geq |\theta_2| \geq \ldots \geq |\theta_N|. \tag{5}$$

Ψ referred to as the sparsifying matrix, is an orthogonal basis of $\mathbf{R}^N$. If $\underline{\mathbf{X}}$ is a k-sparse vector, Ψ is an identity matrix. Otherwise, Ψ can be chosen from another domain such as wavelet, DCT (discrete cosine transform), etc. In order to verify this problem, Figures 3 and 4 are provided. All raw sensor readings from 1000 points of interest are shown in Figure 3. This real data is not sparse but it is highly correlated due to the positions next to each other. DCT Ψ matrix has been chosen to sparsify the real data. Figure 4 shows all the coefficients in DCT domain. We can see that signal energy would be reduced following Equation 5.

At time instant t, mobile sensor l at position i collects sensory data x_i^l $(i = 1, \ldots N)$. All L mobile sensors share their readings (x_i^l), identified by the corresponding position indices, with the other mobile sensors through their neighbors as shown in the graph $G(L,E)$. After a convergence time [1], denoted as I, each mobile sensor has one CS measurement as a sum of L sensor readings

$$y_t = \sum_{l=1}^{L} \phi_{ti} x_i^l + e_t, \tag{6}$$

where e_t is additive noise dependent on the system. ϕ_{ti} is a binary vector (dimensioned $1 \times N$) that represents which positions are sampled at time instant t. It is also the tth row in the measurement matrix $\Phi_{M\times N}$ after M periods of time. The total measurements collected at each sensor are

$$\underline{\mathbf{Y}} = \Phi \underline{\mathbf{X}} + \underline{\mathbf{e}}. \tag{7}$$

According to [22], the number of measurements required to reconstruct perfectly all the raw reading from the network is

$$M = \mathcal{O}(k \log(N/k)), \tag{8}$$

where $k < M \ll N$. So instead of collecting N readings, one for each sampling position in the sensing area, each mobile sensors needs only M measurements and reconstructs all data from the area. If the network is fully connected all the time, each sensor has the same binary measurement matrix Φ with a constant row weight of L.

$$\Phi = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 1 & \ldots & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & \ldots & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & \ldots & 1 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & 0 & 1 & 0 & 0 & 1 & \ldots & 0 \end{bmatrix}_{M\times N} \tag{9}$$

The restricted isometry property (RIP) of the sparse binary matrix has been studied in [23] where it is shown that the matrix can satisfy RIP and therefore can be used as an efficient measurement matrix. This matrix $(\Phi_{M\times N})$ can perform as well as a full Gaussian matrix for the CS recovery process. ϵ is set as $\epsilon < \| \underline{\mathbf{e}} \|_2$. The recovery algorithm is addressed as

$$\hat{\underline{\mathbf{X}}} = arg\,min \| \underline{\mathbf{X}} \|_1, subject\,to \quad \| \underline{\mathbf{Y}} - \Phi \underline{\mathbf{X}} \|_2 < \epsilon. \tag{10}$$

[1] Convergence time is the average number of times each sensor updates data from the neighbors until it achieves a CS measurement collected from all connected sensors. Convergence time is a measure of how fast a group of sensors can create one CS measurement.

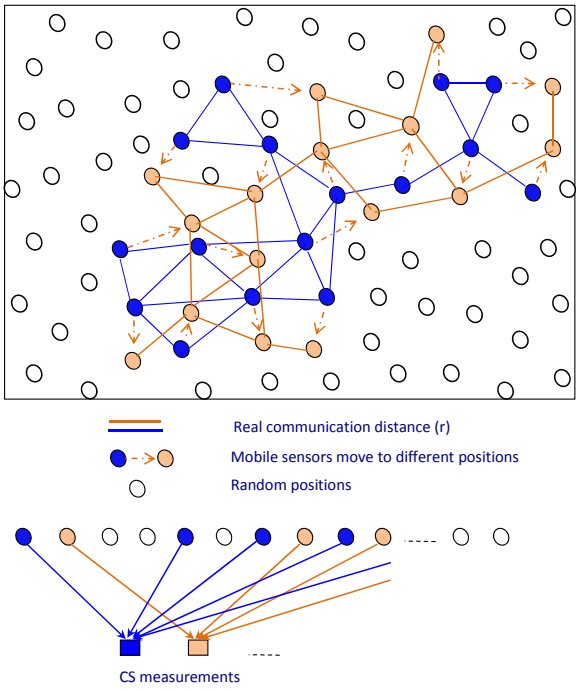


Figure 2: Two CS measurements are created (in blue and yellow) while the mobile sensors are moving in the sensing field.

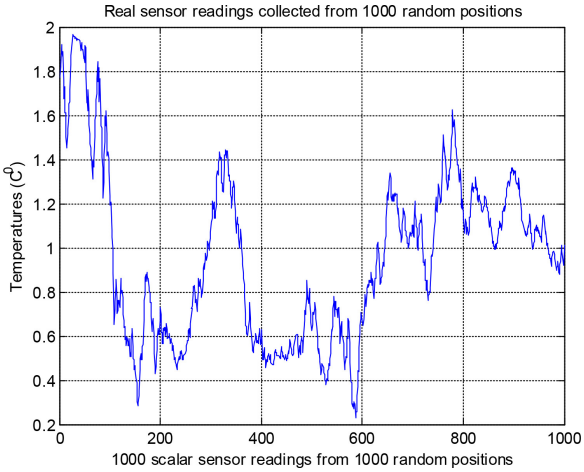


Figure 3: Real temperatures collected from 1000 random positions in a sensing field

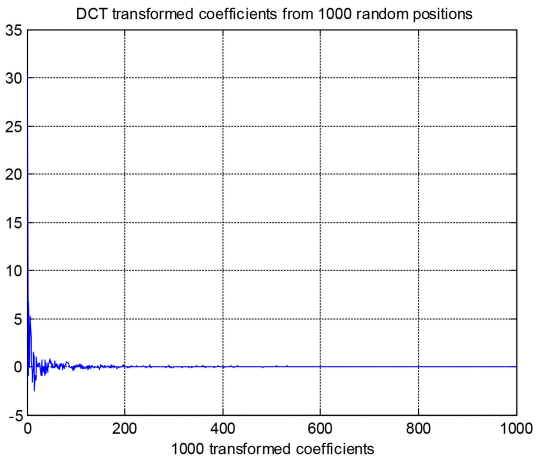


Figure 4: 1000 transformed coefficients using DCT domain

3 Compressive Mobile Sensing Algorithm (CMS)

for *t = 1 to M* **do**
  L distributed mobile sensors move and sample a sensing area at *M* periods of time, and create 1 CS measurement at each time instant *t*
  while *(Number of times sharing data < I)* **do**
    for *i = 1 to L* **do**
      Each sensor out of *L* mobile sensors has two main activities:
      - send/receive data to/from neighbors
      - add/forward new received data from/to neighbors
    end
    New data is detected by its attached indices
  end
  One CS measurement is created following Eq. 6
end
M CS measurements are created for the CS data reconstruction following Eq. 10

Algorithm 1: CMS Algorithm

As mentioned, each distributed mobile sensor must collect M measurements to be able to reconstruct all readings from N positions in the sensing area. With L distributed mobile sensors fully connected as given by the graph $G(L, E)$ and based on the sensor transmission range R, the proposed algorithm is summarized as Algorithm 1.

Due to the limitation of R and sensor motion, some mobile sensors maybe disconnected from others and separated at time instant $t + k$. This means that the measurements y_{t+k} collected at each sensor are not from all L sensors. This does not affect CS performance when the measurement matrix has different row weights [22]. By applying CS, each mobile sensor only has to visit M positions to collect all data from the sensing area ($M \ll N$). This significantly reduces consumed power not only for communications but also for sensor movements.

4 Analysis of Power Consumption for Communications

The total consumed power for communications in the network consists of three main elements: the intra-neighborhood consumed power denoted as P_{nei}, the convergence time I and the number of measurements required (M). The total power consumption can be calculated as follows

$$P_{total} = P_{nei} \times I \times M. \tag{11}$$

P_{nei} represents the communication cost associated with L mobile sensors transmitting data to their neighbors and can be calculated as

$$P_{nei} = \omega \times R^{\alpha} \times L, \tag{12}$$

where ω is the average number of neighbors of each sensor corresponding to the available communication links, and α is the path loss exponent. It is shown in [24] that $\alpha = 2$ and $\alpha = 4$ for free space and multi-path fading channels, respectively. We choose $\alpha = 2$ throughout this paper. In the following sections both circular and square sensing areas are examined and analyzed.

4.1 Circular Sensing Area

If sensors are randomly distributed in a circular sensing area with radius R_0, the average number of sensors deployed in the area covered by each sensor transmission range R can be found as

$$\beta = \frac{L}{\pi R_0^2} \times \pi R^2, \tag{13}$$

where $\frac{L}{\pi R_0^2}$ is the sensor density. We can calculate approximately the mean value of ω as

$$\omega = (\beta - 1) = (L\frac{R^2}{R_0^2} - 1). \tag{14}$$

Hence, the total consumed power for sharing data between L sensors is calculated as

$$P_{total} = (L\frac{R^2}{R_0^2} - 1)R^2 LIM. \tag{15}$$

4.2 Square Sensing Area

We assume a square sensing area with dimensions $H \times H$. From Equation (13), the average number of sensors deployed in the area covered by each sensor transmission range R is $\beta = \frac{L}{H^2} \times \pi R^2$. Hence

$$\omega = (\frac{\pi L R^2}{H^2} - 1). \tag{16}$$

In Figure 6, both the true and the reconstructed data are shown together to illustrate the accuracy of the CS recovery algorithm with the number of CS measurements given by $M = 300$ stored at each sensor. The larger the number of measurements, the better the accuracy of the reconstruction.

The total consumed power for sharing data for L sensors to achieve M converged CS measurements in the square sensing area can be written as

$$P_{total} = (\frac{\pi L R^2}{H^2} - 1)R^2 LIM. \tag{17}$$

The convergence time I depends on both the sensor density ($\frac{L}{\pi R_0^2}$ or $\frac{L}{H^2}$) and the connections between the sensors. The smaller the number of mobile sensors, the smaller the convergence time. However, since the transmission range R determines the connections, increasing the transmission range R to maintain the connections increases the transmission power consumption. This means that increasing or reducing R may

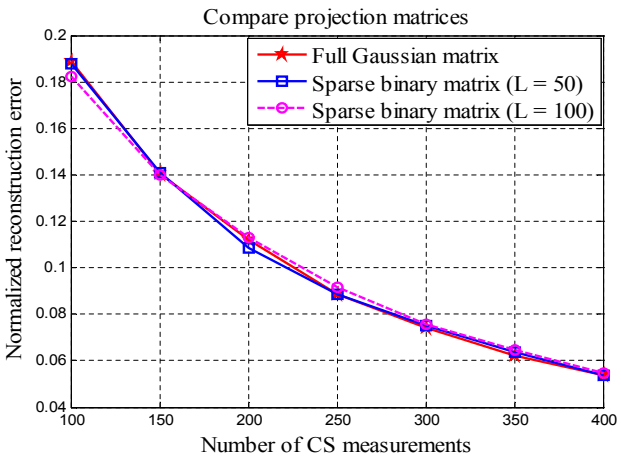


Figure 5: Comparison of three measurement matrices created through the sampling process based on signal reconstruction error.

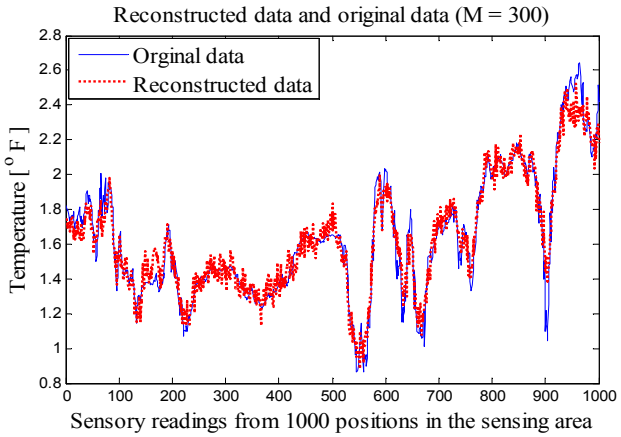


Figure 6: Comparison of reconstructed data and the original data when each sensor stores $M = 300$ CS measurements.

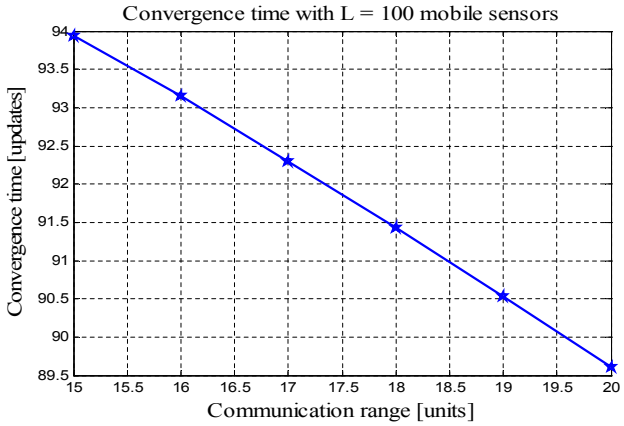


Figure 7: The average convergence time versus different transmission ranges with 100 mobile sensors deployed in a square sensing area.

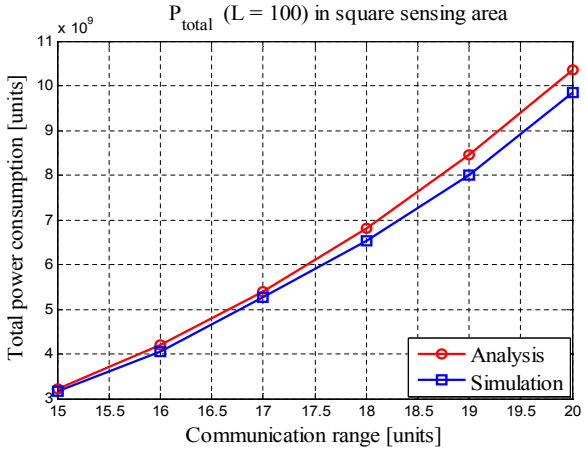


Figure 8: Total power consumption for communications versus different transmission ranges in the square sensing area.

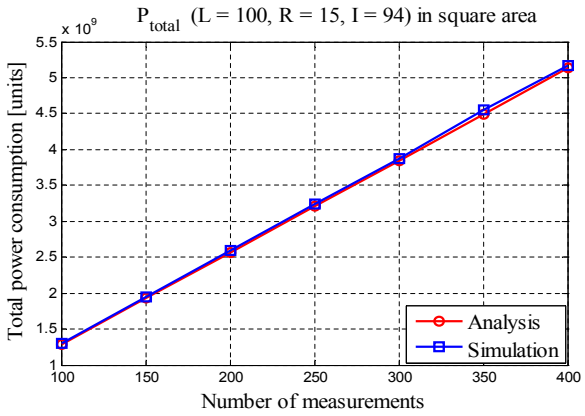


Figure 9: Total communication power consumption with $L = 100$ mobile sensors deployed in the square sensing area and transmission range $R^* = 15$.

increase or reduce the number of connections between the sensors in each neighborhood. Increasing the connections between sensors reduces the convergence time I. Thus a trade-off exists between these various parameters.

5 Simulation Results

We assume a fixed number of 50 or 100 mobile sensors deployed in a square sensing area of 100 square units ($H = 100$). We also assume that there are approximately 1000 positions where temperature needs to be observed. We consider the transmission range of $15 \leq R \leq 20$ units in which the network is always connected. Without loss of generality, it is assumed that the power for transmitting 1 unit of data is 1 unit of power. The simulations were performed using real temperature sensor data from Sensorscope [25]. The reconstruction error related to CS signal recovery is the normalized reconstruction error given as $\frac{\|\mathbf{X}-\hat{\mathbf{X}}\|_2}{\|\mathbf{X}\|_2}$.

Figure 5 compares the ability to reconstruct data in the CS recovery process between three measurement matrices. Both sparse binary matrices with row weight $L = 50$ and 100 are shown to perform as well as the full Gaussian matrix which corresponds to full sampling.

Figure 7 illustrates that as R is increased, the convergence time I is reduced as discussed in Section 4. Figure 8 depicts the corresponding total power consumption in the network, and we see the power consumption is reduced as R is decreased. R cannot be reduced without limit as at some point the L sensors will become disconnected and fragmented resulting in an excessively sparse measurement matrix.

We chose the smallest transmission range $R^* = 15$ and the corresponding convergence time $I^* = 94$ for the optimal total power consumption for sensor communications in the network in both analysis and simulation which are shown in Figure 9.

Finally, in Figure 6, both the true and the reconstructed data are shown together to the show the accuracy of the CS recovery algorithm at the number of CS measurements as $M = 300$ stored at each sensor. The more number of measurements, the better accuracy of the reconstruction.

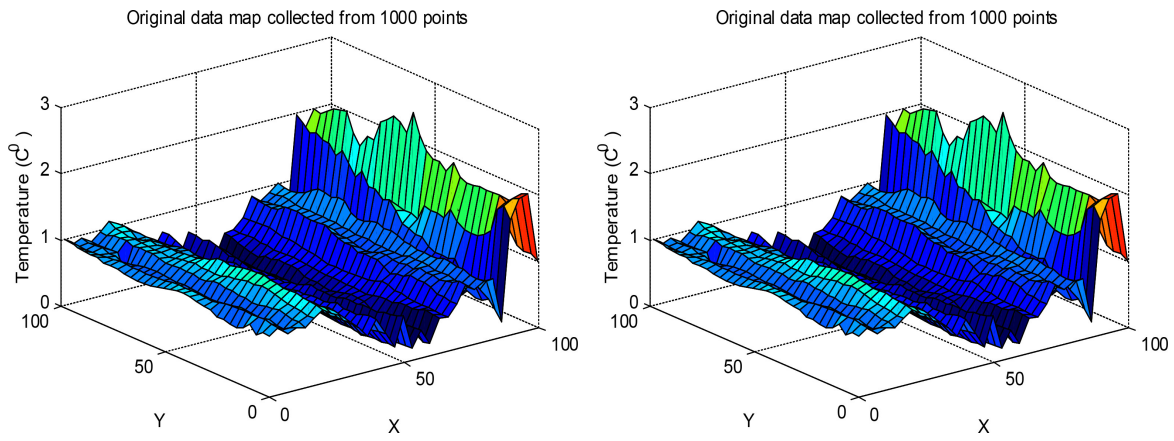


Figure 10: Final result of recovering data in a 3-D map compared to the true map; $M = 300$ CS measurements are collected.

6 Conclusions and Future Work

This paper proposed a framework for distributed MSNs utilizing CS for efficient data sensing and recovery and investigated the dependence of power consumption on various MSN parameters. The algorithm exploits the random mobility of sensors for sparse sampling the sensing area. CS based sampling and reconstruction of the sensory data with a sparse binary measurement matrix were compared with dense Gaussian and shown to be equal under the simulation conditions, illustrating that CS may be used to advantage in MSNs. Expressions for transmission power consumption in the MSN were developed, simulated, and analyzed. The results show the trade-off between the number of sensors, the transmission range, and the convergence time that can further reduce the power consumption for data transmission in the networks.

In future work, we consider the correlation of sensing data in the field. This could help to improve the performance of CS in signal recovery.

Acknowledgment This work is supported by Thai Nguyen University of Technology (TNUT), Viet Nam.

References

[1] B. Liu, O. Dousse, P. Nain, and D. Towsley, "Dynamic coverage of mobile sensor networks," *Parallel and Distributed Systems, IEEE Transactions on*, vol. 24, pp. 301–311, Feb 2013.

[2] C. Tunca, S. Isik, M. Donmez, and C. Ersoy, "Distributed mobile sink routing for wireless sensor networks: A survey," *Communications Surveys Tutorials, IEEE*, vol. 16, pp. 877–897, Second 2014.

[3] D.L.Donoho, "Compressed sensing," *Information Theory, IEEE Transactions on*, vol. 52, pp. 1289 – 1306, 2006.

[4] E. Candes, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," *Information Theory, IEEE Transactions on*, vol. 52, pp. 489 – 509, Feb. 2006.

[5] E. Candes and J. Romberg, "Sparsity and incoherence in compressive sampling," *Inverse problems*, vol. 23, no. 3, p. 969, 2007.

[6] G. Quer, R. Masiero, G. Pillonetto, M. Rossi, and M. Zorzi, "Sensing, compression, and recovery for wsns: Sparse signal modeling and monitoring framework," *Wireless Communications, IEEE Transactions on*, vol. 11, pp. 3447–3461, October 2012.

[7] M. T. Nguyen, K. A. Teague, and N. Rahnavard, "Ccs: Energy-efficient data collection in clustered wireless sensor networks utilizing blockwise compressive sensing," *Computer Networks*, vol. 106, pp. 171 – 185, 2016.

[8] M. Sartipi and R. Fletcher, "Energy-efficient data acquisition in wireless sensor networks using compressed sensing," in *Data Compression Conference (DCC), 2011*, pp. 223 –232, March 2011.

[9] M. T. Nguyen, "Minimizing energy consumption in random walk routing for wireless sensor networks utilizing compressed sensing," in *System of Systems Engineering (SoSE), 2013 8th International Conference on*, pp. 297–301, June 2013.

[10] M. T. Nguyen and K. A. Teague, "Compressive sensing based random walk routing in wireless sensor networks," *Ad Hoc Networks*, vol. 54, pp. 99 – 110, 2017.

[11] M. T. Nguyen and N. Rahnavard, "Cluster-based energy-efficient data collection in wireless sensor networks utilizing compressive sensing," in *Military Communications Conference, MILCOM 2013 - 2013 IEEE*, pp. 1708–1713, Nov 2013.

[12] M. T. Nguyen and K. Teague, "Compressive sensing based data gathering in clustered wireless sensor networks," in *Distributed Computing in Sensor Systems (DCOSS), 2014 IEEE International Conference on*, pp. 187–192, May 2014.

[13] C. Luo, F. Wu, J. Sun, and C. W. Chen, "Efficient measurement generation and pervasive sparsity for compressive data gathering," *Wireless Communications, IEEE Transactions on*, vol. 9, pp. 3728–3738, December 2010.

[14] M. T. Nguyen and K. Teague, "Tree-based energy-efficient data gathering in wireless sensor networks deploying compressive sensing," in *Wireless and Optical Communication Conference (WOCC), 2014 23rd*, pp. 1–6, May 2014.

[15] H. Wang, Y. Zhu, and Q. Zhang, "Compressive sensing based monitoring with vehicular networks," in *INFOCOM, 2013 Proceedings IEEE*, pp. 2823–2831, April 2013.

[16] Y. Mostofi, "Compressive cooperative sensing and mapping in mobile networks," *Mobile Computing, IEEE Transactions on*, vol. 10, pp. 1769–1784, Dec 2011.

[17] Y. Mostofi, "Cooperative wireless-based obstacle/object mapping and see-through capabilities in robotic networks," *Mobile Computing, IEEE Transactions on*, vol. 12, pp. 817–829, May 2013.

[18] S. Huang and J. Tan, "Adaptive sampling using mobile sensor networks," in *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pp. 657–662, May 2012.

[19] M. T. Nguyen, H. M. La, and K. A. Teague, "Compressive and collaborative mobile sensing for scalar field mapping in robotic networks," in *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 873–880, Sept 2015.

[20] M. T. Nguyen and K. A. Teague, "Compressive and cooperative sensing in distributed mobile sensor networks," in *Military Communications Conference, MILCOM 2015 - 2015 IEEE*, pp. 1033–1038, Oct 2015.

[21] M. T. Nguyen, K. A. Teague, and S. Bui, "Compressive wireless mobile sensing for data collection in sensor networks," in *2016 International Conference on Advanced Technologies for Communications (ATC)*, pp. 437–441, Oct 2016.

[22] R. Berinde and P. Indyk, "Sparse recovery using sparse random matrices," MIT-CSAIL Technical Report, 2008.

[23] R. Berinde, A. C. Gilbert, P. Indyk, H. J. Karloff, and M. J. Strauss, "Combining geometry and combinatorics: A unified approach to sparse signal recovery," *CoRR*, vol. abs/0804.4666, 2008.

[24] T. S. Rappaport, *Wireless Communications: Principles and Practice (2nd Edition)*. Prentice Hall, 2 ed., Jan. 2002.

[25] http://lcav.epfl.ch/op/edit/sensorscope en.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 2, No. 3, 254-267 (2017)
www.astesj.com
Special Issue on Recent Advances in Engineering Systems

ASTES Journal
ISSN: 2415-6698

Validity and efficiency of conformal anomaly detection on big distributed data

Ilia Nouretdinov[1,2]

[1]*Information Security Group, RHUL, TW20 OEX, UK*

[2]*Computer Learning Research Center, Department of Computer Science, RHUL, TW20 OEX, UK*

ARTICLEINFO

Article history:
Received: 31 March, 2017
Accepted: 1 May, 2017
On-line: 10 May, 2017

Keywords:
Conformal prediction
Anomaly detection
Distributed computing

ABSTRACT

Conformal Prediction is a recently developed framework for reliable confident predictions. In this work we discuss its possible application to big data coming from different, possibly heterogeneous data sources. On example of anomaly detection problem, we study the question of saving validity of Conformal Prediction in this case. We show that the straight forward averaging approach is invalid, while its easy alternative of maximizing is not very efficient because of its conservativeness. We propose the third compromised approach that is valid, but much less conservative. It is supported by both theoretical justification and experimental results in the area of energy engineering.

1 Introduction

1.1 Topic of the work

This paper is an extension of [1]. The topic of this work is application of **Conformal Prediction** (CP) framework of reliable machine learning to big data. Conformal Prediction was developed in such works as [2, 3]. Its main advantage is producing prediction sets that are valid in weak (i.i.d./exchangeability) assumptions. In the case of supervised learning (classification or regression) the idea of CP is in checking each hypothesis about the new example's label on exchangeability. This means that the exchangeability of the whole data sequence extended by this example is checked.

This has an interpretation in terms of **anomaly detection**: whether the example (supplied with a hypothetical label) is an anomaly with respect to the training ones, or not. As we will discuss, the most questions related to CP can be modelled for the task of anomaly detection first, and then they will be naturally extended to supervised learning or clustering. Therefore in this work we concentrate on unsupervised task of anomaly detection of unlabelled instances. Amongst the previous works on conformal anomaly detection we have to mention[4, 5, 6, 7, 8, 9].

These works did not specifically study the problem of **big data**. Here we model the situation when the data set is so big that CP algorithm can not be calculated directly, the data set has to split into parts. Furthermore, we assume that different parts of the data are collected in different places ('sources'), and these sources are in general case heterogeneous (different in their distribution), and only the mixture of all these distributions is the 'true' distribution that might generate a testing example. Extending Conformal Prediction for this task is not straightforward because of the risks to lose the validity and to informative efficiency of the predictions.

1.2 Application area

Anomaly detection problem can be applied in may areas. In the context of conformal framework the areas mostly considered earlier were vessel trajectories [4, 6, 7] and traces of bots [9].

As a model example, we use a question form the area of energy engineering. Energy Demand Research Project (EDRP) provides the data on customer behaviour (energy consumption) of gas/electricity using households. A data instance here is a summary (profile) of behaviour consumption during a large observation time for an individual household.

This is a kind of a more general problem where the detection of anomalies is understood as an unusual or suspicious behaviour of a complex system or its human users.

[1]i.r.nouretdinov@rhul.ac.uk

https://dx.doi.org/10.25046/aj020335

1.3 Plan of the paper

In the background Section 2 we remind the basic notions of Conformal Prediction, with concentration on validity properties. We discuss both supervised learning and unsupervised learning (anomaly detection) tasks and their connection to each other in more details than in our previous work [1].

In particular, we pay special attention to the choice of Conformity Measure (CM) that is the core element of a CP algorithm. It may be based on any machine learning algorithm, and our preference is k Nearest Neighbours (kNN) approach.

In Section 3 we address the challenge caused by big data split to validity and informative efficiency of Conformal Prediction. Theoretical justification of the suggested solution is extended from [1] with an informal motivating example showing its principal idea.

Sections 4 describes organisation of experiments, which aim to check how the idea works on a set of real data. Here we mostly deal with the questions of computational efficiency of experiments.

Section 5 includes brief description of the data set and its splits, experimental output and its analysis. There are three principal ways of data splits, different in the level of modelled homogeneity/heterogeneity. The part related to the analysis is expanded compared to [1], including more complete series of experiments (using versions of kNN CM with several values of k) and their analysis in several aspects, including influence of different factors.

The paper is concluded with a discussion section including plans for the future work. The most interesting of them is related to potential supervised learning applications.

2 Machine learning background

2.1 Some machine learning problems

In general, the task of machine learning is to say (predict some property) of a new instance z_{l+1} based on existing information about the previous (training) instances $z_1,\dots,z_l$ of the same nature. Learning means that the quality (efficiency) of the predictions grows as the experience (the number of examples analysed during the training) grows. Usually an instance is represented as a vector of features and/or labels.

The most popular machine learning problem is *supervised classification*. Assume that each instance $z_i = (x_i, y_i)$ consists of an object (feature vector) x and a label (usually a scalar) y_i. Everything is known except the label of a new example. So the task is to predict a label for a example x_{l+1} from a given training set of object $x_1, x_2, \dots, x_l \in X$ supplied with labels $y_1, y_2, \dots, y_l \in Y$.

Another large area is unsupervised learning without any labels when there are no labels and $z_i = x_i$. This includes such task as clustering and anomaly detection. The clustering question is to assign some labels to all the objects, based on their similarity to each other. For a new example this will also give a classification, which of the clusters it is likely to belong to. The question of anomaly detection is to detect whether the example is similar to any group of training examples at all or it is a sort of abnormal one. Anomaly detection and clustering tasks together or independently on each other. It can be said that the anomalies are object that remain after clustering of the typical objects while anomaly is a thing such that excluding it from the data set cleans the data makes the clustering task easier. See [9] for more discussion of their connection.

Classification an instance as an anomaly may have several meanings according to the nature of a problem: new tendency in a data stream, some mistakes in data collection, suspicious behaviour of an object or human, reflected in the data. If a normal instance is classified as anomaly, this is a sort of 'false alarm' although may reflect some unusual property of it. Another kind of error is missing a true anomaly (classifying it as a normal instance).

Normally it is better to keep false alarm rate on a decided significance level, and to try decrease the number of second type error, making the method as sensitive for real anomalies as possible. Limitation on the false alarm level allows to talk about reliability of the anomaly detection because otherwise classifying an instance as an anomaly is not a responsible claim.

A reason why we focus our interest in this work on anomaly detection is that the problem of anomaly detection covers the problem of supervised learning in the following sense. If we try to predict a label y_{l+1} for a new example x_{l+1}, this also means that we would like to fill a gap in the data not in an abnormal way. If $(x_{l+1}, \hat{y})$ looks like an anomaly compared to the training examples $(x_1, y_1), \dots, (x_l, y_l)$, then it is unlikely that $y_{l+1} = \hat{y}$ where y_{l+1} is the true label. In the context of supervised learning, the usual understanding of anomaly is the following. A labelled example (a pair (x_i, y_i)) is anomalous as far as y_i is different from the label predicted for x_i by an (underlying) machine learning algorithm.

2.2 Conformal Prediction framework

Conformal Prediction [2] is a framework that can convert practically any machine learning algorithm for classification or regression into a reliable multi-prediction algorithm that produces prediction sets valid in weak (i.i.d.(power) or exchangeability) assumptions about the data generation mechanism.

Conformal anomaly detection was introduced in [4]. Its principal scheme is given by the equations:

$$\alpha_i = A(z_i, \{z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_{l+1}\}). \tag{1}$$

$$p(z_{l+1}) = p(z_1, \dots, z_{l+1}) = \frac{card\{i = 1, \dots, l+1 : \alpha_i \le \alpha_{l+1}\}}{l+1} \tag{2}$$

The principal parameter of the algorithm is a *Conformity Measure (CM) A* that is a measure of information distance of an object z and a set U. Its choice

affects the efficiency of the algorithm which we will discuss further. Then, p-values measures applicability of i.i.d. assumption to the data by testing whether $z_1,\dots,z_l,z_{l+1}$ are likely to be generated by the same distribution, or this is discarded by the last example z_{l+1} being an anomaly.

In other literature such as [2, 6] there is also used the term *Non-Conformity Measure (NCM)* which differs from Conformity Measure in the sign, and correspondingly $\leq$ is replaced by $\geq$ in the equation 2. Usually the CM connects Conformal Prediction framework to one of standard machine learning algorithms such as Support Vector Machines (SVM), Neural Nets, Nearest Neighbours, that is called the *underlying algorithm*.

In the case of supervised learning, p-value is assigned to each possible hypothesis $y \in Y$ about the label of the new object

$$p(y) = p(z_1,\dots,z_l,(x_{l+1},y)).$$

In particular, p-value assigned to the true label y_{l+1} measures exactly the abnormality of the example

$$z_{l+1} = (x_{l+1},y_{l+1}).$$

The main advantage of Conformal Prediction is ability to produce *prediction sets* with guaranteed properties of validity in i.i.d. assumption.

For supervised learning, the prediction set is the set of y such that $p(y)$ is larger than a selected significance level ε. It is proven [2] to cover the true label y_{l+1} with probability at least $1-\varepsilon$.

In anomaly detection, it is possible to talk about prediction set of all z such that $p(z) = p(z_1,\dots,z_l) > \varepsilon$. It is also guaranteed to cover true z_{l+1} in case when it is generated by the same distribution, i.e. it is not an anomaly.

The size of the prediction set is the measure of efficiency: the smaller is the set the more informative is the prediction. It is practically useful that these predictions are individual, so for some examples they may be small (more informative) even if for the majority of testing examples the predictions are not so definite.

Below we will consider these notions in a more formal way.

2.3 Validity

The validity property states the following. If the sequence of examples $z_1,\dots,z_l,z_{l+1}$ is generated by an exchangeable distribution then the probability that $p = p(z_{l+1}) < \varepsilon$ is at most ε. for any significance level ε. An exchangeable distribution means invariance on the permutation of the order of examples. This includes the important case of i.i.d. (power) distribution meaning that the data examples were generated independently by an identical mechanism.

The standard output of Conformal Prediction is in the form of prediction set, where the validity has the meaning that the prediction set is 'large enough' in a statistical sense.

The prediction set in unsupervised task is defined as:

$$R^\varepsilon = \{z \in Z : p(z) = p(z_1,\dots,z_l,z) > \varepsilon\}$$

Validity implies that R^ε covers z_{l+1} with probability at least $1-\varepsilon$. So z_{l+1} can be reported as an anomaly if $p(z_{l+1}) < \varepsilon$ because the probability of this event is below the selected significance level ε. So the validity property actually allows to set a desired bound on the false alarm rate (number of mistakes within normal examples).

In supervised task, the sequence $z_{l+1} = (x_{l+1},y_{l+1})$ is known partially, the task is to predict y_{l+1} by x_{l+1}. The prediction set

$$R^\varepsilon = \{y \in Y : p(y) = p\,(z_1,\dots,z_l,(x_{l+1},y)) > \varepsilon\}.$$

Validity implies that this R^ε covers the true label y_{l+1} with probability at least $1-\varepsilon$. This makes the prediction set reliable: probability of error is limited by ε, if error is understood as a true label left outside the prediction set.

2.4 Informative efficiency

In both supervised and unsupervised tasks informative efficiency is related to size of a prediction set. The smaller it is, the more informative is the prediction.

In supervised case, it means that more hypotheses about the label of a new data instance are rejected, therefore the prediction is more definite. In particular, the prediction of size 1 is called *certain*, and the complement to 1 of the smallest significance level for which this is true is called *individual confidence* of the prediction. This notion is applicable directly to conformal anomaly detection because the predictions test are usually infinite.

In unsupervised case, informativeness means higher sensitivity to abnormality, i.e. classifying more examples as anomalies, (remind that the prediction set includes the instances classified as normals).

In more details, efficiency of supervised learning were discussed in [3] while efficiency of conformal anomaly detection is discussed in detail in [6, 8]. Some ways to measure efficiency numerically were presented in these works.

It this paper we will concentrate just on one aspect: harm to efficiency caused by parallelisation of the framework for big data.

2.5 Deviations from validity

If the formula 2 of p-value calculation is modified or generalised, the first challenge is to keep its validity property, otherwise the outputs of Conformal Prediction would not be reliable.

To measure validity empirically, one has to see how much the empirical distribution of p-values deviates from the uniform distribution on [0,1]. A martingale on-line approach to such testing was presented

in [10] but its focus was on checking exchangeability of data sets.

Now we are focusing on testing the algorithms instead of data sets. The difference form martingale is that for us two kinds of deviations are principally different for us now: if p-values are smaller than expected this leads to invalidity while having too large p-values is a less harmful case of low efficiency.

We consider the probability that $p \le \varepsilon$. Ideally, it should be very close to ε, and this is reached in online smooth version of the standard Conformal Prediction [2].

In general, for a fixed significance level ε there are three practical possibilities:

- Invalidity: $Prob\{p \le \varepsilon\} > \varepsilon$, p-value is not a correct measure of significance.
- Exact validity: $Prob\{p \le \varepsilon\} = \varepsilon$, an ideal situation;
- Conservative validity: $Prob\{p \le \varepsilon\} < \varepsilon$, a cause of undesirable damage to efficiency.

If probability of this event is significantly larger than ε, this means that there is no actual reliability of the results. If this probability is essentially smaller than ε this means the results are reliable but conservative.

Conservativeness is an indirect but important indicator that prediction sets (of non-anomalous objects) are unnecessarily large. If the allowed level ε of normal examples classified as anomalies is not reached, this decreases false alarm level somewhere below the allowed threshold, but it is very possible that some true anomalies are also left undetected.

2.6 Conservativeness as lack of informative efficiency

Note that conservativeness is just one of the possible causes of low information efficiency (too large prediction sets). Another source of damage to informative efficiency than a weak choice of the Conformity Measure (CM) mentioned before.

Studying influence of CM on the efficiency of the prediction is a more complex question. There are no definite answer in the case of anomaly detection, which CM is better, without taking into account the nature of the problem. It may be reflected in another testing set consisting specially of anomalies, or in defining a measure on the object space. See [6] for an example of such comparison.

Another reasons to choice one or another CM may be related to cluster structure of the prediction set (number, purity of clusters), see [9] for details.

Therefore we concentrate on conservativeness now leaving these questions aside this work.

2.7 Assessment of validity and conservativeness

Above we discussed validity and efficiency related to a specific significance level ε. It is more convenient to have a joint parameterless (ε-independent) measure of validity and conservativeness, aggregating information about different levels.

Let us observe the distribution of p-values assigned to testing examples. Ideally (in the case of exact validity) it should be uniform. It was formally proven to be for smoothed version of non-parallelised Conformal Prediction [2]. Smoothing here means assigning a random weight to the cases $\alpha_i = \alpha_{l+1}$ instead of 1. In this work we do not do special smoothing. For the big data size there is usually no practical difference between smoothed and non-smoothed versions, so it is convenient to get rid of this non-deterministic element.

We would like here to assess practically how well this closeness to uniformity is achieved by empirical distribution of p-values of testing examples.

As the first criterion, we measure Average p-Value (APV). Averaging of p-value can be understood as a well as a form of mixing prediction set sizes over different significance levels. It was suggested in [6] as an efficiency measure on a testing set or a measurable space of testing feature vectors. We put it now in a different context: in our check, the testing examples come from the same distribution as the training ones (unlike [6] with a testing set intentionally generated another way). Therefore we know that valid p-values are distributed uniformly, the expected average value of APV is $\frac{1}{2}$.

APV criterion may be considered as a uniform averaging of ε-dependent validity checks by ε. This may be not the best way because in practice only its small values are usually important. Significance levels typically used in statistics are no larger than 0.05, sometimes reaching 10^{-2} or 10^{-4}. Therefore we also consider a modified criterion that is Average Logarithm p-Value (ALPV). It was suggested in [8] as an alternative to APV, applicable in same circumstances. This criterion gives larger weight to small significance levels.For example, if the p-values tend to concentrate around $\frac{1}{2}$ this is not reflected in APV but considered conservative by ALPV. Its expected value of -1 for uniform distribution.

2.8 Selection of Conformity Measure

In general, it is expected that CM approximates the density function of the data distribution. In [3] this principle was justified for supervised classification but can be easily adopted to anomaly detection as well. Further in Sec. 3.8 we will need this property as well.

The problem of CM selection in the context of conformal anomaly detection (not classification) was studied in the paper [6]. In [6] two families of the density approximating functions were compared by their

efficiency.

- A Conformity Measure can be based on kernel density estimation (KDE) by creating a continuous function from the empirical data distribution; a kernel function has to be chosen for smoothing it. Typically the kernel function is the density function of a normal distribution.

- Alternatively, local density at a point can be approximated by nearest neighbours method that shows how close a point is in average to some amount of its nearest neighbours in the data set. Conformity of an example is calculated as the inverse value of the average distance to k nearest neighbours (kNN).

Both methods are actually parametric: either the kernel function needs to be supplied with variance ('degree of smoothing') , or the number of neighbours has to be selected. Experimental comparison in [6] have shown that with the best parameters both methods give the results of similar quality.

However, in the case of k nearest neighbours the method is more robust: the damage to efficiency caused by its imprecise selection is smaller. Therefore we consider nearest neighbours method as the preferable one.

Another intuitive reason is that in context of comparing conformities kNN $k = 1$ may be considered as a 'limit' case of density estimation as the variance parameter of the kernel fucntion tends to 0. At the same time, kNN with $k > 1$ can not be easily explained in terms of kernel density estimation. So this family is in some sense a 'richer' one.

3 Aggregating data sources

3.1 Big data challenges

Processing big data set a limitation caused by limited memory or time does usually appear. This happens because the required time/memory increases in a non-linear way, therefore splitting the data into some parts is the only way to fit it. Also there may be physical or technical reasons requiring parallelisation such as impossibility to collect all the data in the same place simultaneously. Usually we can just assume that, for a concrete algorithm, an upper limit on the data set size processable at once, is somehow determined by the power of one processor/storage, and can not be enlarged.

Parallelising of an algorithm may be exact (producing the same result) or approximate, which happens usually. Some algorithms of machine learning do have special modifications applicable to big data. such as Cascade version of Support Vector Machines [11] which is approximate as well. The principal challenge is that underlying SVM is using a matrix of inter-example similarities, so the load of memory is proportional to the square of the number of training examples, which puts an essential limit on the number of examples which may be processed together.

In our work we also model a situation when only approximate calculation is possible. In the context of conformal framework, the principal challenge is caused by its core detail, calculation of Conformity Measure function A. One of its two arguments is a whole set which may make calculations much complex. In the example which we prefer to consider, CM is based on k Nearest Neighbour algorithm, which also requires a storage of the distance matrix.

In the case of Conformal Prediction, we also wish to see that reliability of results, guaranteed by validity properties, is not affected by parallelisation of the algorithms.

3.2 Modelling of the challenge

In this consideration we try taking into account both causes for data split. The number of data examples processed at once is limited by a number h, and we also assume that data collection was done independently by different groups, and it even may be never stored together. Each of the collecting group may have its own specialising collecting data of a concrete subtype, or related to the collection place.

This means that:

1. in general case it is impossible to make a completely random split;

2. it would be unfair to restrict us just to one of these sources, even as an approximation.

The second note is important. In our modelling we assume that there is no possibility to avoid parallelisation challenge just by using one source of information and expectation that it is not a bad approximation (which in many cases happens in reality).

Formally, we assume that:

- the training set $U = \{z_1, \ldots, z_l\}$ is split into two parts of equal size called U_1 and U_2; this in general case is not always a random split;

- the true data distribution P is a mixture of P_1 and P_2,
$$P = \frac{P_1 + P_2}{2};$$

- the data source U_1 is randomly generated by P_1;

- the data source U_2 is randomly generated by P_2.

- the testing example z_{l+1} is generated by the mixed distribution P;

- simultaneous access to the sources U_1 and U_2 is impossible.

The task is to estimate the $p(z_{l+1})$ with respect to the union of U_1 and U_2 as the training example.

3.3 Averaging Tests for randomness (AT)

Assume that p_1 and p_2 are calculated for the same testing example but for different training sets. So Equation 2 is replaced by two:

$$p_1 = \frac{card\left\{i = 1,\dots,\frac{l}{2}, l+1 : A(z_i, U_1) \le A(z_{l+1}, U_1)\right\}}{\frac{l}{2}+1} \tag{3}$$

$$p_2 = \frac{card\left\{i = \frac{l}{2}+1,\dots,l, l+1 : A(z_i, U_2) \le A(z_{l+1}, U_2)\right\}}{\frac{l}{2}+1} \tag{4}$$

The straightforward way is to approximate p-value by the average of two:

$$p^{AT} = \frac{p_1 + p_2}{2}. \tag{5}$$

This method seems to be a natural way of distributed calculation of p-value but it has no guarantees of validity.

If the split is heterogeneous, then the problem of invalidity can appear as well. This can be shown on the following example. Assume that the sources are completely heterogeneous: P_1 and P_2 distributions having non-overlapping support sets. We can also imagine that they are so disjointed that each example generated by P_1 has the minimal possible p-value $\frac{1}{l+1}$ with respect to the training set generated by P_2 and vice versa. This will lead to averaged p-value of a new P-generated example being approximately uniformly distributed on $\left(0, \frac{1}{2}\right)$ instead of $(0,1)$ that is a strong form of invalidity[1].

3.4 Maximizing Tests for randomness (MT)

The easy way to achieve guaranteed validity is to take the maximum of 'partial' p-value instead of the average:

$$p^{MT} = \max\{p_1, p_2\} \tag{6}$$

It also has a natural meaning: a new example z is typical with respect to the union of two sources if is typical with respect to at least one of them.

However, it such approach is rough and p-values produced this way are likely to be very conservative, especially if the distributions are close to each other. For example, in completely homogeneous case, when each of p_1 and p_2 is lucky to be valid itself as an approximation of p, $\max\{p_1, p_2\}$ will be just larger than p_1 or p_2 of without any use for validity.

3.5 Maximizing Conformities (MC)

The suggested solution is to correct the maximising approach by moving the maximization procedure one step back. Instead of maximizing the p-values themselves, let us maximize the estimates of new example's conformity.

The aggregated p-value is defined by the following equations

$$p^{MC} = \tilde{p} = (\tilde{p}_1 + \tilde{p}_2)/2 \tag{7}$$

$$\tilde{p}_1 = \frac{card\left\{i = 1,\dots,\frac{l}{2}+1 : A(z_i, U_1\}) \le A(z, U_1, U_2)\right\}}{\frac{l}{2}+1} \tag{8}$$

$$\tilde{p}_2 = \frac{card\left\{i = 1,\dots,\frac{l}{2}+1 : A(z_i, U_2\}) \le A(z, U_1, U_2)\right\}}{\frac{l}{2}+1} \tag{9}$$

where

$$A(z, U_1, U_2) = \max\{A(z, U_1), A(z, U_2)\}.$$

A theoretical justification of this approach and is presented further, including a more general from of several sources.

For practical applications, we have to not that this method is not memory-consuming, because it is enough to store conformity scores, not the whole data examples. Also, there is no need to keep them in the original order, this can be replaced by storage of their overall distribution.

3.6 Motivating example

Consider now the following space:

$$X = \{1, 2, 3, 4, 5, 6\}$$

There are two big data sources U_1 and U_2 of equal weight with the following data distributions:

$$P_1 = (0.2, 0.6, 0.2, 0.0, 0.0, 0.0)$$

$$P_2 = (0.0, 0.0, 0.0, 0.2, 0.4, 0.4)$$

The whole distribution is:

$$P = (0.1, 0.3, 0.1, 0.1, 0.2, 0.2)$$

Assume that the new example is $x_{n+1} = z = 3$. If the Conformity Measure corresponds to the local density, the p-values with respect to U_1 and U_2 are

$$p_1 = P_1\{x : A(x, U_1) \le A(z, U_1)\} = P_1\{x : P_1(x) \le P_1(z)\}$$

$$= 0.2 + 0.2 = 0.4$$

$$p_2 = P_2\{x : A(x, U_2) \le A(z, U_2)\} = P_2\{x : P_1(x) \le P_2(z)\} = 0$$

and the true p-value with respect to the whole data set is

$$p = P\{x : P(x) \le P(z)\} = 0.1 + 0.1 + 0.1 = 0.3$$

while its straightforward approximation is

$$\hat{p} = (p_1 + p_2)/2 = 0.2$$

which may lead to a falsely confident rejection of the hypothesis that the new example is not an anomaly.

[1] In particular, it is limited from above by

$$\frac{1 + \frac{1}{l+1}}{2} = \frac{1}{2} + \frac{1}{2l+2}$$

which definitely contradicts the validity for any $\epsilon > \frac{1}{2} + \frac{1}{2l+2}$

3.7 Idea of Correction

How to correct this? The idea is to replace the way of approximation as follows:

$$\tilde{p} = (\tilde{p}_1 + \tilde{p}_2)/2$$

where

$$\tilde{p}_1 = P_1\{x : A(x, U_1) \le A(z, U_1, U_2)\}$$

$$\tilde{p}_2 = P_2\{x : A(x, U_2) \le A(z, U_1, U_2)\}$$

and $A(z, U_1, U_2)$ means the largest of $A(z, U_1)$ and $A(z, U_2)$.

In the example above it would work as follows:

$$A(z, U_1, U_2) = \max\{0.2, 0.0\} = 0.2$$

$$\tilde{p}_1 = 0.2 + 0.2 = 0.4$$

$$\tilde{p}_2 = 0 + 0 + 0 + 0.2 = 0.2$$

$$\tilde{p} = (0.4 + 0.2)/2 = 0.3$$

which is a correct approximation.

3.8 Justification of MC approach

In this section we assume that data U comes from sources $U_1, \dots, U_k$ with weights (contribution percentages) $w_1, \dots, w_k$.

Inspired by the logic of the work [3], we consider true data distribution as a limit (asymptotical) case of the empirical distribution, and the local density as a limit case of (the optimal) conformal density. We are analysing some asymptotical tendency, assuming that the amount of data in each of the sources is representative enough, so the difference between true and empirical distribution is low enough.

We assume that the 'full' distribution is split into sources by weighted formula:

$$P = \frac{w_1 P_1 + \dots + w_k P_k}{w_1 + \dots + w_k}$$

The data in a source U_i is generated by an i.i.d. distribution P_i^* (the star means a power distribution here), while testing examples are generated by P.

We also assume that conformity score $A(x, U_i)$ is an 'optimal' one (in the sense of [3]) i.e. it is an equivalent of the density function $P_i(x)$. Note that CM are equivalent if they are reducible to each other by monotonic transformation. Therefore such methods as k Nearest Neighbours and Kernel Density Estimation are asymptotically suiTable because they are originally created in order to approximate a monotonic transformation of the local density function.

Proposition 1. Assume that the space X is discrete (finite). Each bag U_i is big and representative for P_i and the Conformity Measure $A(x, U)$ is equivalent to the local density so the difference between P_i and the uniform distribution on U_i is negligible. Then a valid p-value is

$$\tilde{p}(z) = \sum_{j=1}^{k} w_j \tilde{p}_j(z)$$

where

$$\tilde{p}_j(z) = P_j\left\{x : w_j P_j(x) \le \max_{i=1}^{k}\{w_i P_i(z)\}\right\}.$$

3.9 Proof of Proposition 1

We will show that this $\tilde{p}(z)$ can be obtained as the resulting p-value of another Conformal Predictor.

We can sssume that an example x is generated in two steps: first, $i(x) \in \{1, \dots, k\}$ is generated according to the distribution $W = (w_1, \dots, w_k)$, then x itself is generated by $P_{i(x)}$.

Set the Conformity Measure to

$$A((i, x), U) = w_i P_i(x)$$

(using our assumption that P_i is recoverable from U_i with required precision). In this case the corresponding conformal p-value is calculated as:

$$p((i, z)) = \sum_{j=1}^{k} w_j P_j\left\{x : w_j P_j(x) \le w_i P_i(z))\right\}$$

$$\le \sum_{j=1}^{k} w_j P_j\left\{x : w_j P_j(x) \le \max_{i=1}^{k}\{w_i P_i(z)\}\right\} = \tilde{p}(z)$$

The last estimate does not depend on i and can be used as a valid p-value for z.

4 Experimental settings

4.1 Data collection

The data are taken from Energy Demand Research Project (EDRP)[2]

representing behaviour (energy consumption) by the customers (households).

We use data for 8,703 households using electricity and gas. For each of the household some social parameters (such as number of rooms, tenants, level of income) are provided. They are summarized by including each of the households into one of 6 Acorn categories. In a part of our experiments we will use these Acorn categories as a way of data split, modelling coming the data from heterogeneous sources.

[2] The data were provided by: AECOM Building Engineering, Energy Demand Research Project: Early Smart Meter Trials, 2007-2010. Colchester, Essex: UK Data Archive [distributor], November 2014. SN: 7591.
Energy Demand Research Project: Early Smart Meter Trials, 2007-2010, UKDA study number:7591.
Principal Investigator: AECOM Building Engineering
Data Collector: Centre for Sustainable Energy.
Sponsor: Department of Energy and Climate Change.
Distributed by: UK Data Archive, University of Essex, Colchester.
November 2014.
http://doc.ukdataservice.ac.uk/doc/7591/mrdoc/UKDA/UKDA_Study_7591_Information.htm

Size	Split	CM		AT	MT	MC
1000	Random	$k = 300$	APV	0.500	0.509	0.511
	Random	kNN	ALPV	-0.992	-0.964	-0.965
1000	Random	$k = 100$	APV	0.500	0.513	0.512
	Random	kNN	ALPV	-0.993	-0.958	-0.971
1000	Random	$k = 30$	APV	0.500	0.520	0.519
	Random	kNN	ALPV	-0.993	-0.944	-0.953
1000	Random	$k = 10$	APV	0.500	0.530	0.529
	Random	kNN	ALPV	-0.989	-0.918	-0.926

Table 1: Results: random (homogeneous) split

The data set provides the consumption data in two forms: detailed information about energy usage in each half-hour during several months, and a brief summary (so called metadata). For the aims of this work, we use only the second representation. 8 numerical features were taken from the metadata summary. They are listed below:

1. Inclusive number of days between first and last advances
2. Lowest advance value
3. Highest advance value
4. Average advance value
5. Number of advances expected based on feature 1
6. Number of available advances
7. Ratio of features 5 and 6
8. Time of use identifier

We will use these features for learning, and in some experiments for data splitting as well.

These features are preprocessed by converting into logarithmic scale and standalising (linear rescaling to equal average and variance).

4.2 Data split

We consider three ways of splitting the data, wish the level of heterogeneity step wisely increasing:

1. Random (or homogeneous) split (Table 1).
2. Split by Acorn category (Table 2).
3. Split by one of the features, using its median value as a threshold (Table 3). [3]

The last way of splitting definitely introduces heterogeneity by splitting the object vector space into two parts. As for Acorn categories, they are expected to have different distributions, but the level of their overlapping is not known initially.

The original sizes of Acorn categories are: 1)2982 2)370 3)2855 4)1162 5)1316 6)18. Therefore some of them are not used, others taken together to have enough number of examples.

4.3 Inductive mode

To make calculations more efficient for big data, we apply so called *Inductive Conformal Predictor* (ICP) [2]. Its scheme is given in Table 4. A part of data called *proper training set* is left aside the learning. It is used only in calculation conformities of the remaining *calibration* examples. In an equivalent interpretation, inductive mode may be considered as a sort of the standard (Transductive) CP where proper training set is a sort of algorithm parameter, while the Conformity Measure depends only on the example itself. However we still may need the limitation of its size. In this version, the most time-consuming set is calculation of the distance matrix, all the others are relatively small.

4.4 Distributed computing details

In our experiments we assume that the data set is split into two sources of equal size. Each of them may have its own distribution, however we assume within one source the order of examples is random. A number h is the limit on the training set size. For convenience, we assume that this number is the limitation both for the proper training set size, for the calibration set size, and for the number of testing examples taken from each of the sources. This setting is summarised in Table 5.

We compare three approaches of aggregation:

1. Averaging Tests (AT), Sec. 3.3;
2. Maximizing Tests (MT), Sec. 3.4;
3. Maximizing Conformities (MC), Sec. 3.5.

In each of them, Inductive Confidence Machine (as presented in Table 4) has to be run twice for each of the testing examples.

[3] Splits by features 2 and 8 are not included because of impossibility of doing split the same way. This is due to their distribution (too many equal values).

Size	Split	CM		AT	MT	MC
500	Acorn category 1	$k = 300$	APV	0.494	0.524	0.525
	Acorn category 3	kNN	ALPV	-0.992	-0.929	-0.942
500	Acorn category 1	$k = 100$	APV	0.495	0.524	0.524
	Acorn category 3	kNN	ALPV	-0.991	-0.926	-0.935
500	Acorn category 1	$k = 30$	APV	0.491	0.532	0.530
	Acorn category 3	kNN	ALPV	-0.998	-0.914	-0.929
500	Acorn category 1	$k = 10$	APV	0.489	0.536	0.534
	Acorn category 3	kNN	ALPV	-1.004	-0.905	-0.917
1000	Acorn categories 1,5	$k = 300$	APV	0.498	0.516	0.517
	Acorn categories 3,4	kNN	ALPV	-0.997	-0.956	-0.957
1000	Acorn categories 1,5	$k = 100$	APV	0.496	0.520	0.519
	Acorn categories 3,4	kNN	ALPV	-1.001	-0.947	-0.961
1000	Acorn categories 1,5	$k = 30$	APV	0.495	0.525	0.523
	Acorn categories 3,4	kNN	ALPV	-1.002	-0.936	-0.947
1000	Acorn categories 1,5	$k = 10$	APV	0.494	0.532	0.531
	Acorn categories 3,4	kNN	ALPV	-0.999	-0.915	-0.924
1000	Acorn categories 1,4	$k = 300$	APV	0.495	0.522	0.522
	Acorn categories 3,5	kNN	ALPV	-1.000	-0.941	-0.947
1000	Acorn categories 1,4	$k = 100$	APV	0.492	0.525	0.523
	Acorn categories 3,5	kNN	ALPV	-1.005	-0.936	-0.950
1000	Acorn categories 1,4	$k = 30$	APV	0.491	0.528	0.527
	Acorn categories 3,5	kNN	ALPV	-1.006	-0.928	-0.938
1000	Acorn categories 1,4	$k = 10$	APV	0.491	0.533	0.532
	Acorn categories 3,5	kNN	ALPV	-1.004	-0.910	-0.918

Table 2: Results: split by categories

4.5 Conformity Measure

For calculation of conformities we use k Nearest Neighbours method (with a large number of neighbours). This choice was discussed earlier in Sections 2.8,3.8.

We try the following values of the parameter k: $10, 30, 100, 300$. As mentioned before, the parameter may reflect natural specific of the problem or an aim of anomaly detection, therefore we consider all of them as equally reasonable. Usually larger k in nearest neighbours anomaly detections reflects more 'centralised' understanding of what is normal (not anomalous) data behaviour, expectation to have only few typical patterns (clusters).

4.6 Measuring deviations from validity

For measuring deviations we use ϵ-free criteria earlier presented in Section 2.5. Table 6 gives a short summary of how they should be interpreted.

Note that validity in sense of this Table does not imply validity for any ε. It just means that invalidity is invisible with this way of its measurement. Therefore, if invalidity is shown just by one of two criteria (APV, ALPV), it is still detected. However, one may consider in this situation APV-invalidity more tolerable, expecting that it is related to rarely used significance levels ε.

Remind also that keeping validity is strictly the first priority. Conservativeness is undesirable but it does not affect the validity, so the goal is only in its minimisation where possible.

5 Experimental results

5.1 Output

The data was selected according to Table 5. All the subsets mentioned above are selected randomly, and the results are averaged over 50 random seeds.

If the 'size' is set to h (1000 or 500) this means that:

- the first ICP learns on h proper training and h calibration examples form the first source;
- the second ICP learns on h proper training and h calibration examples form the second source;
- h testing examples are taken from the first source, and h testing examples are taken from the second source;
- two ICPs assign p-values (p_1 and p_2 respectively) to each of $2h$ testing examples;
- these p-values are aggregated to p by one of three rules (AT, MT, MC).

The output of experiments is presented in Tables 1,2,3 and in Figure 1 in graphical form.

Below we analyse the experimental results from these Tables.

Size	Split	CM		AT	MT	MC
1000	Low feature 1	$k = 300$	APV	0.378	0.542	0.518
	High feature 1	kNN	ALPV	-1.215	-0.812	-0.948
1000	Low feature 1	$k = 100$	APV	0.363	0.528	0.510
	High feature 1	kNN	ALPV	-1.244	-0.840	-0.962
1000	Low feature 1	$k = 30$	APV	0.365	0.529	0.505
	High feature 1	kNN	ALPV	-1.255	-0.848	-0.976
1000	Low feature 1	$k = 10$	APV	0.352	0.522	0.503
	High feature 1	kNN	ALPV	-1.299	-0.879	-0.986
1000	Low feature 3	$k = 300$	APV	0.364	0.518	0.507
	High feature 3	kNN	ALPV	-1.253	-0.888	-0.980
1000	Low feature 3	$k = 100$	APV	0.354	0.512	0.501
	High feature 3	kNN	ALPV	-1.276	-0.902	-0.991
1000	Low feature 3	$k = 30$	APV	0.340	0.508	0.500
	High feature 3	kNN	ALPV	-1.320	-0.925	-0.989
1000	Low feature 3	$k = 10$	APV	0.318	0.504	0.500
	High feature 3	kNN	ALPV	-1.397	-0.954	-0.989
1000	Low feature 4	$k = 300$	APV	0.446	0.522	0.506
	High feature 4	kNN	ALPV	-1.090	-0.895	-0.971
1000	Low feature 4	$k = 100$	APV	0.411	0.510	0.502
	High feature 4	kNN	ALPV	-1.158	-0.924	-0.990
1000	Low feature 4	$k = 30$	APV	0.393	0.507	0.501
	High feature 4	kNN	ALPV	-1.210	-0.945	-0.987
1000	Low feature 4	$k = 10$	APV	0.369	0.503	0.501
	High feature 4	kNN	ALPV	-1.283	-0.971	-0.990
1000	Low feature 5	$k = 300$	APV	0.378	0.542	0.518
	High feature 5	kNN	ALPV	-1.215	-0.812	-0.948
1000	Low feature 5	$k = 100$	APV	0.363	0.528	0.510
	High feature 5	kNN	ALPV	-1.244	-0.840	-0.962
1000	Low feature 5	$k = 30$	APV	0.365	0.529	0.505
	High feature 5	kNN	ALPV	-1.255	-0.848	-0.976
1000	Low feature 5	$k = 10$	APV	0.352	0.522	0.503
	High feature 5	kNN	ALPV	-1.299	-0.879	-0.986
1000	Low feature 6	$k = 300$	APV	0.390	0.560	0.514
	High feature 6	kNN	ALPV	-1.189	-0.769	-0.952
1000	Low feature 6	$k = 100$	APV	0.360	0.535	0.509
	High feature 6	kNN	ALPV	-1.243	-0.815	-0.962
1000	Low feature 6	$k = 30$	APV	0.355	0.532	0.503
	High feature 6	kNN	ALPV	-1.279	-0.827	-0.986
1000	Low feature 6	$k = 10$	APV	0.342	0.524	0.502
	High feature 6	kNN	ALPV	-1.312	-0.856	-0.990
1000	Low feature 7	$k = 300$	APV	0.465	0.556	0.538
	High feature 7	kNN	ALPV	-1.065	-0.861	-0.914
1000	Low feature 7	$k = 100$	APV	0.436	0.520	0.532
	High feature 7	kNN	ALPV	-1.109	-0.922	-0.920
1000	Low feature 7	$k = 30$	APV	0.425	0.516	0.523
	High feature 7	kNN	ALPV	-1.147	-0.940	-0.934
1000	Low feature 7	$k = 10$	APV	0.415	0.515	0.517
	High feature 7	kNN	ALPV	-1.178	-0.942	-0.946

Table 3: Results: split by features

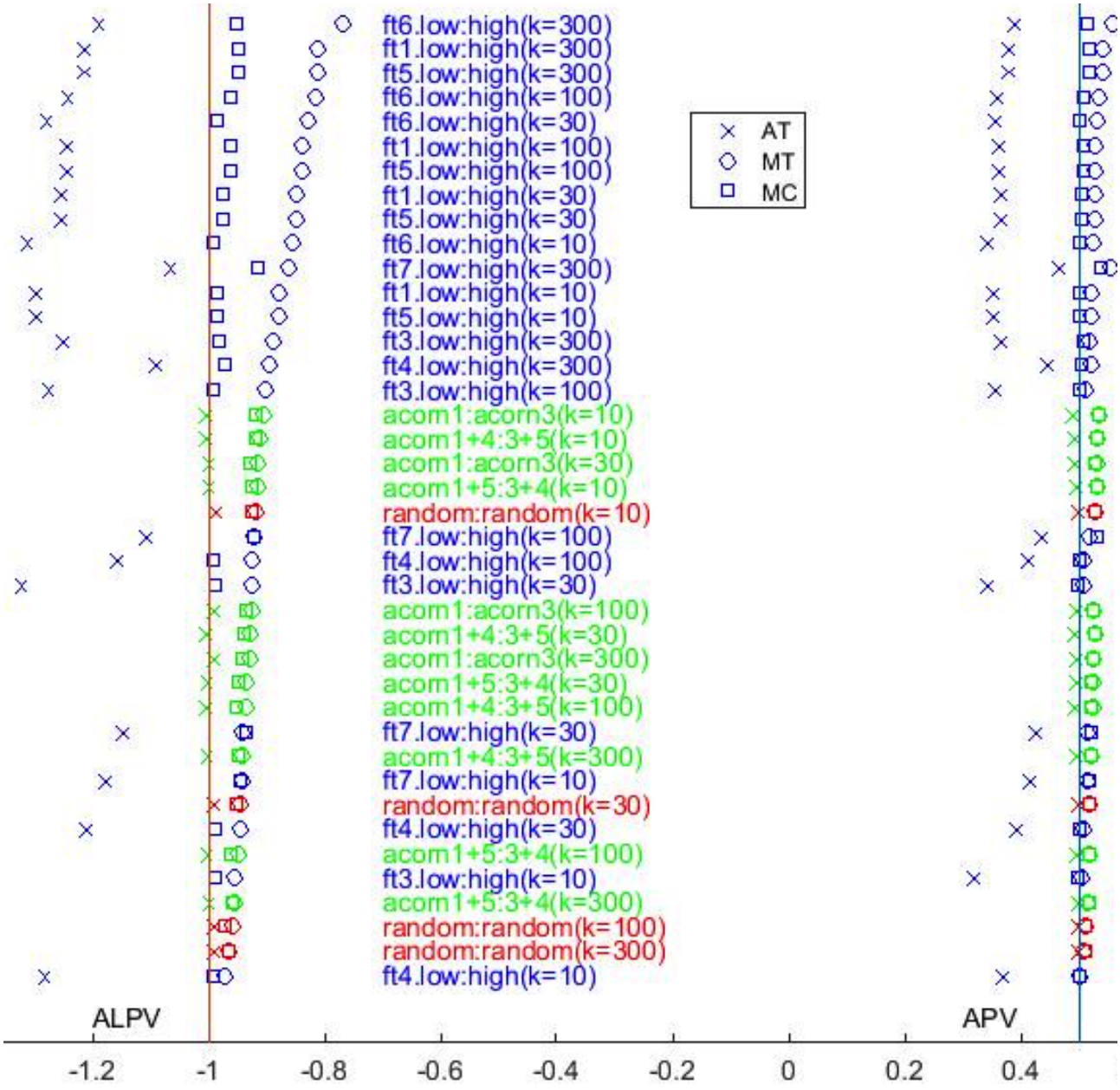


Figure 1: Chart sorted by ALPV for MT conservativeness excess

INPUT: proper training set $w_1,\dots,w_h$
INPUT: calibratlion set $z_1,\dots,z_l$
INPUT: testing example z_{l+1}
INPUT: number k of neighbours
FOR $i = 1,\dots,l+1$
FOR $j = 1,\dots,h$
$d_{ij} = |z_i - w_j|^2$
END FOR
let $d'_{i1},\dots,d'_{ih}$ be $d_{i1},\dots,d_{ih}$ sorted in ascending order
$\alpha_i = \frac{1}{d'_{i1}+\dots+d'_{ik}}$
END FOR
OUTPUT $p(z_{l+1}) = \frac{card\{i=1,\dots,l+1:\alpha_i\le\alpha_{l+1}\}}{l+1}$

Table 4: Scheme of ICP

Set	Size	Generating distribution
Proper training set 1	h	P_1^h
Calibration set 1	h	P_1^h
Proper training set 2	h	P_2^h
Calibration set 2	h	P_2^h
Testing set	$2h$	$\left(\frac{P_1+P_2}{2}\right)^{2h}$

Table 5: Experimental setting for ICP

5.2 Effect of data split

For homogeneous split (Table 1) AT approach is (up to the precision) exactly valid in terms of APV. However AT-ALPV in all cases is slightly smaller than -1, this still shows slight conservativeness concentrated for small levels of significance. This may be the topic of a future special study, concentrated on homogeneous case.

In the other experiments (Tables 2 and 3) we have to analyse invalidity effects for AT. Recall that invalidity means that AT-APVs are smaller than $\frac{1}{2}$ or AT-ALPVs are smaller than -1. In Table 3 these deviations are much larger, because these splits was done in a more radically way, forced to be heterogeneous, while in Table 2 there are more natural, practical splits.

Comparing the splits from Table 3 to each other, we can see that for some of the features (1,2,5) the deviations are larger than for the others. This indirectly shows the importance of these features in terms their high influence on the other features. Features (4,7) with relatively small deviation are more random and isolated from the others.

5.3 Effect of merging approach

As it is seen from Table 1, in case of a purely homogeneous (random) split the best way of aggregating p-values is just averaging them (AT), the two others are more conservative.

Two other Tables 2 and 3) contain 36 comparisons of AT, MT and MC approaches. They show is following:

- AT is invalid in all the experiments. This is caught by all 36 AT-APV being smaller than 0.5, and in 31 experiments is also confirmed by AT-ALPV being smaller than -1.
- MT and MC are conservatively valid in all the experiments.
- the conservativeness of MC is smaller than MT in 30 experiments, equal (up to the precision) in 2 experiments, and larger in 4 experiments (by APV); smaller in 34 experiments and larger in 2 (by ALPV).

5.4 APV and ALPV criteria

As discussed above, ALPV criterion gives preference to small values of the significance level, because such values are usually applied in the practice. So deviations according to ALPV are in principle more critical.

Let us summarise the points where the conclusions slightly differ for these two criteria.

- In homogeneous case, ALPV was more sensitive in catching conservativeness, while APV did not catch any deviation from exact validity.
- On the other hand, in catching invalidity (as seen in heterogeneous case) APV is stronger.
- Being more sensitive for conservativeness, ALPV better shows advantages of MC over MT in decreasing its level.

All this reflects that the damage from conservativeness is more concentrated in small significance levels, when CP is distributed for big data.

Average p-Value	Average Logarithm p-Value	Interpretation
$< \frac{1}{2}$	< -1	invalid
$= \frac{1}{2}$	$= -1$	valid and efficient
$> \frac{1}{2}$	> -1	valid but conservative

Table 6: Understanding the results of evaluation

5.5 Effect of CM choice

This work concentrates on the effects of conservativeness and does not aim to answer in general the question which CM is better for the specific problem of anomaly detection. As mentioned above, the full answer actually can be done based either on an additional testing set of known anomalies, or on a measure on the whole objects space, which may depend on the physical meaning of the data.

Here we just make preliminary notes about influence of the parameter k (number of nearest neighbours) on the conclusion. Usually the larger this number is, the more centered/unified is the notion of normal (not abnormal) behaviour, less splitter into separate behaviour types.

- In homogeneous case (Table 1), conservativeness (of AT) increases for the small values of neighbour number k.
- In heterogeneous case (Tables 2-3), validity (of AT) is also a larger problem for small k.
- Conservativeness of MT is larger for small k when the split is moderately heterogeneous (Table 2) and for large k when the split is more heterogeneous (Table 3).
- Advantage of replacing MT by MC is smaller for large k.

6 Discussion

6.1 Conclusion

In this work we suggest a MC way of aggregation of p-values for Conformal Prediction that is the best middle way between invalid averaging and conservative maximization of partial p-value obtained from different data sources. This method is applicable on big and heterogeneously split data.

It advantage was shown in Section 3 theoretically in an asymptotical sense, and confirmed on the experimental part on real data. The validity property is kept, while the conservativeness is essentially smaller than for aggregating p-values by maximizing them in most of the cases.

6.2 Future work

For the future work, we can mention the following points.

First, the general case presented in Proposition 1 (see Section 3.8) is applicable as well to a more general case of several sources of different size, this also is interesting to be checked practically.

Second, the area of application used in this paper was detection of anomalies behaviour of households. An interesting related area where these ideas may be applied in the future is growing area of internet of things such as medical sensors. Anomaly behaviour of users is an actual thing to be detected. The practice shows that the alarm/alert criteria have to be based on the analysis of essential data collections, otherwise they may be too approximate in a specific patient's situation.

Third, it was discussed in the work that this framework is transferrable from anomaly detection to supervised learning where validity of prediction sets is important as well. This opens many more applications. But there is a challenge of the following type. In a supervised learning task, the validity and conservativeness is determined by p-values assigned to *true* hypotheses about the labels of new data examples. It can be measure by analogy to this paper, but what about hypotheses assigned to *wrong* hypotheses? The conformal supervised learning is the more efficient the smaller are these p-values. The best situation is when p-values for all hypotheses except the true one are small (close to 0). This property may be also affected on the aggregated stage, where the p-values tend to increase, so theoretical and/or experimental validation has to ensure that it is not lost.

Conflict of Interest Authors declare no conflict of interest.

Acknowledgments This work was supported by Technology Integrated Health Management (TIHM) project awarded to the School of Mathematics and Information Security at Royal Holloway as part of an initiative by NHS England supported by InnovateUK. It was also supported by European Union grant 671555 ("ExCAPE"), EPSRC grant EP/K033344/1 ("Mining the Network Behaviour of Bots"), Thales grant ("Development of automated methods for detection of anomalous behaviour"), and by the National Natural Science Foundation of China (No.61128003) grant.

The large help for this work was done to Lars Carlsson. He gave us the initial motivation for the work, stating the problem of conservativeness in the case of homogeneous split. Another motivation came from the area of aggregating results of different pre-

diction algorithms. In particular it was inspired by recent works [12] with application of CP to big data. and [13] covering the theme of merging data splits as well as merging prediction algorithms. We are grateful to Paolo Toccaceli, Vladimir Vovk, Alex Gammerman and Zhiyuan Luo for useful discussions in this area.

TIHM project directed our attention to anomalies in behaviour-related practical problems as an area of application. We are especially grateful to Energy Demand Research Project for providing the data.

References

[1] I. Nouretdinov. "Distributed conformal anomaly detection" in 15th IEEE International Conference on Machine Learning and Applications, Anaheim USA, 2016.

[2] V. Vovk, A. Gammerman, G. Shafer. "Algorithmic learning in a random world" Springer, 2005

[3] V. Vovk, V. Fedorova, I. Nouretdinov, A. Gammerman. "Criteria of efficiency for conformal prediction" in Conformal and Probabilistic Prediction with Applications: 5th International Symposium, Madrid Spain, 2016.

[4] R. Laxhammar, G. Falkman. "Sequential conformal anomaly detection in trajectories based on Hausdorff distance" in 14th International Conference on Information Fusion, Chicago USA, 2011.

[5] R. Laxhammar. "Conformal anomaly detection. Detecting abnormal trajectories in surveillance applications" Ph.D Thesis. University of Skövde, Sweden, 2014.

[6] J. Smith, I. Nouretdinov, R. Craddock, C. Offer, A. Gammerman. "Anomaly detection of trajectories with kernel density estimation by conformal prediction" in Artificial Intelligence Applications and Innovations, Rhodes Greece, 2014.

[7] J. Smith, I. Nouretdinov, R. Craddock, C. Offer, A. Gammerman. "Conformal anomaly detection of trajectories with a multi-class hierarchy" in Statistical Learning and Data Sciences: Third International Symposium, Egham UK, 2015.

[8] J. Smith. "The efficiency of conformal prediction for anomaly detection" PhD thesis. Royal Holloway University of London, UK, 2016.

[9] G. Cherubin, I. Nouretdinov, A. Gammerman, R. Jordaney, Z. Wang, D. Papini, L. Cavallaro. "Conformal clustering and its application to botnet traffic" in Statistical Learning and Data Sciences: Third International Symposium, Egham UK, 2015.

[10] V. Fedorova, A. Gammerman, I. Nouretdinov, V. Vovk. "Plug-in martingales for testing exchangeability on-line" in 29th International Conference on Machine Learning, Edinburgh UK, 2012.

[11] H. Graf, E. Cosatto, L. Bottou, I. Durdanovic, V. Vapnik. "Parallel support vector machines: the cascade SVM" in Advances in Neural Information Processing Systems 17, Vancouver Canada, 2004.

[12] P. Toccaleli, I. Nouretdinov, A. Gammerman. "Conformal predictors for compound activity prediction" in Conformal and Probabilistic Prediction with Applications: 5th International Symposium, Madrid Spain, 2016.

[13] V. Vapnik, R. Izmailov. "Synergy of monotonic rules" J. Mach. Learn. Res., **17**(136), 1–33, 2016.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 2, No. 3, 268-276 (2017)
Special Issue on Recent Advances in Engineering Systems

ASTES Journal
ISSN: 2415-6698

S-Parameters Optimization in both Segmented and Unsegmented Insulated TSV upto 40GHz Frequency

Juma Mary Atieno*, Xuliang Zhang, HE Song Bai

UESTC, No.2006, Xiyuan Ave, West Hi-Tech Zone, 611731, Chengdu, Sichuan, P.R.China

ARTICLEINFO

Article history:
Received: 05 April 2017
Accepted: 28 April 2017
Online: 12 May 2017

Keywords:
Reliability
Segmented TSV
S-parameters
Unsegmented TSV

ABSTRACT

Segmented and unsegmented 3D insulated copper through silicon vias (TSVs) of diameter 10 micro-meter, height 100 micro-meter and silicon of sizes 100 micro-meter by 100 micro-meter by 100 micro-meter are modeled using analysis system (ANSYS), equivalent circuit using advanced design system (ADS) at frequency ranges between 100MHz and 40GHz and 10MHz step size. The segmented via is divided into three parts.Each part modeled separately. Scattering parameters are found. The outputs are optimized to give accurate results. The results show that the outputs reflect the transmission characteristics of an ideal TSV.

1 Introduction

S-Parameters Optimization in both Segmented and Unsegmented Insulated TSV was featured in [1]. This paper advances the work done in [1] by optimizing the S-Parameters but to a higher frequency of 40GHz. S parameters have been given much considerations since they allow a device to be treated as a black box with inputs and resulting outputs, making it possible to model a system without having to deal with the complex details of its actual structure.

However, the bandwidth of todays Integrated Circuits (IC's) keeps increasing. This paper has found it important to characterize their performance over wide frequency ranges up-to a maximum of 40GHz. This is because the traditional low-frequency parameters such as resistance, capacitance, and gain can be frequency dependent and therefore may not fully describe the performance of the IC at the desired frequency. Moreover, it may not be possible to characterize every parameter of a complex IC over frequency therefore system-level characterization using S-parameters may provide better and more reliable data [2, 3, 4].

In the previous paper [1], the scattering parameters especially the S21 which defines power loss in TSVs was given much attention. Here we consider other parameters such as S11 too.

2 S-Parameters

A simple RF relay is used to demonstrate the techniques of high-frequency model verification.It is considered a three-port device with an input, output and a control to switch the circuit on and off. However, since the device performance is independent of the control terminal it is simplified to a two-port device. Therefore, it is characterized by observing the behavior at its input and output terminals. This is elaborated in figure 1.

Where,
a_1 = incident wave on port 1
b_1 = reflected wave from port 1
a_2 = incident wave on port 2 (reflected from the load)
b_2 = reflected wave from port 2
Z_S= source impedance
Z_L= load impedance

From figure 2 notations, S-parameters defined the reflected waves b_1 and b_2 as shown in equation (1) and (2)
.

$$b_1 = S_{11}a_1 + S_{12}a_2 \quad (1)$$

$$b_2 = S_{21}a_1 + S_{22}a_2 \quad (2)$$

*Juma Mary Atieno, UESTC, No.2006, Xiyuan Ave, West Hi-Tech Zone, 611731, Chengdu, Sichuan, P.R.China , majuma84@yahoo.com

https://dx.doi.org/10.25046/aj020336

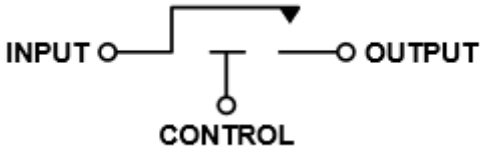


Figure 1: RF relay model [2]

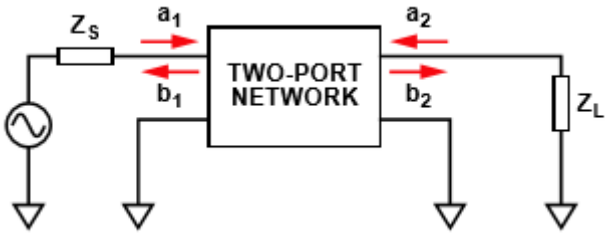


Figure 2: S-parameter model [2]

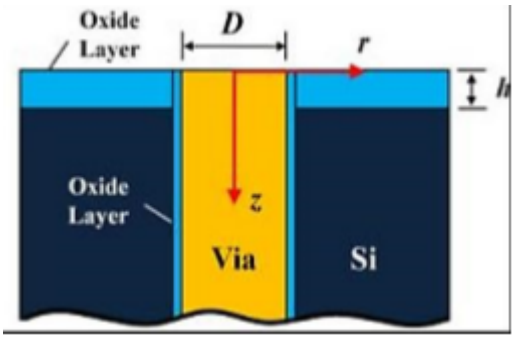


Figure 3: TSV structure [7]

If $Z_S = Z_0$, that is; impedance of two-port input, no reflections occur and $a_1 = 0$. If $Z_L = Z_0$ that is; the impedance of two-port output, no reflections occur and $a_2 = 0$. Therefore we can define the S-parameters, based on the matched condition as follows:

$$S_{11} = \frac{b_1}{a_1} \tag{3}$$

$$S_{22} = \frac{b_2}{a_2} \tag{4}$$

$$S_{21} = \frac{b_2}{a_1} \tag{5}$$

$$S_{12} = \frac{b_1}{a_2} \tag{6}$$

To realize the above parameters in a physical system, Z_S, Z_0, and Z_L must be matched. For most systems this is easily implemented over a wide frequency range [2]. The authors implemented up to a maximum frequency of 40GHz.

3 Through Silicon Via (TSV) Technology

This study contributes to knowledge about parameter distribution, transmission properties, electrical effects and so on. This is the basic foundation of 3D IC [5, 6].

Figure 3 shows a structure of a TSV. It's made up of key areas such as via where the filler material is put. The filler material is copper in most cases since it's a good conductor. The letter z defines the via depth direction, D is the via diameter while r is the via radius direction. The TSV is surrounded by an oxide layer which is silicon dioxide and h defines its height or thickness. The whole model is done in silicon which is a substrate material. Increasing power, performance and financial bottle-necks beyond 3222 nm industry has led to development of TSV [8]. TSVs can be processed in different ways. For instance; Front

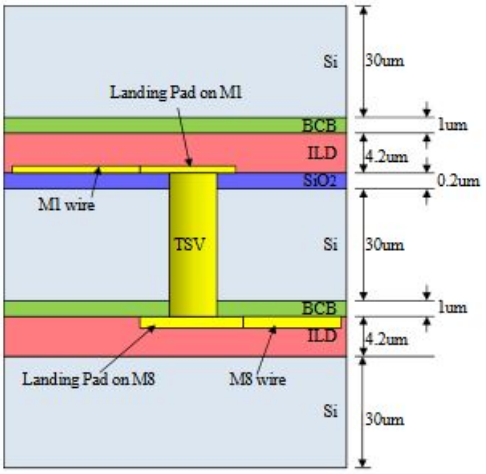


Figure 4: Via First TSV [9]

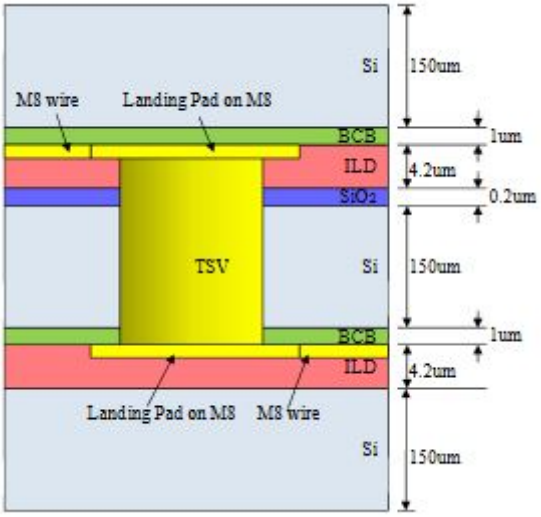


Figure 5: Via Last TSV [9]

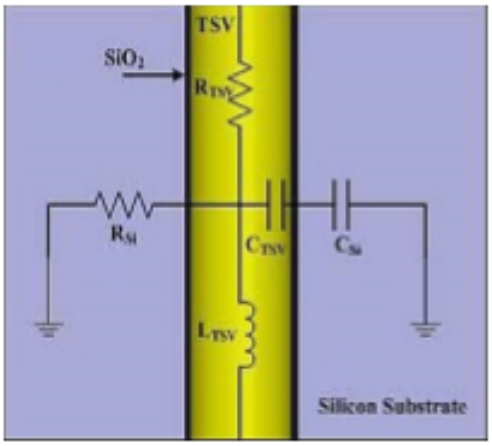


Figure 6: Unsegmented TSV model [6]

End Of Line (FEOL), Back End Of Line (BEOL), Via First (poly-silicon filled via) whereby TSVs are fabricated before transistors. Via Middle (copper filled via) whereby TSVs are fabricated before the copper interconnects and Via Last (copper liner) where TSVs are fabricated after bonding the stack. Figures 4 and

5 show the differences between via first and via last technologies. Potential reliability issues can arise due to during or after these processes. During the via design the stress management can be a problem. This is experienced through via fabrication which may induce tensile stresses in silicon, via may also experience compressive hoop stresses which could lead to via buckling, the quantity and spacing of via may also result in undesirable stress fields and finally the most common thermally induced stresses due to coefficient of Thermal Expansion (CTE) mismatch which may lead to cracking of the silicon wafer. Thermal management is another design problem. This is due to large and non-uniform heat flow within the package. There also exists redundancy in the design used.

Apart from the problems experienced during design of TSVs, some problems are experienced during processing. These include; metal voiding which happens during material filling, uniform via wall material deposition, active Integrated Circuit (IC) surface connectivity and wafer handling problem. Other challenges are experienced during the performance of the TSVs. For instance; limited reliability, thermal cycling, electro migration, silicon depletion at or around the via and also shock and vibration. Noise coupling, fabrication process limits and failure issues have been discussed in [10]. However TSV has several advantages such as reducing the delay, power and area. It also promotes the integration of heterogeneous devices.

4 TSV Liner

Although the TSV has several advantages, these would not be viable unless the TSV is well protected. This is enhanced by TSV liner. In our paper we propose the use of silicon dioxide material as a liner due to its many advantages. Dielectric liner functions as electrical isolation between the TSV metal fill and silicon substrate; therefore it should have zero leakage, high breakdown voltage, low stress, process temperature compatibility, good step coverage and uniformity. Silicon dioxide has these properties. However, other materials such as nitride and polymers are also used as liners for TSV applications [11,12, 13].

5 Methodology

There are several ways of characterizing the ports behavior. One is by admittance matrix Y or impedance matrix Z which defines the relationship between voltages and currents at the ports [14]. Various ways are available and are used to calculate functions that can accurately produce port characteristics within a desired frequency at the same time meet the physical requirements such as symmetry, causality, stability and passivity [15,16, 17 and 18]. The second way of characterizing the port behavior is by use of the incident and reflected waves that are related through the scattering parameter matrix [14], [19], [20]. This second way is often applicable in high speed electronics modeling over the admittance formulation because it gives more accurate results at very high frequencies; it is also useful in applying pure transfer functions models in a simulation. These transfer function capability provides an easy way of observing internal voltages and currents in a subnetwork represented by a Y-, Z- or S-parameter-based port equivalent, and it reduces the computational burden of the model identification process as the number of ports is reduced [14].

As discussed in, [14], Y parameter models can be interfaced with electromagnetic transients program tools through equivalent circuits [21], [22]. However, the circuit experiences inaccuracy hence should be avoided and instead the model can be represented using convolutions [23], [24] which are considered a little bit standard as per [25, 26]. This paper is accessing the variation of S-parameters on a single insulated unsegmented TSV and segmented TSV over a frequency range of 40GHz. The S21 parameter which represents the insertion loss is the key area of research. Insertion loss is expected to be as low as possible on the graphs. The equivalent circuits of the unsegmented and segmented vias are simulated and the results represent the TSV characteristics. The segmented TSV gave a much lower insertion loss as compared to the unsegmented one. This shows that its more reliable and can be used to alleviate the insertion loss problem mainly experienced in the TSV thereby alleviating that reliability problem in 3D packages. The S parameters are converted to Z parameters, Z21 which gave much lower values than the S21 but still maintained the TSV characteristics. S11 characteristics are also listed.

5.1 Unsegmented TSV

Figure 6 represents a vertical TSV model. The R_{TSV}, C_{TSV} and L_{TSV} represent resistance, capacitance and inductance respectively along the TSV. R_{Si} and C_{Si} are the resistance and capacitance respectively in the silicon substrate.

The Advanced Design System (ADS) simulator developed the schematic in figure 7 from the model in figure 6. LTSV, RTSV and C4 are inductance, resistance and capacitance respectively along the TSV. The C1 and C2 are capacitances in the TSV liner, RSi and CSi are the resistance and capacitance respectively along the silicon substrate, S parameter with the terms give the insertion and other loses experienced in TSV while the yield ensures accurate results are obtained. The schematic is an improvement of the work done in [1, 6]. In [6], silicon dioxide layer which act as a TSV liner was not included in the circuit. Moreover, optimization was not done to enhance accurate results. This has been done in this paper by including the yield in the system the frequency has been extended to 40GHz too. A total of 250 trials were made to finally end up with 100 percent yield. The trials were done within the frequency ranges specified in the S-parameter icon. S11 parameters have also been discussed.

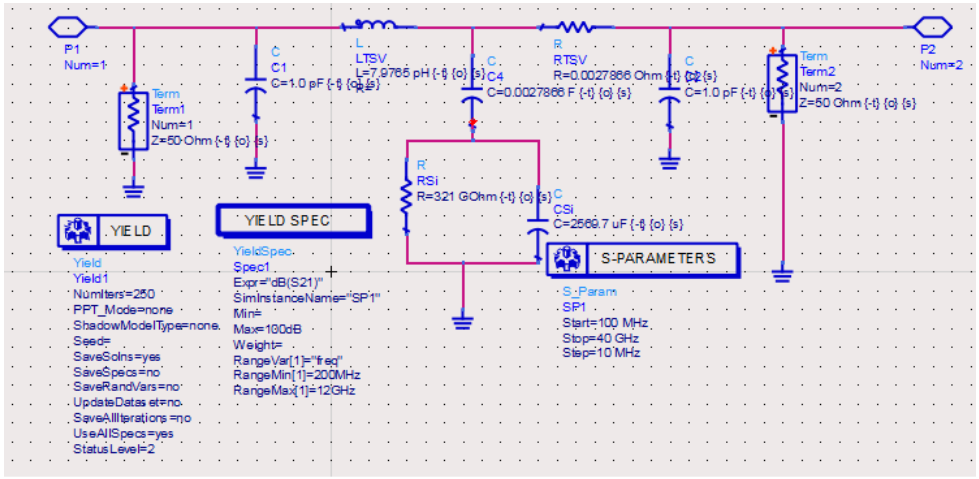


Figure 7: Unsegmented TSV equivalent circuit

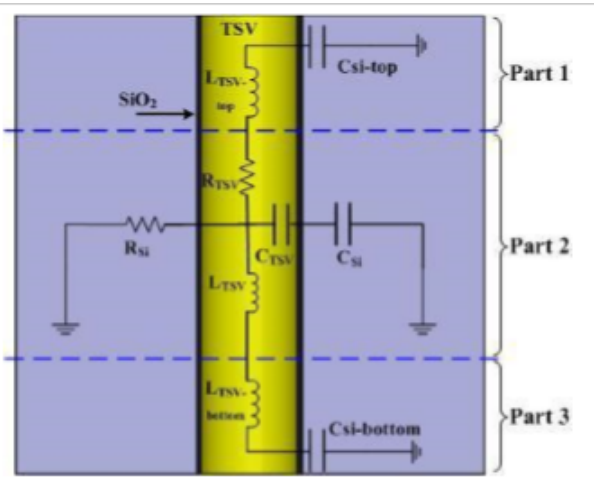


Figure 8: Segmented TSV model

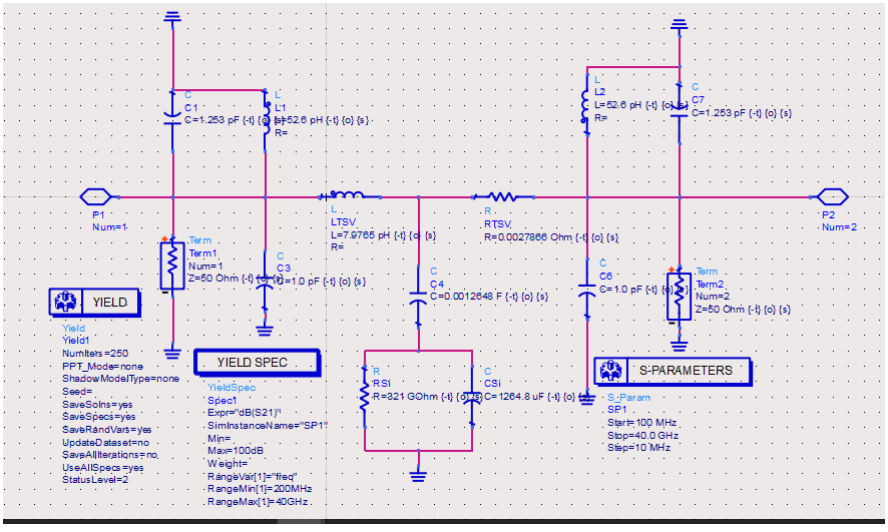


Figure 9: Segmented TSV equivalent circuit

5.2 Segmented TSV

Figure 8 represents a partitioned vertical TSV. It is divided into three main parts namely; Part 1, Part 2 and Part 3. The partitioning of the TSV reduces the insertion loss which is a reliability problem in TSVs. Parasitic resistance, inductance and capacitance exists just like in figure 6.

Figure 8 developed figure 9 whereby; C1 and C7 represent the capacitance developed across the insu-

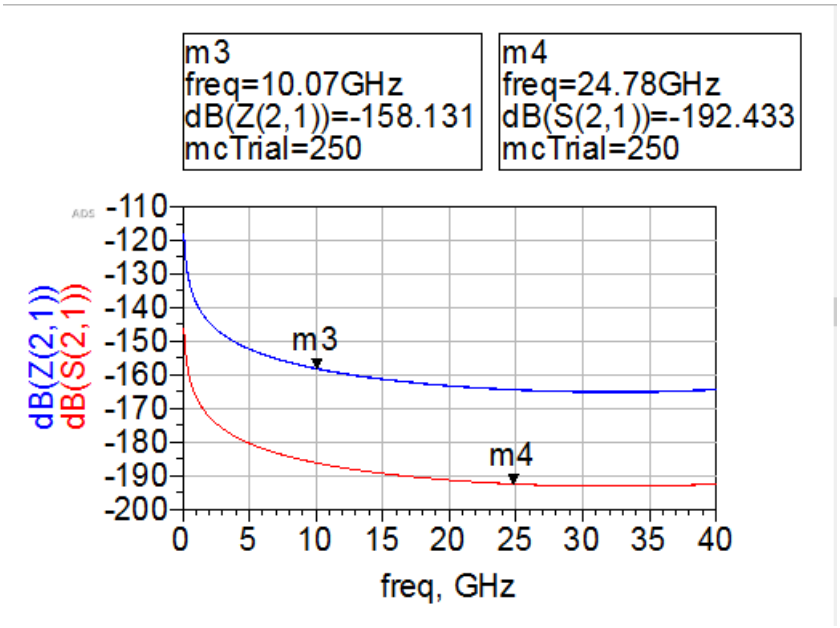


Figure 10: Unsegmented TSV graph

freq	S(1,1)	
	mcTrial=0	mcTrial=250
100.0 MHz	1.000 / 179.989	1.000 / 179.989
110.0 MHz	1.000 / 179.987	1.000 / 179.988
120.0 MHz	1.000 / 179.986	1.000 / 179.987
130.0 MHz	1.000 / 179.985	1.000 / 179.986
140.0 MHz	1.000 / 179.984	1.000 / 179.985
150.0 MHz	1.000 / 179.983	1.000 / 179.984
160.0 MHz	1.000 / 179.982	1.000 / 179.983
170.0 MHz	1.000 / 179.980	1.000 / 179.981
180.0 MHz	1.000 / 179.979	1.000 / 179.980
190.0 MHz	1.000 / 179.978	1.000 / 179.979
200.0 MHz	1.000 / 179.977	1.000 / 179.978
210.0 MHz	1.000 / 179.976	1.000 / 179.977
220.0 MHz	1.000 / 179.975	1.000 / 179.976
230.0 MHz	1.000 / 179.974	1.000 / 179.975
240.0 MHz	1.000 / 179.972	1.000 / 179.974
250.0 MHz	1.000 / 179.971	1.000 / 179.973
260.0 MHz	1.000 / 179.970	1.000 / 179.972
270.0 MHz	1.000 / 179.969	1.000 / 179.970
280.0 MHz	1.000 / 179.968	1.000 / 179.969
290.0 MHz	1.000 / 179.967	1.000 / 179.968
300.0 MHz	1.000 / 179.966	1.000 / 179.967
310.0 MHz	1.000 / 179.964	1.000 / 179.966

Figure 11: S11 list of values in unsegmented TSV

lator layer silicon dioxide. The LTSV, RTSV and C4 are the inductance, resistance and capacitance respectively along the TSV. RSi and CSi are the resistance and capacitance respectively along the silicon substrate. The S-parameter and the yield simulators play the same role as in figure 7.

6 Results and Discussions

6.1 Unsegmented TSV

Figure 10 looks exactly as what was done in [1] but the difference is that the frequency range has been stretched to 40GHz. The dB values are decreasing with the increased frequency. The markers, m3 and m4 define the dB values at those particular frequencies.

Figure 11 is a list of S11 parameters with varying frequencies. As the frequency increases the S11 value decreased as evident in the list.

6.2 Segmented TSV

Figure 12 describes the dB(S(2,1)) with variation in frequency up to 40GHz. The m3 and m4 in this figure defined the parameter at those frequencies. However, the maximum power on the 'YIELD SPEC' was varied to enhance smooth curve; dB(Z(2,1)) was not plotted too, although the graph gave exactly the variations discussed in [1].

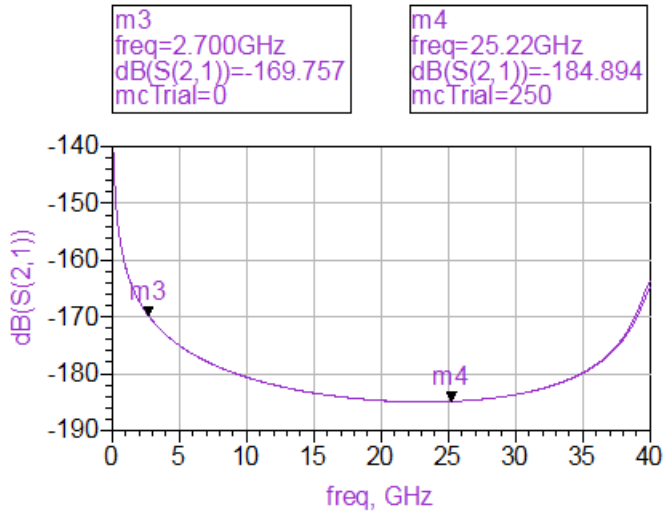


Figure 12: Segmented TSV graph

freq	S(1,1)	
	mcTrial=0	mcTrial=250
100.0 MHz	1.000 / 179.990	1.000 / 179.990
110.0 MHz	1.000 / 179.989	1.000 / 179.989
120.0 MHz	1.000 / 179.988	1.000 / 179.988
130.0 MHz	1.000 / 179.987	1.000 / 179.987
140.0 MHz	1.000 / 179.986	1.000 / 179.986
150.0 MHz	1.000 / 179.985	1.000 / 179.985
160.0 MHz	1.000 / 179.984	1.000 / 179.984
170.0 MHz	1.000 / 179.983	1.000 / 179.983
180.0 MHz	1.000 / 179.982	1.000 / 179.982
190.0 MHz	1.000 / 179.981	1.000 / 179.981
200.0 MHz	1.000 / 179.980	1.000 / 179.980
210.0 MHz	1.000 / 179.979	1.000 / 179.979
220.0 MHz	1.000 / 179.978	1.000 / 179.977
230.0 MHz	1.000 / 179.977	1.000 / 179.976
240.0 MHz	1.000 / 179.976	1.000 / 179.975
250.0 MHz	1.000 / 179.975	1.000 / 179.974
260.0 MHz	1.000 / 179.974	1.000 / 179.973
270.0 MHz	1.000 / 179.973	1.000 / 179.972
280.0 MHz	1.000 / 179.972	1.000 / 179.971
290.0 MHz	1.000 / 179.971	1.000 / 179.970
300.0 MHz	1.000 / 179.970	1.000 / 179.969
310.0 MHz	1.000 / 179.969	1.000 / 179.968
320.0 MHz	1.000 / 179.968	1.000 / 179.967
330.0 MHz	1.000 / 179.967	1.000 / 179.966
340.0 MHz	1.000 / 179.966	1.000 / 179.965
350.0 MHz	1.000 / 179.965	1.000 / 179.964

Figure 13: S11 list of values in Segmented TSV

Figure 13 on the other hand described S11 parameter with varying frequencies. It was noted that just like in the unsegmented TSV, the S11 values decreased with increased frequencies, the differences in S11 are however slight. However, The S11 in the segmented TSV is slightly higher than that in unsegmented one at the same frequency.

7 Conclusions and Recommendations

Both segmented and unsegmented TSV were modeled using ADS software. This work was an improvement of work done by [1 and 6]. In the previous work [1], the maximum frequency range considered was 20GHZ but in this paper, it has been stretched to 40GHz. We recommend extension of the frequency range because by doing so, its application will be possible on more electronic devices made. Optimization of the values also promotes more accurate results. Just like in the previous work we intend to do the modeling of TSV in more portions rather than three parts in future. We will also include the effects of changes in length and thickness of the TSV alongside use of other TSV liner materials such as polymers on the power losses. With more factors put into consideration, we hope to reduce the insertion losses in TSV even further. There has been increasing need of cheaper and high performance devices in the world today. These have made researches seek for solutions to these problems. Figure 14 describes the future of TSVs. Varia-

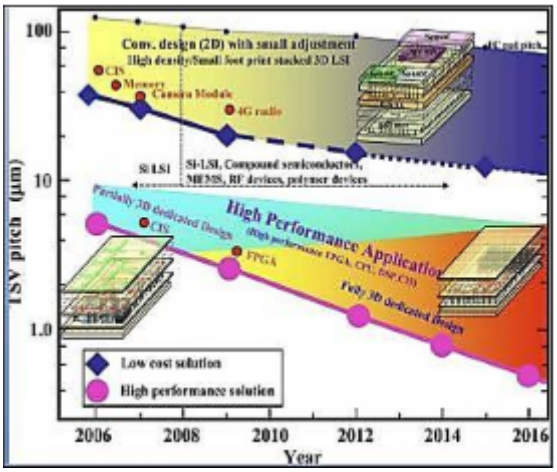


Figure 14: The future of TSV

tion of TSV pitch have greatly influenced the cost and performance of TSV which later influence the devices made between the year 2006 and the current 2017. In the year 2006, the cost of manufacturing TSVs was too high yet the performance of the devices made from the technology was poor due to high TSV pitch. This reduced the reliability of the electronic devices made [1].

Just like in [1],as years went by, research was done and the researchers realized that the TSV pitch size greatly influenced the cost and performance of the TSV modeled devices. They therefore proposed the reduction of the TSV pitch from 100 micro-meters in the year 2006 to less than 1 micro-meter now; 2017. Through the gradual reduction in the TSV pitch size, the cost of production has gone down and the performance of the electronic devices made from the TSV technology has increased. High costs and low performance was evident in 2D conventional designs. However, the use of semiconductors, MEMS, RF devices and polymer devices enhanced development of partially 3D dedicated design to fully 3D dedicated designs which experience low costs and high performance.

Acknowledgment The first author would like to acknowledge the enthusiastic supervision of Prof. HE Song Bai and Zhang Xuliang for their great guidance and advice throughout this paper; they truly inspired me. Their will to motivate me played a major role in this paper. I would also like to thank them for giving me some good examples of some published papers which are related to this work. I would not forget to thank University of Electronic Science and Technology of China (UESTC) for providing a conducive environment and materials to work with on our paper. Our appreciations also go to the laboratory technicians and members who were patient with us all the time. They also contributed to some great ideas in this paper. Finally, we would not like to forget our families and friends for their understanding and support in our work without which this work would not be possible.

References

1. J. M. Atieno, X. Zhang and H. S. Bai, "S-parameters optimization in both segmented and unsegmented insulated TSV," 2016 International Conference on Integrated Circuits and Microsystems (ICICM), Chengdu, 228–232, 2016.

2. Joseph Creech, "S-Parameters Allow High-Frequency Verification of RF Switch Models" Analog Dialogue, **45**(10), 1–4, 2011.

3. Paul, TDR: taking the pulse of signal integrity. EDN, September 3, 2007.

4. Bowick, Chris, John Blyler and Cheryl Ajluni, "RF Circuit Design" Newnes, 2007.

5. W. R. Davis et al., "Demystifying 3D ICs: the pros and cons of going vertical," in IEEE Design and Test of Computers, **22**(6), 498–510, Nov.-Dec. 2005.

6. Z. Ma, Z. Yan, Y. Xiong, M. Zhang and G. Shi, "A novel segmented equivalent circuit modeling method of TSV in 3D-IC," 2013 5th IEEE International Symposium on Microwave, Antenna, Propagation and EMC Technologies for Wireless Communications, Chengdu,646–649, 2013.

7. S. K. Ryu, K. H. Lu, T. Jiang, J. H. Im, R. Huang and P. S. Ho, "Effect of Thermal Stresses on Carrier Mobility and Keep-Out Zone Around Through-Silicon Vias for 3-D Integration," in IEEE Transactions on Device and Materials Reliability, **12**(2) 255–262, June 2012.

8. D. Z. Pan et al., "Design for manufacturability and reliability for TSV-based 3D ICs," 17th Asia and South Pacific Design Automation Conference, Sydney, 750–755, NSW,2012.

9. J. Pak, M. Pathak, S. K. Lim and D. Z. Pan, "Modeling of electromigration in through-silicon-via

based 3D IC," 2011 IEEE 61st Electronic Components and Technology Conference (ECTC), Lake Buena Vista, FL, 1420–1427, 2011.

10. J. Lim et al., "Shielding structures for through silicon via (TSV) to active circuit noise coupling in 3D IC," 2015 10th International Workshop on the Electromagnetic Compatibility of Integrated Circuits (EMC Compo), Edinburgh, 248–251, 2015.

11. S. Ramaswami et al., "Process Integration Considerations for 300 mm TSV Manufacturing," in IEEE Transactions on Device and Materials Reliability, **9**(4), 524–528, Dec. 2009.

12. L. Zhang, H. Y. Li, S. Gao and C. S. Tan, "Achieving Stable Through-Silicon Via (TSV) Capacitance with Oxide Fixed Charge," in IEEE Electron Device Letters, **32**(5) 668–670, May 2011.

13. Q. Chen, C. Huang, Z. Tan and Z. Wang, "Low Capacitance Through-Silicon-Vias With Uniform Benzocyclobutene Insulation Layers," in IEEE Transactions on Components, Packaging and Manufacturing Technology, **3**(5) 724–731, May 2013.

14. B. Gustavsen and H. M. J. De Silva, "Inclusion of Rational Models in an Electromagnetic Transients Program: Y-Parameters, Z-Parameters, S-Parameters, Transfer Functions," in IEEE Transactions on Power Delivery, **28**(2), 1164–1174, April 2013.

15. B. Gustavsen and A. Semlyen, "Rational approximation of frequency domain responses by vector fitting," in IEEE Transactions on Power Delivery, **14**(3), 1052–1061, Jul 1999.

16. B. Gustavsen, "Improving the pole relocating properties of vector fitting," in IEEE Transactions on Power Delivery, **21**(3), 1587–1592, July 2006.

17. D. Deschrijver, M. Mrozowski, T. Dhaene and D. De Zutter, "Macromodeling of Multiport Systems Using a Fast Implementation of the Vector Fitting Method," in IEEE Microwave and Wireless Components Letters, **18**(6), 383–385, June 2008.

18. P. Triverio, S. Grivet-Talocia, M. S. Nakhla, F. G. Canavero and R. Achar, "Stability, Causality, and Passivity in Electrical Interconnect Models," in IEEE Transactions on Advanced Packaging, **30**(4), 795–808, Nov. 2007.

19. K. Kurokawa, "Power Waves and the Scattering Matrix," in IEEE Transactions on Microwave Theory and Techniques, **13**(2), 194–202, Mar 1965.

20. Z. Zhongyuan, L. Fangcheng and L. Guishu, "A High-Frequency Circuit Model of a Potential Transformer for the Very Fast Transient Simulation in GIS," in IEEE Transactions on Power Delivery, **23**(4), 1995–1999, Oct. 2008.

21. A. S. Morched, J. H. Ottevangers and L. Marti, "Multi-port frequency dependent network equivalents for the EMTP," in IEEE Transactions on Power Delivery, **8**(3), 1402–1412, July 1993.

22. B. Gustavsen, "Computer code for rational approximation of frequency dependent admittance matrices," in IEEE Transactions on Power Delivery, **17**(4), 1093–1098, Oct 2002.

23. T. Noda, "Identification of a multiphase network equivalent for electromagnetic transient calculations using partitioned frequency response," in IEEE Transactions on Power Delivery, **20**(2), 1134–1142, April 2005.

24. C. Dufour, J. Mahseredjian and J. Belanger, "A Combined State-Space Nodal Method for the Simulation of Power System Transients," in IEEE Transactions on Power Delivery, **26**(2), 928–935, April 2011.

25. A. Semlyen and A. Dabuleanu, "Fast and accurate switching transient calculations on transmission lines with ground return using recursive convolutions," in IEEE Transactions on Power Apparatus and Systems, **94**(2), 561–571, Mar 1975.

26. M. Motoyoshi, "Through-Silicon Via (TSV)," in Proceedings of the IEEE, **97**(1), 43–48, Jan. 2009.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 2, No. 3, 277-283 (2017)
www.astesj.com
Special Issue on Recent Advances in Engineering Systems

ASTES Journal
ISSN: 2415-6698

Synthesis of Important Design Criteria for Future Vehicle Electric System

Lisa Braun[*,1], Eric Sax[2]

[1]*Ph.D. Candidate, Institute for Information Processing Technologies, Karlsruhe Institute of Technology, Germany*

[2]*Professor, Institute for Information Processing Technologies, Karlsruhe Institute of Technology, Germany*

ARTICLEINFO

Article history:
Received: 05 April, 2017
Accepted: 02 May, 2017
Online: 14 May, 2017

Keywords:
e/e-architecture
Design criteria
Vehicle electric system design

ABSTRACT

In this article, we give a brief overview of future trends and corresponding stakeholder requests regarding the design of automotive electric and electronic (e/e)-architectures. While today's optimization is mainly based on the static design criteria of unit costs and weight, three dynamic design criteria are considered decisive factors in the future: voltage stability, vehicle energy consumption and total cost of ownership. It is shown, how quantitative values for each criterion can be derived, on basis of which different designs of a potential future power supply system could be evaluated. Subsequently, their potential influences on the future system design is discussed.

1 Introduction

This paper is an extension of work originally presented in the 4th International Conference on Electrical Systems for Aircraft, Railway, Ship propulsion and Road Vehicles and International Transportation Electrification Conference [1]. The focus of [1] is to summarize the results of guided expert interviews on future trends, associated hurdles and requirements of the electric and electronic (e/e)-architecture.

The design of the e/e-architecture of vehicles faces many challenges. The information and communication system has to manage and distribute an amount of data that is rapidly growing due to digitization and new functionalities, while the energy distribution system has to provide sufficient energy to securely supply the growing number of electrical components. This growing number of components, functions and connections makes system complexity ever-increasing [1]. This also means that development expenditures are growing significantly. Over 40% of the development expenses of a luxury vehicle is related to e/e [2].

The expert interviews were designed to find out which trends the e/e-architecture will shape in the future, and whether and how the design criteria for development will change. Considering these results, we are expanding our work by synthesizing design criteria that will be important in the future, focusing on the power supply system as part of the e/e-architecture. These criteria are based on the trends and the overview given in [1]. We also discuss the effects of such a change in design criteria on the development of vehicles, and in particular the design of the energy distribution system in the future.

This paper is organized as follows: section two discusses current trends concerning the e/e architecture. Section three summarizes important stakeholder requests on the design of e/e-architectures. While design criteria that are considered important for future power supply systems are synthesised in section four. Examples show their possible impact on design. Subsequently section 5 discusses their influence on future system design. The final section shows conclusions and future works.

2 Trends Concerning the Vehicle Electric System

Today, autonomous driving is one of the most important trends in the automotive industry. Last year, the number of companies testing autonomous vehicles in California doubled. By March 2017, 27 companies received Autonomous Vehicle Testing Permits from California Department of Motor Vehicles.

What is more, in recent years, a change in the mobility behaviour can be observed in all the countries of the Triad [3]. In 2015, the number of registered car sharing users in Germany amounted to 1.26 million already, equating to a six-fold increase since 2011

[*]Lisa Braun, lisa.braun@partner.kit.edu

https://dx.doi.org/10.25046/aj020337

[4]. This strong growth might even be reinforced even more by the introduction of autonomous vehicles. A possible combination of apps like SMILE, GoogleNow, and e-hailing with autonomous vehicles promises to create ideal robot taxis as part of a multimodal mobility concept to implement mobility as a service [5].

To allow autonomous manoeuvres, an exact environment model and thus a large number of sensors for environment detection on board the vehicle are necessary. Moreover, networking with other vehicles and the infrastructure is a precondition to avoid dangers on the route [1]. Both lead to yet another significant increase in the amount of data on board [6]. In addition, they must not fail during driving in order to ensure safe operation of the vehicle. Therefore, the power supply needs to provide at least two independent sources, so that the supply of these safety-relevant components is at no time endangered.

To allow for the provision of this growing power demand, a multi voltage power supply is seen as certain. With the 12 V electrical system limited to 3.6 kW maximum power, it became increasingly difficult to provide the growing power demand of new functions such as active roll stabilization (ARS) [7]. The introduction of an additional 48 V level shall solve that problem [8].

Further, it is cheaper as high-voltage systems in hybrid vehicles as it is still below 60 V, which means no additional isolation is necessary. The higher voltage also allows a reduction of cable cross sections while the efficiency of the distribution increases. The weight of the cable harness could be further reduced by using semiconductors instead of fuses and relays [9].

3 Stakeholder Requests

About 90% of the surveyed experts mention costs as an important design feature of the established e/e-architecture, followed by 80% naming weight. Thus, these two criteria are clearly mentioned most frequently, see also Figure 1 [1].

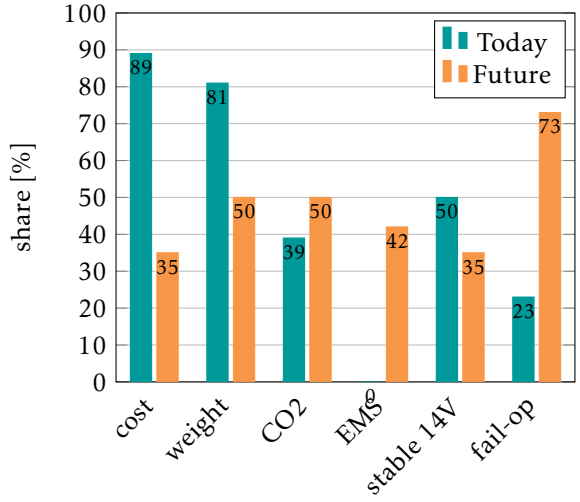


Figure 1 Significance of today's and future requirements concerning the e/e-architecture.

Concerning future design criteria, however, this picture changes. While the importance of cost and weight sink to 35% and 50%, over 73% of the experts think a fail-operational design is important. Another criterion for safe operation, the stability of the voltage, is considered important in future design by at least one third of the experts. The reference to the reduction of carbon emissions increases from 39% to 50%. In addition, about 40% of the experts are asking for an energy management system (EMS) for the future, while it does not seem to matter at all [1].

The focus on fail-operational is due to current developments in driver assistance with the aim of integrating autonomous driving. Failures are not acceptable and must therefore be prevented by a corresponding design of the system, mostly realized by redundancy mechanisms. An example how this could be realized for the information and communication architecture is the developed centralized platform computer of the RACE project [10]. A fail-operational power supply is realized through at least two independent power sources [11]. While this applies to an electric vehicle with DC/DC-converter and one battery in the low-voltage network, sailing with switched off combustion engine or start-stop mode would lead to a system with only one source, if only one battery is installed additionally to the generator [7].

What is more is that voltage stability must be guaranteed even with fast load changes and high power peaks. An EMS could therefore contribute to the stabilization of the voltage as well as the reduction of the consumption and associated emissions.

However, the experts' answers do not allow any conclusion as to whether these changes in the values are due to actual changes in the prioritization or due to the current trends and associated expected future problems. In the case of costs in particular, it seems questionable whether their significance in the future will actually decrease so clearly. Nevertheless, these figures allow conclusions to be drawn on expected changes in the development. In the following, important future design criteria of the power supply system are derived from this selection.

4 Synthesis of Important Design Criteria

Not technical aspects are the main problem for achieving future solutions but legacy restrictions and fixed corporate structures, which go along with the structure of the vehicle's e/e-architecture [1]. Design criteria that consider future changes can help to weigh up possible solutions in the early stages of development. Therefore, the aim of the derived design parameters is to provide the advantages and disadvantages of the various designs of the power supply system even in an early development phase. For this reason, they must provide measurable target variables, with the aid of which different designs can be evaluated and an optimal solution can be derived.

Firstly, to guarantee the safe operation the voltage stability must be guaranteed. This is an essential constraint for the integration of safety-relevant functions. Although fail-operational is of great importance for the design of future e/e-architectures due to the results in [1], no corresponding evaluation criterion is developed in the following. In addition to the requirement of a redundant power source, an isolated view of the power supply system does not seem to make sense. Rather, an intelligent combination of measures in the communication and power supply network can save required redundancy effort, e.g. through a combination of powerline communication and power over dataline for safety-relevant components [12].

Secondly, at the latest since the diesel-gate, vehicle emissions have been the focus of attention. Weight reduction as well as higher efficiency of the power distribution or reduction of the power consumption e.g. through EMS can help to reduce the energy consumption of the vehicle as well as the corresponding emissions. There is a conflict, however, between a light weight wire harness with small cable cross-sections and the achievement of minimal power losses through the distribution, as small cross-sections increase the cable resistance. In order to find an optimum compromise, the vehicle consumption is therefore chosen as a further design criterion.

Finally, today the costs per function must be minimal. We expect this to change in the future due to the growing volume of car sharing and stricter emission limits. Operators of car sharing fleets are less interested in the purchasing prize than in the total cost of ownership. Simultaneously, the operating time and thus the influence of the travel costs on the total cost of ownership increases. For this reason, not only development and production costs but also operating costs must be taken into account in minimizing costs in the future.

4.1 Voltage Stability

For safe operation of e/e-systems, it is important that the voltage limits of 11 V and 16 V are maintained at the sinks, otherwise the functionality is not, or only partially, provided. Via the buffer of the voltage to these limits, for example, in an extreme driving scenario, it is possible to compare different configurations of the power supply system [13].

The disadvantage of this method is that it does not provide any information about the temporal behaviour of the voltage. The minimally determined value could be anything between an outreach and a constant curve or a broadband noise within the determined range.

Conclusions on the scattering can, for example, be drawn with the so-called quantiles. This is a method of statistics in which an ordered data series is split into two equal halves by the median. The box plot can thus be formed over the interquartile distance between the 25% quantile and the 75% quantile. The median indicates which voltage level the curve varies, and the size of the box between the 25% value and the 75% value shows how strong this variation is.

Alternatively, the standard deviation s based on the empirical variance σ can be used as a further method of statistics for evaluation. It indicates the average deviation of the measured values x_i from the mean value $\bar{x}$ of n data points. The smaller the standard deviation, the more stable is the curve of the voltage for the observed time interval. The standard deviation is calculated as follows [14]:

$$s = \sqrt{\sigma(x)} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}} \quad (1)$$

Figure 2 shows voltage curves with varying volatility. The blue curve shows a strong fluctuation around 12V. The green curve fluctuates considerably less by 14V. The orange curve includes a voltage jump after 300s. Before the jump, it is comparatively stable at 11V, then stable at 14V. Figure 3 shows that for the blue and green curves the fluctuation and its corresponding voltage level can be represented well with the box plot. However, the jump of the orange curve is not evaluated correctly. The values before the jump are considered as outreach and therefore this curve appears incorrectly as the most stable.

If the mean value is replaced by the system voltage of 14.3 V the standard deviation in figure 4 can as well consider the voltage level in the evaluation. Figure 4 shows that the standard deviation, in contrast to the box plot, correctly reflects the stability of the voltages in figure 2 and the green voltage curve is correctly recognized as the most stable followed by the orange. Therefore, the standard deviation appears to be a suitable parameter to evaluate voltage stability.

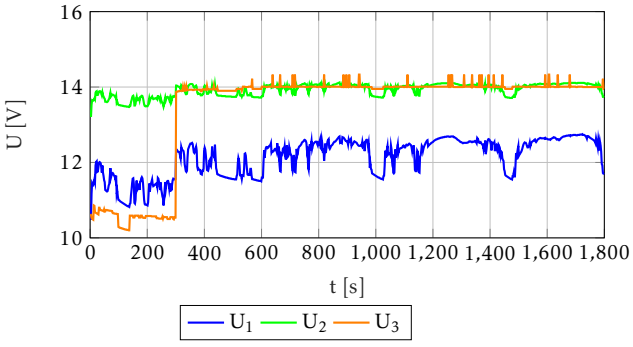


Figure 2 Voltage curves with variing volatility.

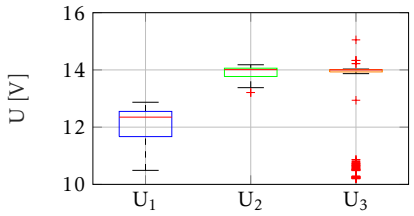


Figure 3 Box plots of shown voltage curves.

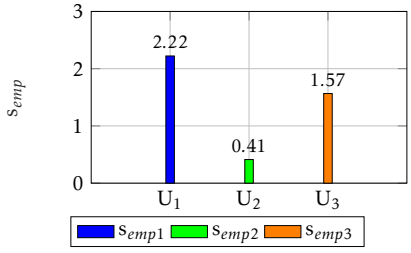


Figure 4 Standard deviation of shown voltage curves.

4.2 Vehicle Energy Consumption

In Europe, the allowed CO_2-emissions will be reduced to 95$\frac{gCO_2}{km}$ by 2020, while the full fines will reach 95$\frac{€}{g}$ from 2019 onward [15]. What makes it even more difficult is the fact that from 2020 onwards a new test method with a realistic driving cycle, the Worldwide harmonized Light duty driving Test Cycle (WLTC), for the determination of the emissions takes effect. It is expected that the discrepancy between the actual consumption and the calculated standard values will drop from an average of about 45% in 2015 to 23% in 2020 [16]. Conversely, this means that if the actual consumption is constant, the measured standard emissions are increased by 22%.

One reason for this is the fact that the consumption optimization of the vehicles to the unrealistic, static speed profile of the previous New European Driving Cycle (NEDC), no longer takes effect. Without the untypically high dead times and frequent stops, the previous high influence of consumption-reducing measures such as start-stop-automatic and downsizing is reduced [17]. Other factors such as weight reduction and the reduction of transmission losses in the power supply network are instead brought into the focus.

The specific vehicle consumption B_{sV} per 100 km can be calculated as a function of the distance travelled s_c and the needed propulsion energy W_{pro} and recuperation energy W_{rek}. If no recuperation is integrated, W_{rek} is zero.

$$B_{sV} = \left(W_{pro} + W_{rek}\right) \cdot \frac{100}{s_c} \tag{2}$$

The energy required by the propulsion or the recuperation energy which may be fed back depends of the energy requirement W_c on the tires. It depends on the total drive resistor F_t and the travelled distance s_c which can be derived from the velocity profile v_c of the drive cycle.

$$W_c = F_t \cdot s_c = F_t \cdot \int_0^t v_c(\tau)\mathrm{d}\tau. \tag{3}$$

In addition to the time profile of the energy requirement on the wheels during the cycle, the energy required by the propulsion or the recuperation energy to be fed back depends on the efficiency η_{dt} of the drive train.

$$W_{pro} = \int_0^t \frac{1}{\eta_{dt}(\tau)} \cdot F_t(\tau) \cdot s_c(\tau)\mathrm{d}\tau \qquad \text{for}(F_a > 0) \tag{4}$$

$$W_{rek} = \int_0^t \eta_{dt}(\tau) \cdot F_t(\tau) \cdot s_c(\tau)\mathrm{d}\tau \qquad \text{for}(F_a < 0). \tag{5}$$

The total drive resistance F_t is composed of several values: the rolling resistance F_r, the slope resistance F_s, the air resistance F_l and the acceleration resistance F_a which are calculated as follows [18]:

$$F_t = F_r(m_{veh}) + F_s(m_{veh}) + F_l(v_c(t)) + F_a(m_{veh}, a(t)) \tag{6}$$

$$F_r = m_{veh} \cdot g \cdot c_r \tag{7}$$

$$F_s = m_{veh} \cdot g \cdot sin(\beta(t)) \tag{8}$$

$$F_l = \frac{1}{2} \cdot \rho_l \cdot c_W \cdot A_{veh} \cdot v_c(t)^2 \tag{9}$$

$$F_a = F_{a.trans} + F_{a,rot} = a(t) \cdot \left(m_{veh} + \frac{J_{Wh} + i^2 \cdot J_{En}}{r_{stat} \cdot r_{dyn}}\right) \tag{10}$$

The rolling resistance F_r and the slope resistance F_s depend on the vehicle weight m_{veh} and the gravity g. The coefficient c_r takes into account the walk losses which occur on the road. β indicates the slope angle of the street. The air resistance F_l is calculated from the front surface A_{veh}, the air density ρ_l, the air drag coefficient c_W, and the air inflow velocity v_c(t) of the vehicle. The resistance F_a, which is dependent on the acceleration a, consists of a weight-dependent translatory part and a rotational part. The latter depends on the moment of inertia of the wheels J_{Wh} and engine J_{En}, the translation i and the static r_{stat} and dynamic r_{dyn} free radius.

The energy consumption of the electrical components and the power distribution W_E can be determined either by measurement in a test stand or, especially in early development phases, by means of a model-based analysis. In both cases, an operating model of the loads $P_i(t)$ is required to determine the energy consumption, as it fixes the current flow and thus the losses $P_{V,dst}(t)$ during transmission. What is more, in a vehicle with an internal combustion engine (ICE), the efficiency of the drive η_d and the efficiency of the generator η_g must be considered. In the case of an electric vehicle, this would instead be the efficiency of the high voltage battery and the DC/DC-converter. In order to determine the specific consumption B_{sE}, the values have to be converted to a distance of 100 km analogous to the driving consumption

$$W_E = \int_0^t \frac{1}{\eta_d(\tau) \cdot \eta_g(\tau)} \cdot \left(\sum_{i=1}^{n} P_i(\tau) + P_{V,dst}(\tau)\right)\mathrm{d}\tau \tag{11}$$

$$B_{sE} = W_E \cdot \frac{100}{s_c} \tag{12}$$

The sum of both specific consumption values provides the total consumption of the vehicle, which is to be optimized.

Table 1 Vehicle data of a small car (A), a SUV (J) and an electric luxury car (F).

Parameter	unit	A	J	F
propulsion		ICE	ICE	electric
weight	[kg]	885	2,435	2,000
standard consumption	$[\frac{kWh}{100km}]$	38.7 (=4.5l)	80.84 (=9.4l)	22.04
standard emission	$[\frac{gCO_2}{km}]$	93	218	0

The given standard values in table 1 are measured with the established New European Driving Cycle (NEDC). Assuming that both ICE-vehicles would be 50 kg lighter, the power consumption per 100 km of the small car could be reduced by 0.89 kWh and the standard emissions by 2.4 g. For the SUV, the consumption would drop by 0.91 kWh which would lead to 2.45 g less emission per km. In comparison, the high efficiency of the electric drive train with 3.71 kWh enables a significantly higher reduction in consumption by weight reduction. These values are calculated in MATLAB with the given equations (2) to (10) and the cycle data of the NEDC.

To calculate the equivalent electrical energy consumption during the NEDC, an average speed of 33.6 $\frac{km}{h}$ is assumed. It is further assumed that the efficiency of the ICE is 35%, the efficiency of the generator is 75%, while the efficiency of the high-voltage battery and the DC/DC-converter is 95%. Under these circumstances a reduced electrical energy consumption of 78 W for the small car, 80 W for the SUV and 1125 W for the electric car would have the same effect.

4.3 Total Cost of Ownership

Since, on the one hand, the permissible emission limit values $G_{CO_2,0}$ fall, and on the other hand, higher standard values are to be expected by the more realistic determination in the future, the deviating development costs of alternative designs are estimated with the help of CO_2-emission penalties.

The allowed emission limit G_{CO_2} of a vehicle is calculated depending on its weight m_{veh} and the weight correction factor m_0 [15]:

$$G_{CO_2} = G_{CO_2,0} + 0.0457 \cdot (m_{veh} - m_0). \qquad (13)$$

For each automobile manufacturer its fleet limit value is calculated with the individual emission limit of their different vehicle models and their sales figures. If the fleet emission exceeds this value, a penalty of €95 will be due for each gram from 2019 onwards.

The production costs can be estimated from the sum of the material costs of the power supply system. Furthermore, the operating costs can be determined from the specific driving consumption and the respective energy costs of the vehicle. Taxes for petrol vehicles in Germany are also dependent on the specific CO_2-emissions and the engine displacement of the vehicle.

Table 2 Emission penalties and annual travel costs of the example vehicles at a running distance of 36,500km a year.

Parameter	unit	A	J	F
propulsion		ICE	ICE	electric
allowed emission limit	$[\frac{gCO_2}{km}]$	72.74	143.58	123.7
emission penalty per car in 2019	[€]	1,735	7,070	0
cost energy	$[\frac{€}{kWh}]$	0.1605	0.1605	0.2869
specific cost	$[\frac{€}{100km}]$	6.21	12.97	6.32
Travel expenses per year	$[\frac{€}{a}]$	2,267	4,736	2,308
taxes (Germany)	$[\frac{€}{a}]$	198	496	0

Table 2 shows the associated costs of the three vehicles given in table 1. First of all, table 2 shows that considering the weight correction factor of 1,372 of the year 2016, the emission limit of 95 gCO_2 per km, and today's standard emission, both vehicles with internal combustion engine are significantly above the emission limit. Considering the new test procedure mentioned above, the high penalty per sold vehicle could rise even further. The annual taxes calculated in Germany are also dependent on CO_2-emissions and therefore are the highest for the SUV.

For the comparison of different alternatives of the power supply network, an additional statistical calculation of the service costs is dispensed with since it is assumed that due to the low failure rates of electrical systems these will be similar for all alternatives.

5 Influence on Future System Design

While existing evaluation criteria such as unit costs and weight are static and can therefore be determined directly from the system design [19], all design criteria defined here depend on the dynamic operation of the vehicle and the electrical loads. Consequently, an operational model is essential for the determination of these criteria.

In order to evaluate the stability of the voltage, a worst-case scenario appears to be fitting, as is currently used when designing the energy source. However, for an accurate determination of the energy consumption and operating costs, the operation has to be taken into account over a longer period of time, as the ambient conditions such as light, weather and temperature change substantially over the course of a year and thus the use of the electrical loads varies. The created operating models of the loads must allow to adjust the environmental conditions, the route, and the driver in order to simulate the energy consumption of

the electrical network in a realistic manner.

In the established determination of the standard consumption, electrical loads are switched off if possible. Thus, they are not considered when measuring the standard consumption and emission.

If for the implementation of safety, critical functions voltage stability is an indispensable constraint, it must be fulfilled at all costs. In particular, in the case of speed-dependent power generation in a combustion vehicle, a minimum battery capacity is required to maintain the voltage limits at all times. The standard deviation allows to evaluate different designs. It can for example be used to define a maximum allowed deviation value for a design to be considered stable. If a stable architecture is to be selected, designs with high storage capacity are preferred. Considering the fail-operational request for a redundant power source for internal combustion vehicles, the power supply network will likely contain at least two batteries independent of the other design criteria. In addition, bus or ring topologies are preferable to the existing star topology since these allow a simple distribution of the energy sources in the network in order, for example, to ensure the supply of the components in the event of a fault in the supply line [20].

The defined cost analysis also shows that the unit costs over the emission penalty as well as the operating costs depend on the consumption and consequently on the energy efficiency of a vehicle. Consequently, future architectures follow a trade-off between an energy-efficient design and small unit cost. A reduction of the consumption of the energy supply system by 200 W of the SUV discussed in the previous section by means of a system more expensive by € 570 would be cost neutral due to the 6 g lower emissions and therefore reduced penalties. But an equivalent reduction of the emission by a weight reduction of 125 kg would only result in a reduction of the penalties by € 50, since the permissible emission limit also drops to 138.1 gCO_2 per km.

This shows that in this combination of evaluation criteria the optimization of the efficiency of the energy distribution is preferred to the optimization of the weight. In contrast, in the case of an electric vehicle, the optimization of the weight has a greater effect on the consumption, because of the high efficiency levels in the supply of the electrical network as well as the drive train. To minimize the operating costs, it does not matter how the reduction of the energy consumption is realized.

In summary, a redesign of the established power supply system to reduce the power consumption would not only lead to lower operating costs, but could also reduce the product price of ICE vehicles. In contrast to of today's cost optimization, higher costs per function through higher development or production costs might be acceptable as long the saved emission penalty costs are even or higher. These assessment criteria therefore call into question the established optimization approaches of the cable harness based on unit costs and weight.

Today, the introduction of a power bus with semiconductor switches instead of existing fuses in established vehicle models seems to be economically unprofitable due to the high development and material costs. If, however, the design criteria determined here are applied, the additional costs associated with the introduction would be acceptable if weight or efficiency savings in the electrical grid lead to equal or higher savings in emission penalties.

6 Conclusion

Electrification and digitalisation combined with new functions challenge the established e/e-architecture. Hence, it must change to meet future requirements. To be able to compare different alternatives of the power supply network, important design criteria are derived from prioritized stakeholder requests, taking into account the trends identified.

The stability of the voltage is an important boundary condition for the increasing integration of safety-relevant components into the vehicle. It is shown that using the standard deviation not only the voltage limits, but the temporal fluctuations in the assessment of the voltage stability of different design alternatives can be considered. This allows to evaluate different designs concerning their expected stability.

Moreover, consideration of the energy consumption of the whole vehicle allows to find an optimum compromise between minimal weight and minimal electrical losses. On the one hand, the energy consumption of the vehicle combined with the usual energy costs of the different drive alternatives determine the specific operating costs. On the other hand, the energy consumption will significantly influence the selling price of a vehicle in the future by means of the emission penalties if the emission limit values are exceeded. This means that higher development or production costs for a redesign of the established power supply system are acceptable as long as they are equal or below the saved penalty costs. An additional bonus is that a higher efficiency lead to lesser operating costs which are of high importance for fleet vehicles e.g. for car sharing especially.

Since a weight reduction also reduces the permissible emission limit values of a vehicle, an equivalent reduction of the emission over a more efficient energy distribution system has a higher effect on the penalty savings. Therefore, the found optimization criteria call into question the established optimization approaches of the cable harness based mostly on unit costs and weight.

In future research, we will investigate the extent to which different freedoms in shaping the energy network can influence and facilitate the fulfillment of these criteria. For this purpose, a modelling environment in MATLAB/SIMULINK is created, which allows to examine alternative designs of the power supply network with regard to the dynamic design criteria defined here and to derive optimal combinations.

References

[1] L. Braun, M. Armbruster, E. Sax, "Stakeholder Issues Concerning the Automotive E/E-Architecture" in 4th International Conference on Electrical Systems for Aircraft, Railway, Ship propulsion and Road Vehicles and International Transportation Electrification Conference, Toulouse France, 2016.

[2] E. Frank, R. Wilhelm, R. Ernst, A. Sangiovanni-Vincentelli, M. Di Natale, "Methods, Tools and Standards for the Analysis, Evaluation and Design of Modern Automotive Architectures" in Design, Automation and Test in Europe, 659-663, 2008.

[3] W. Schade, C. Zanker, A. Kuehn, S. Kinkel, A. Jaeger, T. Hettesheimer, T. Schmall, "Zukunft der Automobilindustrie", Buero fuer Technikfolgen-Abschaetzung beim deutschen Bundestag, AB152 (working report), 2012.

[4] Bundesverband CarSharing, "CarSharing-Jahresbilanz 2015: Wachstum und Konsolidierung im deutschen CarSharing-Markt", press release, 29.02.2016

[5] F.-J. van Audenhove, L. Dauby, O. Korniichuk, J. Pourbaix. "The Future of Urban Mobility 2.0: Imperatives to shape extended mobility ecosystems of tomorrow", Arthur D Little and UITP, 2014.

[6] M. Klauda, M. Schaffert, A. Lagospiris, G. Piel, S. Kappel, and M. Ihle, "Position of Points 2020 - Paradigm Shift in E/E Architectures", ATZ Elektron Worldw, **10**(02), 4-11, 2015.

[7] L. Braun, M. Armbruster, F. Gauterin, "Trends in vehicle electric system design: State-of-the Art Summary" in Vehicular Power and Propulsion Conference, Montreal Canada, 2015.

[8] T. Doersam, S. Kehl, A. Klinging, A. Radon, O. Sirch, "The New Voltage Level - 48V for Vehicle Power Supply", ATZ Elektron Worldw, **7**(01), 10-15, 2012.

[9] A. Graf, J. Kositza, T. Raithel, "Leading Edge Automotive Power Distribution", ATZelektronik wordlwide **8**(03), 14-17, 2013.

[10] S. Sommer, A. Camek, K. Becker, C. Buckl, A. Zirkler, L. Fiege, M. Armbruster, G. Spiegelberg, A. Knoll, "RACE: A Centralized Platform Computer Based Architecture for Automotive Applications", in International Electric Vehicle Conference, Santa Clara CA USA, 2013.

[11] G. Weiss, P. Schleiss, C. Drabek "Fail-operational E/E Architecture for Highly-automated Driving Functions", ATZ Elektron Worldw **11**(03), 16-21, 2016.

[12] W. Langhoff, M. Ernst, "Living wiring system bionics as a model for E/E architectures of the future", ATZ Elektron Worldw **9**(04), 40-45, 2014.

[13] J. Kloetzl, D. Gerling, "Stability in automotive power nets: Definitions, algorithms and experimental validation", in 14th European Conference on Power Electronics and Applications, Birmingham England, 2011.

[14] T. Cleff, Deskriptive Statistik und moderne Datenanalyse: Eine computergestuetzte Einfuehrung mit Excel, PASW (SPSS) und STATA, Gabler Verlag, 2012.

[15] EUROPEAN PARLIAMENT AND OF THE COUNCIL of 11 March 2014 amending Regulation (EC) No 443/2009 to define the modalities for reaching the 2020 target to reduce CO2 emissions from new passenger cars, Regulation (EC) No 443/2009

[16] U. Tietge, S. Daz, P. Mock, J. German, A. Bandivadekar, N. Ligterink, "From Laboratory to Road: A 2016 Update of Official and Real-World Fuel Consumption and CO2 Values for Passenger Cars in Europe", International Council on Clean Transportation, 2016.

[17] A. Steward, A. Hope-Morley, P. Mock, and U. Tietge, "Quantifying the impact of real-world driving on total CO2 emissions from UK cars and vans", element energy, 2015.

[18] S. Pischinger, U. Seiffert, Vieweg Handbuch Kraftfahrzeugtechnik, ATZ/MTZ-Fachbuch, Springer Vieweg, 2016.

[19] L. Brabetz, M. Ayeb, O. Baumgarten, "Tool-based Optimization of the Topology of an Electrical Distribution System (EDS)", SAE Int. J. Engines **9**(3),2016.

[20] U. Dieckhoefer, V. Hoffmann, U. Hornfeck, "Decentralised Wiring System Architecture Power Distribution Next Generation", ATZ Elektron Worldw **11**(01), 38-41, 2016.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 2, No. 3, 284-290 (2017)
www.astesj.com
Special Issue on Recent Advances in Engineering Systems

ASTES Journal
ISSN: 2415-6698

Gestural Interaction for Virtual Reality Environments through Data Gloves

G. Rodriguez*, N. Jofré, Y. Alvarado, J. Fernández, R. Guerrero

Laboratorio de Computación Gráfica (LCG), Universidad Nacional de San Luis, 5700, Argentina

ARTICLEINFO

Article history:
Received: 06 April, 2017
Accepted: 20 April, 2017
Online: 14 May, 2017

Keywords :
Human Computer Interaction (HCI)
Natural User Interface (NUI)
Non-Verbal Communication (NVC)
Virtual Reality (VR)
Computer Graphics

ABSTRACT

In virtual environments, virtual hand interactions play a key role in interactivity and realism allowing to perform fine motions. Data glove is widely used in Virtual Reality (VR) and through simulating a human hands natural anatomy (Avatar's hands) in its appearance and motion is possible to interact with the environment and virtual objects. Recently, hand gestures are considered as one of the most meaningful and expressive signals. As consequence, this paper explores the use of hand gestures as a mean of Human-Computer Interaction (HCI) for VR applications through data gloves. Using a hand gesture recognition and tracking method, accurate and real-time interactive performance can be obtained. To verify the effectiveness and usability of the system, an experiment of ease learning based on execution's time was performed. The experimental results demonstrate that this interaction's approach does not present problems for people more experienced in the use of computer applications. While people with basic knowledge has some problems the system becomes easy to use with practice.

1 Introduction

As a result of technological transition, VR has allowed to get over physical tools' limitations by triggering the evolution of traditional devices into new more simple and intuitive ones, thus fostering the emergence to *Natural User Interfaces* (NUI). These enhance the natural impression of living in a simulated reality by improving the interaction between users & VR environments.

NUI set a new HCI paradigm exploiting the skills that human being has developed throughout his life for social interaction. Though most social interaction among humans takes the form of conversation, there is a large sub-text to any interaction that is not captured by a literal transcription of the words that are said. All of the factors going beyond the verbal aspects of speech are called *Non-Verbal Communication* (NVC) and takes many forms, or modalities like: *facial expression*; *gaze* (eye movements and pupil dilation); *gestures*, *bodily movements*; *body posture*; *bodily contact*; *spatial behaviour gesture*; *non-verbal vocalizations*, among others.

Social interaction is a key element of modern virtual environments. VR systems that enhance the presence in social interactions must collect gestural, positional, sound and biometric user information. As a solution, most VR developers propose a NUI with virtual characters simulating real human social interactions (personifying user and computer). Additionally, as a result of a physical and mental need in user's interaction with computer games/applications, many games are designed to test the physical skills of its users setting gestures as the most common means to establish communication. In particular, studies on cognitive science state that hand gestures are which best express key information.

When developing a NUI, two different approaches are usually addressed: based on *tracking devices* or based on *vision*. In robotics, research focus on controlling a war robot or drones in high-risk areas or in remote-control surgery by tracking a robot hand. In medicine, virtual environments allow objectively quantify surgical performance, and accelerate the acquisition of baseline surgical skills without risk to real patients based on motion capture. These approaches have also greatly aided in sign language recognition research, to training of children to write, play instruments, to control objects, 3D modelling, among others. This paper's work presents a VR system built on NUI principles, allowing the simulation of gestures and animations of human hands based on the recogni-

*Ejército de los Andes 950, Tel: +54 (0266) 4520300 & gbrodriguez@unsl.edu.ar

https://dx.doi.org/10.25046/aj020338

tion of their movements using a data glove as an input device.

2 Related Work

From bibliography, most of VR and NUI related works try to increment the immersion level by replacing classical devices (such as mouse, keyboard, joysticks) for new technological device tracking or image processing [1, 2]. From these works, it has been identified the following specified shortcomings [3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15]:

- Study of manipulation in VR environments results in a gap for vision-based works.
- Related research to gesture recognition are mostly vision-based.
- Vision-based systems that perform gesture recognition are usually limited to a tracking zone, in addition to noise and high processing requests.
- Regarding gestures recognition, the combination of two or more gestures to get a new complex gesture for a complex task has not yet been achieved.
- Works based on data gloves are scarce where the main VR primitives are only studied and those existing works only emphasize in some of them.
- In many of mentioned works the *User Experience* has not been analyzed.

The present work is an extension of a previous work originally presented in [16] covering the following topics:

First of all, there is a virtual character that exhibits a correlational behaviour with a user (*Avatar*). It is explicitly controlled by commands specified by the user through a device as an external animator. This is the most common option while considering that VR applications based on NUI concepts and Gesture Interaction should provide certain facilities to the user for a more bearable interaction [17].

Secondly, during the interaction through the avatar each gesture is a control command associated to a process. The involved process tries to match incoming data from data glove with the modelled avatar [18, 19]. The avatar's virtual behaviour is controlled by the following standard primitives [20, 21]:

- **Navigation**, the displacement of the user in the virtual space and the "cognitive map" he/she builds on it.
- **Selection**, the action of user pointing to an object.
- **Manipulation**, the user modification of an object's state; and
- **System control**, the dialogue between the user and the application.

In this context, this work tries to meet the provided demands by mentioned gaps. Therefore, in addition to developing an HCI environment, the system provides a natural way to interact freely on a scenario through a data glove. Mainly, the system focus on the use of natural movements and gestural protocols (in HCI, set of gestures to establish communication between human and system) performed by hand [22, 23]. As an effective way to prove how positive can be this kind of interaction, it is worth to perform a user's experiment that shows the richness of the virtual experience. The next section show the offered features by system.

2.1 Natural User Interface based on Data Glove

The system uses a data glove (integrated sensor textile glove) as a gestural interface that allows a real-time interaction where the user communicates through an *avatar*. In consequence, according to the mentioned multi-dimensional control tools, excepting the selection tool, it has been developed the next basic features:

- **Navigation**. Most of the known and implemented navigation options used by game and VR applications developers are: "walk", "stop", "run", "jump", change different views of the scenario, among others. Particularly, this system has implemented the "walk", "stop" and change of view of the scenario (As shown in Figure 1 (a)). All of them are associated with a set of gestures (protocol) that performs each action. As an extra functionality of the system, it is possible to remotely control an air-plane toy included in the scenario (As shown in Figure 1 (b)). As it is possible to do two different types of navigation (avatar navigation and air-plane navigation), it is necessary to perform a gestural protocol to swap between them.
- **Manipulation**. While a user interacts naturally on the scenario, also has the need to perform handling actions in it. Basically, these actions are to be able to touch the virtual elements and grab them (As shown in Figure 1 (c)). Some involved factors for implementing this functionality are: physics treatment, proximity to objects (to grab) and avatar hand position.
- **System Control**. This control tool relates to the user-application dialogue. In this approach, user-application dialogue implies to control the system by using simple and natural gestures. This feature is achieved through navigation and manipulation tools.

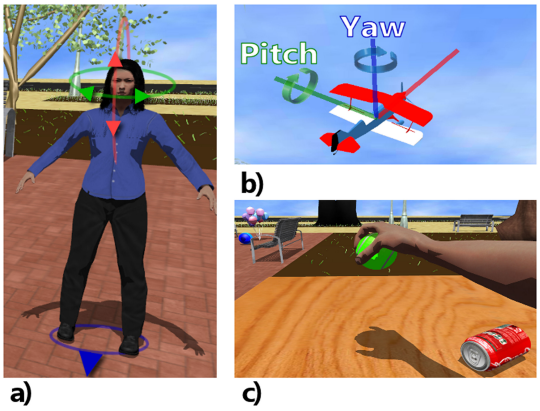


Figure 1: System's Features: a) Change of View Avatar, b) Airplane Navigation, c) Grasping in a work's table.

As a consequence of the correlational behaviour between user and avatar it would be nice to be able to visualise user hands and avatar hands as synchronised as possible. Thus, this system implements the next feature:

Real-time Imitation. This involves that avatar's hands can mimic user's hands through a refined model management, physics treatment and rendering tuning (frame rates). Realistic visualisation of the replicated gestures needs the model representation of hands and mapping of human hands' motion into the model which is not straight-forward due to different sizes and joint structures of bones for the human hands. It is necessary to consider that some movements are not feasible by the user then the avatar should not be able to do it.

3 System Overview

For a proper and efficient operation of the system is necessary to consider certain particular operating features (As shown in Figure 2). It is worth mentioning that the used input device was an integrated sensor textile glove *DG5 Vhand Data Glove 2.0* distributed by *DGTech Engineering Solutions* [24, 25]. This device has five embedded bend sensors which enables to accurately measure the finger movements, while an embedded 3 axes accelerometer allows for sensing both hand movements and hand orientation (roll and pitch).

The interaction between user and system implies that, at first, the direct **acquisition** must be performed by different sensors capturing performer's actions. So, to use the system, the user should go through an implemented **calibration** process that will allow obtaining greater sensitivity. In order to this, it is necessary to establish the appropriate user maximum and minimum flexion degrees of fingers with the glove. The corresponding values will be provided with gestural protocols that will allow flexion limits to be recorded. These protocols were named "Extended Hand " and "Closed Hand".

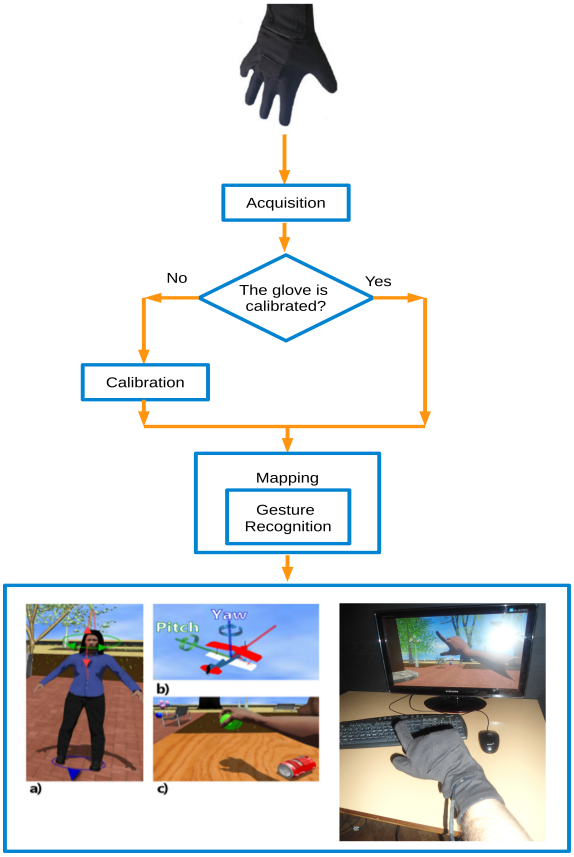


Figure 2: System's Flowchart

Thereafter, the new bending, orientation, and acceleration values after calibration should be mapped to values within the avatar model domain (**mapping**), i.e, are matched to model values getting a transition from physically detected movement (real domain) to avatar model image movement (virtual domain).

In relation to **gesture's recognition**, each gestural protocol was defined by a set of gestures, so each gesture consisted of the degrees, orientation and acceleration values previously stored in a database. The developed gestural protocols are *forward walking*, *backward walking*, *stop walking*, *revolver*, *grasp*, among others. Finally, to generate a particular event such as "walk" on the scenario, user must execute a forward walking or a backward walking. The system is responsible for controlling, monitoring and supervising in real-time all the user's actions/requests.

4 Evaluation and Results

McNamara and Kirakowski proposed a model of factors for understanding the interactions between humans and technology: *User Experience* (UX) and *Usability* [26]. The UX can be academically summarized as every aspect of the interaction between a user and a product. According to the most famous usability expert Jakob Nielsen, usability is a quality attribute that assesses how easy user interfaces are to use [27][28]. Also states that usability has five attributes which are: *ease of learnability*, *ability to remember*, *efficiency*, *error*

rate and *user satisfaction*.

On the other hand, it is known that there are two method types to evaluate an HCI experiment: qualitative and quantitative form.

Thus, knowing that gestures allow executing predefined actions very quickly and naturally, it is interesting to develop a test based on wasted time while executing main actions on the virtual world. On this way, an ease of learnability test was designed and aimed to measure the quality of NUI system. Basically, the test analyzes how well user uses the interface to arrive at a specific scenario's point and how well achieves to manipulate/modify the elements on a virtual work table (see Figure 1(c)).

4.1 Evaluation Test Design

An evaluation test was applied to a thirty participants' group between fifteen and thirty-nine years old. Before test, the system's goal and implemented gestural protocols were explained to each participant. Based on collected answers from a priori questionnaire about user's computer experience, participants were divided into two subgroups: *Basic* and *Intermediate-to-Advanced* users.

There were fifteen participants by subgroup. People that only use cellphone as a communication way and computer for their work were gathered together into *basic* experience group (Group 1). The *intermediate-to-advanced* experience group (Group 2) gathered digital natives (people who have grown up using technology like the Internet, computers and mobile devices). This group included girls and boys that have ever played video games.

During the experiment, each participant must execute two proposed tasks (ten repetitions per task). The main goal is to accomplish an evaluated challenge in time by timing the task. The following tasks were evaluated:

- **Task 1 (Navigation feature)**: Initially, beginning at a specified point on the scenario, user must move to another already specified point through the defined gestural commands.
- **Task 2 (Manipulation feature)**: The goal is to arrive at the virtual work table, to grasp an object and move it to another table's place.

4.2 Tasks Performances

Figure 3 shows the recorded average time from each performed repetition for the execution of task 1 to both evaluated groups. Curves behave similarly but with different initial average time (first repetition). In *basic* group, the average time remained approximately constant for all replicates, showing that the execution's ease was minimal. Particularly in the seventh repetition it arises the maximum local progress. In *intermediate-to-advanced* group, the execution time decreased progressively as increasing the repetitions, but in the sixth, seventh and eighth repetition there were certain regressions that are due to users' attempts to improve their execution times (it might be because Murphy's law).

Similarly, Figure 4 shows the recorded times for the execution of task 2. As long as, in both cases, the average time decreased significantly as repetitions increase, it reveals an implicit incremental curve of ease of learnability. In *basic* group, excepting the first and second repetitions, the execution's time decreased progressively in each repetition. In *intermediate-to-advanced group*, in first repetitions a high progress was gained, whereas in last repetitions the progression was constant.

While considering each step of curve as a line it is possible to obtain a gradient. Table 1 and Table 2 shows all gradients for each group in each task, meaning that when a gradient is zero, there is no variation in wasted execution time, so there is no progress and no regression at ease; when a gradient is under zero, wasted execution time was reduced, so there is a progress at ease; and finally, when a gradient is over zero, wasted time was increased, so there is a regression at ease.

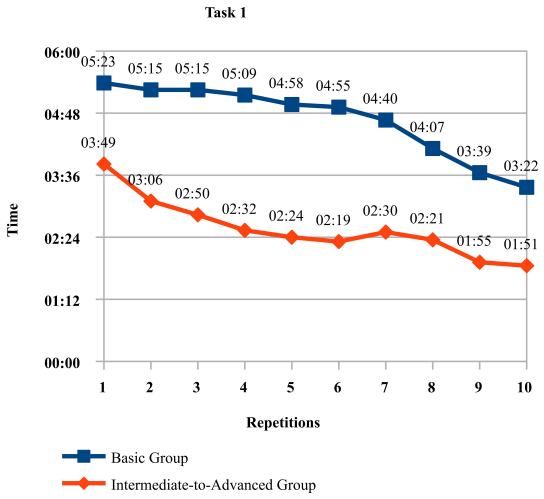


Figure 3: Task 1 Results.

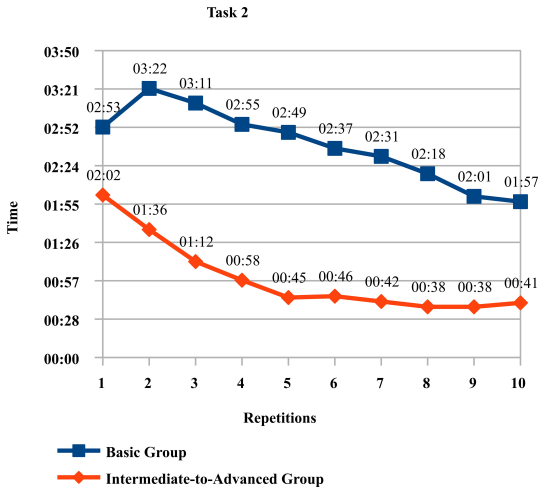


Figure 4: Task 2 Results.

Gradient (seconds) to task 1	
Group 1	Group 2
-00:08	-00:43
-00:00	-00:16
-00:06	-00:18
-00:11	-00:08
-00:03	-00:05
-00:15	00:11
-00:33	-00:09
-00:28	-00:26
-00:17	-00:04

Table 1: Gradient between repetitions for task 1

Gradient (seconds) to task 2	
Group 1	Group 2
00:29	-00:26
-00:11	-00:24
-00:16	-00:14
-00:06	-00:13
-00:12	00:01
-00:06	-00:04
-00:13	-00:04
-00:17	00:00
-00:04	00:03

Table 2: Gradient between repetitions for task 2

Average Time in seconds (standard deviation)	
Group 1	Group 2
-13,4 (11,1)	-13,1 (15,2)

Table 3: Average and Standard Deviation of gradients for task 1

Average Time in seconds (standard deviation)	
Group 1	Group 2
-6,2 (13,9)	-9 (10,7)

Table 4: Average and Standard Deviation of gradients for task 2

According to this established statements, *intermediate-to-advanced* group shows a more accelerating and increasing ease of learnability than *basic* as was expected; the average and standard deviation of the gradient (Tables 3 and 4) clarify this affirmation.

These results show that when user experience is basic, ease of learnability execution time tends to reflect constant values till a specific point of repetitions were gradients decreases exponentially showing mayor ease of learnability execution time.

5 Conclusion and Future works

In HCI, virtual worlds are developed for users to interact with a computer, and a solution to provide this is through of developing of avatars; an efficient, realistic and immersive alternative.

The process of designing 3D human correlated avatars must consider verbal and non-verbal communication. As well as non-verbal behaviour appears to be a primary factor in human-human interaction it is also considered for human-computer interaction.

Particularly, gestures are first tools of human communication. Therefore, with the intention of harnessing and enhancing the natural abilities of the human, at the present time, there exist works involved in adding user technology to implement Natural User Interfaces into multi-RVMedia systems [29].

This work focuses on use of hands as interaction's media. The development is based on the insertion of appealing devices such as data gloves within a virtual world, mostly, the research is addressed to real-time hand movements and gesture recognition providing the following features: hand tracking, navigation on stage and object manipulation, where for each feature, gestures play an important role and should be easy to learn and do not require any special attention from the user.

Presented preliminary user experiments consist in the evaluation of usability, in particular, the *ease of learnability* based on execution time where results are quantitative. From graphics it can be inferred that experienced users have no problems in adapting to new interface paradigms while basic users reflect a notable drawbacks till a threshold (60% repetitions). It will be a challenge to reduce this threshold by reducing system' s shortcomings. Nevertheless, the computer experience and previous practice do not have to determine the system simplicity and comfortability.

Due to some systems aspects are beta version, future works will be oriented to improve the environment issues, such as hardware platform. We consider that the system should provide the necessary multi-VR media and multi-platform structure and optimise the handling of 3D objects in the scenario using new libraries of physics or reprogramming of existing ones.

Gestural communication in virtual environments continues to be a major challenging but we think that are on the path to new and interesting research.

References

[1] Daniel Wigdor and Dennis Wixon. *Brave NUI World: Designing Natural User Interfaces for Touch and Gesture*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1st edition, 2011.

[2] James J. Park, Han-Chieh Chao, Hamid R. Arab-

nia, and Neil Y. Yen. *Advanced Multimedia and Ubiquitous Engineering: Future Information Technology - Volume 2*. Springer Publishing Company, Incorporated, 1st edition, 2015.

[3] G. Tofighi, H. Gu, and K. Raahemifar. Vision-based engagement detection in virtual reality. In *2016 Digital Media Industry Academic Forum (DMIAF)*, pages 202–206, July 2016.

[4] David Demirdjian, Teresa Ko, and Trevor Darrell. Untethered gesture acquisition and recognition for virtual world manipulation. *Virtual Reality*, 8(4):222–230, 2005.

[5] H. Cheng, L. Yang, and Z. Liu. Survey on 3d hand gesture recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(9):1659–1673, Sept 2016.

[6] Siddharth S. Rautaray and Anupam Agrawal. Vision based hand gesture recognition for human computer interaction: A survey. *Artif. Intell. Rev.*, 43(1):1–54, January 2015.

[7] Rick Kjeldsen and Jacob Hartman. Design issues for vision-based computer interaction systems. In *Proceedings of the 2001 workshop on Perceptive user interfaces*, pages 1–8. ACM, 2001.

[8] Frol Periverzov and Horea T. Ilieş. 3d imaging for hand gesture recognition: Exploring the software-hardware interaction of current technologies. *3D Research*, 3(3):1, 2012.

[9] Piyush Kumar, Jyoti Verma, and Shitala Prasad. Hand data glove: a wearable real-time device for human-computer interaction. *International Journal of Advanced Science and Technology*, 43, 2012.

[10] Shitala Prasad, Piyush Kumar, and Kumari Priyanka Sinha. A wireless dynamic gesture user interface for hci using hand data glove. In *Contemporary Computing (IC3), 2014 Seventh International Conference on*, pages 62–67. IEEE, 2014.

[11] Nik Isrozaidi Nik Ismail and Masaki Oshita. Motion selection and motion parameter control using data gloves. In *Games Innovation Conference (IGIC), 2011 IEEE International*, pages 113–114. IEEE, 2011.

[12] Xiaowen Yang, Qingde Li, Xie Han, and Zhichun Zhang. Person-specific hand reconstruction for natural hci. In *Automation and Computing (ICAC), 2016 22nd International Conference on*, pages 102–107. IEEE, 2016.

[13] Shitala Prasad, Piyush Kumar, and Kumari Priyanka Sinha. A wireless dynamic gesture user interface for hci using hand data glove. In *Contemporary Computing (IC3), 2014 Seventh International Conference on*, pages 62–67. IEEE, 2014.

[14] Inrak Choi and Sean Follmer. Wolverine: A wearable haptic interface for grasping in vr. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, UIST '16 Adjunct, pages 117–119, New York, NY, USA, 2016. ACM.

[15] Farid Parvini, Dennis Mcleod, Cyrus Shahabi, Bahareh Navai, Baharak Zali, and Shahram Ghandeharizadeh. An approach to glove-based gesture recognition. In *Proceedings of the 13th International Conference on Human-Computer Interaction. Part II: Novel Interaction Methods and Techniques*, pages 236–245, Berlin, Heidelberg, 2009. Springer-Verlag.

[16] L. N. Jofré, G. B. Rodríguez, Y. M. Alvarado, J. M. Fernández, and R. A. Guerrero. Non-verbal communication interface using a data glove. In *2016 IEEE Congreso Argentino de Ciencias de la Informática y Desarrollos de Investigación (CACIDI)*, pages 1–6, Nov 2016.

[17] Konstantinos Cornelis Apostolakis and Petros Daras. *Natural User Interfaces for Virtual Character Full Body and Facial Animation in Immersive Virtual Worlds*, pages 371–383. Springer International Publishing, Cham, 2015.

[18] Yikai Fang, Kongqiao Wang, Jian Cheng, and Hanqing Lu. A real-time hand gesture recognition method. In *2007 IEEE International Conference on Multimedia and Expo*, pages 995–998. IEEE, 2007.

[19] MA Moni and ABM Shawkat Ali. Hmm based hand gesture recognition: A review on techniques and approaches. In *Computer Science and Information Technology, 2009. ICCSIT 2009. 2nd IEEE International Conference on*, pages 433–437. IEEE, 2009.

[20] William R Sherman and Alan B Craig. *Understanding virtual reality: Interface, application, and design*. Elsevier, 2002.

[21] Marcio S Pinho, Doug A Bowman, and Carla M Freitas. Cooperative object manipulation in collaborative virtual environments. *Journal of the Brazilian Computer Society*, 14(2):54–67, 2008.

[22] Prashan Premaratne. *Human Computer Interaction Using Hand Gestures*. Springer, Singapore, 2014.

[23] Mohamed Alsheakhali, Ahmed Skaik, Mohammed Aldahdouh, and Mahmoud Alhelou. Hand gesture recognition system. *Information & Communication Systems*, 132, 2011.

[24] L. C. Pereira, R. V. Aroca, and R. R. Dantas. Flexdglove: A low cost dataglove with virtual hand simulator for virtual reality applications. In *2013 XV Symposium on Virtual and Augmented Reality*, pages 260–263, May 2013.

[25] H. Teleb and G. Chang. Data glove integration with 3d virtual environments. In *2012 International Conference on Systems and Informatics (ICSAI2012)*, pages 107–112, May 2012.

[26] Niamh McNamara and Jurek Kirakowski. Functionality, usability, and user experience: Three areas of concern. *interactions*, 13(6):26–28, November 2006.

[27] Regina Bernhaupt and Florian Floyd Mueller. Game user experience evaluation. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 940–943. ACM, 2016.

[28] Ali Trkyilmaz, Simge Kantar, M. Enis Bulak, and Ozgur Uysal. User experience design: Aesthetics or functionality? pages 559–565. ToKnowPress, 2015.

[29] Y. Alvarado, N. Moyano, D. Quiroga, J. Fernández, and R. Guerrero. *Augmented Virtual Realities for Social Developments. Experiences between Europe and Latin America*, chapter A Virtual Reality Computing Platform for Real Time 3D Visualization, pages 214–231. Universidad de Belgrano, 2014.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 2, No. 3, 291-301 (2017)
www.astesj.com
Special Issue on Recent Advances in Engineering Systems

ASTES Journal
ISSN: 2415-6698

Solving the Capacitated Network Design Problem in Two Steps

Meriem Khelifi[*,1], Mohand Yazid Saidi[2], Saadi Boudjit[2]

[1]*Laboratoire Réseaux et Systèmes, Université Badji Mokhtar, Algérie*

[2]*L2TI/Institut Galilée, Université Paris 13, 93430 Villetaneuse, France*

ARTICLEINFO

Article history:
Received: 06 April, 2017
Accepted: 30 April, 2017
Online: 14 May, 2017

Keywords :
Network Design Problem
Genetic algorithm
Modular capacity
Multicommodity flow problem
Linear programming

ABSTRACT

In this paper, we propose a two steps-based algorithm to solve the modular link version of the Capacitated Network Design Problem (CNDP) which consists to determine the optimal network that guarantees the routing of a set of commodities. In our proposition, CNDP is divided into two sub-problems: Network Design problem (NDP) and Network Loading Problem (NLP). In the first step, we solved NDP by using the genetic algorithms which select sets of network topologies. In the second step, NLP is solved with the use of Linear programming to evaluate and validate the best network topologies. Simulation results on three real network instances (Atlanta, France and Germany) show that the proposed algorithm is better and more efficient than the Iterative Local Search algorithm.

1 Introduction

To save resources (routers, optical fibers, etc.) networks should be efficiently designed. Diverse networks models were then defined and used to represent a wide range of issues in transportation, telecommunications, logistics, production and distribution networks. All these models consider a graph composed of nodes and edges (optical fibers, cables, etc). For a better use of these resources, networks designers should solve the modular Capacitated Network Design Problem (CNDP) which consists of selecting edges and allocating optimal capacities to route a set of commodities between source and destination pairs. Each edge of the graph has a potential set of module capacities with their associated costs, a fixed cost that is incurred only if the edge is selected, and a routing cost which is proportional to the amount of flows along the edge. Each commodity is defined by origin and destination nodes, and the amount of flow to be routed. The objective is to minimize three criteria: edge cost, modules and routing. These capacitated network design problems are NP-hard and very difficult to solve in practice. The CNDP is a particular case of the well known Multicommodity Network Design problem (MNDP), in which we distinguish an important number of special cases and extensions [1]. The most studied ones are:

- The unsplittable variant where the flow of each commodity is required to follow one route between the origin and the destination. This variant increases the difficulty of the problem [2].
- The expansion variant, where some edges already have an existing capacity.
- The fixed charge MNDP [3][4] in which the link capacities are known. Solving MNDP consists to determine the set of edges that should be opened in the final topology.
- The capacitated MNDP, where the number of modules to install on the edges are modeled by integers [1]
- The Network Loading Problem (NLP), where the number of module types is limited, each one with a given unit cost and capacity.

Various heuristics and exact approaches have been developed for designing capacitated networks. However, the heuristic approaches are more likely to be trapped in local optima, while the exact approaches are applied only to small or medium size problems. Due to the weaknesses of the two approaches and the increasing popularity of metaheuristic approaches, we have witnessed many metaheuristics being applied to network optimization problems. In this paper, we propose a novel metaheuristic that combines linear programming with Genetic Algorithms (GAs) to solve

[*]Meriem Khelifi, Laboratoire Réseaux et Systèmes, Université Badji Mokhtar, Algérie, khelifi.meriem@lrs-annaba.net

https://dx.doi.org/10.25046/aj020339

the CNDP problem. We recall that the genetic algorithms were extensively used to solve many difficult combinatorial optimization problems in industrial engineering and operation research. Genetic algorithms are one of the most powerful and broadly applicable stochastic search and optimization technique. They have achieved great advancement in related research fields, such as network optimization, combinatorial optimization, multi-objective optimization, etc. Our contribution consists in an efficient two steps-based heuristic that extends the approach in [5] by combining the GAs and Linear Programming (LP) to solve CNDP.

The remainder of this paper is structured as follows: related work is introduced in section 2, notations and mathematical formulation of the addressed problem are given in section 3. Section 4 describes and explains in details our proposed heuristic. Experimental results are discussed in section 5 where we compare our proposition against Iterative Local Search (ILS) algorithm. Finally, section 6 concludes the paper.

2 Related work

Capacitated Network Design Problem is one of the major research area in network optimization. It is related to two issues: Network Design Problem (NDP) and Network Loading Problem (NLP). In the NDP, the goal is to identify the network topology by selecting routers and links that interconnect them. Thus, the objective function aims to minimize the total constructive cost under some topological constraints. In this class of problems, the flow is not modeled and considered as uncapacitated. In the NLP, it is assumed that the topology is already established. Thus, solving NLP consists to search for the set of resources to allocate for the network components. These problems are complementary. Generally, NDP and NLP are solved separately though they should be combined together to optimize the resources.

One can say that most of network optimization problems can be seen as a kind of (1) NDP, (2) NLP or (3) a combination of both where the objective and the constraints may differ from one problem to another: connectivity [6] [7], limited budget [8], hop limit [2] [9], delay [10] [11], reliability [9] [10], and survivability [11]. The NLP in capacitated or uncapacitated case and with both single or multiple facilities is a special case of the well known Multicommodity Network Design Problem (MCND). Previous works on this problem can be classified as:

- Uncapacitated network design problems where on each network link, it is only possible either to open the link with an infinite capacity and a given fixed cost, or the capacity and cost are nil [12].
- Single facility capacitated network loading problem where, the capacity can be done by installing on each link an integer unit of a given basic facility [13].
- Two facilities capacitated network loading problems where the capacity can be achieved by means of two types of modules, each capacity has a specific cost [14].
- Multi-type facility capacitated network loading problems where various types of capacities can be installed on each link, each facility has a specific cost [8].

The early works on capacitated modular network problems were focused on the approximation methods. These methods define residual capacity and cut-set inequalities for single commodity and multicommodity cases on directed, undirected and bidirected link models [15][16]. Since these works consider that the underling network is established, they focus only on the determination of the facilities allowing the accommodations of flow demand. Their effectiveness depends on the size of the problem instance.
With the appearance of metaheuristics, both the NLP and the NDP have attracted some attention. The authors benefit from their efficiency to deal with more complex variants with real size instances. In [17], the author compared several neighborhood structures to solve the uncapacitated facility location problem. In [11], the authors proposed an evolutionary approach for capacitated network design considering cost, performances and survivability. The objective is to minimize network cost and packet delay. Kleeman et al. [10] used an evolutionary algorithm to solve multicommodity capacitated network design problem with an objective function optimizing costs, delay, robustness, invulnerability and reliability. A tabu search heuristic algorithm with real costs on facilities is developed in [18]. A firefly algorithm is proposed by Ragheb et al [8], they combined facility location and network design problem with multi-type of capacitated link and limited budget on facilities. Contreras et al. [19] presented a unified framework of general network design problems which combine location decision and network design decision.

3 Mathematical Formulation

Let $G = (V, E)$ be an undirected network where V is the set of vertices and E is the set of undirected edges. Let K be the set of commodities. Each one $k \in K$ is associated with a flow demand d_k and has a source denoted by the function $s(k)$ and a target $t(k)$. Let f_{ij} be the fixed cost of including edge (i, j) in the network, r_{ij} the unit variable flow cost on (i, j), and p_{ij} the pre-installed capacity on the edge (i, j).
The formulation of CNDP is shown below:

$$Min \;\; z(x,y,n) = \sum_{(i,j)\in E} r_{ij} \sum_{k\in K} \sum_{(i,j)\in E} x_{ij}^k + \sum_{(i,j)\in E} f_{ij} y_{ij} + \sum_{(i,j)\in E} \sum_{l\in L} c_{ij}^l n_{ij}^l$$

$$\sum_{(i,j)\in E} x_{ij}^k - \sum_{(j,i)\in E} x_{ji}^k == \begin{cases} 1, & i = s(k) \\ -1, & i = t(k) \\ 0, & \text{otherwise} \end{cases}, \forall k \in K \quad (1a)$$

$$\sum_{k\in K} (x_{ij}^k + x_{ji}^k)\, d^k \le p_{ij} + R_{ij}, \quad \forall (i,j) \in E \quad (1b)$$

$$R_{ij} \le \sum_{k\in K} d^k \, y_{ij}, \quad \forall (i,j) \in E \quad (1c)$$

$$R_{ij} \le \sum_{l\in L} m_l \, n_{ij}^l, \quad \forall (i,j) \in E \quad (1d)$$

$$x_{ij}^k \ge 0, \quad \forall (i,j) \in E, k \in K \quad (1e)$$

$$R_{ij} \ge 0, \quad \forall (i,j) \in E \quad (1f)$$

$$y_{ij} \in \{0,1\}, \quad \forall (i,j) \in E \quad (1g)$$

$$n_{ij}^l \in Z^+, \quad \forall l \in L, \forall (i,j) \in E \quad (1h)$$

This formulation is a mixed integer linear program which uses four types of variables: the first type is a binary design variable y_{ij} which is defined as $y_{ij} = 1$ if (i,j) is included in the network and $y_{ij} = 0$ otherwise. The second type is a continuous path flow variable x_{ij}^k, which represents the amount of flow of commodity k routed on link (i,j). The third type is an integer allocation module variable n_{ij}^l which represents the number of module type $l \in L$ (L is the set of potential modules) that should be allocated on edge (i,j). Each module l is characterized by a capacity m_l and an installation cost c_{ij}^l. The fourth type is a continuous variable R_{ij} representing the required flow capacity on link (i,j). A positive capacity $p_{ij} + R_{ij}$ on edge (i,j) implies that it is used to route demands in the two directions: from i to j or from j to i. This formulation corresponds to a general model that can deal with several variants of capacitated network design problems.

The objective function corresponds to the sum of the flow costs, the fixed costs of edges and the allocated module costs. These costs are relative to the problem that we deal with and are not all aggregated in some cases. For instance, the fixed charge problem MNDP includes only the edge costs. The modules and routing costs on edges are nil. Constraints (1a) consist of flow conservation equations, for each commodity k. Constraints (1b) specifies that the required capacity R_{ij} on each link (i,j) should be greater than the cumulated flows traversing the link minus the pre-installed capacity. Constraints (1c) forces the installation of link (i,j) (i.e, $y_{ij} = 1$) if the required capacity R_{ij} is positive (otherwise, $y_{ij} = 0$). Constraints (1d) specify that the total capacity of the allocated modules should be greater or equal to the required capacity on each link. (1e) and (1f) state out non-negativity constraints for the decision variables X and R. Constraints (1g) express the binary nature of the variables Y whereas constraints (1h) show that the module facility are allocated in discrete amounts.

If we assume that the edge and module vectors (y,n) are fixed and known, the main task in the above formulation becomes the search of a feasible flow that satisfies all the demands with the use of limited and fixed link capacities. Hence, the NP-hard CNDP problem can be reduced to the Capacitated Multicommodity Flow Problem (CMFP) which is solvable in polynomial time and formulated as follows:

$$Min \;\; z(x(\bar{y},\bar{n})) = \sum_{(i,j)\in E(\bar{y})} r_{ij} \sum_{k\in K} \sum_{(i,j)\in E(\bar{y})} x_{ij}^k$$

$$\sum_{(i,j)\in E(\bar{y})} x_{ij}^k - \sum_{(j,i)\in E(\bar{y})} x_{ji}^k == \begin{cases} 1, & i = s(k) \\ -1, & i = t(k) \\ 0, & \text{otherwise} \end{cases}, \forall k \in K \quad (2a)$$

$$\sum_{k\in K} (x_{ij}^k + x_{ji}^k)\, d^k \le p_{ij} + R_{ij}, \quad \forall (i,j) \in E(\bar{y}) \quad (2b)$$

$$R_{ij} \le \sum_{k\in K} d^k \, \bar{y}_{ij}, \quad \forall (i,j) \in E(\bar{y}) \quad (2c)$$

$$R_{ij} \le \sum_{l\in L} m_l \, \bar{n}_{ij}^l, \quad \forall (i,j) \in E(\bar{y}) \quad (2d)$$

$$x_{ij}^k \ge 0, \quad \forall (i,j) \in E(\bar{y}), k \in K \quad (2e)$$

$$R_{ij} \ge 0, \quad \forall (i,j) \in E(\bar{y}) \quad (2f)$$

The multicommodity flow problem formulation presented in the linear programming (2) is obtained by replacing the vectors (y,n) in the original CNDP formulation (1) by their fixed values $(\bar{y},\bar{n})$. In this reformulation, the objective function aims to minimise the routing of flows. Constraints (2a) are flow conservation equations. For simplicity and efficiency, we delete the variables R by replacing the constraints (2b), (2c), (2d) and (2f) by the following constraints (3a):

$$\sum_{k\in K} (x_{ij}^k + x_{ji}^k)\, d^k \le p_{ij} + \sum_{l\in L} m_l \, \bar{n}_{ij}^l \, \bar{y}_{ij} \quad \forall (i,j) \in E(\bar{y}) \quad (3a)$$

The previous constraints (3a) force the use of limited capacities by guaranteeing that the flows do not exceed the installed capacity (i.e, sum of the pre-installed capacity and module capacities). Constraints (2e) state out continuity and non-negativity of the decision variables X.

A solution to the CNDP can be viewed as (1) a binary assignment $(\bar{y})$ to each design variable, (2) an integer vector assignment $(\bar{n})$ to the allocation module

design variables and (3) an optimal solution for the multicommodity minimum cost flow problem $x^*(\bar{y}, \bar{n})$. In this way, the CNDP objective function value associated to a solution $(\bar{y}, \bar{n}, x^*(\bar{y}, \bar{n}))$ is the sum of the fixed cost of the open edges in $(\bar{y})$, the cost of the modules allocated $(\bar{n})$ and the objective function value of the CMFP associated to $x(\bar{y}, \bar{n})^*$. Thus, to solve the CNDP problem, we propose here to separate it into two sub-problems:
(1) link and resource selection which consists to choose the links, modules and their numbers by making decision on the vectors y and n,
(2) solve the CMFP problem (formulated in (2)) with respect of decisions made in (1) to determine the optimal routing that flows all the demands.

Note that sub-problem (2) can be solved in polynomial time. Thus, the challenge consists to solve the sub-problem (1) by identifying the best decision vectors (y) and (n) which allow to solve the flow sub-problem in (1). In other words, the decision vectors (y) and (n) should be selected so that the overall routing, link and module costs are optimal.

In the next section, we show how to solve jointly and efficiently sub-problems (1) and (2). In the first step of our proposition, we used the genetic algorithms to explore different potential solution areas for the decision vectors (y, n). Each solution $(\bar{y}, \bar{n})$ is then evaluated in the second step by solving CMFP problem.

4 Genetic Algorithm for CNDP

Genetic algorithms introduced by Goldberg et al. [20], are based on the mechanics of natural selection and genetic. They start with an initial set of random solutions, called a population. Each individual in the population, called a chromosome, represents a solution to the problem. The initial population evolves through successive iterations, called generations. A measure of fitness defines the quality of an individual chromosome. In each generation, chromosomes are evaluated by a fitness function, also called an evaluation function. After a number of generations, highly fit individuals, which are analogous to good solutions to a given problem, will emerge. Genetic algorithms consist of five components:

1. A method for encoding potential solutions into chromosomes;
2. A means of creating the initial population;
3. An evaluation function that can measure the fitness of chromosomes;
4. Genetic operators that can create the next generation population;
5. A way to set up control parameters; e.g., population size, the probability of applying a genetic operator, etc.

4.1 Individual representation

In the design of genetic algorithms, the encoding is the most important task. There are some methods to encode each individual in a population, such as binary encoding, integer encoding, etc. In this paper, we define a new encoding method called IME (Implicit Modular Encoding) that is relative to our modular case. An individual built by IME is shown in Figure 1. Each individual I is a matrix $I_{n,m}$, where n and m corresponds to the number of modules and to the number of edges respectively. Hence, $I[l_i][e_j]$ gives the number of module types l_i allocated on edge e_j.

Our encoding represents the decision vector n and implicitly the decision vector y. For example, $T[l_1][e_4] = 2$ (see Figure 1) means that we should allocate two modules on the edge e_4. Thus, we implicitly deduce that edge e_4 exists in the final topology.
When multiple types of modules are allowed, the edge exists if at least one module is allocated on it, i.e:

$$\begin{cases} x_e = 1 & if \ \sum_{l_{i=1}}^{n} T[l_i][e] > 0 \\ x_e = 0 & otherwise \end{cases}$$

In our example in Figure 1, edge e_1 will not be opened in the final network since $\forall i : T[l_i][e_1] = 0$.

4.2 Initial population

The choice of the initial population is a very important aspect of the whole search procedure. If it is too specific, then the search will be limited to a small region of the solution space leading to a local optimum. On the other hand, if the initial population is very

I

	e_1	e_2	e_3	e_4	.....			.....	.....	.....	e_m
l_1	0	1	4	3	.....			.....	1	1	0
l_2	0	2	1	0	.....			.....	3	0	2
⋮	⋮	⋮								⋮	⋮
l_n											

Figure 1: Individual with Implicit Modular Encoding

InitialPopulation

Inputs: S_0 , K
Local variables: P_0
$P_0 \leftarrow \emptyset$
Generate initial individual I_0 throughout encoding ILS solution with IME encoding
$P \leftarrow I_0$
foreach *module* $\{l_n \in L\}$ **do**
  foreach *link* $\{e_m \in E\}$ **do**
    if $I_0[l_n][e_m] > 0$ **then**
      Generate new individual $I \leftarrow I_0$
      $I[l_n][e_m] \leftarrow I_0[l_n][e_m] - 1$
      if $(CplexSolver(I, K) == True)$ **then**
        Add individual I to P_0
      end
    end
  end
end
Return P_0

Figure 2: S_0 is the ILS solution, K is the set of flow demands. P_0 is the initial population. L is the set of capacity modules. E is the set of links. I_0 is the initial individual. I is the new individual. $CplexSolver()$, that returns $True$ if it finds a feasible flow, is a procedure that solves the $CMFP$ on the individual (network) transmitted as a parameter.

diverse then the algorithm will spend valuable computational resources exploring a variety of promising areas of the search space.

There are two ways to generate an initial population: random initialization and heuristic initialization. Here, we applied the Iterative Local Search (ILS) heuristic [21] that provides one approximated solution for the CNDP problem.

To create a diverse initial population (c.f. algorithm depicted in Figure 2), we applied various perturbations on the individuals. After encoding the solution given by ILS heuristic according to IME, we obtain the first individual I_0. To form the rest of the initial population, we apply some perturbations on I_0. This results in the creation of new individuals, wherein there are some ones corresponding to unfeasible solutions. Obviously, only individuals corresponding to feasible solutions are added to the initial population. We recall that an individual corresponds to a feasible solution to CNDP problem if its decision vectors $\bar{n}$ and $\bar{y}$ allow the flowing of all the demands. Thus, an individual corresponds to a feasible solution $(\bar{n}, \bar{y})$ if the linear program in (2) has solutions. Otherwise, the individual corresponds to an unfeasible solution.

Creation of the initial population's individuals

We recall that our individual representation is based on edges and modules (see Figure (1)), where the

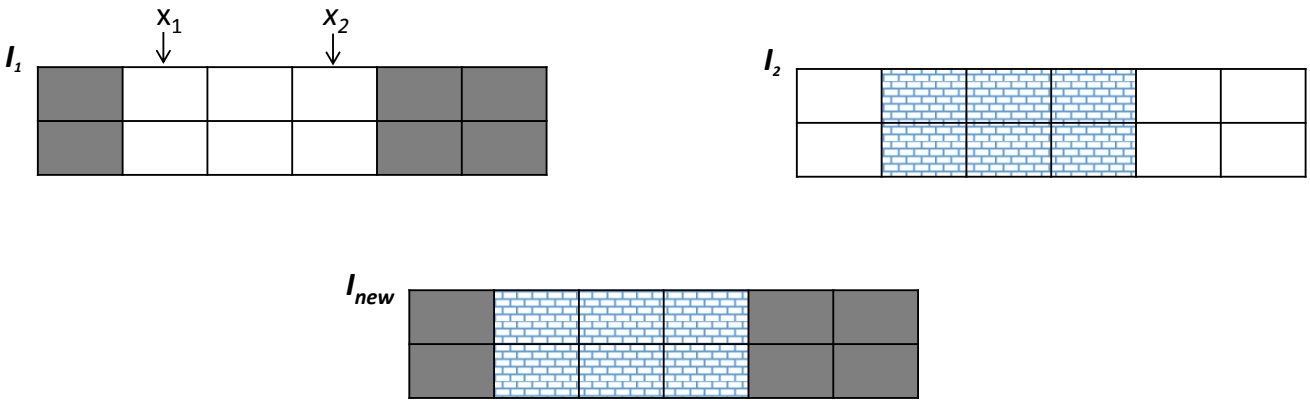


Figure 3: Crossover operator

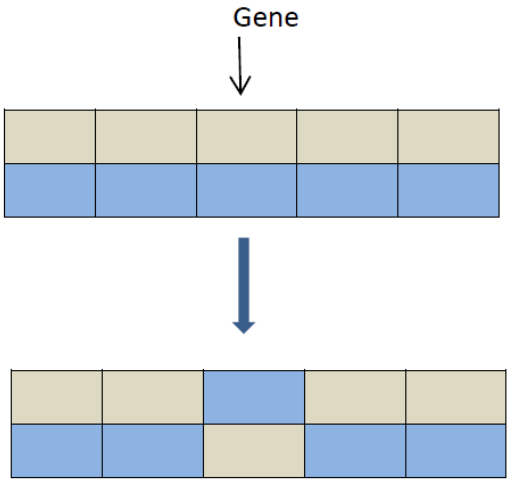


Figure 4: Mutation operator

columns represent the edges, and the rows represent the module types allocated on edges.

At initial stage, our population contains only one individual I_0. The other Individuals in the initial population are determined by applying some perturbations on the individual I_0.
A perturbation consists to delete one module from the I_0's edges. The complete procedure of generating the initial population is described in the following steps:

- for each module type $l_n \in L$,
- for each edge e from the set E

Genetic Algorithm

Inputs: S_0, K
Local variables: i, P , $Gbest$
$P \leftarrow InitialPopulation(S_0, K)$
$Gbest \leftarrow$ best individual in the population P
$Termination \leftarrow false$
while ! $Termination$ **do**
  $i \leftarrow$ size of population P
  while $i < Max\text{-}Size$ **do**
    $(I_1, I_2) \leftarrow$ RandomSelection(P)
    $I_{new} \leftarrow Crossover(I_1, I_2)$
    $I_{new} \leftarrow Mutation(I_{new})$
    if $(CplexSolver(I_{new}, K) == True)$ **then**
      Add individual I_{new} to population P
      Update Gbest
      $i \leftarrow i$+1
    end
  end
  P $\leftarrow$ CleanPop (P)
  Update(Termination)
end
Return $Gbest$

Figure 5: S_0 is the initial solution, K is the set of flow demands. $Gbest$ is the best value . P is the current population. $MaxSize$ is the fixed size of the population. I_1, I_2 and I_{new} are individuals. $CplexSolver()$, that returns $True$ if it finds a feasible flow, is a procedure that solves the $CMFP$ on the individual (network) transmitted as a parameter.

Problem Instance	$\|V\|$	$\|E\|$	$\|K\|$	Nbr	L	$MCost$	p_{ij}	r_{ij}
Atlanta	15	22	210	TF	1000, 4000	variant	yes	yes
France	25	45	300	SF	2500	fixed	no	no
Germany50	50	88	662	SF	40	variant	no	no

Table 1: The Instance Setting Parameters

- create a new individual I that corresponds to a copy of I_0,
- decrement the number of modules l_n on the edge e of individual I
- add the new individual to the population if it allows a feasible flow of demands.

4.3 Fitness function

The fitness function corresponds to the objective function of CNDP. It is computed as the sum of the allocated module costs, the fixed edge costs and the routing costs. Note that the first two costs are deduced from the individual representation whereas the routing costs are given by solving the CMFP linear program (2).

4.4 Genetic operators

The next populations are generated by applying the crossover and mutation operators described below.

4.4.1 Crossover

The crossover operator is responsible for combining two chromosomes so that a new offspring chromosome can be generated. In our proposition, we used a two point crossover: we randomly chose two integers in the individual length interval ($0 < x_1 \leq x_2 \leq m$) and two individuals (I_1 and I_2) in the current population, then we apply the two-point crossover operator to generate a new individual as shown in Figure 3. Typically, a new individual I_{new} is generated by selecting the modules of edges in $(e_1, e_{x_1}] \cup [e_{x_2}, e_m)$ from I_1 and (e_{x_1}, e_{x_2}) from I_2.

4.4.2 Mutation

An edge (that corresponds to a gene) and two modules types are randomly chosen in the chromosome. Then, the numbers of modules relating to the selected types of modules are exchanged on the chosen edge (see Figure 4).

4.5 The genetic algorithm

After explaining and detailing the basic components of our proposed genetic algorithm, we describe below its operation (see Figure 5 for instructions). In our algorithm, we first initialize the population through $InitialPopulation()$ procedure (see Figure 2). Then N successive populations are generated by applying the two-point crossover and mutation operators ($Crossover()$ and $Mutation()$).

As said previously, only individuals allowing a feasible multicommodity flow solution are added to the current population. This is verified by the running of $CplexSolver()$ procedure that solves the linear program in (2). The best solution $Gbest$ is updated at each generation and returned when the termination condi-

Instance	BS	ILS	$Gap\%$	GA	$Gap\%$
Atlanta	86492550	92904547	7.41	87959303	1.69
France	20200	21400	5.94	20600	1.98
Germany50	645520	719060	11.39	667840	3.45

Table 2: The ILS and GA solutions

Demand model	Undirected demand (U)
Link model	Undirected links (U)
Link capacity model	Modular link capacities(M)
Fixed-charge model	No fixed-charge cost (N)
Routing model	Continuous (C)
Admissible path model	All paths (A)
Hop limit model	No hop-limits (N)
Survivability model	No survivability (N)

Table 3: The model filter

Instance	Atlanta			France			Germany50		
	BS	*ILS*	*GA*	*BS*	*ILS*	*GA*	*BS*	*ILS*	*GA*
Total installed link capacities	294000	307000	300000	252500	270000	257500	7200	8000	7440
Total working flow	282338.5	281188	284503	246938	237952	240351	7140	7024	7265.83
Total Unused flow	11661.5	25812	15497	5562	32048	17149	60	976	174.17

Table 4: Working and unused capacities

tion is satisfied.
CleanPop() procedure allows to switch from one population to another by selecting individuals from the first population. It is based on elitist strategy, i.e., better is the fitness of the individual, greater is the probability of keeping that individual in the next population.

The algorithm stops its running after a fixed number of generations or when the result is not improved after a certain number of generations.

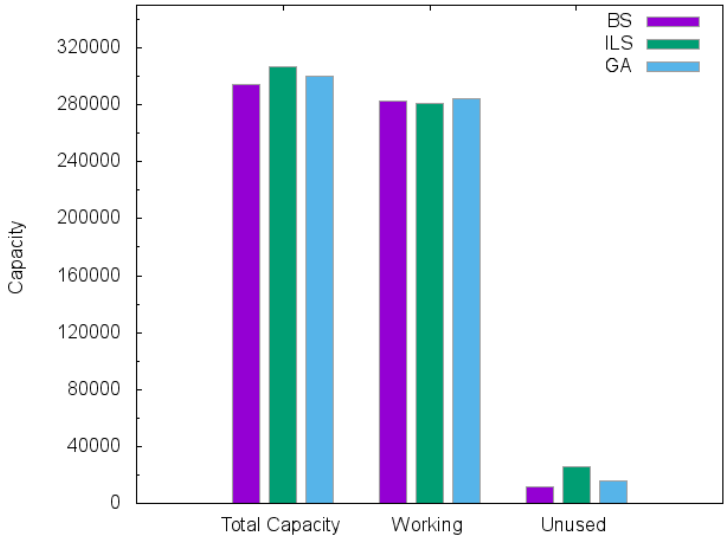


Figure 6: Atlanta network

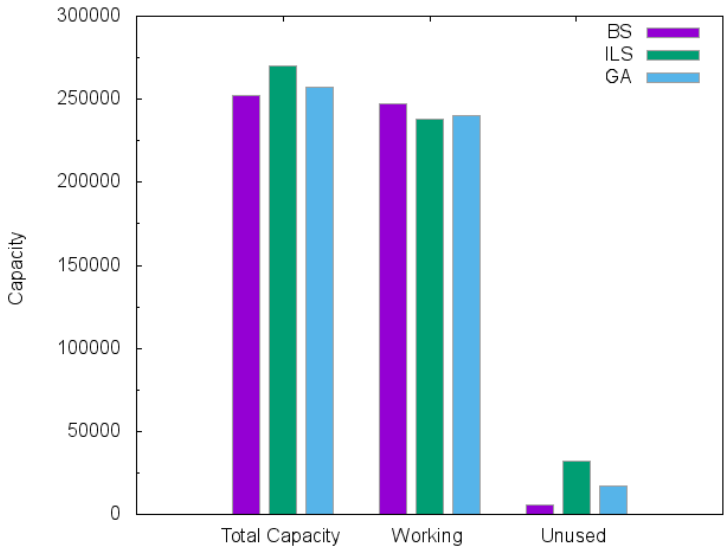


Figure 7: France network

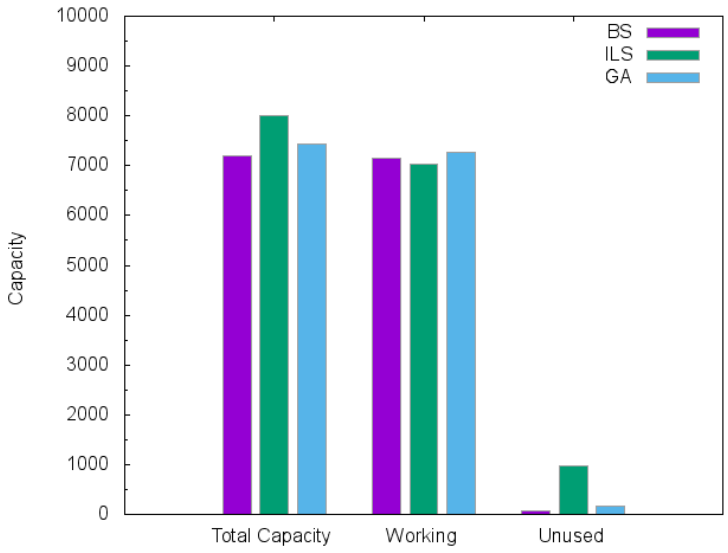


Figure 8: Germany50 network

5 Experiments and Results

In our experiments, we used three real world instances of network topologies including Atlanta, France and Germany50. All can be downloaded from $http://sndlib.zib.de$ [22]. We followed the model filter specified in Table 3. The population size is 50 and the number of generations is 15.
Each instance is characterized by the number of nodes $|V|$, the number of potential links $|E| = m$ and the number of traffic demands $|K|$. Table 1 summarizes the instance specification details. We classify them into two categories; instances with Single Facility allocation (SF) and instances with Two Facilities allocation (TF). The set of capacity modules L differs from one network instance to another. The allocation cost $MCost$ is variant on links except in France instance. Atlanta instance assumes a Pre-installed capacities p_{ij} on their potential links with a unit routing cost r_{ij}. See [22], for more details on the filter model and on the setting parameters.

In Table 2, ILS and GA correspond to the solutions obtained by iterative local search algorithm and our genetic algorithms-based heuristic respectively. We examine the quality of a given algorithm A (A could be GA or ILS) by computing its optimality gap (see equality 10) that is defined as the ratio between the difference of the A's cost and the Best Solution (BS) cost. Note that BS corresponds to the best solutions published in [22].

$$GAP(A) = \{Cost(A) - Cost(BS)\}/Cost(BS) * 100 \quad (4)$$

As depicted in Table 4, GA is better than ILS since it determines solutions more close to the best solutions than those of ILS. Concretely, the mean gap obtained with ILS is 3.5 times higher than the mean gap obtained with GA. This can be explained by the exploration of multiple solution areas with GA while ILS determines only a local optimum.

Figures 6, 7 and 8 show the allocated capacities and their usage for Atlanta, France and Germany50 networks respectively. We compared the total installed link capacities, the total working capacities and the total unused capacities for BS, ILS and GA solutions. We remark that the total installed link capacities in ILS and GA are more larger than the BS ones. This justifies the cost gap. On the other hand, ILS uses fewer working capacities than BS because ILS wastes and over-allocates module resources, leading to efficient routing. Indeed, instead of splitting flows and exploring the small unused capacities on links, ILS routes the majority of demands on shortest paths. With GA, the CPLEX optimizer tries to exploit the residual quantities on the allocated modules to route the flows. This leads to a bifurcation of demands on multiple paths that could be quite long, though routing costs slightly limit the path lengths.

6 Conclusion

In this paper, we proposed a two steps-based algorithm to solve the modular version of the Capacitated Network Design Problem (CNDP). In the first step, we applied genetic algorithms to select sets of promising networks (i.e., links with their modules) which are checked, evaluated and validated in the second step with the use of linear programming.

To explore the promising areas in the solution space, the genetic algorithms we used were adapted to treat and design efficient network topologies by: (1) defining a flexible and meaningful encoding scheme IME (Implicit Modular Encoding), (2) two point crossover and mutation operators and (3) an elitist population strategy. We generated the initial population by combining an iterative local search al-

gorithm with a heuristic procedure that modifies the number of modules on the network links. Successive populations were then generated by applying two point crossover operator which creates a new network from two other ones by selecting some edges from the first network and the rest from the second network. A mutation operator is possibly applied on links to exchange the module types. Finally, a probability-based elitist population strategy chooses promising network topologies that we combine to determine the next population. In our proposition, the networks are associated with probabilities in a way they guarantee an offspring often coming from the best individuals.

To measure the fitness of a solution, polynomial time linear program which determines the routes by searching for feasible flows on the corresponding network is used.

Simulations results confirm that the combination of genetic algorithms and linear programming in two steps is satisfactory and efficient. Indeed, the results show clearly that our proposition outperforms the iterative local search heuristic and determines solutions close to the known best ones.

References

[1] A. Frangioni and B. Gendron, "0-1 reformulations of the multicommodity capacitated network design problem," *Discrete Applied Mathematics*, vol. 157, no. 6, pp. 1229 – 1241, 2009.

[2] B. Thiongane, J.-F. Cordeau, and B. Gendron, "Formulations for the nonbifurcated hop-constrained multicommodity capacitated fixed-charge network design problem," *Computers & Operations Research*, vol. 53, no. 0, pp. 1 – 8, 2015.

[3] I. Rodríguez-Martín and J. J. Salazar-González, "A local branching heuristic for the capacitated fixed-charge network design problem," *Computers & Operations Research*, vol. 37, no. 3, pp. 575–581, 2010.

[4] D. C. Paraskevopoulos, T. Bektaş, T. G. Crainic, and C. N. Potts, *A Cycle-Based Evolutionary Algorithm for the Fixed-Charge Capacitated Multi-Commodity Network Design Problem*. Tech. rep., Centre interuniversitaire de recherche sur les réseaux d'entreprise, la logistique et le transport, Montreal, 2013.

[5] M. Khelifi, S. Boudjit, and M. Y. Saidi, "Two level evolutionary algorithm for capacitated network design problem," in *13th IEEE Annual Consumer Communications & Networking Conference, CCNC16, Las Vegas, NV, USA*, January 2016, pp. 299–300.

[6] A. Balakrishnan, M. Banciu, K. Glowacka, and P. Mirchandani, "Hierarchical approach for survivable network design," *European Journal of Operational Research*, vol. 225, no. 2, pp. 223–235, 2013.

[7] M. Abd-El-Barr, "Topological network design: A survey," *Journal of Network and Computer Applications*, vol. 32, no. 3, pp. 501 – 509, 2009.

[8] R. Rahmaniani and A. Ghaderi, "A combined facility location and network design problem with multi-type of capacitated links," *Applied Mathematical Modelling*, vol. 37, no. 9, pp. 6400–6414, 2013.

[9] T. Liu, W. Yang, and J. Huang, "Reliable network design problem under node failure with benders decomposition," *Applied Mathematics*, vol. 5, p. 241, 2014.

[10] M. Kleeman, G. Lamont, K. Hopkinson, and S. Graham, "Solving multicommodity capacitated network design problems using a multiobjective evolutionary algorithm," in *IEEE Symposium on Computational Intelligence in Security and Defense Applications*, 2007, pp. 33–41.

[11] A. Konak and A. Smith, "Capacitated network design considering survivability: an evolutionary approach," *Engineering Optimization*, vol. 36, no. 2, pp. 189–205, 2004.

[12] T. Magnanti, P. Mireault, and R. Wong, *Tailoring Benders decomposition for uncapacitated network design*. Springer, 1986.

[13] F. Barahona, "Network design using cut inequalities," *SIAM Journal on optimization*, vol. 6, no. 3, pp. 823–837, 1996.

[14] H. Kerivin, D. Nace, and T.-T.-L. Pham, "Design of capacitated survivable networks with a single facility," *IEEE/ACM Trans. Netw.*, vol. 13, no. 2, pp. 248–261, 2005.

[15] A. Atamtürk, "On capacitated network design cut–set polyhedra," *Mathematical Programming*, vol. 92, no. 3, pp. 425–437, 2002.

[16] A. Atamtürk and D. Rajan, "On splittable and unsplittable flow capacitated network design arc–set polyhedra," *Mathematical Programming*, vol. 92, no. 2, pp. 315–333, 2002.

[17] D. Ghosh, "Neighborhood search heuristics for the uncapacitated facility location problem," *European Journal of Operational Research*, vol. 150, no. 1, pp. 150–162, 2003.

[18] O. Brun, J. Garcia *et al.*, "A tabu search heuristic for capacitated network design," in *IEEE Symposium on Computers and Communications,ISCC*, 2009, pp. 461–467.

[19] I. Contreras and E. Fernández, "General network design: A unified view of combined location and network design problems," *European Journal of Operational Research*, vol. 219, no. 3, pp. 680–697, 2012.

[20] D. E. Goldberg and J. H. Holland, "Genetic algorithms and machine learning," *Machine learning*, vol. 3, no. 2, pp. 95–99, 1988.

[21] M. Khelifi, M. Babes, S. Ghanemi, M. Y. Saidi, and S. Boudjit, "Hybrid heuristic for capacitated network design problem," *J. High Speed Networks*, vol. 21, no. 4, pp. 313–330, 2015.

[22] S. Orlowski, R. Wessäly, M. Pióro, and A. Tomaszewski, "Sndlib 1.0 survivable network design library," *Networks*, vol. 55, no. 3, pp. 276–286, 2010.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 302-320 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

A Computationally Intelligent Approach to the Detection of Wormhole Attacks in Wireless Sensor Networks

Mohammad Nurul Afsar Shaon, Ken Ferens*

University of Manitoba, Electrical and Computer Engineering, R3T2N2, Canada

ARTICLE INFO

Article history:
Received: 06 March, 2017
Accepted: 26 April, 2017
Online: 14 May, 2017

Keywords:
Artificial Intelligence
Wormhole attack
Non-uniform Sensor distribution
In-band wormhole attack
Out of band wormhole attack
Neighborhood counts
Average residual energy pattern

ABSTRACT

A wormhole attack is one of the most critical and challenging security threats for wireless sensor networks because of its nature and ability to perform concealed malicious activities. This paper proposes an innovative wormhole detection scheme to detect wormhole attacks using computational intelligence and an artificial neural network (ANN). Most wormhole detection schemes reported in the literature assume the sensors are uniformly distributed in a network, and, furthermore, they use statistical and topological information and special hardware for their detection. However, these schemes may perform poorly in non-uniformly distributed networks, and, moreover, they may fail to defend against "out of band" and "in band" wormhole attacks. The aim of the proposed research is to develop a detection scheme that is able to detect all kinds of wormhole attacks in both uniformly and non-uniformly distributed sensor networks. Furthermore, the proposed research does not require any special hardware and causes no significant network overhead throughout the network. Most importantly, the probable location of the malicious nodes can be identified by the proposed ANN based detection scheme. We evaluate the efficacy of the proposed detection scheme in terms of detection accuracy, false positive rate, and false negative rate. The performance of the proposed algorithm is also compared with other machine learning techniques (i.e. SVM and regularized nonlinear logistic regression (LR) based detection models). The simulation results show that proposed ANN based algorithm outperforms the SVM or LR based detection schemes in terms of detection accuracy, false positive rate, and false negative rates.

1. Introduction

Wireless sensor networks (WSNs) consist of self-directed devices (i.e., sensor nodes), which are used in a collective manner to monitor environmental or physical phenomena in a remote and/or hostile environment. Spatially distributed autonomous sensor nodes are capable of communicating amongst themselves in order to transfer sensed data to a base station. The WSN is a type of ad-hoc network that has gained popularity for its versatile application in military and civil domains such as battlefield monitoring, tracking objects, healthcare, and home automation[1].

In WSN, a larger number of sensor nodes perform an assigned task in a hostile environment without any human intervention. Since sensor nodes use a known in-band radio channel for communication and are usually deployed in a hostile or remote environment, therefore, WSNs are prone to various security threats like a Sinkhole attack, Sybil attack, and wormhole attack. WSN has several vulnerabilities that an attacker can exploit to obtain access to the network. Implementing security measures like data encryption is not practical solution for the most WSNs. However, research on security measures has made certain progress in secure localization algorithm, lite weight routing protocols, and secure data aggregation scheme. However, those security mechanisms don't provide any protection against any attack from a legitimate node. If an attacker gains control over a few legitimate nodes, full access would be gained to the data traveling through these compromised nodes. The attacker may achieve the capability to modify the contents of the control

*Corresponding Author: Ken Ferens, University of Manitoba, Electrical and Computer Engineering, R3T2N2, Canada | Email: Ken.Ferens@umanitoba.ca

packets later by extracting the cryptographic contents from the captured data packets.

Wormhole attack is recognized as one of the most detrimental security threats for any routing protocol of WSNs [2-3]. The wormhole attack can be easily launched by taking over at least two legitimate nodes from the two distant parts of the network or deploying two nodes with superior capabilities (e.g. directional antenna, larger radio range) in two distant places of the sensor field. Wormhole nodes are connected through a virtual tunnel which can be implemented in numerous ways (e.g. high-quality channel, packet encapsulation or packet relay and high powered transmission) [4]. This direct low latency tunnel is known as a *wormhole link* [5]. A *wormhole link* creates an illusion in the network that these two colluding nodes are located within their communication range, but, in fact, their physical locations are very far apart.

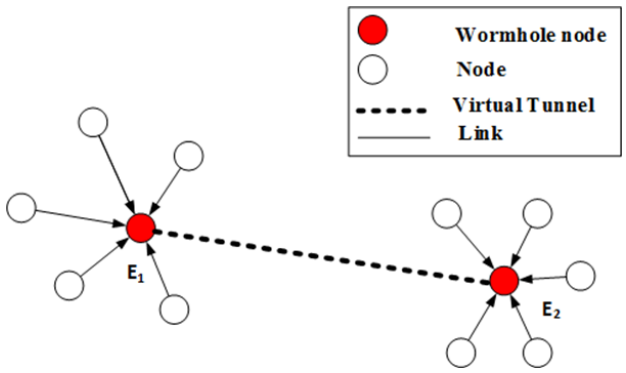


Fig. 1 Depiction of the structure of a wormhole attack.

By creating this unauthorized link, wormhole nodes gain the ability to circulate false route information into the network that they are few hops away from the base stations. This illusion drives other sensor nodes to transmit data packets to the base station through the wormhole nodes. Wormhole attack disrupts the existing network data flow in order to monitor and capture the data packets passing through it. As shown in Fig. 1, the E_1 and E_2 wormhole nodes, connected by a *wormhole link*, capture the data packets from one terminal of the virtual tunnel and retransmit them to another terminal of that link.

Subsequently, this wormhole attack becomes so severe that it might destroy the network or hamper the usual operation of the network by the selective dropping of packets, manipulation of traffic, or modify data packets without revealing their identities.

Therefore, detection of wormhole nodes is an essential task for ensuring the security of wireless sensor networks. It is a very simple task to implement wormhole attack, but a very difficult task to detect an infected network since wormhole nodes retransmit valid packets into the network. Most of the existing countermeasures use the distance bounding technique, direction, and location abnormality among claimed neighbor nodes as detection attributes to fight against wormhole attack. To gain a certain level of accuracy, some existing schemes uses complex and highly advanced devices such as directional antenna[6], GPS[7], or ultrasound for distance measurement [8]. However, those special devices are very costly for practical deployment. A few statistical wormhole detection schemes based on hop count [9], node connectivity [5], or neighborhood count [10-12] are proposed that do not need any special hardware. However, they usually include a hardware supported scheme as a secondary approach. Furthermore, centralized statistical wormhole detection [10] may cause significant network and communication overhead in contrast to a distributed statistical approach [11]. In the network connectivity based wormhole attack detection schemes [5,13-14], the positions of neighboring nodes are estimated from the received signal strength (RSSI) by each node, which sends this information to the base station. By doing this, the network layout is determined by the base station and compared with the given network layout. This approach also causes a significant amount of control packets flow to the base station. Moreover, it is prone to distance estimation errors. Furthermore, neighborhood-based wormhole detection schemes [10,12] may not detect wormhole attack if the wormhole nodes are located in a sparsely populated area and caused the significant flow of packets to the base station. In addition, their performance in non-uniform sensor distribution is in question.

In recent years, network anomaly detection schemes have been increasingly using artificial intelligence to improve detection accuracy. An artificial neural network (ANN) is a very simplified information processing model that aims to grossly imitate the human brain function. An ANN consists of interconnected processing units and works in a parallel fashion to find a solution to a particular non-linear problem. The adaptive and self-learning ability of an ANN help to increase the competence of an anomaly detection model [15]. Moreover, ANNs have been widely deployed to deal with pattern recognition and classification problems [16].

In this paper, we introduce a novel detection scheme based on an ANN using '*neighborhood count*' and '*average residual energy pattern of neighbors (AREPN)*' as detection attributes. The proposed detection scheme is able to detect wormhole attacks in both uniform and non-uniform sensor distributions and does not need any special hardware. Here, we have introduced a mobile node, called as detector node (D_N) that visits randomly chosen locations within the region of interest and collects two featured data samples, along with coordinate for each site visited. When the detector node D_N moves into a wormhole infected zone, this paper *theorizes* that the collected number of neighbors increases *abnormally* (uniform network scenario) or slightly *abnormally* (non-uniform network scenario), compared to a non-infected zone, in which the counts change *normally*. This paper also introduces a new detection attribute, named AREPN that significantly

decreases at the wormhole infected zone compare to non-infected zone. D_N captures these features as evidence of the presence or absence of a wormhole attack. The gathered dataset is used for training in the proposed ANN based detection scheme. After the training phase, the test data samples are fed into the ANN, and the scheme decides if there is any wormhole attack in the network or not.

We have studied in detail and compared the performance of the proposed algorithm with SVM and LR based detection schemes through simulations. Our simulation results confirm that an ANN based wormhole detector can detect wormhole attacks with higher precision and accuracy as compared to the non-linear logistic classification algorithm.

The rest of this paper is arranged as follows: Section 3 presents a detail of wormhole attack and its classification. We discuss literature review, ANN, SVM and LR in Section 3, 4, 5, and 6, respectively. The proposed ANN-based detections scheme described in detail in Section 7. The evaluation results are discussed in Section 8. Section 9 concludes the paper and provide a scope of future work.

2. Wormhole attacks

A wormhole can be a severe attack against any packet routing protocol, especially in ad-hoc and wireless sensor networks. It is very difficult to detect and to take preventive measures against wormhole attack since the malicious nodes behave as legitimate nodes and initially do not perform any illegal activity in the network[2]. The word 'wormhole' means the creation of any shortcut path between two far apart points in the space-time [17]. Thus, the concept of 'wormhole' is used as a tool to launch this attack aiming to spoil the existing routing protocol.

The wormhole attack starts by compromising at least two nodes from the sensor network by hacking or, deploying two nodes with superior capabilities (e.g. directional antenna, larger radio range) in two distant places of the network [18]. This attack would be more devastating if the aggressor launches the attack with multiple nodes. Although, wormhole attack can be launched with the single node by broadcasting received packets with high power level [19-20]. Those malicious nodes are known as '*wormhole node'*. Furthermore, to launch the attack, the wormhole nodes connect themselves through a low latency communication link, which is called the wormhole link. In some literature, this link is also named as wormhole tunnel [11]. The wormhole nodes gain unprecedented access to the network by forming this low latency link. This wormhole link can be formed in numerous ways, such as packet encapsulation, high power transmission, wired link and out of band radio link [4]. Packet encapsulation is the most prominent way to establish wormhole link in the network where smallest hop count is used as metric to select the best route.

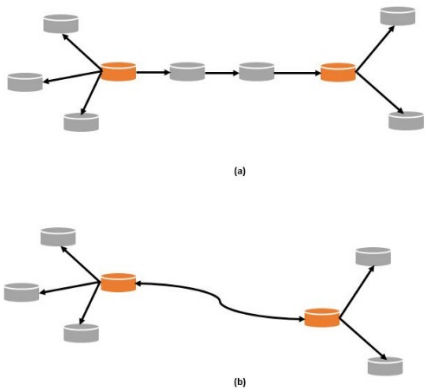


Fig. 2 The depiction of (a) 'in-band' and (b) 'out-band' wormhole attack.

The wormhole attack can be categorized into two classes based on the wormhole link formation. In the '*out band' wormhole* [21], two external nodes are deployed with higher communications and computational power than deployed sensor nodes. This kind of attack is hard to establish due to the requirement of the special hardware. In this case, the adversary connects two separate regions of the network using a directional antenna or wired link. Similarly, in the *'in band'* wormhole, the adversary compromises at least two legitimate nodes from the different zone of the network. One of these is usually located close to the base station so that it could disguise its adjacent nodes by advertising fabricated routing information. In contrast to '*out of band'* wormhole attack, adversary uses packet encapsulation technique to create wormhole node rather than a directional antenna or wired link [9]. In this circumstance, two compromised nodes are connected through several legitimate nodes between them.

Once the wormhole link is functional, one of the colluding nodes transmits the data packets, collected from the one part of the network, towards another malicious node. The other malicious node broadcasts those received packets into its radio range [22]. The wormhole node influences those nodes, who normally multiple hops away from it, to send data packets via wormhole by convincing them that they are few hops away from the base station [23]. In the other words, due to the high-speed wormhole link, the received data packets (by wormhole nodes) would travel faster from one part of the network to another part of the network than a usual multi-hop route. This illusion would disrupt existing packet routing mechanisms.

At the initial stage, wormhole node eavesdrops or captures the packets passing through it for further analysis and retransmits them to another wormhole node. When the wormhole attack begins, malicious nodes do not know about the cryptographic keys are being used in the network. If this malicious node starts dropping the packets without knowing the content of the packets, the target of compromising the integrity and confidentially would not be achieved. On the other hand, the dropping of the packets might rise suspicion of those nodes who have relayed the packet

through wormhole node. Therefore, there is a chance of being detected as a malicious node in the network. In fact, Wormhole node tries to place itself in most of the route without revealing its identity. However, the initial phase of wormhole attack is called *hidden wormhole attack* [20]. In the *hidden mode of attack*, the wormhole does not appear in the routing table. Still, wormhole node is able to establish the denial of service (DOS) [24] attack by dropping packets in the *hidden mode of attack*.

Once the wormhole node breaks the encryption technique by analyzing the recorded packets, then the aggressor takes the attack to the new level. In the *active mode*, the malicious node takes part in the routing mechanism as legitimate node [20] and starts modifying or dropping the packets passing through it [4]. In some cases, wormhole node drops the selected or critical packets to interrupt the usual operation of the network [21].

2.1. The impact of wormhole attack:

Wormhole attack is considered as server threats for routing protocols of WSNs. This attack usually occurs in the network layer and immune to encryption techniques. The wormhole attack is able to degrade the performance of the routing protocol and compromises the integrity and confidentiality of the data packets traveling throughout network [25]. Once wormhole nodes get the access to the network, the adversary can drop the packet selectively or delay the transmission of critical packets for the system in order to destabilize the system performance [21]. The aggressor tries to establish the denial of service attack (DOS) and attempts to compromise the integrity and confidentiality of the network.

During the active mode of this attack, the malicious nodes become the sinkholes [4]. However, other nodes around wormhole send data packets without aware of the fact that they are the victims of the wormhole attack. Hence, a significant amount of data traffic is passing through the malicious node and attacker can control and monitor the packets without having multiple observation points in the network. Some literature also suggests that wormhole node also lessens the throughput of the network by the selective dropping of the packets [4]. Furthermore, wormhole node also can turn on and off the wormhole link randomly [4]. This event creates instability in the routing service and causes a significant amount of control packets flow throughout the network.

In general, the routing protocol of the wireless sensor network can be categorized into two classes, such as 'pro-active' routing protocol and 'on-demand' routing protocol. Routing updates are transmitted periodically in the pro-active routing protocol, whereas on-demand routing protocol searches the route to a specific destination when it is necessary. However, the wormhole attack is successful to invade the network accessibility for the both class of wormhole attack [26]. Some literature on wormhole attack mentions that two wormhole nodes are able to attract more than 50% data traffic towards them directed to the base station[6][27].

3. Related works

As the wormhole attack can be launched from legitimate nodes and relays the valid packets into the network; therefore, it is reasonably difficult to detect it from the infected network. Furthermore, lightweight cryptographic solutions are incorporated into most of the routing protocol to prevent false data packets injection into the network. As the legitimate nodes are compromised by the adversary and the wormhole node doesn't change the packet's content at the initial stage, so the wormhole node easily passes the cryptographic test. The wormhole attack is a simple task to launch, but very difficult to identify. Many researchers have been working on this field to develop efficient wormhole detection schemes based on the geographical locations, transmission time, connectivity graph, neighborhood counts and radio fingerprint.

3.1. Distance consistency based approach:

Most of the researchers in this field try to detect the wormhole attack by distance bounding techniques. In these approaches, two communicating nodes are allowed to determine the distance between them; based on message traveling time information, transmission time information, and geographical information. Sometimes sensor nodes are equipped with specialized hardware like the directional antenna, GPS, ultrasound[8] to measure the distance between two adjacent sensor nodes. However, these schemes are considered impractical due to the addition of the special hardware and their performance in the sparse sensor network. These schemes may not perform well in the presence of *'out of band'* wormhole nodes.

3.2. Time information based solutions:

The most popular detection methods of wormhole attack use the packet traveling time between two consecutive nodes as a detection attribute. Most of the cases, the data packets traveling time is calculated from the measured round trip time (RTT) [28-31], the authors proposed RTT based solutions to confront against wormhole attack. In [32], delay per hop is determined by measured RTT whereas, for each successive hop, RTT is measured to detect wormhole attack [33]. In these schemes, the distance between two adjacent nodes is measured using RTT and determine if the two communicating sensor nodes are apart by the valid possible distance. If the measured RTT is higher than a defined threshold value, the presence of wormhole node will be declared. However, the RTT based solutions require the cooperation of the sensor nodes on the path and don't work properly on the DSR[34] and DSDV routing protocols [35-36]. These RTT based solutions may not perform well if the wormhole is created by protocol deviation. The wormhole nodes forward the packets without waiting for the certain time in order to reach the destination earlier than legitimate multiple hop path. Beside this, in the *'out of band'* wormhole attack, the wormhole nodes use the high-speed low latency link for transferring packets between them. In this type of attack, data packets reach to another wormhole node more quickly than *'In-band'* wormhole attack. In fact, the

time required to forward the packets to another wormhole node is significantly reduced in '*out of band'* wormhole attack. Therefore, the performance of RTT based detection scheme is in question with the presence of *'out of band'* wormhole attack. Some works of literature [37] also suggested that the RTT based detection scheme cannot detect *'active'* mode of wormhole attack and may not detect wormhole attack in the sparse sensor network.

In [38], the authors proposed a detection scheme based on neighborhood count and RTT. The authors considered the two facts. First, as the malicious nodes establish the new link in the network, so the neighbors around the malicious node increases significantly. Second, the measured RTT between two malicious nodes increases significantly in a similar fashion. However, these proposed solutions may not detect *'out of band'* wormhole attack from the infected network.

In [37], the authors proposed a detection scheme based data transmission time of each successive hop in a route assuming that the wormhole nodes are in hidden mode. In this scheme, the data transmission time of each successive hop is calculated from the measured RTT. All sensor nodes in the route transmit the recorded transmission time to the source. The source is solely responsible for checking if there is any wormhole node exist in the route. This scheme may perform poorly in the sparse sensor network. In fact, the performance of this scheme may be degraded when malicious nodes use high-speed communication channel.

3.3. Special Hardware-based scheme:

The directional antennas are incorporated with sensor nodes for adding access restriction [6,39-41] and finding legitimate neighboring nodes in the network. The entire communication region of a sensor node is divided into several zones. Moreover, each zone is defined by a directional antenna. When a sensor node receives the signal first time from its peering sensor node, then the probable location of the sender in terms of the zone is determined by the directional antennas. According to the authors, if a sensor node sends a packet from a particular zone, the recipient will get the signal from the opposite direction (zone). After that, the recipient sensor node verifies the legitimacy of the sender through receiving direction of the packet from unknown sensor nodes. The recipient node also cooperates with its neighbors to find out whether this node is known to other legitimate neighboring nodes or not. The incorporation of directional antennas to each sensor node makes this scheme expensive and inappropriate for practical deployment.

Another protocol named SECTOR was introduced in [29], that mostly depends on special hardware. The main concept of this detection scheme is to measure the distance between two communicating nodes based on the data transmission speed. The proposed model does not need any clock synchronization and geographical information of communicating nodes. In this algorithm, the mutual authentication with distance bounding (MADB) protocol [42] was implemented to measure the distance between nodes at the time of the encounter. According to the authors in [42], the proposed scheme permits one node to measure the mutual distance between two nodes and compares with maximum possible upper bounding distance. In this scheme, each node is incorporated with the special device that can give feedback to the sender without any delay. Initially, a node sends the one-bit challenge to its neighbors before sending any packets. Then recipient node responds back through a special device to the sender at the time it receives the one-bit challenge. When the sender transmits the one-bit challenge to a node, it turns on the clock and measures the time till it gets the response back from that node. After that, the sender estimates the mutual distance based on the measured time assuming that the data transmission speed is as same as the speed of light. However, in the [30], the authors slightly modify the schemes described in [42]. In [30], both parties are allowed to measure the mutual distance and validated the authenticity of the adjacent nodes around their communication range. The addition of the special device to each sensor node makes this scheme expensive and inappropriate for practical deployment.

3.4. Geographical information-based schemes:

In [43] and [7], the author proposed a detection scheme that assigns a maximum traveling distance to each data packet. The authenticity of a data packet is ensured by using the concepts of geographical and temporal packet leashes that would help to minimize the effectiveness of the wormhole nodes in the network. In the geographical leash (GL), when a sender sends packets to any sensor node, the sender node incorporates sending time and its own location information to the packet. When the recipient node receives the packet, it will estimate the approximate distance between the sender and receiver based on the geographical leashes. The temporal leash includes the maximum lifetime to each packet. In temporal leash (TL), a sender adds either transmitting time or expiration time of packet so that the recipient can verify if the packet has made a journey too far based on maximum transmission speed and time. A predetermined time threshold is set between any two neighbors based on their position in the network. For a specific two neighboring nodes, the data packet would be discarded if it violates the time boundary set by these specific nodes. This scheme can perform better if strict time synchronization and the additional device like GPS are provided. Beside this, this proposed scheme may be performed poorly when the wormhole node is in active mode.

3.5. Trust Based Solutions:

Trust information of neighboring nodes is used as an attribute to detect wormhole attack in the WSN. Each sensor node monitors the data packet forwarding pattern of its neighbors and rates them accordingly. In the trust base scheme, a sensor node selects the most trustworthy path to reach the destination based on the trustworthiness of its neighbors. In this scheme, the researchers consider the fact that the wormhole node drops all the received packets coming from its adjacent nodes. It is expected that the system would rate the least trust level to the malicious nodes. By using this trust level scheme, the wormhole node can

be avoided during the selection of a path to the destination. This also helps to reduce the effectiveness of the wormhole nodes in the infected network. The source node would forward the packets to the neighbor which possess maximum trustworthiness.

In [44], a new detection scheme was proposed to detect wormhole attack in the network based on both time and trust. In this scheme, both data travel time and trustworthiness of the sensor are used as potential attributes to detect the malicious nodes from the infected network. These two modules of this proposed scheme run in parallel. Here, the time-based module works in three stages. Firstly, each sensor node identifies its neighbors and update the neighbor list accordingly. Secondly, each node finds the best path toward the base station based on the trustworthiness of neighbors. Finally, the proposed algorithm determines the presence of the wormhole nodes in the selected path. According to the authors, the malicious nodes deceive the time-based solutions in many ways. Hence, the trust-based model is incorporated with the time base scheme. In this scheme, a sensor node monitors the neighboring node continuously and select the best route to the destination based on the rating.

Another trust based wormhole attack detection scheme was proposed in [45].To execute the model, the deployed sensor node must operate in a particular mode, named as 'promiscuous mode'. This trust-based scheme is applied to the DSR protocol and inherent features of DSR routing protocol are used to measure the trust level of the neighboring nodes. Here, the algorithm must be executed in each sensor node and each node must measure the trust level of its neighboring nodes by monitoring the packet transmission pattern stated by the system. The source node verifies in several stages whether the forwarding node passes the IP packets or not through the sequence of integrity check. If the neighboring node for a particular source forwards all the packets, the trust level of the node would be increased. Similarly, if the opposite happens, the trust level of that node would be decremented.

The success of this module lies on the packet dropping criteria of the malicious nodes. However, the wormhole nodes do not drop packets in the hidden mode of this attack. Hence, the trust base detection model is not capable of detecting the hidden wormhole attack. On the other hand, each node monitors the packets forwarding pattern of its neighbors. As we know, sensor nodes have some constrained-on power and energy resources. It would be a burden for the system which leads to excessive energy dissipation of the nodes.

3.6. Graph-based solution:

Multi-dimensional scaling-visualization of the wormhole (MD-VOW) [46], was proposed based on the graph theory. The multi-dimensional scaling analysis of the constructed connectivity graph was used as an analysis tool to detect malicious node in the network. For the static sensor network, the connectivity graph is not supposed to change frequently. Hence, the authors have considered the fact that the presence of malicious nodes introduces the anomalies in the connectivity graph. In the network connectivity based wormhole attack detection schemes, the positions of the neighboring nodes are estimated from the received signal strength (RSSI) by each node and send this information to the base station. By compiling the received information, the base station determines the network layout and compared with the current connectivity graph. Then the presence of the wormhole node can be detected if any forbidden structure is found in the constructed network layout. However, it is prone to the distance estimation errors, especially for the sparse network. The surface smoothing technique is applied to the constructed network layout graph to compensate distance error. This approach also causes a significant amount of control packets flow to the base station. Similarly, if the wormhole node is located in the sparsely populated area, then the wormhole attack may not be identified by this network visualization based algorithm.

3.7. Radio fingerprinting based scheme:

In [47], the authors explored the potentiality of the fingerprinting device as a tool to validate the legitimacy of the node in a wireless sensor network. Most important goal of this research is to extract the features from the radio signals, radiating from the nearby sensor nodes; by which the legitimacy of a node can be evaluated. In this research, each node has to be equipped with the radio fingerprinting device. The radio fingerprinting device captures the radiated radio signal from the nearby sensor nodes. After that, the fingerprinting device converts the radio signals into digital format. The transient part of the signal is located and the important features are extracted. After that, those extracted features are taken into account as fingerprints so that recipient sensor node can identify the neighboring node. The authors also implemented the incorporation of the radio fingerprinting device with a sensor node [47]. In this research, the authors also showed that the fingerprinting approach identified the nearby node while the message contents and the identification of the nearby devices were hidden. There are few issues are required to be investigated in this approach. The radio fingerprinting devices' performance in a noisy environment, and in the mobile platform is in question. By using this detection method, only the 'out of band' wormhole attack can be detected. This proposed algorithm cannot detect 'in-band' wormhole attack as it can be launched from the legitimate node.

3.8. Neighborhood counts based detection scheme

No. of neighbors around a sensor node are used in [10-12,19] as detection attributes in the neighbor based detection scheme. In the centralized method, each sensor node finds the number of neighbors within its communication region and sends this information to the base station. As the distribution of the sensor node is known, the base station computes the hypothetical distribution of the number of neighbors along with the true distribution of the neighborhood counts. This process also creates a significant amount of control data packet flow throughout the network and leads to the unexpected energy dissipation of sensor node. This process is also used as secondary approach with the

distance-based scheme. In another neighborhood count based approach [12], detector node takes the count of neighbors at the each site it visited. In this approach, ANN-based detection scheme may not able to detect the wormhole attack if the wormhole nodes are located in the sparsely populated area. Another important factor is that by this scheme, the probable location of the wormhole node cannot be identified. Some authors also use signal processing tool like Fourier transform, Discrete wavelet transform [48] and, variance fractal dimension [49] on the data series of the neighborhood counts. However, these proposed algorithms only work on the acquired data series while deployed sensors are distributed uniformly. However, these schemes may not able to locate the probable location of the wormhole node.

4. Artificial neural Network (ANN)

The Artificial Neural Network (ANN) is a network based stochastic learning model that has evolved from the study, characteristics, organization, and decision-making ability of the unit cell of a human brain called a neuron [19]. In other words, it aims to imitate the most simplified and basic function of the human brain. Analogous to the unit cell of a human brain, an ANN contains several interconnected information processing units, called a neuron, learning the underlying process of the data samples presented to the network. The significant phenomena of ANN are the ability to estimate the non-linear complex relationship between inputs and outputs without any prior knowledge of dataset like a black box. ANN is usually observed as a model of interconnected neurons that maps the input and outputs through information exchange among neurons[50].

Furthermore, in a multi-layer perceptron (MLP), there is an input layer, followed by one or more hidden layers, and an output layer [51]. In each layer, several neurons are employed, which are fully connected with other neurons of an adjacent layer, and they are connected to different random weight values. In other words, neurons are fully attached to the neurons of the following layer; but in the same layer, neurons are not connected with each other. The number of neurons in the output layer depends on the type of the problem that we want to solve by the ANN.

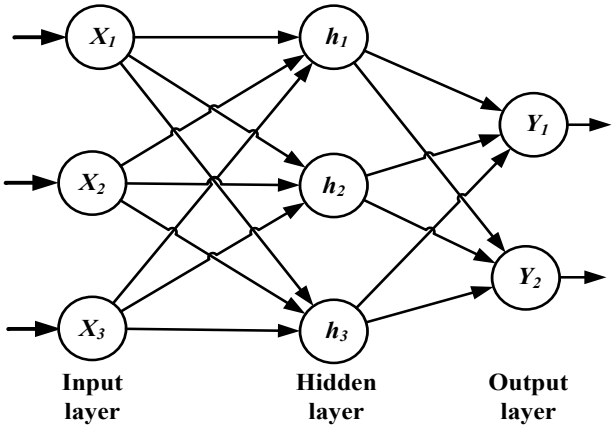


Fig. 3 The structure of an ANN.

4.1. Forward Propagation of the ANN:

Input features from the input layer are shared with an adjacent hidden layer through unidirectional branches[50]. Those input features are multiplied by random weights associated with the unidirectional branches; summed up, and passed through the activation function of the neuron (e.g. sigmoid function). The bias term is also added to each layer except the output layer to activate the artificial neurons. This bias term is also connected to the neurons of the adjacent layer with unidirectional branches associated with the weight values ($b_i^{(l)}$) so that the summed inputs exceed the predefined threshold.

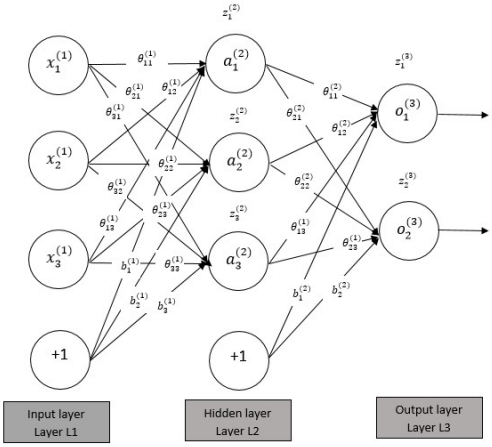


Fig. 4 Forward propagation of ANN

In the forward propagation, the output of each neuron of the prior layer is considered as the input to all neurons of the following layer. As shown in the Fig. 4, the three layers are connected with the weight value $\theta_{ij}^{(l)}$. $\theta_{ij}^{(l)}$ represents the weight value going to the unit i in layer $(l+1)$ and coming from the unit j in layer l. According to the Fig. 4, the net output of ith unit (including bias term) of the hidden layer ($a_i^{(2)}$) is,

$$Z_i^{(2)} = \sum_{j=1}^{n} \theta_{ij}^{(1)} x_j^{(1)} + b_i^{(1)} \quad (1)$$

$$a_i^{(2)} = g(z_i^2) = g(\sum_{j=1}^{n} \theta_{ij}^{(1)} x_j^{(1)} + b_i^{(1)}) \quad (2)$$

In the equation (1) and (2), $Z_i^{(2)}$ represents the total input of the ith unit of layer 2 .Similarly, the net output of kth unit of the output layer can be shown as equation (3),

$$o_k^{(3)} = g(\sum_{j=1}^{n} \theta_{kj}^{(2)} a_j^{(2)} + b_k^{(2)}) \quad (3)$$

The cost function,

$$J(\theta) = \sum_{k=1}^{n} \frac{1}{2m} \sum_{i=1}^{m} (o_k^m - y_k^m)^2 \quad (4)$$

In the equation (4), $J(\theta)$ refers to the average error occurred over m training samples during the training procedure, and k defines the number of neurons in the output layer.

The transfer function of a neuron plays an important role in the training procedure. The purpose of the transfer function is to replicate the activation mechanism of the biological neuron. Usually, the net output of a unit in a layer is calculated from the net received input through the transfer function. The transfer function must be non-linear, continuous and differentiable at any point in order to apply gradient descent learning algorithm. There are several types of transfer function such as sigmoid function, hyperbolic tangent function, and rectified linear function. Some literature also suggests that the rectified linear function was found to accelerate the convergence rate of stochastic gradient descent due to its non-linear and non-saturating form. However, in our research, the sigmoid function was used as the transfer function of each unit in any layer (except input layer).

$$g(z) = \frac{1}{1 + e^{-z}} \quad (5)$$

$$g(z) = \tanh(z) = \frac{e^{z} - e^{-z}}{e^{z} + e^{-z}} \quad (6)$$

$$g(z) = \begin{cases} 0 & if\ z \leq 0 \\ z & if\ z > 0 \end{cases} \quad (7)$$

4.2. Back Propagation of the ANN

In the back propagation, the outcome of the output layer is compared with desired output. The error between the network output and desired output is measured and propagated backward to adjust the branch weights in order to minimize error that would occur due to the estimation of output. In other words, we minimize the cost or energy of the error function, $J(\theta)$ by passing back the error occurred during the training phase. In the back propagation, the gradient descent algorithm is applied to learn the network parameters $(\theta_{ij}^{(l)}, b_{ij}^{(l)})$ from the training set $\{x_i^m, y_i^m\}$.

However, there are many ways to update the weights: batch mode, online mode. In the batch mode, each weight is updated once by the total error occurred in an epoch (i.e. after one training cycle). In contrast, in the online mode, a training sample is drawn randomly from the training set and passes through the network. The network parameters are modified m times per training cycle if there are m training examples in the training set. This online mode of weight updating is also known as stochastic gradient descent[50].

Let's consider the multilayer feed forward network shown in Fig. 4. The impact of the change in the weight $(\theta_{ij}^{(2)})$ on error $(E_k^{(3)})$ occurred in the output layer is,

$$\frac{\partial E_k^{(3)}}{\partial \theta_{ij}^{(2)}} = \left(o_k^{(3)} - y^{(t)}\right).o_k^{(3)}\left(1 - o_k^{(3)}\right).a_j^{(2)} \quad (8)$$

In the equation (8), $o_k^{(3)}$ represents the output of the kth unit of the output layer and $y^{(t)}$ refers to the desired output of a training example. Similarly, the impact of the change in weight $(b_k^{(2)})$, associated with bias and the neurons of the output layer, on error $(E_k^{(3)})$ occurred in output layer is,

$$\frac{\partial E_k^{(3)}}{\partial b_k^{(2)}} = \left(o_k^{(3)} - y^{(t)}\right).o_k^{(3)}\left(1 - o_k^{(3)}\right) \quad (9)$$

According to the above equations (8) and (9), the update formulas for the weights associated with output layer are given below.

$$\theta_{kj}^{(2)} = \theta_{kj}^{(2)} - \alpha\left(o_k^{(3)} - y^{(t)}\right).o_k^{(3)}\left(1 - o_i^{(3)}\right).a_j^{(2)} \quad (10)$$

$$b_k^{(2)} = b_k^{(2)} - \alpha\left(o_k^{(3)} - y^{(t)}\right).o_k^{(3)}\left(1 - o_i^{(3)}\right) \quad (11)$$

In the equations (10) and (11), α refers to the learning rate of the network. Now, we can calculate the rate of the change in the error $(E_k^{(3)})$ over the weight $(\theta_{ij}^{(1)})$, associated with the input layer and the hidden layer shown in the equation (12).

$$\frac{\partial E_k^{(3)}}{\partial \theta_{ij}^{(1)}} = \delta_k^{(3)}.a_i^{(2)}.x_j^{(1)}.\left(1 - a_i^{(2)}\right).\theta_{kj}^{(2)} \quad (12)$$

Considering

$$\delta_k^{(3)} = \left(o_k^{(3)} - y^{(t)}\right).o_k^{(3)}\left(1 - o_k^{(3)}\right) \quad (13)$$

Similarly, impact of the changes in weights $(b_i^{(1)})$, associated with bias and the neurons of the hidden layer, on error $\left(E_k^{(3)}\right)$ occurred in output layer can be calculated as

$$\frac{\partial E_k^{(3)}}{\partial b_i^{(1)}} = \delta_k^{(3)}.a_i^{(2)}\left(1 - a_i^{(2)}\right).\theta_{kj}^{(2)} \quad (14)$$

Therefore, according to the equations (12) and (14), the weights associated with input layer and the hidden layer can be modified by the following equations during the training procedure.

$$\theta_{ij}^{(1)} = \theta_{ij}^{(1)} - \alpha\delta_k^{(3)}.a_i^{(2)}.x_j^{(1)}.\left(1 - a_i^{(2)}\right).\theta_{kj}^{(2)} \quad (15)$$

$$b_i^{(1)} = b_i^{(1)} - \alpha\delta_k^{(3)}.a_i^{(2)}\left(1 - a_i^{(2)}\right).\theta_{kj}^{(2)} \quad (16)$$

These weights adjacent procedure continues recursively until the network reaches either to the minimum tolerance level or maximum epochs at the training.

5. Support Vector Machine

The support vector machine (SVM) is one of the reliable and widely used supervised learning models that analyze the presented data samples to perform classification and regression analysis[50]. The SVM learns the given data samples, each sample marked with a specific class, builds a model that can assign a class to a new data sample. SVM learning model set a hyperplane in an optimum position in the data space so that Euclidian distance from the decision surface for all training data samples would be maximized. SVM has been emerged to provide the generalized performance to solve a wide range of classification and pattern recognition problems such as face detection, pedestrian detection, and text categorization etc.

Let's consider a training data set, $D = \{(x^m, y^m)\}$ contains m training data samples where $x^m \in R^n$ and $y^m \in \{-1, +1\}$, represents the label of the mth data sample. The hyperplane in form of decision surface can be defined as

$$\sum_{i=1}^{m} \boldsymbol{w}^T x^{(i)} + b = 0 \quad (17)$$

In the equation (17) $\boldsymbol{w}$ and b represent a weight vector and a bias term that can be determined through the training process. Through these parameters (w, b), the decision surface places itself in the optimum position in the data space. As we know, the SVM places the decision surfaces in the data space in such way that it maximizes the geometric margin of all training data samples. In this circumstance, the optimization problem is

$$\min_{w,b} \frac{1}{2} \parallel W \parallel^2 \quad (18)$$

Subject to

$$\sum_{i=1}^{m} y^{(i)}(W^T x^{(i)} + b) - 1 \geq 0$$

The concept of the Lagrange multiplier is implemented to solve this optimization problem with constraint boundary stated in the equation (19). Therefore, the Lagrange function is

$$L(w, b, \alpha) = \frac{1}{2} \parallel W \parallel^2 - \sum_{i=1}^{m} \alpha(y^{(i)}(W^T x^{(i)} + b) - 1) \quad (19)$$

In the equation (19) α is the multiplication factor and $\alpha \geq 0$. If we differentiate the Lagrange function with respect to w, b and ; then the optimization problem mention in the equation (18) can be formulated as

$$\max_{\alpha} L(\alpha) = \max_{\alpha} \Big(\sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y^{(i)} y^{(j)} x^{(i)} x^{(j)} \Big) \quad (20)$$

Subject to

$$\sum_{i=1}^{m} \alpha_i y^{(i)} = 0$$

$$\alpha_i \geq 0, i = 1,2, \ldots\ldots, m$$

The solution of the equation (20) drives to get optimum decision surface that is able to separate the positive and negative training data samples. If the presented data samples are not linearly separable, then the non-linear kernel trick can be implemented to deal with this problem. The Gaussian kernel is widely used as the kernel function. Because this function can increase the dimensionality of the data samples infinitely[52]. The expression of the Gaussian kernel is shown as equation (21).

$$K\left(x^{(i)}, x^{(j)}\right) = \exp\left(\frac{-\parallel x^{(i)} - x^{(j)} \parallel^2}{2\sigma^2}\right) \quad (21)$$

6. Non-linear Logistic regression

Logistic regression is a statistical model in which the category, from a predefined list of categories, of a new observation or data sample, is predicted, by estimating the probability of the class using a logistic function (e.g. email sorting). It is evolved from linear regression and designed to solve classification problems. That's why, in some literature, it is referred as logistic classification[50].

Logistic classification can be shown as a special case of linear regression, but we can draw a distinct line between these two statistical models. In the linear regression, the predictor predicts the continuous value by a fitting curve to the given training input data samples [53]. In contrary, in logistic classification, a predictor learns how to classify the data through a training phase and predicts a discrete value for the corresponding new data sample. Like linear regression, the same decision boundary equation is used for logistic classification.

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 {x_2}^2 + \cdots + \theta_n {x_n}^n = \theta^T \mathbf{x} \quad (22)$$

Here, in the equation (22), $\boldsymbol{x}$ represents a given input vector that contains n input features. The θ refers to the model parameter that is required to be optimized by using the training input data samples. This decision boundary can be either linear or nonlinear that depends on the application of the logistic classification.

$$Z = \theta^T \boldsymbol{x} \quad (23)$$

$$g(z) = \frac{1}{1+e^{-z}} \quad (24)$$

In logistic regression, a sigmoid function is used to measure the certainty level of the class that new observation or data sample belongs to. The main advantage of using sigmoid function is that it is continuous, differentiable at any point and monotonically increasing. The probabilistic interpretation can be done in a simple way; as the value of the $g(z)$ in greater than 0.5, the predictor would be more certain about the class of the given data sample [54]. If the same cost function of linear regression is used, $J(\theta)$ yields non-convex cost function. Consequently, special kind of the cost function is used in logistic classification to learn model from the training data which is convex.

$$J(\theta) = -\left[\frac{1}{m}\sum_{i=1}^{m} y^i \log\left(h_\theta(x^i)\right) + (1-y^i)\log\left(1 - h_\theta(x^i)\right)\right] + \frac{\lambda}{m}\sum_{j=1}^{n}\theta_j^2 \quad (25)$$

$J(\theta)$ denotes the cost function of the logistic classification with regularization term, where m refers to the total training examples of a given training set, λ is the regularization parameter, y^i represents the target output for i^{th} training example, and x^i represents the ith training example. There are two learning methods used in the training procedures: Gradient decent and Newton's method. In this research, the newton's method is applied as the learning model. Because it converges to the optimal solution within few iterations.

7. Propose detection scheme

The proposed algorithm is a network based approach, in which neighborhood counts are used as the detection feature to detect a wormhole attack.

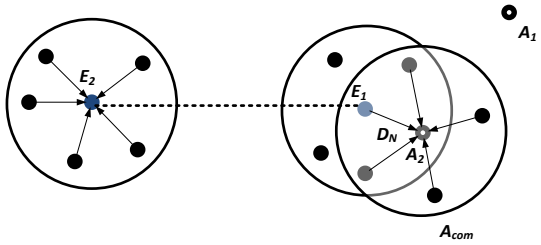


Fig. 5 Impact of wormhole attack on neighborhood count.

A mobile sensor node, known as detector node (D_N), is deployed in an area where sensor nodes could be uniformly or non-uniformly distributed. The detector node D_N moves around this sensor field and collects a neighborhood count and the coordinate at each site it visited. When it reaches to the wormhole infected zone, the neighborhood population would increase abnormally sharply or abnormally lightly, depends on the sensor distribution and the position of the wormhole nodes. For instance, Fig. 5 shows the impact of wormhole nodes on the neighborhood counts. Let's say, the detector node D_N is moving spontaneously around the area where the sensor nodes are deployed. At the time t_1, D_N moves from a location A_1 to another location A_2. Since the detector node collects the neighborhood count at each site, it transmits neighbor discovery message (NDM) to the adjacent sensor nodes within its communication range. According to the Fig. 5, the wormhole node (E_1) also hears the broadcast as it is located in the transmission range of D_N.Since wormhole node doesn't read the content of the packet, it encapsulates and forwards the packets along the virtual tunnel to another malicious node (E_2). Furthermore, E_2 retransmits the received packets towards its neighbors. The neighbors of the distant malicious nodes (E_2) respond back with valid neighbor ID (NID) through the wormhole link. After that, E_1 unicasts the received responses to the originator, D_N. E_1 compels D_N to think that the responses have come from the sensors located within its radio ranges. Though some of those responses have traveled long distant within the network. That's how the neighborhood counts from the infected zone increases sharply due to the wormhole node shown in the Fig. 6. This may be true for uniform sensor distribution.

As we know, the sensor density in the non-uniform sensor distribution is inconsistent over the area. If the wormhole nodes are placed in the sparsely populated area, then the number of neighbors would not increase abruptly or lightly as expected. In this circumstances, it would be difficult for ANN based detection scheme to detect wormhole attack based on the collected neighborhood counts (also true for the sparse network).

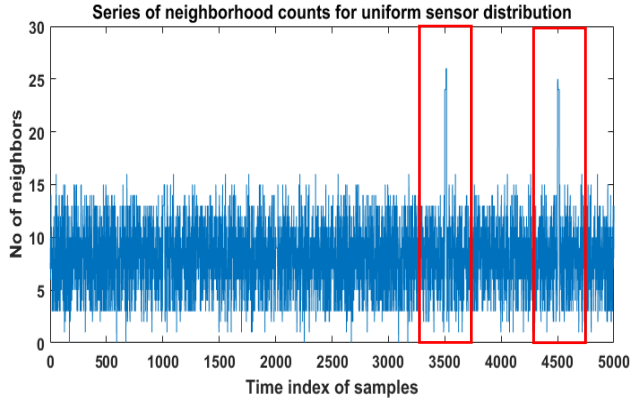


Fig. 6 The series of neighborhood count (Uniform sensor distribution)

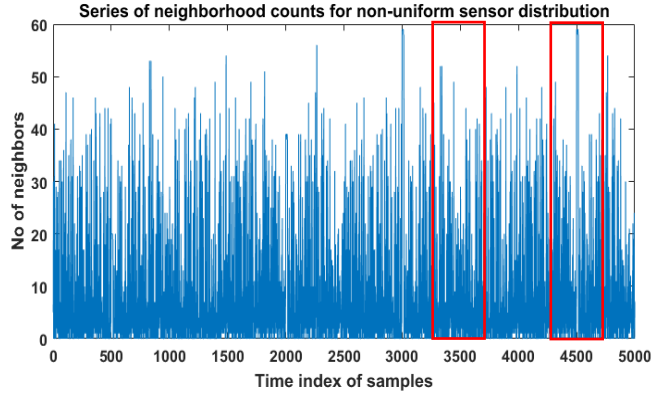


Fig. 7 The series of neighborhood count (Non-uniform sensor distribution)

Sometimes the neighborhood counts taken from the infected zone are same as, or far smaller than the counts taken from the

non-infected zone shown in the Fig. 7. In this case, the ANN-based detector may suffer to distinguish between positive and negative data samples. Therefore, another detection feature along with neighborhood count is required in order to detect wormhole attack from the infected network (either uniform or non-uniform) more precisely and confidently.

In this paper, we introduce a new detection feature called 'average *residual energy pattern of the neighbors (AREPN)'*. As we know, wormhole nodes circulate false route information into the network that the base station is multiple hops away from the wormhole nodes. Therefore adjacent sensor nodes of the wormhole get influenced and transmit their data packets to the base station through wormhole nodes. However, the wormhole nodes force them to hand over the collected data packets to one of the wormhole nodes. That means the wormhole nodes receive more data packets after the base station and its neighbors. Now, the question is how the wormhole nodes get those data packets coming from the adjacent nodes. Apparently, the data packets are coming through its neighbors.

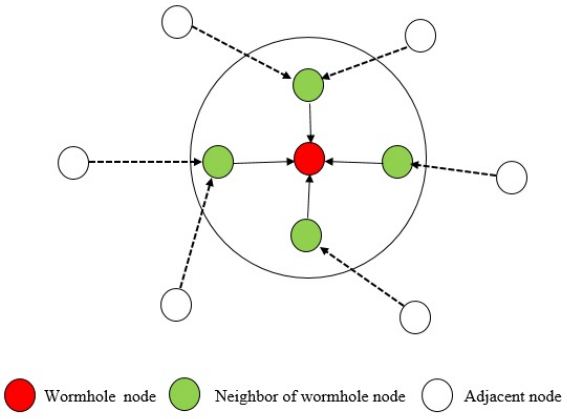


Fig. 8 Impact of wormhole nodes on adjacent nodes.

More packets arrive at the wormhole node means, more packets have been received and transferred by its neighbors. As we know, a node dissipates a significant amount of energy due to the transmission and reception of the data packets. Since the wormhole nodes are getting the numerous data packets from the adjacent nodes, so the neighbors of the wormhole nodes are also losing energy more quickly by retransmitting packets intended for the wormhole node. Now it can be said that the sensors located in the infected zone lose more energy than other nodes located in the non-infected zone.

In our proposed scheme, the detector node broadcasts NDM to the adjacent sensor nodes at each site it visited. The adjacent nodes reply back by transmitting a data packet incorporating valid NID and information of residual energy. By this way, D_N is able to calculate the number of neighbors and average residual energy of the neighbors (AREPN). D_N captures this evidences as a two-featured sample and stores it. It is expected that AREPN taken from the infected zone is much smaller than the AREPN taken from the non-infected zone. As we know, the adjacent nodes of the base station dissipate energy quicker than the normal node. That's why, D_N doesn't cover a certain geographic area based on the location of the base station. Once the mobile node reaches close to the base station, it transfers all the collected samples for further analysis.

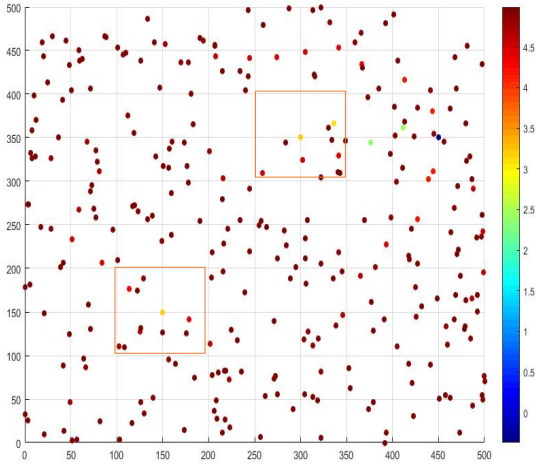


Fig. 9 Wormhole infected network.

As shown in the Fig. 9, the sensors of the wormhole infected zone lose energy (red marked box) more quickly than the non-infected zone. In that case, the initial energy of each sensor node is 5 units.

The performance of an artificial neural network highly depends on the method of training and the dataset containing potential features. The Base station gathers (by the help of the detector node) a data set (D_{set}) consists of N data samples. In this first half, D_N gathers two featured K data samples ($K \in N$) from the non-infected zone which are called negative data samples. Similarly, the same amount of data samples is collected from the wormhole infected zone, known as positive data samples. After that, those two types of data samples are mixed up randomly so that training can be performed without any bias. Then, M data samples ($M \in N$) are drawn from the D_{set} and stores in a training dataset D_{train} ($D_{train} \subset D_{set}$). At the same time, rest of the P data samples ($P \in N$) are stored in D_{test} ($D_{test} \subset D_{set}$) to evaluate the learning performance of the trained neural network.

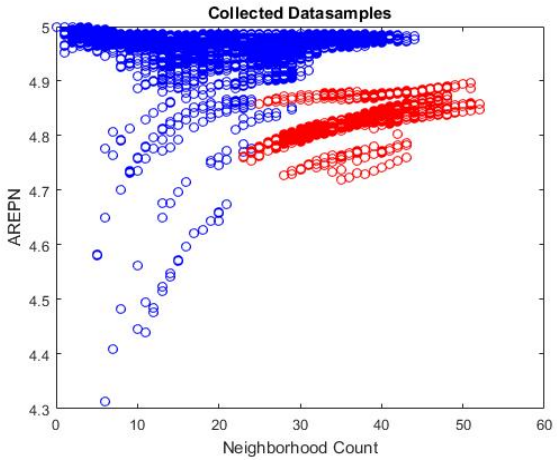


Fig. 10 Collected two featured data samples by D_N (red marked samples are positive data samples) .

Proposed Algorithm:

1. Collect two featured K negative data samples from the non-infected zone
2. Collect two featured K positive data samples from the wormhole infected zone
3. Store the both type of data samples in D_{set} which consists of N data samples
4. Select M data samples from D_{set} and store in D_{train} where $(D_{train} \subset D_{set})$
5. Rest of the P data samples are stored in D_{test} where $(D_{test} \subset D_{set})$
6. Train the neural network with the data set D_{train} and appropriate network parameters up to T epochs
7. Test the neural network by using the samples of D_{test}
8. If output for a specific sample ≥ 0.5 , then this sample represents wormhole attack and probable location of the malicious node are identified.
9. If output for a specific sample < 0.5 then this sample doesn't represent wormhole attack
10. Update D_{train} by D_{test} for further training
11. Reset D_{test} and update with new data samples gathered by D_N

Furthermore, data samples of D_{train} are fed into the input layer of the ANN. The training procedure is performed repeatedly until it reaches a predefined maximum number of training cycles i.e. T epochs. The testing procedure involves checking the learning progress of the ANN-based detector. In the testing part, each data sample passes through the trained neural network. If the output for a specific data sample is greater than 0.5, then this sample represents wormhole attack. Since, D_N stores the coordinate of the locations along with the two featured data samples, so the probable position of the wormhole node can also be identified by the detection scheme.

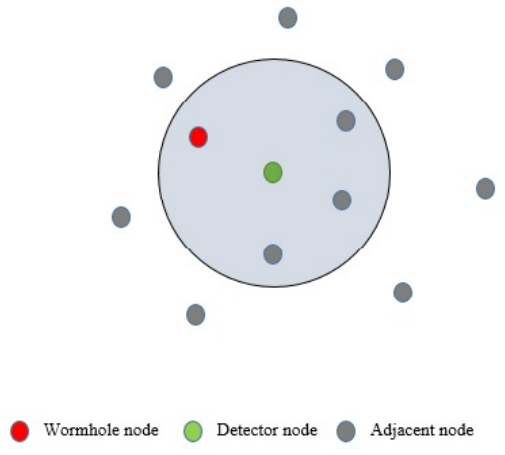


Fig. 11 Probable location of a wormhole node

After that, D_{train} is updated by the D_{test} for further training. This would minimize the error level that has achieved in the training phase. The D_{test} entries are cleared up and updated with new data samples collected by the D_N in real time.

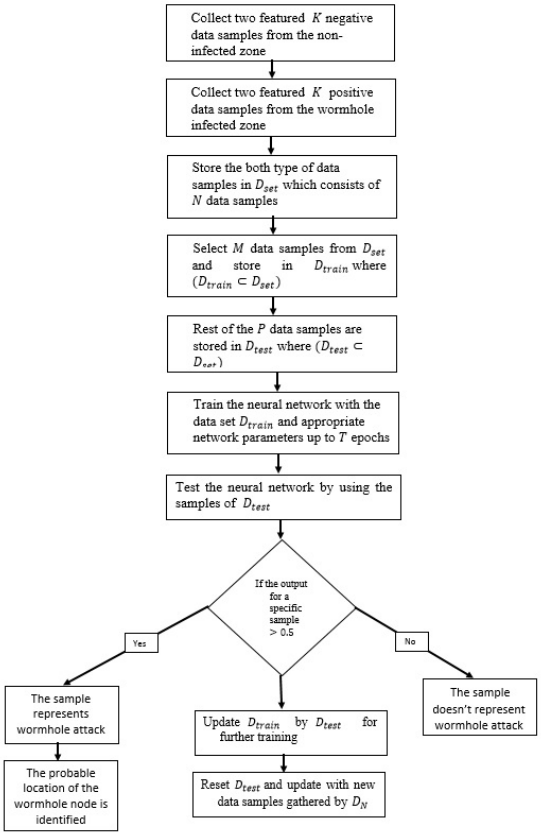


Fig. 12 Flowchart of the ANN-Based proposed algorithm

8. Simulation and results

In this section, extensive research experiments are conducted under various network scenarios to assess the effectiveness of the proposed algorithm in detecting wormhole attack from the affected sensor network. The first phase of the experiments is conducted to see if the proposed scheme is able to classify the malicious data samples that represent wormhole attack. Furthermore, we evaluate the performance of the detection scheme in terms of detection accuracy, false positive rates, and false negative rates. In the second phase, the efficacy of the proposed algorithm is explored considering single featured data samples and two featured (i.e. the number of neighbors and AREPN) data samples. As the energy dissipation of the sensor node depends on the routing protocol; hence, the proposed algorithm is also tested considering 'On demand' based routing protocol (AODV) and cluster based routing protocol (LEACH). Afterward, we record and analyze the efficacy of the proposed algorithm in detecting wormhole attack from different sensor distributions. Finally, the performance of the proposed algorithm is compared with the performance of other machine learning technique based detection schemes like support vector machine (SVM) and regularized non-linear logistic regression (LR). In addition, all experiments have been performed in MATLAB 2015a.

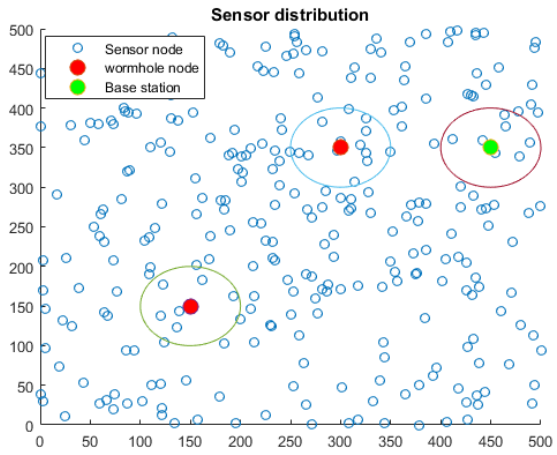


Fig. 13 The depiction of the simulation set up (uniform sensor distribution)

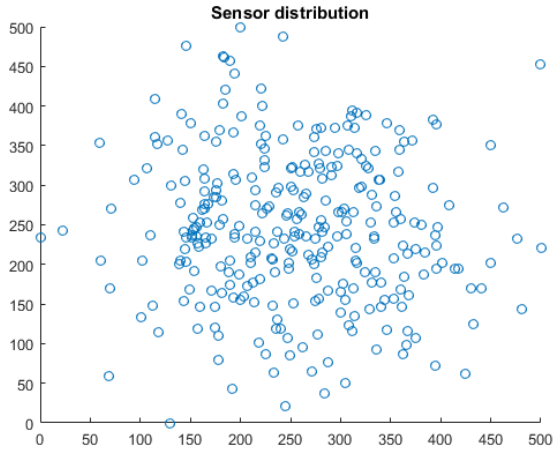


Fig. 14 Non- uniform sensor distribution (Gaussian)

In the simulation, 300 sensor nodes are distributed (uniform or non-uniform) within the square field of $500m \times 500m$. The deployed sensor nodes and the base station are static in nature. Each sensor node has the communication resources of which they use to communicate with the adjacent sensor nodes. Radio range of each sensor node equals to $50\ m$. Initially, each sensor node has five (05) units of energy that would be used in sensing and transferring packets. A detector mobile sensor node, D_N is deployed as a mobile observation point of the network. The basic task of the D_N is to collect neighborhood counts and AREPN of neighboring population at each site it visited within this deployed area. The radio range of the detector node is same as deployed sensor nodes. We assume that the detector node is fully aware of its position, the boundary of the targeted area, and any obstacles in the area that may restrain its movement. A pair of wormholes node is placed at the locations $150m \times 150m$, and $300m \times 350m$. Random waypoint model is implemented in the simulation as the mobility model for the D_N.Similarly, the base station is positioned at a location of$350m \times 450m$ .In these experiments we only consider the data communication between a node to the base station. Moreover, the experiments have been conducted on thirty (30) different instances for each sensor distribution (e.g. uniform sensor distribution) in order to get valid (average) performance matrices of the proposed detection scheme.

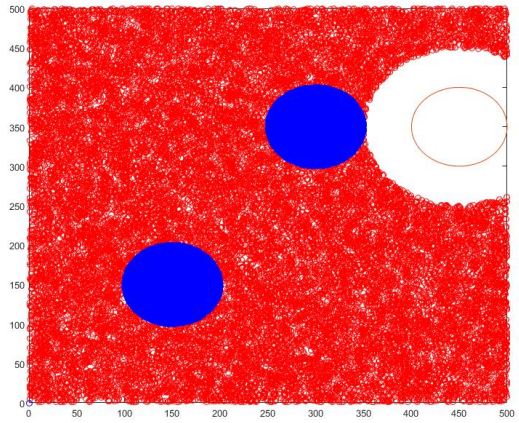


Fig. 15 The location visited by the D_N (red ,blue and white indicate non-infected zone, infected zone and the area not covered by D_N).

In this simulation, the detector node collects two featured data sample (i.e. number of neighbors and AREPN) at each site it visited. For each instance of each sensor distribution (uniform or non-uniform sensor distribution), the detector node,D_N collects 50000 data samples from deployed area in which 25000 samples are negative data samples, and 25000 samples are positive data samples. The collected data samples are stored at the D_{set} in the base station. After that, 49000 randomly selected samples are stored in the D_{train} for the training purpose. Rest of the data samples is used to test the learning performance of the proposed detection scheme.

8.1. The performance of Proposed ANN based detection scheme:

A multi-layer, feed forward network (with back propagation algorithm) is implemented for the experiments. The input layer contains one or two neurons considering the type of data samples (one featured or two featured) used to train the network. Moreover, the hidden layer consists of 100 neurons and the output layer has only one (01) neuron. We used a sub data set, D_{train} comprised of 49000 randomly selected data samples from D_{set} for training collected by D_N.

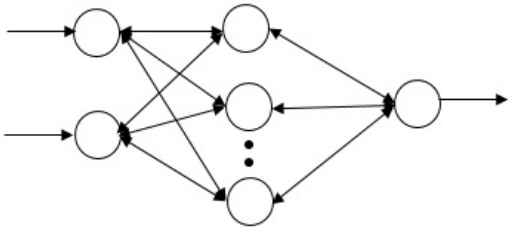


Fig. 16 The structure of the ANN for two featured data samples

At first, the neural network was trained by the single featured the training examples from D_{train} and evaluated learning performance of ANN by the rest of the data samples stored in the D_{test} . After that, two featured training samples are

fed into the neural network and acquired results are compared with the results obtained by using single feature data samples. During the training period, the minimum error tolerance level is set to 10^{-5}. The Table 1 represents the parameters which are used during the training phase.

Table 1 Parameters used for ANN.

Parameter	Value
No of attributes	2
No of Data points (training)	49000
No of Data points (testing)	1000
Architecture (for one feature)	[1,100,1]
Architecture (for two features)	[2,100,1]
Performance	0.00001
Learning rate, α	0.001
Epoch	500
CPU time (for one instance)	2.17 mins

In the testing phase, the test data set, D_{test} is fed into the input layer. Then the output of the ANN based model is observed if it could detect the existence of wormhole attack in the given network. Fig. 17 shows that, the ANN based wormhole attack detection scheme can classify the samples that represent wormhole attack from the both single featured and two featured data samples successfully

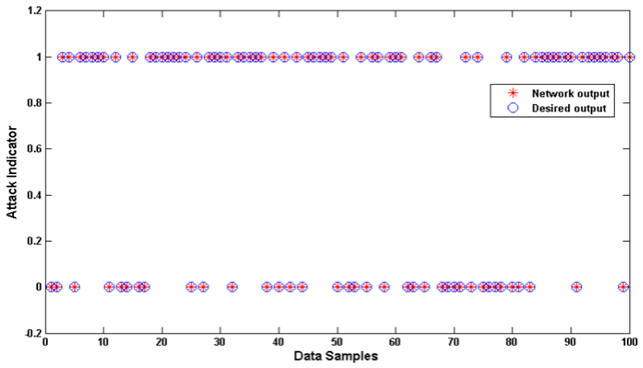


Fig. 17 Classification of wormhole attack

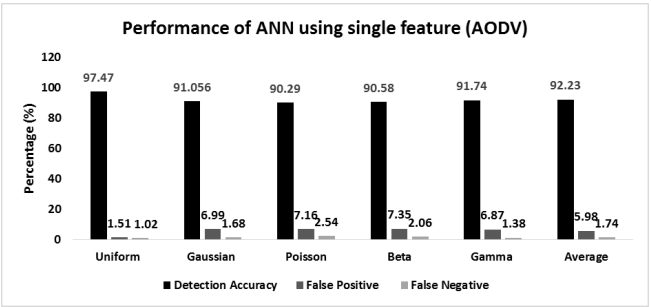


Fig. 18 Performance of the ANN-based detection scheme using single featured data samples (AODV)

Fig. 18 shows the performance of the ANN-based detection model in detecting wormhole attack using single featured data samples (i.e. neighborhood count) when the network uses AODV routing protocol. In this graph, the highest detection accuracy is recorded as 97.47% when sensors are distributed uniformly in the square field, whereas the lowest detection accuracy is measured 90.29% for poisson sensor distribution. Furthermore, detection accuracy for Gaussian distribution is almost same as the gamma distribution. Accordingly, 91.056%, 91.74%, and 97.873% detection rates are measured for Gaussian, Gamma, and Beta sensor distribution. However, the average detection accuracy is calculated as 92.23%. The false positive rate and false negative rate entirely follow the same trend of detection accuracy. The lowest false positive rates and false negative rates are measured for uniform sensor distribution. The average false positive rates and false negative rate are accordingly 5.98% and 1.74%.

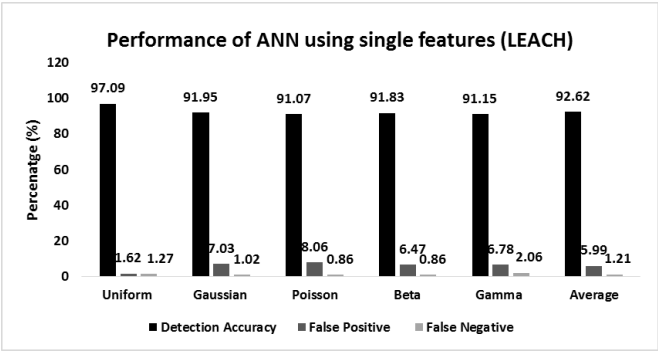


Fig. 19 Performance of the ANN-based detection scheme using single featured data samples (LEACH)

Fig. 19 shows the results obtained by applying single featured data samples on ANN based detection model considering LEACH routing protocol. It is observed that this algorithm performs better for uniform sensor distribution compares to non-uniform sensor distribution. 97.63% detection accuracy, 1.62% false positive rate, and 1.23% false negative rate are achieved for the uniform sensor distribution. After that, it has performed better for Gaussian sensor distribution among all non-uniform sensor distributions. The detection accuracy, false positive rate and false negative rate for Gaussian sensor distribution are 91.95%, 7.03% and 1.02% accordingly. The similar trend in detection accuracy is observed for Poisson, beta, and gamma sensor distribution. However, considering all the sensor distributions, the false negative rates are lower than the false positive rates. The average detection accuracy and false positive rate for all sensor distributions are accordingly 92.62% and 5.99%.

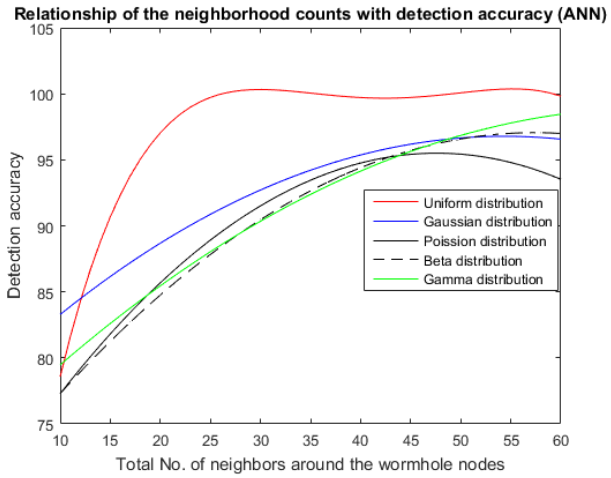


Fig. 20 The relationship between neighborhood count and detection accuracy (ANN)

Fig. 20 represents the relationship between the neighborhood counts and the detection accuracy. Undeniably, the detection accuracy of the proposed algorithm (considering single featured data samples) increases as the total number of neighbors around wormhole nodes increase for all sensor distributions. The detection accuracy of the system has a non-linear relationship with the total number of the neighbors around the wormhole nodes. For the uniform sensor distribution, we can model this relationship as 4th degree polynomial. For the other non-uniform sensor distribution, the relationship can be modelled as 2nd .It means that the proposed scheme may not perform well considering single featured data samples if the wormhole nodes are located in the sparsely populated area in the network.

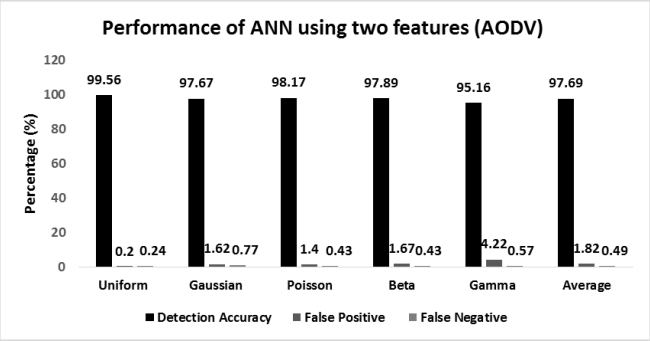


Fig. 21 Performance of the ANN-based detection scheme using two featured data samples (AODV)

The two featured data samples are applied to the ANN-based detection scheme. Like prior, the performance of the ANN is evaluated considering both AODV and LEACH routing protocol. Fig. 21 presents the performance of the ANN-based model considering two featured samples and AODV routing protocol. The proposed algorithm gives a better performance to detect the wormhole attack for the uniform sensor distribution. 99.56% detection accuracy, 0.2% false positive rates, and 0.24 % false negative rate are achieved for the uniform sensor distribution. For all non-uniform sensor distribution, the detection accuracy varies in between 95% to 97%. In this circumstance, the average detection accuracy and false positive rate for all the sensor distributions are 97.69% and 1.82%. Most importantly, the detection accuracy, considering all sensor distribution, is significantly increased for the application of the two featured data samples.

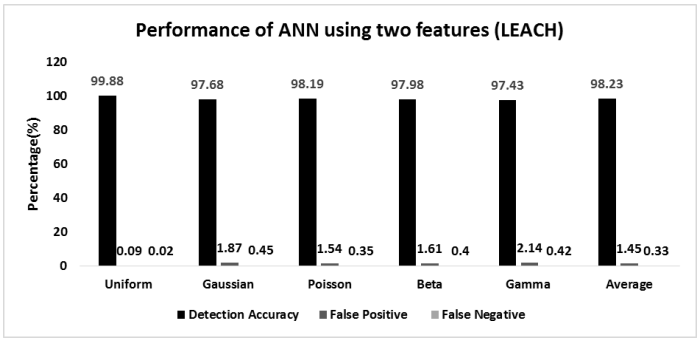


Fig. 22 Performance of the ANN-based detection scheme using two featured data samples (LEACH)

As shown in the Fig. 22, the performance of the ANN-based detection scheme improves significantly considering LEACH routing protocol and the application of two featured data samples on the ANN. Evidently, the ANN-based model performs better for the uniform sensor distribution compares to all non-uniform sensor distribution. The highest achieved detection accuracy for the uniform sensor distribution is 99.88%. The detection accuracy for the non-uniform sensor distributions fluctuates from 97.43% to 97.98%. Furthermore, the lowest false positive rates and false negative rates are achieved for the poisson sensor distribution which is 1.54% and 0.35% respectively. In this circumstance, the average detection accuracy, false positive rate, and the false negative rate for all the sensor distribution are accordingly 98.23%, 0.45%, and 0.33%. In summary, the application of the two featured data samples on the ANN enhances the performance of the proposed scheme for both AODV and LEACH protocol.

A. The performance of the SVM-based detection scheme

For the different sensor distributions, one and two featured training examples are applied to the SVM accordingly. The average result of detection accuracy, false positive rates, and false negative rates are calculated to measure the efficacy of the SVM-based algorithm for all network scenarios. The Table 2 represents the parameters which are used during the training phase.

Table 1 The parameters used for SVM

Parameter	Value
No of attributes	2
No of Data points (training)	49000
No of Data points (testing)	1000
Kernel	Gaussian
Sigma	default
Tool	MATLAB
CPU time (for one instance)	3.17 mins

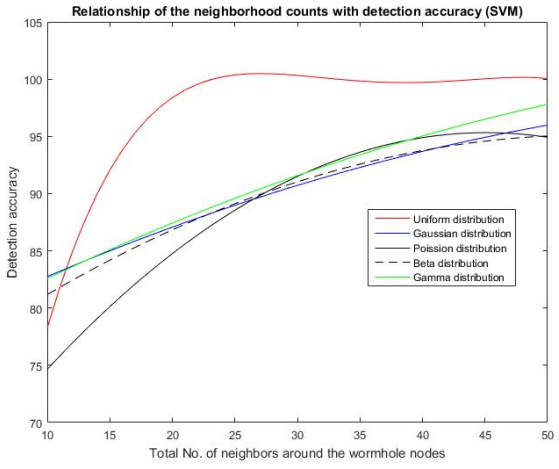


Fig. 23 The relationship between the neighborhood count and the detection accuracy (SVM).

Fig. 23 represents the relationship between the counts of neighbors around wormhole nodes and the detection accuracy considering one featured data samples applied to the SVM-based detection scheme. Like the previous case, the detection accuracy follows the non-linear relationship with the total number of neighbors for all sensor distribution. As the neighborhood count increases, the detection accuracy of the SVM-based algorithm also increases. Similar to the ANN-based detection scheme, it may suffer to detect wormhole attack if the wormhole nodes are placed in the sparsely populated area.

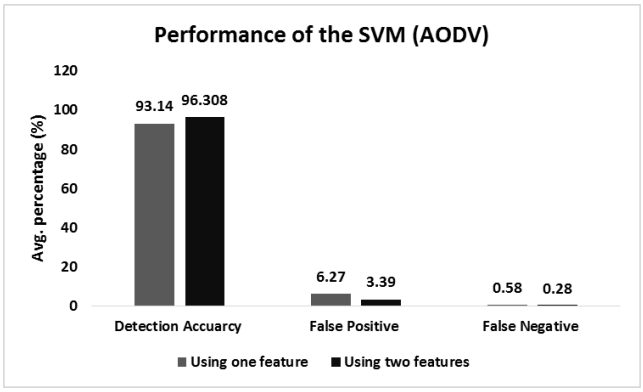


Fig. 24 Performance of the SVM-based detection scheme (AODV).

Fig. 24 shows that the performance of the SVM-based algorithm when the network uses AODV as the routing protocol. According to the Fig. 24, SVM-based detection scheme gives better the performance considering two featured data samples like the ANN-based model. SVM-based algorithm reaches approximately 96.30% on an average for all sensor distribution. Similarly, it achieves lowest false positive and false negative rates for two featured data samples which are 6.27% and 0.58% approximately. Using two featured data samples, the performance of the SVM-based algorithm enhances around 3.39%.

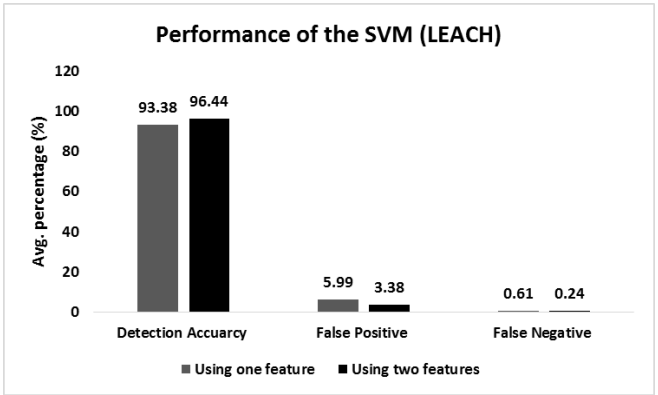


Fig. 25 Performance of the SVM-based detection scheme (LEACH).

Fig. 25 shows the performance of the SVM-based detection scheme considering LEACH protocol. Similar kind of trend in detection accuracy, shown in Fig. 24, is observed for the LEACH protocol. Like Fig. 24, SVM-based detection scheme performs better with the two featured data samples. It achieves averagely 96.44% detection accuracy for all sensor distribution, whereas, 93.38% detection accuracy is measured for the one featured data samples. In this case, the detection accuracy increases approximately 3.27% for the two featured data samples. Inversely, the false positive rates and false negative rates declined significantly for the two featured data samples.

B. The performance of the non-linear logistic regression:

Non-linear logistic regression (LR) algorithm is used to measure the performance on this classification problem and compare its outcomes with the proposed algorithm. Table 3 shows the parameters which are used during the training phase of the non-linear logistic classification algorithm.

Table 2 Parameters used for logistic linear classification

Parameter	Value
No of attributes	2
No of Data points (training)	49000
No of Data points (testing)	1000
Iteration	07
Regularized parameter (λ)	default
Learning method	Newton's methods
CPU time (for one instance)	3.17 mins

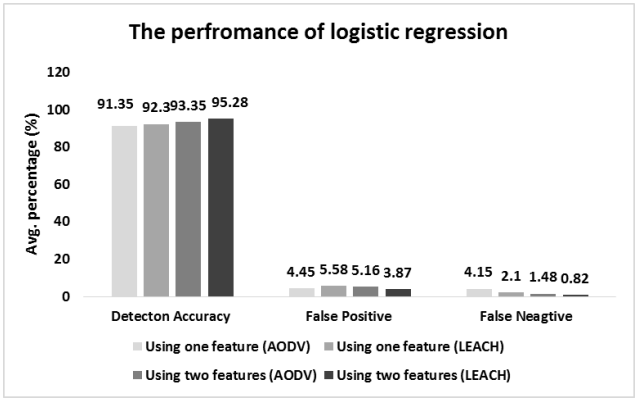


Fig. 26 Performance of the LR.

Fig. 26 represents the overall performance of the LR to detect the wormhole attack considering one featured and two featured data samples. According to the Fig. 26, LR based detection scheme performs better with two featured data samples when the network uses LEACH as the routing protocol. The detection accuracy reaches to 95.28 % considering two featured data samples. In this case, lowest false positive rate and false

negative rate are achieved for the LEACH routing protocol, which is 3.87% and 0.82% respectively. Most importantly, this LR based detection scheme gives better performance with two featured data samples like ANN or SVM based detection scheme.

C. Performance comparison

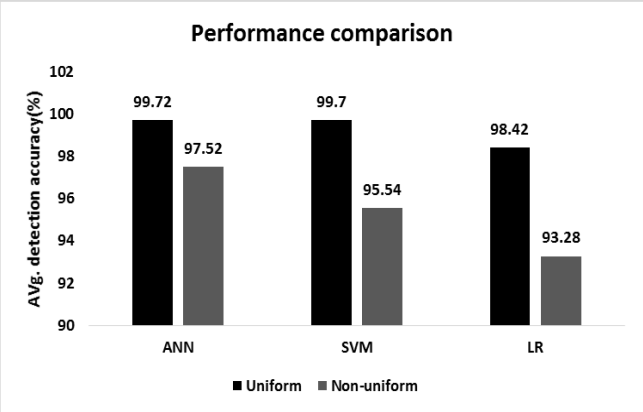


Fig. 27 Performace analysis on the uniform and the non-uniform sensor distribution.

Fig. 27 represents the performance of the detection schemes on the uniform and the non-uniform sensor distributions considering two featured data samples. Clearly, for the uniform sensor distribution, three (03) detection schemes achieve higher detection accuracy in contrast to the non-uniform sensor distributions. As the sensor density of the area is uniform; the neighborhood counts increase significantly with the presence of the wormhole nodes in the network. In contrary to the non-uniform sensor distribution, the neighborhood counts increase slightly or significantly depends on the position of the wormhole nodes in the network. Sometimes the neighborhood counts are smaller than the count taken from the wormhole non-infected zone (wormhole nodes can be placed in the sparsely populated area in the network). Since the uniform sensor distribution has consistency in the node density, the total neighbor population of the wormhole nodes is much smaller than the total neighborhood counts in non-uniform sensor distribution. As the neighbors of the wormhole nodes are the only way to reach the wormhole nodes; therefore, the neighbors of wormhole nodes in uniform distribution dissipate more energy than the neighbors of the malicious node in non-uniform sensor distribution. Hence, positive data samples are easily distinguished from the presented data set. That's why the detection accuracy of these schemes is much better for uniform sensor distribution.

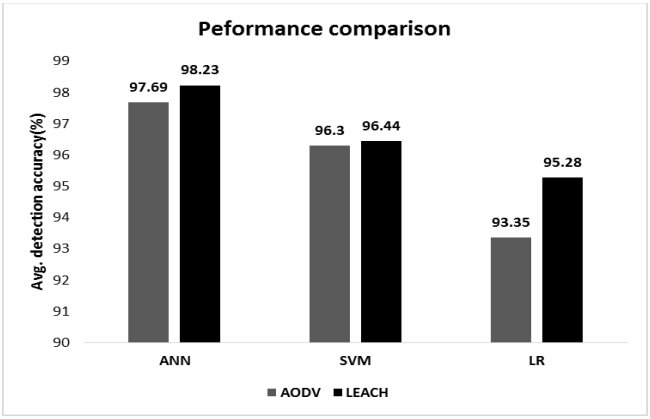


Fig. 28 Performance comparison among ANN, SVM, and LR.

If we analyze the performance of the three detection schemes (shown in the Fig. 28), ANN outperforms SVM and LR in detecting the malicious samples that represent wormhole attack considering both AODV and LEACH routing protocol. ANN achieves 97.69% and 98.23% detection accuracy respectively for AODV and LEACH routing protocol. SVM based detection scheme performs better than LR based detection scheme (achieves approximately 96% detection accuracy) in view of AODV and LEACH. Considering LEACH routing protocol, the LR based detection scheme attains 95.28% detection accuracy which is the lowest among all the detection schemes. As we know, the performance of the machine learning techniques highly depends on the dataset. In this case, ANN distorts the two featured data samples in the higher dimension better than other two machine learning techniques; and efficiently put dynamic decision surface in between positive and negative data samples. That would help the ANN-based detection scheme to gain highest detection accuracy.

Another important observation of this research is that the three detection schemes give better performance in classifying positives data samples for LEACH protocol. In the cluster based routing protocol like LEACH, a node is nominated as a cluster head among the members of a cluster for a specific round of data transmission to the base station. Once the cluster head is selected, other members send their data packets to the cluster head. Afterward, selected cluster head transfers the data packets to the base station. For the next round, a new member is randomly chosen as cluster head from the other members who is not selected as cluster head before for any round. However, wormhole node manipulates the cluster head selection process in a particular cluster of nodes. Furthermore, the 'in-band' wormhole node can be selected as a cluster head. If it happens, the wormhole node gets a bunch of data packets easily from the other members through its neighbors. As we know, wormhole nodes create shortcut path to the base station in the network. Therefore, the cluster head is deceived throughout the data transmission process to the base station. If wormhole nodes are placed in two different clusters, the members of the cluster may also be deceived while

they are selecting closest cluster head to transmit their data packets. Hence, the neighbors around wormhole node dissipate more energy in LEACH compared to AODV. The extra dissipation of the energy reflects in the feature called AREPN collected by detector node. That would help in separating the positive data samples and negative data samples when two featured data samples are presented to the detection schemes.

9. Conclusions

Wormhole attack is one of the detrimental network layer attacks for wireless sensor networks. This thesis presents a novel detection model based on neighborhood count and AREPN using an ANN for wireless sensor networks. The goal of this proposed detection scheme is to detect wormhole attacks (In-band, out of band, hidden or active mode of wormhole attack) with higher precision and accuracy, especially in a non-uniform network environment. The experimental results confirm that the proposed detection scheme is able to identify the existence of the wormhole node without requiring any special hardware both in uniform and non-uniform sensor distribution. Another important aspect of this detection model is it doesn't increase the significant amount of network overhead flow throughout the network. The simulation results also validate that our ANN-based approach performed better than SVM or LR based detection scheme. This ANN based detection scheme achieves around 99.72% (average) and 97.52% (average) accordingly for uniform and non-uniform sensor distribution. Most significantly, the probable location of the wormhole node can be identified by this scheme. Future works are required to enhance the performance of the detection scheme and to investigate in different directions so that we can evaluate the efficacy of the proposed detection scheme perfectly.

Proposed future works are noted as follows.

a) Nowadays, deep neural networks and convolutional neural networks are used to enhance the performance internet threat detection schemes. We want to apply this advanced machine learning algorithms on the acquired data set to evaluate their performance. After that, we will compare the results with the Proposed ANN based scheme.

b) In this research, the relationship between detection accuracy and the total count of neighbors around wormhole nodes is investigated. In the future work, we want to investigate the impact of changes in radio range of sensor nodes (60m,70m, and 80m etc.) and number of the sensor nodes (deployed) and position of the wormhole nodes on the detection accuracy of the proposed model.

c) If we use more than one mobile node, the two featured data samples will be gathered more quickly (detection time) rather than using single detector node. Multiple detector nodes will help to locate the positions of the wormhole nodes by triangulation algorithm. Therefore, in the future work, we want to investigate the performance of the proposed detection scheme by deploying multiple detector nodes in the different sensor network environments.

d) As we know, the performance of the ANN depends on its own architecture. In the future work, we want to apply the acquired data set on different network architecture (Varying number of neurons in a layer, varying number of the hidden layer) to optimize the performance of the proposed ANN-based detection scheme.

e) In this detection scheme, probable location of the wormhole node is identified through D_N. We want to work further in this direction to find the exact location of the malicious nodes in the infected sensor network.

f) The energy dissipation of a sensor node depends on the routing protocol that being used in the network. In this research, the performance of the proposed detection model is verified considering on demand based and cluster based routing protocol. In the future, we want to the test the proposed algorithm on the hybrid routing protocol such as zoned based routing protocol (ZRP) and wireless ad-hoc routing protocol (WARP).

Reference

[1] J. Yick, B. Mukherjee, and D. Ghosal, "Wireless sensor network survey," *Comput. Networks*, vol. 52, no. 12, pp. 2292–2330, 2008.

[2] S. K. Jangir and N. Hemrajani, "A comprehensive review on detection of wormhole attack in MANET," in *2016 International Conference on ICT in Business Industry & Government (ICTBIG)*, 2016, pp. 1–8.

[3] M. Sharma, A. Jain, and S. Shah, "Wormhole attack in mobile ad-hoc networks," in *2016 Symposium on Colossal Data Analysis and Networking (CDAN)*, 2016, pp. 1–4.

[4] M. E.-S. Marianne Azer, Sherif El-Kassas, "A Full Image of the Wormhole Attacks Towards Introducing Complex Wormhole Attacks in wireless Ad Hoc Networks," *Int. J. Comput. Sci. Inf. Secur.*, vol. 1, no. 1, pp. 41–52, 2009.

[5] Y. Xu, G. Chen, J. Ford, and F. Makedon, "Detecting wormhole attacks in wireless sensor networks using connectivity information," *Crit. Infrastruct. Prot.*, vol. 2006, 2007.

[6] L. Hu and D. Evans, "Using Directional Antennas to Prevent Wormhole Attacks," *Netw. Distrib. Syst. Symp. NDSS*, no. February, pp. 1–11, 2004.

[7] Y.-C. Hu, a. Perrig, and D. B. Johnson, "Packet leashes: a defense against wormhole attacks in wireless networks," *IEEE INFOCOM 2003. Twenty-second Annu. Jt. Conf. IEEE Comput. Commun. Soc. (IEEE Cat. No.03CH37428)*, vol. 3, no. C, pp. 1976–1986, 2003.

[8] N. Sastry, U. Shankar, and D. Wagner, "Secure verification of location claims," *Proc. 2003 ACM Work. Wirel. Secur. WiSe 03*, vol. 0, no. Section 2, pp. 1–10, 2003.

[9] N. Song, L. Qian, S. Ning, Q. Lijun, and L. Xiangfang, "Wormhole attacks detection in wireless ad hoc networks: a statistical analysis approach," *Parallel Distrib. Process. Symp. 2005. Proceedings. 19th IEEE Int.*, p. 8 pp., 2005.

[10] L. Buttyán, L. Dóra, and I. Vajda, "Statistical Wormhole Detection in Sensor Networks," *Secur. Priv. Adhoc Sens. Networks*, pp. 128–141, 2005.

[11] S. Song and H. Wu, "Statistical Wormhole Detection for Mobile Sensor Networks," pp. 322–327, 2012.

[12] Mohammad Nurul Afsar Shaon and Ken Ferens, "Wireless Sensor Network Wormhole Detection using an Artificial Neural Network," *ICWN*, pp. 115–120, 2015.

[13] M. a Azer, S. Member, S. M. El-kassas, M. S. El-soudani, and S. Member, "An Innovative Approach for the Wormhole Attack Detection and Prevention In Wireless Ad Hoc Networks," pp. 366–371, 2010.

[14] Y. Zhou, L. Lamont, and L. Li, "Wormhole attack detection based on distance verification and the use of hypothesis testing for wireless ad hoc networks," *2009 IEEE Mil. Commun. Conf. MILCOM 2009*, 2009.

[15] J. Tian, M. Gao, and F. Zhang, "Network Intrusion Detection Method Based on Radial Basic Function Neural Network," *2009 Int. Conf. E-bus. Inf. Syst. Secur.*, pp. 1–4, 2009.

[16] S. Chattopadhyay and G. Bandyopadhyay, "Artificial neural network with backpropagation learning to predict mean monthly total ozone in Arosa, Switzerland," *Int. J. Remote Sens.*, vol. 28, no. 20, pp. 4471–4482, 2007.

[17] T. Okamoto, "Introduction to wormholes," pp. 1–14, 1988.

[18] M. Meghdadi, S. Ozdemir, and I. Gueler, "A Survey of Wormhole-based Attacks and their Countermeasures in Wireless Sensor Networks," *Iete Tech. Rev.*, vol. 28, no. 2, pp. 89–102, 2011.

[19] C. Stergiou and Siganos Dimitrios, "Neural Networks," 2015. [Online]. Available:https://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html. [Accessed: 15-Apr-2017].

[20] M. Khabbazian, H. Mercier, and V. K. Bhargava, "NIS02-1: Wormhole Attack in Wireless Ad Hoc Networks: Analysis and Countermeasure," *IEEE Globecom 2006*, vol. 8, no. 2, pp. 736–745, 2006.

[21] P. Lee and S. Member, "A Passivity Framework for Modeling and Mitigating Wormhole Attacks on Networked Control Systems," *Tac*, vol. 59, no. 12, pp. 3224–3237, 2014.

[22] A. Rasheed and R. Mahapatra, "Mobile sink using multiple channels to defend against wormhole attacks in wireless sensor networks," *2009 IEEE 28th Int. Perform. Comput. Commun. Conf.*, pp. 216–222, 2009.

[23] T. Giannetsos, T. Dimitriou, and N. R. Prasad, "State of the art on defenses against wormhole attacks in wireless sensor networks," *Wirel. Commun. Veh. Technol. Inf. Theory Aerosp. Electron. Syst. Technol. 2009. Wirel. VITAE 2009. 1st Int. Conf.*, pp. 313–318, 2009.

[24] A. D. Wood and J. A. Stankovic, "A Taxonomy for Denial-of-Service Attacks in Wireless Sensor Networks," *Handb. Sens. Networks Compact Wirel. Wired Sens. Syst.*, pp. 739–763, 2004.

[25] S. Ughade, R. K. Kapoor, and A. Pandey, "An Overview on Wormhole Attack in Wireless Sensor Network : Challenges , Impacts , and Detection Approach," vol. 2, no. 4, pp. 105–110, 2014.

[26] H. Yih-Chun and A. Perrig, "A survey of secure wireless ad hoc routing," *IEEE Secur. Priv. Mag.*, vol. 2, no. 3, pp. 28–39, 2004.

[27] J. S. Atel and Hinal, "Wormhole Attack Detection in Wireless Sensor Networks: A Survey.," *Int. J. Innov. Res. Dev.*, pp. 545–550, 2014.

[28] T. Korkmaz, "Verifying physical presence of neighbors against replay-based attacks in wireless ad hoc networks," *Int. Conf. Inf. Technol. Coding Comput. - Vol. II*, p. 704–709 Vol. 2, 2005.

[29] S. Capkun, L. Buttyan, and J.-P. Hubaux, "SECTOR: Secure Tracking of Node Encounters in Multi-hop Wireless Networks," *Proc. 1st ACM Work. Secur. ad hoc Sens. networks*, vol. 67322, no. 5005, pp. 21--32, 2003.

[30] F. Hong, L. Hong, and C. Fu, "Secure OLSR," *Proc. - Int. Conf. Adv. Inf. Netw. Appl. AINA*, vol. 1, pp. 713–718, 2005.

[31] P. Sarkar, C. Kar, B. Sen, and K. Sharma, "Sensitivity analysis on AODV with Wormhole attack," in *2016 2nd International Conference on Next Generation Computing Technologies (NGCT)*, 2016, pp. 803–807.

[32] C. Hon Sun and L. King-Shan, "DelPHI: wormhole detection mechanism for ad hoc wireless networks," *1st Int. Symp. Wirel. Pervasive Comput.* , no. 852, p. 6 pp., 2006.

[33] P. Van Tran, L. X. Hung, Y. K. Lee, S. Lee, and H. Lee, "TTM: An efficient mechanism to detect wormhole attacks in wireless ad-hoc networks," *2007 4th Annu. IEEE Consum. Commun. Netw. Conf. CCNC 2007*, pp. 593–598, 2007.

[34] S. Ali and P. Nand, "Comparative performance analysis of AODV and DSR routing protocols under wormhole attack in mobile ad hoc network on different node's speeds," in *2016 International Conference on Computing, Communication and Automation (ICCCA)*, 2016, pp. 641–644.

[35] D. Johnson and D. Maltz, "The Dynamic Source Routing Protocol (DSR) for Mobile Ad Hoc Networks for IPv4," pp. 1–107, 2007.

[36] K. U. R. Khan, R. U. Zaman, a. V. Reddy, K. A. Reddy, and T. S. Harsha, "An Efficient DSDV Routing Protocol for Wireless Mobile Ad Hoc Networks and its Performance Comparison," *2008 Second UKSIM Eur. Symp. Comput. Model. Simul.*, pp. 506–511, 2008.

[37] T. Van Phuong, N. T. Canh, Y. K. Lee, S. Lee, and H. Lee, "Transmission time-based mechanism to detect wormhole attacks," *Proc. 2nd IEEE Asia-Pacific Serv. Comput. Conf. APSCC 2007*, pp. 172–178, 2007.

[38] Z. T. and A. H. Maw., "Wormhole attack detection in wireless sensor networks," in *Proceedings of World Academy of Science Engineering and TechnologyEngineering and Technology*, 2008, vol. 46, no. 3, pp. 545–50.

[39] S. Yi, Y. Pei, and S. Kalyanaraman, "On the capacity improvement of ad hoc wireless networks using directional antennas," *Proc. 4th ACM Int. Symp. Mob. ad hoc Netw. Comput. - MobiHoc '03*, pp. 108–116, 2003.

[40] M. Takai, J. Martin, R. Bagrodia, and A. Ren, "Directional virtual carrier sensing for directional antennas in mobile ad hoc networks," *Proc. 3rd ACM Int. Symp. Mob. ad hoc Netw. Comput. - MobiHoc '02*, pp. 183–193, 2002.

[41] R. Ramanathan, "On the performance of ad hoc networks using beamforming antennas," *Proc. ACM MOBIHOC, Oct.*, pp. 1453–1458, 2001.

[42] S. Brands and D. Chaum, "Distance bounding protocols.pdf," in *In Theory and Application of Cryptographic TechniquesTheory and Application of Cryptographic Techniques*, 1993, pp. 344–359.

[43] Y. Hu, A. Perrig, and D. B. Johnson, "Wormhole Detection in Wireless Ad Hoc Networks," pp. 1–15, 2001.

[44] Ý. G. Özdemir and S. M. Meghdadi, "A time and trust based wormhole detection algorithm for wireless sensor networks," *3rd Inf. Secur. Cryptol. Conf.*, vol. 94, no. 20, pp. 1–5, 2008.

[45] A. A. Pirzada and C. McDonald, "Circumventing sinkholes and wormholes in wireless sensor networks," *Conf. Wirel. Ad Hoc Networks*, 2005.

[46] W. Wang and B. Bhargava, "Visualization of wormholes in sensor networks," pp. 51–60, 2004.

[47] K. B. Rasmussen and S. Capkun, "Implications of radio fingerprinting on the security of sensor networks," *Proc. 3rd Int. Conf. Secur. Priv. Commun. Networks, Secur.*, pp. 331–340, 2007.

[48] M. N. A. Shaon and K. Ferens, "Wormhole Attack Detection in Wireless Sensor Network using Discrete Wavelet Transform," *ICWN*, pp. 29–35, 2016.

[49] M. N. A. Shaon and K. Ferens, "Wormhole Attack Detection in Wireless Sensor Network using Variance Fractal Dimension," *SAM*, pp. 55–62, 2016.

[50] K. Ferens, "Applied Computational Intelligent (Lecture notes)," 2017. [Online]. Available: http://ece.eng.umanitoba.ca/graduate/ECE7650ACI/. [Accessed: 15-Apr-2017].

[51] W. Huang and L. Ju, "A novel intrusion detection method based on conjugate gradient neural network," in *2010 International Conference On Computer Design and Applications*, 2010, vol. 2, no. Iccda, pp. V2-470-V2-472.

[52] J. Ren, "ANN vs. SVM: Which one performs better in classification of MCCs in mammogram imaging," *Knowledge-Based Syst.*, vol. 26, pp. 144–153, 2012.

[53] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities.," *Proc. Natl. Acad. Sci.*, vol. 79, no. 8, pp. 2554–2558, 1982.

[54] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *Bull. Math. Biophys.*, vol. 5, pp. 115–133, 2008.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 321-326 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Real Time Advanced Clustering System

Giuseppe Spampinato[1*], Arcangelo Ranieri Bruna[1], Salvatore Curti[1], Viviana D'Alto[2]

[1]*STMicroelectronics, Advanced System Technology, Catania, Italy*

[2]*STMicroelectronics, Advanced System Technology, Agrate Brianza, Italy*

ARTICLE INFO

Article history:
Received: 30 March, 2017
Accepted: 24 April, 2017
Online: 16 May, 2017

Keywords:
Automotive
Control traffic alert
Optical flow
Clustering
Moving object detection

ABSTRACT

This paper describes a system to gather information from a stationary camera to identify moving objects. The proposed solution makes only use of motion vectors between adjacent frames, obtained from any algorithm. Starting from them, the system is able to retrieve clusters of moving objects in a scene acquired by an image sensor device. Since all the system is only based on optical flow, it is really simple and fast, to be easily integrated directly in low cost cameras. The experimental results show fast and robust performance of our method. The ANSI-C code has been tested on the ARM Cortex A15 CPU @2.32GHz, obtaining an impressive fps, about 3000 fps, excluding optical flow computation and I/O. Moreover, the system has been tested for different applications, cross traffic alert and video surveillance, in different conditions, indoor and outdoor, and with different lenses.

1. Introduction

Clustering is the task of grouping a set of information in such a way that information in the same group (called a cluster) are more similar to each other than to those in other groups. A clustering system, in particular regarding detection and tracking of moving objects, is a really common and important task, widely used in computer vision as low-level task for different applications like: camera surveillance [1], Cross Traffic Alert (CTA) [2], Intelligent Transport System (ITS) [3], object detection and recognition [4], people detection and recognition [5], people behavior analysis (e. g. anti-theft, falling detection, etc.) [6] and user interfaces by gesture [7]. Efficiency and robustness of clustering, considered as low-level task, is really important because it influences the performance of all higher level tasks of the aforementioned applications.

To implement the clustering task, there are in general two types of approaches: region based and boundary based. For what concerns region based, different approaches have been proposed: combined color based and region based particle filters, using multiple hypotheses for tracking objects [8], motion-based segmentation and a region-based Mean-shift tracking approach, fused by a Kalman filter, to extract object independent motion trajectory under uncontrolled environment [9], and so on. However the most popular and recent region based approaches is background subtraction [10-14]. Once the background is removed from the scene, through different kind of filters applied on the foreground, it is possible to identify moving object. Since these methods are pixel-based, they are not suitable for real time applications, also because they require a long time for estimating the background models.

In boundary based approaches, different features have been proposed: color feature descriptors, based on the spectral power distribution of the illuminant and object's surface reflectance property [15], multi-feature descriptors [16], and so on. However, many recent approaches use edge based optical flow [17-20]. Optical flow estimation gives a two-dimensional motion vector, which represents the movement of some point of an image in the following one of the sequence. Basic idea in boundary based approaches is to consider vectors with similar spatial and motion information belong to the same object [21]. Since these approaches are feature-based, they are a good choice for all real time applications, even if usually the do not reach high performances.

In this work, we propose a very low-cost clustering system, working with a generic optical flow and reaching high performances in different application scenarios. This paper is organized as follows: Section II describes the proposed system, block by block, also showing visual impact on a CTA example; Section III shows the experimental results for different application

*Corresponding Author: Giuseppe Spampinato, STMicroelectronics, Advanced System Technology, Catania, Italy | Email: giuseppe.spampinato@st.com

cases; finally, conclusion and future work are exposed in Section IV.

2. Proposed System

Since we are interested in light systems, we started our investigation with optical flow based algorithm, which system [22,23] is described in Fig. 1. The standard system takes as input the optical flow, that is the list of motion vectors between adjacent frames. First step is usually a pre-filtering to eliminate noise and/or spikes. Afterwards, labelling step inserts same label for each "similar" vector. The role of clustering step is to group vectors with the same label into clusters. At the end, "similar" clusters are merged together to obtain the final list of identified clusters. In this system, some steps (in particular pre-filtering) can be computationally heavy and no previous vector information is used. Moreover, no temporal cluster information is used to enhance clustering, but the previous clustering information is usually used only for tracking.

To overcome the aforementioned drawbacks of typical optical flow based system, we propose a novel low-cost system only based on optical flow, described in Fig. 2. The new blocks have been added to increase performances of the system. Moreover, previous frame is not necessary, but just a list of previous vectors and clusters are needed for processing. Detailed description of each block is explained in the next sub-paragraphs. Moreover, to better understand the peculiarity of various blocks, example images are shown.

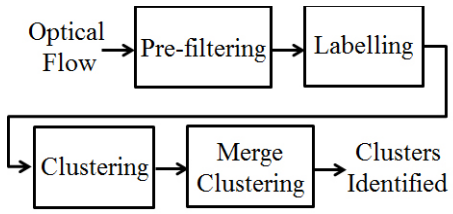


Fig. 1. Typical optical flow based system.

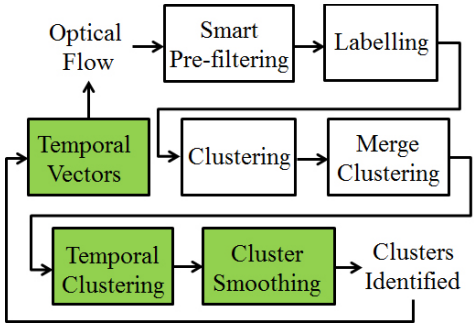


Fig. 2. Proposed system

2.1. Smart Pre-filtering

This starting block takes as input the optical flow, the set of motion vectors generated in any manner, and its role is to remove from this set all vectors which can be considered as noise. Usually complex filtering is used as for example median filters [22] and so on. We propose a new simple filtering with two different options, applied depending on the quality of incoming optical flow.

First option, called *brute-force*, just eliminates motion vectors with small movements, which are considered as noise. This option is indicated just for low quality optical flow. In particular, a motion vector $v=(dx,dy)$ is removed from optical flow if the following condition is satisfied:

$$(ABS(dx) \leq MAX\_DX) \text{ and } (ABS(dy) \leq MAX\_DY) \quad (1)$$

Where MAX_DX and MAX_DY are thresholds chosen depending on the precision of the optical flow and ABS is the absolute value.

Second option, called *smart*, apart applying (1), do not remove in the output optical flow, the motion vector v with the following characteristics: it lies in the area around a cluster C of previous frame and its orientation is similar to the orientation of C, that is:

$$ABS((orientation(v) - orientation(C)) \leq MAX\_ORIENT1) \quad (2)$$

Where *MAX_ORIENT1* threshold should be opportunely chosen, depending on specific application. In particular, for rigid objects, as vehicles, a small value of *MAX_ORIENT1* is a good choice, while for not rigid objects, such as people, we need to increase its value.

About *orientation*, we have to distinguish between orientation of a vector $v=(dx,dy)$ and orientation of a cluster C. In the first case, the formula is the following:

$$orientation(v) = tan^{-1}(dy/dx) \quad (3)$$

Instead, orientation of a cluster C, it is defined as the mean of the orientation of all motion vectors included in C. To execute the tan^{-1} function at real-time a fixed point arctangent algorithm was implemented (http://www.coranac.com/documents/arctangent/).

Figure 3 shows an input optical flow example. In this example we have three cars moving: the middle one to the left, the other two to the right. We can easily notice in the scene some noisy vectors in particular around the leafs of the surrounding trees, which is a typical problem in these kind of applications.

Fig. 3. CTA example: input optical flow

2.2. Labelling

This block takes as input the filtered optical flow. Its role is to insert a label for each motion vector, assuring to insert the same label for motion vectors witch potentially belong to the same cluster. Two motion vectors $v_1=(dx_1,dy_1)$, spatially situated in point $p_1=(x_1,y_1)$, and $v_2=(dx_2,dy_2)$, spatially situated in point $p_2=(x_2,y_2)$, have the same label if the following three conditions are satisfied:

$$(ABS(x1-x2) + ABS(y1-y2)) \leq MAX\_DIST\_POINT \quad (4)$$

$$(ABS(dx1-dx2) + ABS(dy1-dy2)) \leq MAX\_MOD \quad (5)$$

$$ABS((orientation(v1) - orientation(v2)) \leq MAX\_ORIENT2) \quad (6)$$

Where the thresholds should be opportunely chosen: *MAX_DIST_POINT*, depending on specific application and typical objects distance; *MAX_MOD* depending on precision of optical flow; *MAX_ORIENT2* depending on specific application.

The above formulas indicate that two motion vectors are considered to belong to the same cluster if they are spatially near (4), have got similar module (5) and similar orientation (6). While (4) and (5) are commonly used for labelling [23], (6) has been added for robustness of the system.

2.3. Clustering

This block takes as input the filtered optical flow, obtained by *smart pre-filtering* block and the related labels, coming from *labelling* block elaboration. The aim of this step is to group vectors with the same label into clusters. To avoid having clusters with dimensions too small, it is important to extend the cluster found to a minimum dimension N×*M*. This dimension depends on various factors: typical object size we want to cluster, typical distance of the object from the camera and the type of the camera (liner or fish-eye).

Figure 4 shows in boxes the nine clusters, obtained from *clustering* step in the CTA example of Fig. 3. From the center of each cluster, its *orientation* is displayed to better understand the next *merge clustering* step, which will reduce the number of clusters.

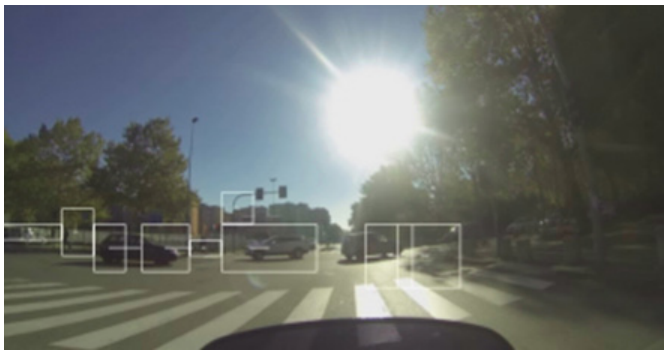


Fig. 4. CTA example: clustering step output

2.4. Merge Clustering

This block takes as input a list of clusters, coming from previous *clustering* block, and merges together similar clusters to obtain a new list of identified clusters. Two clusters C_1 and C_2 are considered similar and then merged if the following conditions are satisfied:

$$distance\_X(C_1,C_2) \leq MAX\_DIST\_CLUSTER\_X \quad (7)$$

$$distance\_Y(C_1,C_2) \leq MAX\_DIST\_CLUSTER\_Y \quad (8)$$

$$ABS((orientation(C_1) - orientation(C_2)) \leq MAX\_ORIENT3) \quad (9)$$

Where the thresholds should be opportunely chosen: *MAX_DIST_CLUSTER_X* and *MAX_DIST_CLUSTER_Y*, depending on application and typical objects distance; *MAX_ORIENT3*, depending on application.

In (7) and (8), the distance for clusters concept is analogous to distance for points concept in (4) in *labelling* step; while (9) is analogous to (6). The above formula indicates that two clusters are similar if they are spatially near (7-8) and have got similar orientation (9). Since for each cluster we have its orientation, it is easy to extend the similarity also to the module, in analogous way of (5).

Looking at Fig. 4, from left to right, the first car is spit in two cluster and these clusters have got the similar orientation, so the *merge clustering* step will join these two clusters. The middle car is well detected, but since there are near clusters with similar orientation, these clusters will be joined together. In the last car, we can notice two near clusters, but since they have opposite orientation, they will be not joined together. Figure 5 shows, marked by ellipses, similar clusters that *merge clustering* step will join, while Fig. 6 shows the merged clusters, now reduced to six.

Fig. 5. CTA example: similar clusters to merge

Fig. 6. CTA example: merge clustering step output

2.5. Temporal Clustering

This block takes as input two lists of clustering: current frame list and previous frame list. The roles of this block are two: the first one is similar to the *merge clustering* block, that is to merge similar clusters in current and previous frame; while the second role is to eliminate clusters in current frame which do not have similar clusters in the previous frame. Similarly to the method described in Sub-section D, two clusters are similar if condition (7), (8) and (9) are satisfied. The difference, compared to *merge clustering* block, are the following: for *temporal clustering* block C_1 and C_2 are two clusters coming from two adjacent frames; *MAX_DIST_CLUSTER_X* and *MAX_DIST_CLUSTER_Y* thresholds can be different in value. Of course, as said before, in this block we can use the module as similarity condition. Another important note is that, to increase reliability of the system, assuming constant speed of objects to identify, the clusters of previous frame can be motion compensated.

In the CTA example, looking Fig. 6 from left to right, the first two and the last clusters will be removed, since they do not have similar clusters in previous frame. Figure 7 shows the resulting clusters, reduced to three, as the real moving objects in the scene.

Fig. 7. CTA example: temporal clustering step output

2.6. Cluster Smoothing

This block takes as input the list of clusters, coming from the previous *temporal clustering* block, and the previous frame clustering list. Clusters marked as similar, in the temporal clustering block, to avoid big discrepancy in dimensions, should be tuned according to the similar clusters in the previous frame. In particular, a weighted average between previous C_1 and current C_2 similar clusters should be calculated as follows:

$$C_2 = (C_2 * CURRENT\_W + C_1 * PREVIOUS\_W)/SUM\_W \quad (10)$$

With:

$$SUM\_W = CURRENT\_W + PREVIOUS\_W \quad (11)$$

Fig. 8. CTA example: cluster smoothing output

The weights *CURRENT_W* and *PREVIOUS_W* should be opportunely chosen, usually giving more weight on the previous cluster. As said for *temporal clustering* step, to increase reliability of the system, the clusters of previous frame can be motion compensated.

The final effect of this block is shown in Fig. 8, where we can note that dimensions of the three clusters are more similar to the real dimensions of objects. This is the final output of the proposed schema.

2.7. Temporal Vectors

It is important, in particular for not dense optical flow, at the end of the proposed system, to insert in the optical flow of the next frame all vectors contained in the current cluster, to limit clusters disappearing in the scene. Of course, these inserted vectors can be motion compensated, assuming constant speed.

2.8. Tracking

Since in *temporal clustering* step, we obtain similarity and then relationship between previous and current frame clusters, it is easy to calculate the trajectory of the identified object in the scene. The cluster centroid can be obtained and, according to relationship between clusters, displayed in the scene.

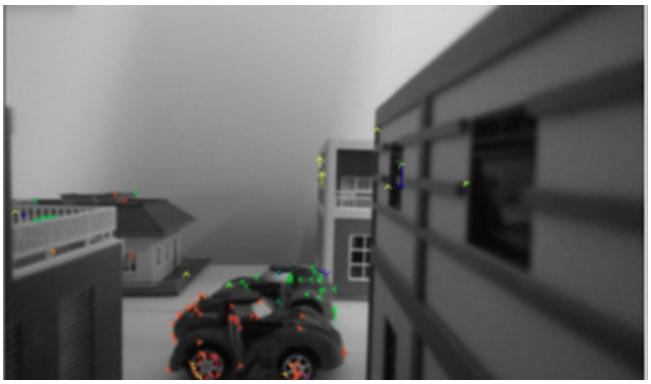


Fig. 9. Simulated example: input optical flow.

Fig. 10. Simulated example: proposed system output.

3. Experimental Results

Lots of tests have been executed with different scenarios, as CTA and video surveillance, and different cameras at different resolutions, with both linear and fish-eye lenses, obtaining really good visual results.

Speed testing has been made calculating computation time for VGA sequences on a 2.3GHz processor (ARM Cortex-A15, using a single core). About 3000 frame/sec were obtained, executing only the described ANSI-C algorithm (excluding the optical flow and I/O calculation).

Considering that the memory required by the system is really low, about 96K considering 4K optical flow for frame, and considering the aforementioned speed test, we can assure that the system is suitable also for really low-cost devices.

An example of simulated CTA with partially occluded objects is shown in Fig. 9 and Fig. 10. In particular the proposed system, with the input motion vectors in Fig. 9, correctly identifies and well spots the two partially occluded crossing cars, as shown in Fig. 10.

An example of real CTA with fish-eye lens, with multi objects moving, is shown in Fig. 11 and Fig. 12. In particular the proposed system, with the input motion vectors in Fig. 11, correctly identifies the three crossing cars shown in Fig. 12.

Fig. 11. Fish-eye example: input optical flow.

Fig. 12. Fish-eye example: proposed system output.

An example of video surveillance is shown in Fig. 13 and Fig. 14. In particular, the proposed system, with the input motion vectors in Fig. 13, correctly identifies the person in Figure 14. It is to underline that the proposed system not only correctly identifies the man leaving the bag, but also bag itself.

The last case can be useful also to recognize the unattended luggage application, really important for example for airport security.

Fig. 13. Video surveillance example: input optical flow.

Fig. 14. Video surveillance example: proposed system output.

4. Conclusion and future work

The proposed approach has been experimentally tested on a representative dataset of scenes obtaining effective results in terms of accuracy. A very reliable and low-cost clustering system has been developed with the characteristic to be very low power consumption, since it works only with optical flow information. The proposed system is also very flexible: can be used with any algorithm which estimates motion vectors between adjacent frames. An important characteristic is the consideration of previous frame clustering and vectors information for a more robust clustering determination in the current frame. Moreover, average of current and previous related cluster, to maintain more graceful clustering output for the user.

A limitation of this work is that we consider just stationary camera, so can be used for some specific applications, like CTA and video surveillance. Another limitation is related to input optical flow: if the image is too dark or blurred, or moving object is too speed, poor optical flow will be produced and the proposed method will fail.

Future work will be orientated on overcoming this limitations, leaving the system reliable also in the case of moving cameras, introducing robust filtering to remove from input optical flow the global motion vector, opportunely calculated.

Conflict of Interest

The authors declare no conflict of interest regarding the publication of this paper.

References

[1] D. Coppi, S. Calderara, R. Cucchiara, "Transductive People Tracking in Unconstrained Surveillance" in IEEE Transactions on Circuits and Systems for Video Technology, 2016.

[2] V. Balisavira, V. K. Pandey, "Real-time Object Detection by Road Plane Segmentation Technique for ADAS" in Eighth International Conference on Signal Image Technology and Internet Based Systems (SITIS), 2012.

[3] N. Seenouvong, U. Watchareeruetai, C. Nuthong, K. Khongsomboon, N. Ohnish, "A computer vision based vehicle detection and counting system" in Eighth International Conference on Knowledge and Smart Technology (KST), 2016.

[4] V. Kalogeiton, V. Ferrari, C. Schmid, "Analysing domain shift factors between videos and images for object detection" in IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016.

[5] A. García-Martín, J. M. Martínez, "People detection in surveillance: classification and evaluation" in IET Computer Vision, 2015.

[6] K. Ozcan, S. Velipasalar, "Wearable Camera- and Accelerometer-Based Fall Detection on Portable Devices" in IEEE Embedded Systems Letters, 2016.

[7] X. Dang, W. Wang, K. Wang, M. Dong, L. Yin, "A user-independent sensor gesture interface for embedded device" in IEEE Sensors, 2011.

[8] S. Kumar, M.S. Narayanan, P. Singhal, J.J. Corso, V. Krovi, "Product of tracking experts for visual tracking of surgical tools" in IEEE International Conference on Automation Science and Engineering (CASE), 2013.

[9] X. Wu, X. Mao, L. Chen, A. Compare, "Combined Motion and Region-Based 3D Tracking in Active Depth Image Sequence" in IEEE International Conference on Cyber, Physical and Social Computing, 2013.

[10] D. K. Panda, S. Meher, "Detection of Moving Objects Using Fuzzy Color Difference Histogram Based Background Subtraction" in IEEE Signal Processing Letters, 2016.

[11] M. Anandhalli, V. P. Baligar, "Improvised approach using background subtraction for vehicle detection" in IEEE International Advance Computing Conference (IACC), 2015.

[12] O. E. Harrouss, D. Moujahid, H. Tairi, "Motion detection based on the combining of the background subtraction and spatial color information" in Intelligent Systems and Computer Vision (ISCV), 2015.

[13] T. Huynh-The, O. Banos, S. Lee, B. H. Kang, E. Kim, T. Le-Tien, "NIC: A Robust Background Extraction Algorithm for Foreground Detection in Dynamic Scenes" in IEEE Transactions on Circuits and Systems for Video Technology, 2016.

[14] S. Battiato, G. M. Farinella, A. Furnari, G. Puglisi, A. Snijders, J. Spiekstra, "An Integrated System for Vehicle Tracking and Classification" in Expert Systems with Applications, Vol. 42, Issue 21, pp. 7263–7275, 2015.

[15] A. Yilmaz, O. Javed, M. Shah, "Object Tracking: A Survey" in ACM Computing Surveys, 2006.

[16] H. Yang, L. Shao, F. Zheng, L. Wangd, Z. Song, "Recent advances and trends in visual tracking: A review" in Elsevier Neurocomputing, 2011.

[17] C. Huang, M. Hung, "Target motion compensation with optical flow clustering during visual tracking" in IEEE 11th International Conference on Networking, Sensing and Control (ICNSC), 2014.

[18] C. Wong, W. C. Siu, S. Barnes, P. Jennings, B. Fong, "Shared-use motion vector algorithm for moving objects detection for automobiles" in IEEE International Conference on Consumer Electronics (ICCE), 2016.

[19] J. Hariyono, K. Jo, "Detection of pedestrian crossing road" in IEEE International Conference on Image Processing (ICIP), 2015.

[20] C. Liang, C. Juang, "Moving Object Classification Using a Combination of Static Appearance Features and Spatial and Temporal Entropy Values of Optical Flows" in IEEE Transactions on Intelligent Transportation Systems, 2015.

[21] W.-C. Lu, Y.-C. F. Wang, C.-S. Chen, "Learning dense optical-flow trajectory patterns for video object extraction" in IEEE International Conference on Advanced Video & Signal-based Surveillance (AVSS), 2010.

[22] S. Aslani, H. Mahdavi-Nasab, "Optical Flow Based Moving Object Detection and Tracking for Traffic Surveillance" in World Academy of Science, Engineering and Technology International Journal of Electrical, Robotics, Electronics and Communications Engineering, 2013.

[23] M. Yokoyama, T. Poggio, "A Contour-Based Moving Object Detection and Tracking" in Proceedings of the 14th International Conference on Computer Communications and Networks (ICCCN), 2005.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 327-337 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Indoor Mobile Robot Navigation in Unknown Environment Using Fuzzy Logic Based Behaviors

Khalid Al-Mutib[1], Foudil Abdessemed[2*]

[1]*Department of Software Engineering, King Saud University, Saudi Arabia*

[2]*Department of Electronics, Faculty of technology, Batna University, 05000, Algeria*

ARTICLE INFO

Article history:
Received: 03 April, 2017
Accepted: 27 April, 2017
Online: 17 May, 2017

Keywords:
Wall following
Mobile robot
Target reaching
Obstacle avoidance
Interval Type-2 fuzzy control

ABSTRACT

This paper presents the development and design of a methodology based on fuzzy logic to control an indoor mobile robot for a complete navigation in an unknown environment. The methodology incorporates two basic behaviors, namely: reaching the goal and avoiding obstacles. The obstacle avoidance behavior is treated using wall-following scheme based on an Interval Type-2 Fuzzy Inference technique. .This helps in handling data uncertainties to produce a better performance. The mobile robot control mechanism uses some sort of knowledge base arranged in a set of fuzzy-rule-base to implement the wanted behavior that makes the mobile robot follow the boundary of an obstacle or a wall. A constant distance to the obstacle/wall is maintained while the robot tries successfully to get around this difficulty. Once the path is clear, the obstacle avoidance behavior is inhibited and reaching the goal behavior is activated using a secondary fuzzy controller. This methodology was successfully tested on a real mobile robot for different sort of scenarios. In order to provide better insight into the work's objective, a comparison work with another method, which uses a Partial Swarm Optimization-Fuzzy method, is carried out based on some defined criteria. The experiment shows a better improvement in the results of the proposed method.

1. Introduction

Mobile robots have a special place in robotics. Their interest lies in their mobility, which covers applications in many fields (nuclear, navy, space, firefighting, surveillance, load transportation ...). The particular aspect of mobility requires a methodological and technological complexity that adds to the general problems faced by mobile robots. Solving these problems requires the use of all available resources both in technology (sensors, motor, and energy) and in information processing by using artificial intelligence techniques or special processors. The autonomy of a mobile robot is a faculty, which enables it to adapt or make a decision in order to perform a task despite a lack of preliminary or possibly erroneous information. In other cases, such as vehicle exploration of planets [1], garbage collector [2], foraging [3], autonomy is a fundamental issue because remote control is impossible due to the propagation delay. The main focus of such research directions is how to let a mobile robotic body navigate in an unstructured environment without collisions; and how to combine these two behaviors in order to achieve the required robot tasks. In fact, if we arrive to control the robot to achieve the first behavior, it would be a difficult task for the robot to navigate in a cluttered environment. One of the solutions which has been an interesting research topic over the past few years is wall following. This behavior is one of the multi-issues that can be used to avoid obstacles. In fact, the robot moves along the object while keeping a certain safe distance. This behavior may be used to follow a real wall, concave or convex objects, and labyrinths. Wall-following strategy could be very helpful when the mobile robot is being trapped in deadlocks caused by local minima. Despite the number of publications appeared in this field of research, mobile robot navigation is still considered as a topic of extensive research.

Many papers have dealt with the topic of mobile robot navigation using different approaches. Conventional approaches have limitations and cannot respond optimally to non-linear dynamics and the change of system parameters. In the literature, one can find new approaches based on artificial intelligence

*Corresponding Author: Foudil Abdessemed, Department of Electronics, Faculty of technology, Batna University, 05000, Algeria
Email: f.abdessemed@univ-batna2.dz

developed to cope with these limitations and fuzzy logic theory was used extensively for this purpose [4]. The obstacle avoidance problem is also considered as a challenge when designing an autonomous mobile robot, and this issue has been tackled using many approaches [5]. On the other hand, signals captured by sensors, when an obstacle is detected, are contaminated with noise from different sources and probably associated with other unwanted signals. Therefore, the problem of detecting the obstacles and positioning its emplacement poses the problem of uncertainties, precision and accuracy. In this case, new control methods based on artificial intelligence, such as fuzzy logic, were introduced. Liu *et al* [6] analyzed the relationship between the motion characteristics and installation position of sensors of mobile robots and proposed a self-convergence mathematical model. The purpose of their work is to execute the wall-following activity with only a single distance proximity switch. Based on a spiking neural network controller, Wang *et al*. [7] used sonar sensory readings to guide their mobile robot to follow the wall. The idea of using fuzzy rules to implement a wall-following-based obstacle controller has been proposed in [8], a fuzzy system based on the concept of general perception was proposed to control the mobile robot navigating along walls of arbitrary shape and around obstacles. The autonomous mobile robot (MR) in [9] is equipped with two IR proximity sensors. The wall following is based merely on sensorial information given by these two IR sensors. In order to improve noise resistance ability, some researchers applied control laws based on interval type-2 fuzzy logic [10-13]. Hsu et al, [10] proposed an evolutionary wall-following control of a mobile robot using an interval type-2 fuzzy controller with Species-Differential-Evolution activated Continuous Ant Colony Optimization (SDE-CACO). All fuzzy rules are generated online using a clustering-based approach during the evolutionary learning process. . Other works investigates a control based on Inverted Ant System IAS, such as [4] and other investigated Cellular Automata and Inverted Ant System to control a swarm of robots in foraging task [3]. The control and other aspects implementation, such as, coordination and synchronization are described in [11]. An interesting work has been proposed in [14], where the authors suggested a reinforcement and optimized fuzzy controller method for mobile robot wall following. The method was successfully tested on the mobile robot Pioneer 3-DX. However, the methodology requires a learning procedure, which requires gathering data and a lot of computation time.

This paper proposes a new fuzzy logic methodology to control an indoor mobile robot for a complete navigation in an unknown environment. This methodology is based on two basic behaviors, reaching-the-goal while avoiding unforeseen obstacles. The two behaviors are designed using fuzzy logic tools inspired from human capacity reasoning with perception-based information. Type-1 fuzzy sets are precise but unable to handle data uncertainties. To incorporate these uncertainties, we need to make an extension. The idea is to consider the membership grade of an element to be a fuzzy set of type-1 rather than a number within the interval [0 1]. In this work, we propose a new scheme for obstacle avoidance based on wall following and to our knowledge, the manner we used in getting around the obstacle by partitioning the obstacle-near-region into three regions has not been used before. The mobile robot is equipped with US sensors that provide range reading, determining the distances between the robot and the obstacle. The inputs characterizing the information on the environment are fuzzified and presented to the inference engine for generating control outputs. The interesting part in this work is the way the obstacle avoidance using wall following is treated. This methodology, up to the author's knowledge, offers a new idea of tackling this problem when the obstacle neighborhood is divided into three regions. For each region, a linguistic fuzzy rule based system is defined.

1. The proposed general navigation problem.

The complete navigation problem has to deal with two important different behaviors. The main behavior is to reach the goal starting from any initial robot configuration. The second behavior is the mobile robot ability to avoid obstacles. Given these two behaviors, we should find a way to combine them in order for the mobile robot to execute the given mission successfully. In this case, one can either build a mode that switches between the two behaviors or have a blended action that combines between the two. Switching between the two behaviors can guaranty the success of the mission. However, the rapidly transitions between these two behaviors might result in a non-practical or a non-feasible action. On the other hand, the blended behavior might result in a smooth action. However, it does not guaranty the success of the task. One solution consists of weighting these two actions in order to successfully achieving the mission. Given the fact that the real world is composed of convex and concave obstacles, walls, labyrinths, etc., we need to elaborate a control action that guides the mobile robot along the obstacle boundaries toward its final destination. To detect the obstacles, we make use of the ultrasonic or infrared sensors that are provided with the wheeled mobile robot Khepera developed at the LAMI laboratory, Lauzane Switzerland.

Khepera is an automated differential drive guided vehicle specially designed and equipped for autonomous navigation utilized in many research laboratories. The execution of the task is related to the two-evocated main behaviors:

- Reaching the goal
- Avoiding obstacles

In this case, one can picture the robot as if it is subjected to two forces: an attractive force $\boldsymbol{F_a}$ due to the goal action and a repulsive force $\boldsymbol{F_r}$ due to the obstacle action as it is shown in Figure 1. The resultant force that acts on the robot may be determined by properly weighting the distance and direction. However, this would necessitate more tuning due to imperfect sensors and practical considerations. Therefore, to deal with obstacles, we choose the wall following approach while switching between the two cited behaviors with some added precautions to avoid the swinging effect. This method, which finds its application mainly to execute the corridor and wall following activity, is used in this work to avoid unforeseen obstacles. It could also help the robot reaching its goal in a blind fashion when the mobile robot makes the wall its main guidance. Tracking the wall would be an interesting application when dead-reckoning methods suffer from cumulative errors [15]. Wall following approach is as well very helpful in situations like being trapped in a cul-de-sac (Figure. 2-a) or when the mobile robot finds itself in a labyrinth-like situation (Figure. 2-b).

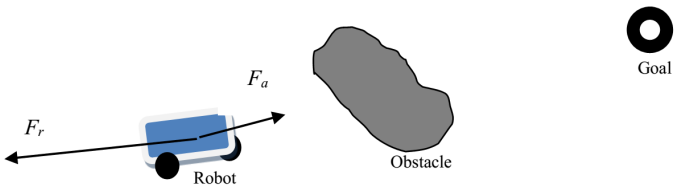


Figure 1: Attractive and repulsive Forces Acting on the mobile robot

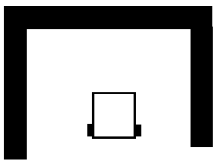


Figure 2a: cul-de-sac

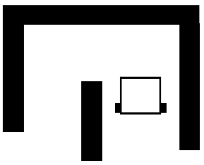


Figure 2b: Labyrinthe

3. Obstacle avoidance using wall following approach

3.1 Problem formulation

Many papers have dealt with wall following based on sensorial information such as IR sensors [16], US sensors [17], or visual navigation [18]. Ultrasonic and Infrared sensors are used for their relatively low cost, reliability and widespread availability. If the mobile robot evolves along a corridor, it will face two choices. Depending on its current situation with respect to the walls, the autonomous mobile robot should make a decision. Obviously, the robot follows the wall on its left side if the wall following task is chosen given only left sensor inputs, and likewise for the wall following on the right side. This decision may, for instance, be taken depending on the side that is closer to the wall. Hence, in case of the left wall following behavior, we use the two sensors on the left side of the mobile robot that give us the distances d_1 and d_2 as it is shown in Figure 3. But, if the mobile robot chooses to follow the wall being in its right side, the right wall following behavior is activated instead.

These two behaviors are mutually exclusive, which means that the activation of the first prevents the second to happen. This activation occurs when there are not obstacles on the front of the robot or there is only a detection of the wall that is to be followed. The implementation of the fuzzy logic controller is based on the sensorial information obtained from the two lateral US sensors, which give us the distance to the wall. In fact, this information is essential in determining the error angle between the actual current heading and the wall orientation. Our aim is to make this error equal to zero, such that the robot will always be in parallel to the wall.

3.2 The adopted strategy

Our idea lies on the following strategy. Consider the configuration of the mobile robot with respect to the wall and consider the distances d_1 and d_2 measured by the two lateral US sensors as it is shown in Figure 3. It is clear that the error angle α between the actual current robot heading and the wall orientation is given by equation (1)

$$\alpha = \alpha_l = atan\left(\frac{d_1 - d_2}{L}\right) \quad (1)$$

Where L is the length of the mobile robot. In this scheme, the desired position of the mobile robot is defined relative to the wall emplacement such that the robot maintains a desired distance. The distance that should be maintained between the mobile robot and the wall is chosen to satisfy an arbitrary security distance range $D + 2\varepsilon$, such that ε is a chosen tolerable offset error to avoid swinging or chattering behavior. In this case, the robot may be found in three possible regions, as it is clear from Figure 4.

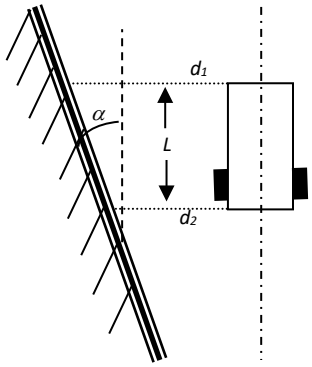


Figure 3: Robot configuration with respect to the wall

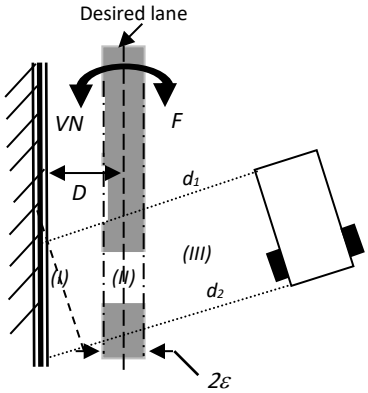


Figure 4: Desired lane following with respect to the wall

From Figure 4, one can see the inner region (I), the desired region (II) and the outer region (III). If the robot lies in the inner or outer regions, an action should be provided to drive it toward the desired region II. Considering the above stated requirement, our goal is to propose a robust fuzzy controller that maintains the mobile robot within region II.

4. Mobile robot Type-2 fuzzy motion control

The motion of the mobile robot may be controlled and adequately adjusted by many types of controllers. In this work, we choose the fuzzy logic control because we believe in its robustness for controlling nonlinear systems and its immunity in face to parameter variations and uncertainties. Introduced in 1965 by L. A. Zadeh from the University of California at Berkley [19], Fuzzy logic proved to be a strong tool for controlling complex systems. Using knowledge and experience, one can easily derive the rule base on which the fuzzy controller strongly depends. However, the presence of uncertainties in any system control requires thinking on extending type-1 fuzzy sets to type-2. In fact, the basic blocks used for designing the type-2 fuzzy controller are the same as those used with type-1; these include fuzzification, inference engine and defuzzification. The only difference lies on output processing block, where the defuzzifier is not considered as the solely block but contains a type reducer and defuzzifier blocks. The type

reducer is added because of its association with the nature of the membership grades of the elements that are no more a number within the interval [0 1], but a fuzzy set of type-1.

4.1 Interval Type-2 fuzzy logic system (IT2FLS)

The interval type-2 fuzzy set is completely described by the footprint of uncertainty (FOU). The FOU is a bounded region that uses an upper and lower type-1 membership function. By definition [14], an interval type-2 fuzzy set denoted by $\tilde{A}$ is expressed as

$$\tilde{A} = \{(x,y), \mu_{\tilde{A}}(x,y) | \forall x \in X, \forall u \in J_x \subseteq [0\ 1] \quad (2)$$

The upper bound and the lower bound of the FOU($\tilde{A}$) are expressed respectively as $\overline{\mu}_{\tilde{A}}(x)$ and $\underline{\mu}_{\tilde{A}}(x)$. Hence, J_X is just the interval $[\underline{\mu}_{\tilde{A}}(x), \overline{\mu}_{\tilde{A}}(x)]$. A type-2 Fuzzy Logic System FLS (T2FLS) is characterized by IF-THEN rules, where the premises and the consequences sets of the rules are of type-2.

More details regarding the fundamentals of fuzzy control and design can be found in many textbooks such as Mendel [20], Wang [21], Passino and Yurkovich [22].

4.2 Design of a type-2 fuzzy obstacle avoidance behavior

In this subsection, we give the appropriate design of a Type-2 fuzzy controller based obstacle avoidance behavior. This fuzzy controller is composed of primary and secondary Interval Type-2 fuzzy controllers.

4.2.1 Primary Interval Type-2 fuzzy controller

Type-2 fuzzy logic controller is similar to Type-1. The structure of the rules is the same as for that of type-1, but uncertainties are added in both the antecedent and consequent parts of each rule to account for uncertainties that come especially from sensors and/or when the circumstances are so fuzzy.

a) Fuzzifier: This input block maps a real-valued variable x to an interval Type-2 fuzzy set $\tilde{A}_x$. All the values of all variables representing the input and the output of the system are fuzzified. We consider three fuzzy membership functions for the distance d with labels $\widetilde{VN}$ standing for Very Near, $\tilde{N}$ for Near and $\tilde{F}$ for Far, as it could be seen from Figure 5. In the same manner, we consider four membership functions for the mobile robot orientation with labels $\tilde{N}$ for Negative, $\widetilde{NZ}$ for Negative Zero, $\widetilde{PZ}$ for Positive Zero and $\tilde{P}$ for Positive as it is depicted in Figure 6. The geometries of the membership functions are chosen arbitrary. The outputs of the fuzzy controller are the left and right velocities of the driving wheels, implemented with three linguistic variables defined as: Velocity: v_l, v_r= {$\tilde{S}$- Slow, $\tilde{M}$- Medium, $\tilde{H}$- High} on a normalized universe of discourse (Figure 7). Note that the symbol tilda above each label indicates that the fuzzy membership functions are of type-2. The two lateral US-sensors provide the sampled distance data d_1 and d_2, which are used to determine the two crisp input variables: d, such that $d=min(d_1,d_2)$, and the error angle α given by equation 1. These crisp inputs are converted into a fuzzy singleton and are mapped to the fuzzy sets $\tilde{A}_i^j$ with an interval degree $[\underline{\mu}_{\tilde{A}_i^j}(x), \overline{\mu}_{\tilde{A}_i^j}(x)]$.

b) Fuzzy inference engine: Once the input and output variables are defined, fuzzy inference engine is used to design the rule-base composed of IF-THEN rules to convert the inputs into output membership functions. It accomplishes the intersection and union of type-2 sets and performs compositions of type-2 relations.

The set of linguistic rules describe the desired behavior. Since the robot has the possibility to follow the wall on its right or the wall in its left side, we carefully have to set the rules adequately. The knowledge bases related to the wall on the left and the right sides of the robot are reported in Tables 1 and 2 respectively. For instance, we can have rules of the form:

$$If\ (d\ is\ \widetilde{VN} \wedge \alpha\ is\ \widetilde{PZ})\ Then\ (v_r\ is\ \tilde{S} \wedge v_l\ is\ \tilde{M})$$

$$If\ (d\ is\ \tilde{F} \wedge \alpha\ is\ \widetilde{NZ})\ Then\ (v_r\ is\ \tilde{H} \wedge v_l\ is\ \tilde{S})$$

For an Interval Type-2 fuzzy logic with p inputs and one output, the *i-th* rule might have the following form

$$Rule\ R^i: x_1\ is\ \tilde{A}_1^i\ AND\ \dots\ x_n\ is\ \tilde{A}_n^i\ THEN\ y\ is\ \tilde{B}^i, i = 1, \dots, M.$$

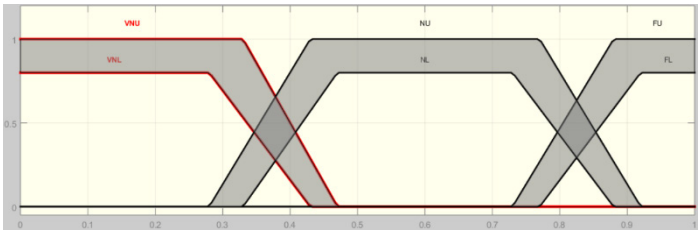


Figure 5: Membership functions of the distance d

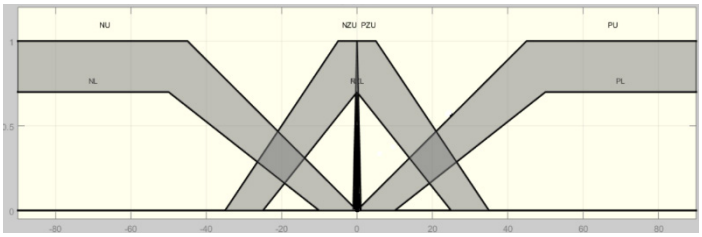


Figure 6: Membership functions of the error angle α

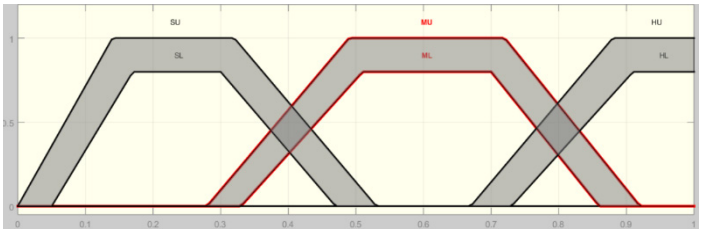


Figure 7: Membership functions of the left and right velocities

Under this bloc, the fuzzy *min t-norm* operation is implemented. For a crisp input vector, i.e., $x=x'$, the rule firing strength $F^i(x')$ is given by type-1 fuzzy set [23].

$$F^i(x') = [\underline{f}^i(x'), \overline{f}^i(x')] \quad (3)$$

Where

$$\underline{f}^i(x') = min\left(\underline{\mu}_{\tilde{A}_i^1}(x_1'),\ \underline{\mu}_{\tilde{A}_i^2}(x_2'), \dots, \underline{\mu}_{\tilde{A}_i^n}(x_n')\right) \text{ and}$$

$$\overline{f}^i(x') = min\left(\overline{\mu}_{\tilde{A}_i^1}(x_1'),\ \overline{\mu}_{\tilde{A}_i^2}(x_2'), \dots, \overline{\mu}_{\tilde{A}_i^n}(x_n')\right) \quad (4)$$

However, one can notice that no decision is taken when the robot enters the desired region II. We have marked this situation by a cross in the second row of Tables 1 and 2 to show that no decision is taken for these cases. In fact, whenever this situation is met, a secondary Interval Type-2 fuzzy controller is called to maintain the mobile robot in the desired region while keeping its orientation parallel to the wall. The details of this controller will be discussed latter.

c) Type reducer: This block found in the output processing is present in order to map the Interval T2FLS into a T1FLS. May be the most widely used type reducer among the many existing ones is the center of sets type reducer proposed by Karnik and Mendel [24]. It is characterized by the interval set determined by its left and right end points y_l and y_r and described by the following expression

$$Y_{cos} = [y_L, y_R] = \int_{y^1} \dots \int_{y^M} \int_{f^1} \dots \int_{f^M} 1 / \frac{\sum_{i=1}^{M} f^i y^i}{\sum_{i=1}^{M} f^i}, \; i=1, \dots, M \quad (5)$$

Where M is the number of rules and $f^i \in [\underline{f}^i, \overline{f}^i]$.

$y^i \in [y_L^i, y_R^i]$, is the centroid of the type-2 interval consequent set. For any value $y \in Y_{cos}$, y can be expressed as [25]

$$y = \frac{\sum_{i=1}^{M} f^i y^i}{\sum_{i=1}^{M} f^i} \quad (6)$$

To determine the left and right limits, y_l and y_r [25], the Karnik-Mendel Algorithm is used. The left and right limits are determined by the following expressions

$$y_l = \frac{\sum_{i=1}^{L^*} \overline{f}^i y_l^i + \sum_{i=L^*+1}^{M} \underline{f}^i y_l^i}{\sum_{i=1}^{L^*} \overline{f}^i + \sum_{i=L^*+1}^{M} \underline{f}^i} \quad (7)$$

$$y_r = \frac{\sum_{i=1}^{R^*} \underline{f}^i y_r^i + \sum_{i=R^*+1}^{M} \overline{f}^i y_r^i}{\sum_{i=1}^{R^*} \underline{f}^i + \sum_{i=R^*+1}^{M} \overline{f}^i} \quad (8)$$

Where $L^* = arg \min_{0 \le L \le M} \{y_l^L\}$ and $R^* = arg \max_{0 \le R \le M} \{y_r^R\}$,

$\underline{f}^i = min\left(\underline{\mu}_{\tilde{A}_i^1}(d), \underline{\mu}_{\tilde{A}_i^2}(\alpha)\right)$ and $\overline{f}^i = min\left(\overline{\mu}_{\tilde{A}_i^1}(d), \overline{\mu}_{\tilde{A}_i^2}(\alpha)\right)$

d) Defuzzifier: This second block constituting the output processing converts the T1FLS into a crisp number; i.e., the fuzzy variables from the type reducer are converted into a real valued variable. Hence the centroid of the type-1 fuzzy set contain two fuzzy singletons, and the final output is their average written as

$$y = \frac{y_l + y_r}{2} \quad (9)$$

Table. 1. Rule base for left wall following fuzzy controller

α / d	$\widetilde{N}$	$\widetilde{NZ}$	$\widetilde{PZ}$	$\widetilde{P}$
$\widetilde{VN}$	v_r=$\widetilde{S}$	v_r=$\widetilde{M}$	v_r=$\widetilde{S}$	v_r=$\widetilde{S}$
	v_l=$\widetilde{H}$	v_l=$\widetilde{H}$	v_l=$\widetilde{M}$	v_l=$\widetilde{M}$
$\widetilde{N}$	X	X	X	X
$\widetilde{F}$	v_r=$\widetilde{H}$	v_r=$\widetilde{H}$	v_r=$\widetilde{H}$	v_r=$\widetilde{H}$
	v_l=$\widetilde{M}$	v_l=$\widetilde{S}$	v_l=$\widetilde{S}$	v_l=$\widetilde{S}$

Table. 2. Rule base for right wall following fuzzy controller

α / d	$\widetilde{N}$	$\widetilde{NZ}$	$\widetilde{PZ}$	$\widetilde{P}$
$\widetilde{VN}$	v_r=$\widetilde{H}$	v_r=$\widetilde{H}$	v_r=$\widetilde{M}$	v_r=$\widetilde{M}$
	v_l=$\widetilde{S}$	v_l=$\widetilde{M}$	v_l=$\widetilde{S}$	v_l=$\widetilde{S}$
$\widetilde{N}$	X	X	X	X
$\widetilde{F}$	v_r=$\widetilde{M}$	v_r=$\widetilde{S}$	v_r=$\widetilde{S}$	v_r=$\widetilde{S}$
	v_l=$\widetilde{H}$	v_l=$\widetilde{H}$	v_l=$\widetilde{H}$	v_l=$\widetilde{H}$

In our case the *i-th* rule is written as

Rule R_i: d *is* $\tilde{A}_1^i$ AND α *is* $\tilde{A}_2^i$ *THEN* v_r *is* $\tilde{B}_1^i$ *AND* v_l*is* $\tilde{B}_2^i$,

i=1, …, 12

The speed of the right and left wheels are computed using the centroids for each aggregated Lower Membership Function (LMF) and Upper Membership Function (UMF). The speed for the left wheel v^L is

$$v^L = \frac{1}{2}(v_l^L + v_r^L) \quad (10)$$

where

$$v_l^L = \frac{\sum_{i=1}^{L^*} \overline{f}^i v_l^i + \sum_{i=L^*+1}^{M} \underline{f}^i v_l^i}{\sum_{i=1}^{L^*} \overline{f}^i + \sum_{i=L^*+1}^{M} \underline{f}^i} \quad (11)$$

$$v_r^L = \frac{\sum_{i=1}^{R^*} \underline{f}^i v_r^i + \sum_{i=R^*+1}^{M} \overline{f}^i v_r^i}{\sum_{i=1}^{R^*} \underline{f}^i + \sum_{i=R^*+1}^{M} \overline{f}^i} \quad (12)$$

The speed for the right wheel, v^R, is determined analogously.

4.2.2 Secondary Interval Type-2 fuzzy controller

The goal of the secondary Interval Type-2 fuzzy controller is to maintain the mobile robot within region II. For better compromise between good performances and reduced rules, the knowledge base is constructed using 25 rules as they are formulated in Table 3, which represent instructions to the robot to maintain its error angle to zero value. Input variables for this fuzzy logic controller

are the error angle α and its time derivative $d\alpha$. The outputs are again the right and left velocities of the driving wheels. For the design of the fuzzy logic controller, we assume that the cardinality of the input and output sets are the same and equal to 5. We propose to use the following labels for the three sets: $\{\widetilde{NL}$- Negative Large, $\widetilde{N}$- Negative, $\widetilde{Z}$- Zero, $\widetilde{P}$- Positive, $\widetilde{PL}$- Positive Large$\}$. The membership functions corresponding to the inputs α and $d\alpha$, are shown in Figure 8 and those corresponding to the outputs are shown in Figure 9. The symbol tilda above each label indicates again that the fuzzy membership functions are of type-2. For the right wall following, we either construct a new rule base or keep the previous one by making a small change in equation (1) while permuting the distances d_1 and d_2 as in equation (13).

$$\alpha = \alpha_r = atan\left(\frac{d_2 - d_1}{L}\right) \tag{13}$$

The design of the secondary Interval Type-2 fuzzy controller follows exactly the same steps as those presented in sub-section 4.2.1, for the two fuzzy inputs: the angle error α and its variation $d\alpha$.

The velocities responsible for maintaining the mobile robot within region II are equally given by equations (8) and (9), such that

$$\underline{f}^i = min\left(\underline{\mu}_{\tilde{A}_i^1}(\alpha), \underline{\mu}_{\tilde{A}_i^2}(d\alpha)\right) \text{ and}$$

$$\overline{f}^i = min\left(\overline{\mu}_{\tilde{A}_i^1}(\alpha), \overline{\mu}_{\tilde{A}_i^2}(d\alpha)\right) \tag{14}$$

On the other hand, the mobile robot should know the orientation it should take in case it is trapped in a local minimum as it is shown in Figure 2.

The obstacle avoidance algorithm must determine the sign of the orientation angle given by equation (15).

$$\theta = atan2\left(\frac{y_t - y_0}{x_t - x_0}\right) \tag{15}$$

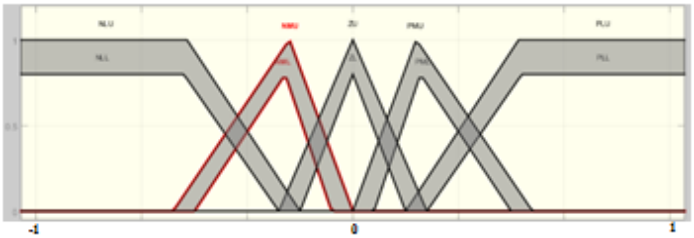


Figure 8: Membership functions of the angle error and the angle error variation

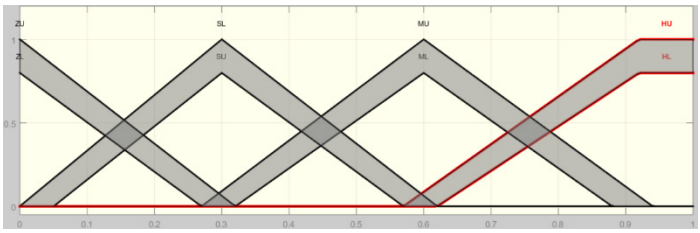


Figure 9: Membership functions of the left and right velocities

Table. 3. Rule base for the lane following fuzzy controller

α / $d\alpha$	$\widetilde{NL}$	$\widetilde{NM}$	$\widetilde{Z}$	$\widetilde{PM}$	$\widetilde{PL}$
$\widetilde{NL}$	$v_r=\tilde{S}$	$v_r=\tilde{S}$	$v_r=\tilde{S}$	$v_r=\widetilde{M}$	$v_r=\widetilde{H}$
	$v_l=\widetilde{H}$	$v_l=\widetilde{M}$	$v_l=\tilde{S}$	$v_l=\tilde{S}$	$v_l=\tilde{S}$
$\widetilde{NM}$	$v_r=\widetilde{M}$	$v_r=\widetilde{M}$	$v_r=\widetilde{M}$	$v_r=\widetilde{H}$	$v_r=\widetilde{H}$
	$v_l=\widetilde{H}$	$v_l=\widetilde{H}$	$v_l=\widetilde{M}$	$v_l=\widetilde{M}$	$v_l=\widetilde{M}$
$\widetilde{Z}$	$v_r=\widetilde{M}$	$v_r=\widetilde{M}$	$v_r=\widetilde{H}$	$v_r=\widetilde{H}$	$v_r=\widetilde{H}$
	$v_l=\widetilde{H}$	$v_l=\widetilde{H}$	$v_l=H$	$v_l=\widetilde{M}$	$v_l=\widetilde{M}$
$\widetilde{PM}$	$v_r=\widetilde{M}$	$v_r=\widetilde{M}$	$v_r=\widetilde{M}$	$v_r=\widetilde{H}$	$v_r=\widetilde{H}$
	$v_l=\widetilde{H}$	$v_l=\widetilde{H}$	$v_l=\widetilde{M}$	$v_l=\widetilde{M}$	$v_l=\widetilde{M}$
$\widetilde{PL}$	$v_r=\tilde{S}$	$v_r=\tilde{S}$	$v_r=\tilde{S}$	$v_r=\widetilde{M}$	$v_r=\widetilde{H}$
	$v_l=\widetilde{H}$	$v_l=M$	$v_l=\tilde{S}$	$v_l=\tilde{S}$	$v_l=\tilde{S}$

4.3. Reaching the Goal

Reaching the goal behavior, allows the mobile robot to navigate toward the target. Meanwhile, if the robot encounters an obstacle, the control switches immediately to wall-following behavior. "Reaching the Goal" and "Wall-Following" are two mutually exclusive behaviors. The activation of the first behavior prevents the second to happen. Figure 10, illustrates this operation using a flow chart. In contrast to the obstacle avoidance type-2 fuzzy controller, reaching the goal behavior is designed using a Type-1 fuzzy controller. The input variables are the distance, from the current mobile robot location to the target, and orientation angle, which measures the heading of the mobile robot with respect to the target point. For our convenience, we define seven fuzzy membership functions for the input distance with labels *Z* for Zero, *NS* for Near Small, *S* for Small, *M* for Medium, *F* for Far, *VF* for Very Far, and *VVF* for Very Very Far. Figure 11 illustrates the corresponding membership functions. In the same manner, we define seven membership functions for the orientation angle error with labels *LN* for Large Negative, *SN* for Small Negative, *N* for Negative, *Z* for Zero, *P* for Positive, *SP* for Small Positive and *LP* for Large Positive as it is depicted in Figure 12. The shape of the membership functions are chosen arbitrary and their emplacement along the universe of discourse has been carefully and adequately fixed after some experimental trials leading to a smooth navigation. The fuzzy outputs are again the right and left velocities of the driving wheels, whose shapes are trapezoidal on a normalized universe of discourse, as shown in Figure 13. We define for this application the following labels: *Z* for Zero, *NZ* for Near-Zero, *S* for Small, *M* for Medium, *H* for High, *VH* for Very-High and *VVH* for Very-Very-High.

Table 4 illustrates the fuzzy rules responsible of guiding the mobile robot to its final destination. The implementation of the fuzzy controller follows the steps of a conventional fuzzy controller, where one can find the main three blocks: the fuzzifier, the inference engine and the defuzzifier.

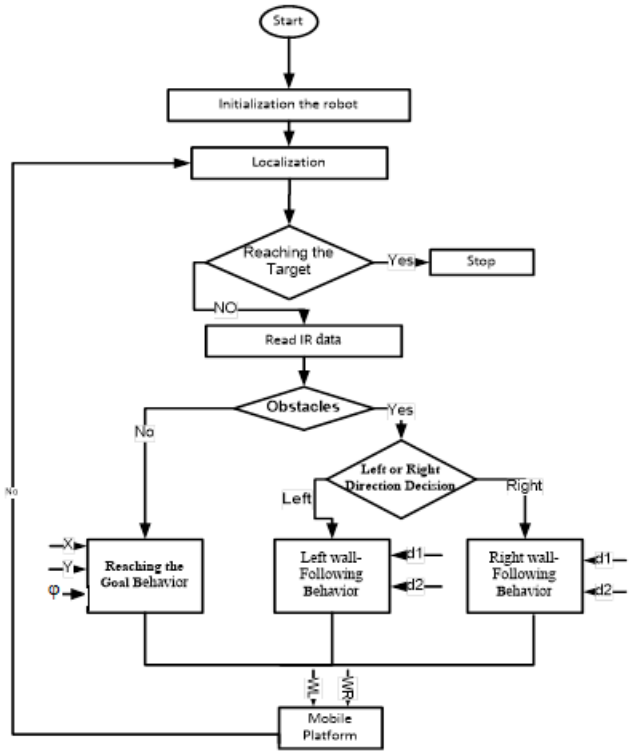


Figure 10. Mobile robot navigation control algorithm

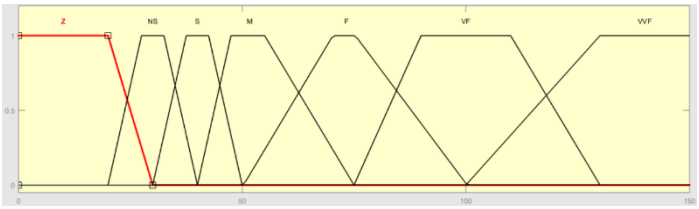


Figure: 11 Membership functions of the distance "d" in cm

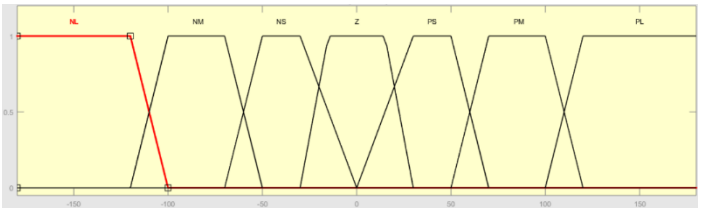


Figure 12: Membership functions of the orientation angle "φ".

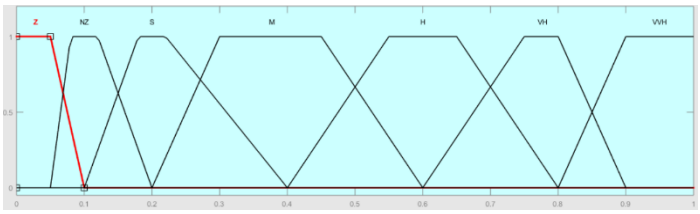


Figure 13: functions of the left and right velocities

Table. 4. Rule base for the "*Reaching the Goal*" behavior

d \ φ	*NL*	*NM*	*NS*	*Z*	*PS*	*PM*	*PL*
Z	L^{Z}	L^{Z}	L^{Z}	L^{Z}	L^{S}	L^{S}	L^{M}
	R^{M}	R^{S}	R^{S}	R^{Z}	R^{Z}	R^{Z}	R^{Z}
NS	L^{NZ}	L^{NZ}	L^{Z}	L^{NZ}	L^{M}	L^{VH}	L^{VH}
	R^{VH}	R^{VH}	R^{M}	R^{NZ}	R^{Z}	R^{NZ}	R^{NZ}
S	L^{NZ}	L^{NZ}	L^{NZ}	L^{S}	L^{VH}	L^{VH}	L^{VVH}
	R^{VVH}	R^{VH}	R^{VH}	R^{S}	R^{NZ}	R^{NZ}	R^{NZ}
M	L^{NZ}	L^{NZ}	L^{NZ}	L^{M}	L^{VH}	L^{VH}	L^{VVH}
	R^{VVH}	R^{VH}	R^{VH}	R^{M}	R^{NZ}	R^{NZ}	R^{NZ}
F	L^{NZ}	L^{NZ}	L^{S}	L^{VH}	L^{VH}	L^{VH}	L^{VVH}
	R^{VVH}	R^{VH}	R^{VH}	R^{VH}	R^{S}	R^{NZ}	R^{NZ}
VF	L^{NZ}	L^{NZ}	L^{M}	L^{VH}	L^{VH}	L^{VH}	L^{VVH}
	R^{VVH}	R^{VH}	R^{VH}	R^{VH}	R^{M}	R^{NZ}	R^{NZ}
VVF	L^{NZ}	L^{NZ}	L^{S}	L^{VVH}	L^{VH}	L^{VH}	L^{VVH}
	R^{VVH}	R^{VH}	R^{VH}	R^{VVH}	R^{S}	R^{NZ}	R^{NZ}

5. Experimental results

5.1 The mobile robot description

The mobile robot used in this work is Khepera III, a wheeled mobile robot from K TEAM Corporation. It is a mini mobile robot with functionality similar to larger robots used in research and education. It is an automated differential drive guided vehicle designed and equipped for autonomous and intelligent tasks as shown in Figure 14. The platform operates as a client in a client-server environment. Khepera is provided with Infrared and US-range-finder. It comes standard with five US transmitter/receiver pairs covering the front and the sides of the robot. We depict in Figure 15 the Khepera US sensors. To calculate the pose information of the robot we used the dead-reckoning equations.

5.2 The real experimentation Scenarios and Results

Various scenarios were set-up to test our proposed method in a real environment with different scenarios. These experimental results will determine the robustness, accuracy, adaptability and efficiency of the proposed method. For each scenario, we use the Khepera mobile robot platform, and matlab as a programming platform language. We tested the proposed system by using tens of scenarios; in this section, we are going to illustrate six of them. We selected six scenarios with various configuration, different shapes, and unknown environment. For all scenarios, the Khepera moves from the initial ***O***(0mm, 0mm) point to the target point ***T***(0mm, 1000 mm). Note that, the *x-y* data measurements are stored during the movement of the robot and plot by Matlab to get the graphs. In the first scenario, the Khepera robot navigates in a simple environment with one obstacle, as illustrated in Figure 16. During

its motion to the target, the mobile robot encounters an obstacle. In this situation, the robot stops moving, determines the angle θ from equation 15 and turns 90° clockwise due to the negative sign of the angle. The robot activates the obstacle avoidance behavior using the wall following approach. When the path is clear, the obstacle avoidance is inhibited and the Go-to-target behavior is re-activated. Figure 17 illustrates the real experimentation of this scenario.

In a second analog scenario, we simply change the length of the obstacle by considering a taller one. We remark that the mobile robot exhibits the same behavior as it did previously. We show in Figure 18 the plot of the experimental results, while Figure 19 illustrates the real experimentation.

In the third scenario, the Khepera robot navigates in a somewhat difficult environment with two long obstacles, as illustrated in Figure 20. In this scenario, during the movement of the robot toward its target, it faces the first obstacle of size (300 mm x 20 mm) and a second one of size (700 mm x20 mm). The robot determines the angle θ and turns 90° clockwise. The robot makes a first call to the wall-following algorithm to avoid the first obstacle. After passing the obstacle, the Go-to-Target algorithm is called in order for the robot to navigate toward its final configuration. However, in its path to the target a second obstacle is encountered. Again, the go-to-target is inhibited and the wall following behavior is activated. Immediately after passing the obstacle, the mobile robot regains control on the go-to-target behavior. The plot of this scenario is shown in Figure 20 and Figure 21 illustrates the real experimentation.

In the fourth scenario, the mobile robot navigates until a cul-de-sac situation is met, as it is can be seen from Figure 22. In this scenario, the mobile robot succeeds in escaping from this situation to join its final configuration using the same procedures. Figure 23 illustrates the real experimentation of this scenario.

In the fifth scenario, the robot is used to navigate in an environment embedded with many obstacles. The purpose of this manipulation is to show the validity of the approach in case of difficult situations. As we can notice from Figure 24, the mobile robot navigates successfully around the obstacles until the target is reached. Figure 25 illustrates the real experimentation of this scenario.

Figure 26 depicts the sixth scenario, where we repeated the fifth scenario using different obstacles shapes. During its motion, the robot faces the first circle shaped obstacle. In this case, the avoid obstacle behavior is executed using the wall-following approach, immediately followed by the Go-to-Target when the path is clear. Figure 27 illustrates the real experimentation of this scenario.

Figure 14: Khepera side view

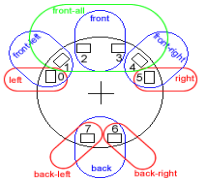


Figure 15: Khepera US sensors.

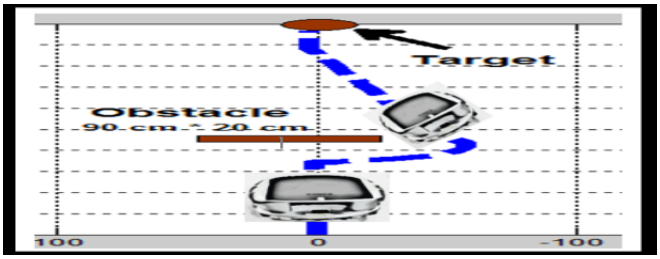


Figure 16: First scenario.

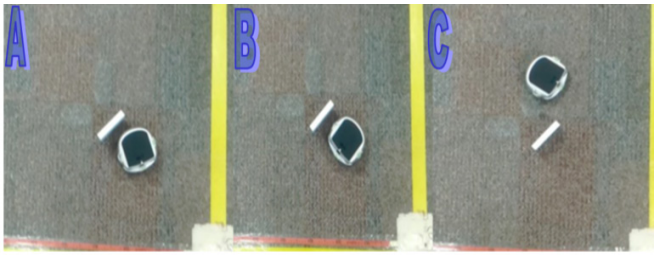


Figure 17: Real pictures of the first scenario.

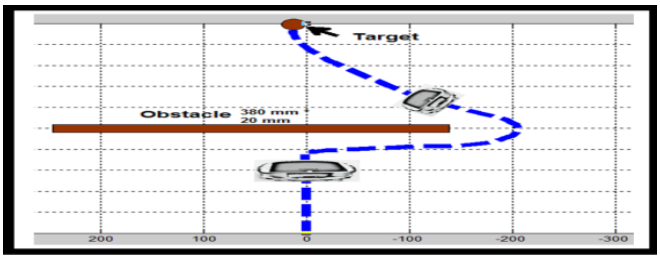


Figure 18: Second scenario

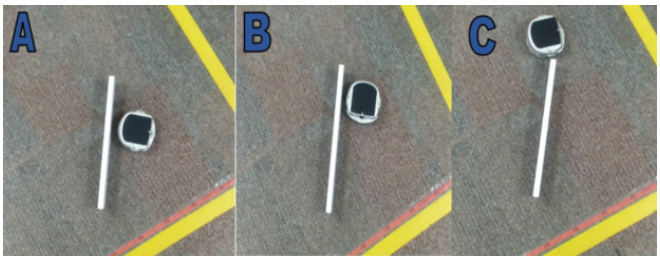


Figure 19: Real pictures of the second scenario

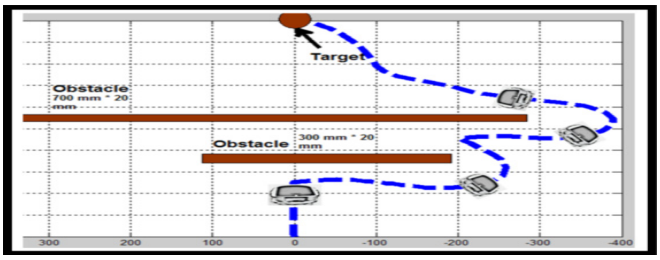


Figure: 20 Third scenario.

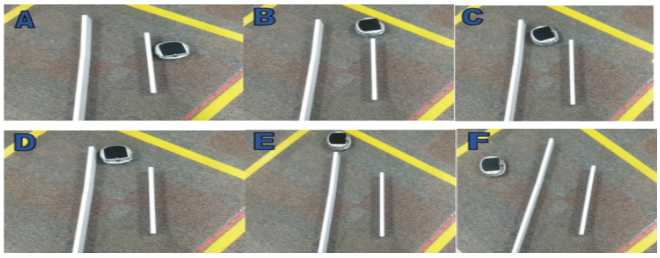


Figure: 21 Real pictures of the third scenario.

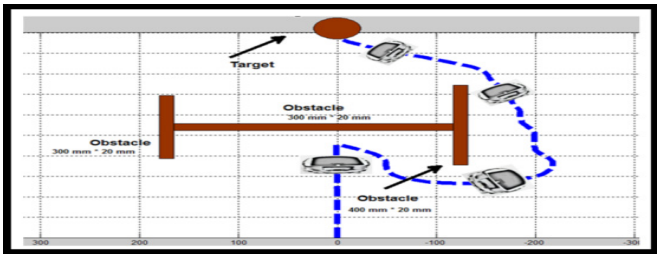


Figure 22: Fourth scenario

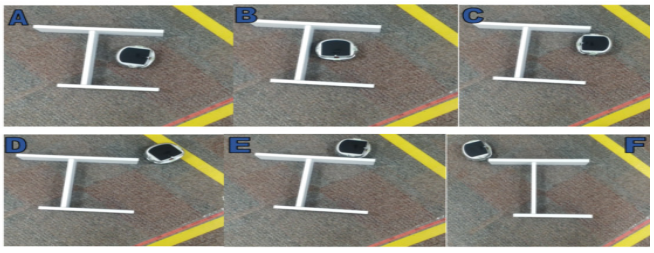


Figure 23: Real pictures of the fourth scenario

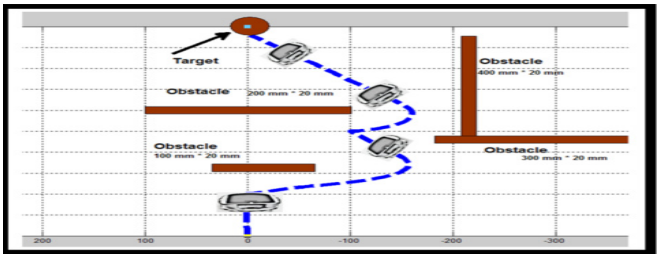


Figure: 24 Fifth scenario.

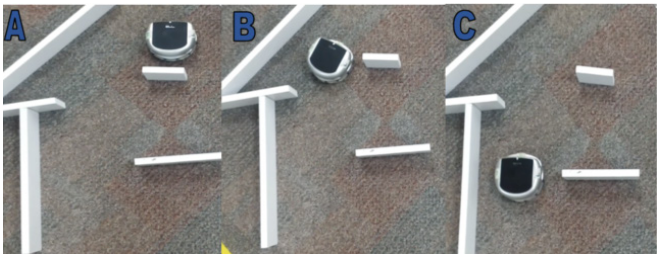


Figure: 25 Real pictures of the fifth scenario.

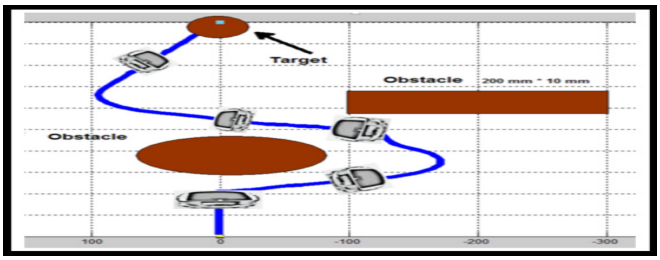


Figure 26: Sixth scenario.

Figure 27: Real pictures of the sixth scenario.

6. A Comparison study with PSO-Fuzzy methodology

In this section, we aim to compare the proposed method (Obstacle Avoidance Using Wall-Following Strategy) with another work, which uses a Partial Swarm Optimization-Fuzzy (PSO-Fuzzy) for navigation and obstacles avoiding [26].

From an experimental point of view, and to compare the performance of the proposed method with PSO-Fuzzy, each method was executed in two scenarios. The first scenario is simple and the second is somehow complicated. Each method was executed from the same startup point ***O***(0 cm, 0 cm) to the target point ***T***(0 cm, 100 cm). The comparison will include the traveled distance and the execution time. Plots of the results of the first and second scenarios of the experiments are shown in Figures 28 and 29 successively. The performance indexes are listed in Table 5. In fact, we have used the distance travelled by the mobile robot to go from the starting point to the destination point, whereas the second metric used is the time spent by the mobile robot to finish its travelling distance.

In these scenarios, we applied the two methods to moving the robot from the initial point to the target with the obstacle avoidance strategy. When comparing the two methods, it is obvious from the results reported in Table 5 that the proposed method performed better than the PSO-fuzzy one.

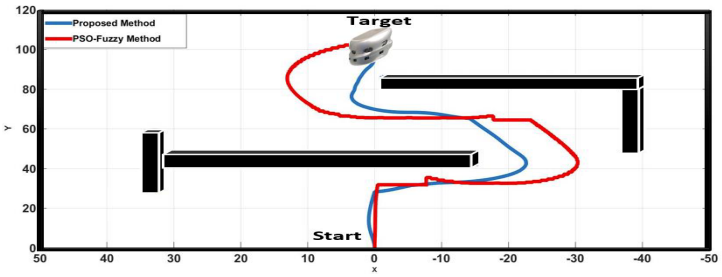


Figure 28: comparison between proposed method and PSO-Fuzzy

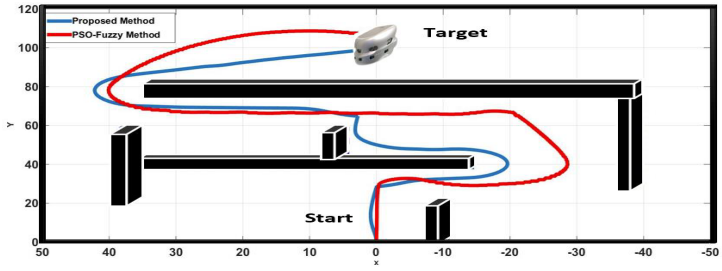


Figure 29: comparison between proposed method and PSO-Fuzzy

Table 5: Performance indexes

Parameters		Proposed Method	PSO-Fuzzy Method
First scenario	Execution time	33.83 sec	46.46 sec
	Traveled distance	124 cm	163 cm
Second scenario	Execution time	48.12 sec	69.35 sec
	Traveled distance	184 cm	219 cm

7. Conclusion

In this paper, a complete navigation mobile robot control is used to reach a given target while avoiding unforeseen obstacles. To satisfy these two requirements, a fuzzy logic design has been implemented for the two behaviors. The first behavior is achieved by designing a fuzzy controller that uses the distance and the orientation as inputs. The controller generates the required wheels velocities to drive the mobile robot to its destination along a smooth path. In case the mobile robot encounters an obstacle, it switches to the obstacle avoidance behavior designed using a wall following approach. A new fuzzy controller approach based on Type-2 fuzzy sets has been presented for obstacle avoidance using wall following methodology. To achieve this requirement, we divided the robot environment into three regions, two transient regions and one stable region. Whenever the mobile robot finds itself in the transient regions, the first fuzzy controller is activated to pull the mobile robot to the lane that is parallel to the wall. Once trapped in this region, the first fuzzy controller is inhibited and the second is activated such that the mobile robot remains within the limits of the lane. The mobile robot keeps moving within the stable region of displacement until a new situation appears, otherwise the go-to-target behavior is activated with Type-1 fuzzy controller and the mobile robot heads for the target point. Experimental works are carried out on the mobile robot Khepera III. The experimental results obtained are very satisfactory and prove the validity of the proposed approach, which is simple, and does not require complicated and burden computations compared to other methods. As an extension to this work, we plan to adapt this methodology to the mobile robot PowerBot from Adept Corporation.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This work was supported by the NPST programme of King Saud University, Project No: 08-ELE-300-02.

References

[1] Crawford, Ian A., and Katherine H. Joy. "Lunar exploration: opening a window into the history and evolution of the inner Solar System." Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences 372.2024: 20130315, 2015.

[2] G. Pessin, D. O. Sales, M. A. Dias, R. L. Klaser, D. F. Wolf, J. Ueyama, FS Osorio & Vargas, P. A., "Swarm intelligence and the quest to solve a garbage and recycling collection problem". Soft Computing, 17(12), 2311-2325, 2013.

[3] L. Danielli and G. MB de Oliveira. "A cellular automata ant memory model of foraging in a swarm of robots." Applied Mathematical Modelling, Vol. 47, pp.551-572, 2017.

[4] C. Rodrigo, M. Figueiredo, and E. A. Antonelo. "Evolutionary fuzzy system for architecture control in a constructive neural network." Computational Intelligence in Robotics and Automation, 2005. CIRA 2005. Proceedings 2005, IEEE International Symposium on. 2005.

[5] F. M. Marchese, "A directional diffusion algorithm on cellular automata for robot path-planning." Future Generation Computer Systems Vol. 18, issue7 pp. 983-994, 2002.

[6] Y. Liu, R. Fu, J. Wang, Y. Ou, X: Wu, A: Peng, "A Wall-Following Strategy for Mobile Robots Based on Self-Convergence"; Proceedings of the 2011 IEEE, International Conference on Robotics and Biomimetics, December 7-11, Phuket, Thailand, 2011.

[7] X. Wang, Z. G. Ho, M. Tan, Y. W., L. Hu, "The wall-following controller for the mobile robot using spiking neurons", Int. Conf. on Artificial Intelligence and Computational Intelligence, 2009.

[8] R. Braunstingl, P. S.anz – J. M. Ezkerra, "Fuzzy Logic Wall Following of a Mobile Robot Based on the Concept of General Perception", ICAR '95, 7th Int. Conf. on Advanced Robotics, Sant Feliu De Guixols, Spain, pp .367-376, Sept., 1995.

[9] I. Gavrilut, V. Tiponut, A. Gacsadi, L. Tepelea, "Wall-following Method for an Autonomous Mobile Robot using Two IR Sensors", 12th WSEAS International Conference on Systems, Heraklion, Greece, July 22-24, 2008.

[10] C. H. Hsu and C. F. Juang, Evolutionary Robot "Wall-Following Control Using Type-2 Fuzzy Controller with Species-DE Activated Continuous ACO", Fuzzy Systems, IEEE Transactions Vol. 21 Issue: 1, pp. 100-112, Feb. 2013.

[11] D. Lima, C. Tinoco, J. Viedman and G. Oliveira, "Coordination, Synchronization and Localization Investigations in a Parallel Intelligent Robot Cellular Automata Model that Performs Foraging Task" .In Proceedings of the 9th Int. Conf. on Agents and Artificial Intelligence ISBN 978-989-758-220-2, pp. 355-363, 2017.

[12] N. Baklouti1, R. John2, A. M. Alimi, Interval Type-2 Fuzzy Logic Control of Mobile Robots, Journal of Intelligent Learning Systems and Applications, vol.4, pp. 291-302, 2012.

[13] H. Hagras, "A Hierarchical Type-2 Fuzzy Logic Control Architecture for Autonomous Mobile Robots," IEEE Transactions on Fuzzy Systems, Vol. 12, No. 4, pp. 524-539, 2004.

[14] C-F Juang, and C-H Hsu, "Reinforcement Ant Optimized Fuzzy Controller for Mobile-Robot Wall-Following Control", IEEE Transaction on Industrial Electronics, Vol. 56, no. 10, October 2009.

[15] I. Dumitrache, M. Dragoicea, “Some Problems of advanced mobile robot control”, J. of Control. Engineering and Applied Informatics, Vol. 7, No. 4, pp. 11-30, 2006.

[16] H. Park, S. Baek and S. Lee, “IR Sensor Array for a Mobile Robot”, Proceedings of the 2005 IEEE/ASME Int. Conf. on Adv. Intell. Mecha., Monterey, California, USA, 24-28 July, 2005.

[17] D. P. Stormont, C. T. Abdallah, R. H. Byrne and G. L. Heileman, “ A Survey of Mobile Robot Navigation Methods Using Ultrasonic sensors”, Space and Robotics, ISBN (print): 978-0-7844-0337-2, 1998.

[18] F. Bonin-Font, A. Ortiz and G. Oliver, " Visual Navigation for Mobile Robots: a Survey, “Journal of Intelligent and Robotic System”, Vol. 53, issue 3, pp. 263-296, Nov. 2008.

[19] L.A. Zadeh,"Fuzzy Sets," Information Control, vol. 8, pp. 338-353, 1965.

[20] J. M. Mendel, Uncertain Rule-Bused Fuzzy Logic Systems: Introduction and new directions. Prentice Hall. NJ_ 2001.

[21] L. X. Wang, A course in Fuzzy systems and Control, Prentice-Hall, Englewood Cliffs, NJ, 1997.

[22] K. M., Passino and S. Yurkovich, Fuzzy Control, Addison Wesley Longman, Menlo Park, CA, 1998.

[23] Q. Liang and J. Mendel, “Interval type-2 fuzzy logic systems: Theory and design,” IEEE, Trans. Fuzzy syst., vol. 8, no. 5, pp. 535-550, Oct. 2000.

[24] N. N. Karnik and J. M. Mendel,, “Type-2 fuzzy logic systems: Type-reduction,” in IEEE Syst., Man, Cybern. Conf., San Diego, CA, Oct. 1998.

[25] Dongrui Wu, Jerry M. Mendel, “Enhanced Karnik-Mendel Algorithms,” IEEE Transactions on Fuzzy Systems, 17(4), 923-934, 2009.

[26] Algabri Mohammed, Hedjar Ramdane, Hassan Mathkour, Khalid Al-Mutib, and Mansour Alsulaiman. "Optimization of Fuzzy Logic Controller Using PSO for Mobile Robot Navigation in an Unknown Environment." Applied Mechanics and Materials 541: 1053-1061, 2014.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 338-344 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ
ISSN: 2415-6698

Validity of Mind Monitoring System as a Mental Health Indicator using Voice

Naoki Hagiwara*, Yasuhiro Omiya[1], Shuji Shinohara[2], Mitsuteru Nakamura[2], Masakazu Higuchi[2], Shunji Mitsuyoshi[2], Hideo Yasunaga[3], Shinichi Tokuno[2]

[1]*Research and Product Development, PST Inc., 231-0023, Japan*

[2]*Verbal Analysis of Pathophysiology Graduate School of Medicine, The University of Tokyo, 113-0033, Japan*

[3]*Department of Clinical Epidemiology & Health Economics School of Public Health, The University of Tokyo, 113-0033, Japan*

ARTICLEINFO

Article history:
Received: 06 April, 2017
Accepted: 28 April, 2017
Online: 17 May, 2017

Keywords :
Voice analysis
Mental health care
Mobile phone

ABSTRACT

We have been developing a method of evaluating the mental health condition of a person based on the sound of their voice. Currently, we have applied this technology to create a smartphone application that shows the vitality and the mental activity as mental health condition indices. Using voice to measure one's mental health condition is a non-invasive method. Moreover, this application can be used continually through a smartphone call. Unlike a periodic checkup every year, it could be used for monitoring on a daily basis. The purpose of this study is to compare the vitality index to the widely used Beck depression inventory (BDI) and to evaluate its validity. This experiment was conducted at the Center of Innovation Program of the University of Tokyo with 50 employees of one corporation as participants between early December 2015 and early February 2016. Each participant was given a smartphone with our application that recorded his/her voice automatically during calls. In addition, the participants had to read and record a fixed phrase daily. The BDI test was conducted at the beginning of the experimental period. The vitality index was calculated based on the voice data collected during the first two weeks of the experiment and was considered as the vitality index at the time when the BDI test was conducted. When the vitality and the mental activity indicators were compared to BDI score, we found that there was a negative correlation between the BDI score and these indices. Additionally, these indices were a useful method to discriminate a participant of high risk of disease with a high BDI score. And the mental activity index shows a higher performance than the vitality index.

1. Introduction

This paper is an extension of the work originally presented in IEEE BIBE 2016 (The 16th International Conference on Bioinformatics and Bioengineering) [1].

In recent years, mental health disorders have become an issue in many developed countries and the cost of dealing with such disorders is enormous [2, 3]. There is a need for a screening technology that allows for a quick and easy way to detect depression and stress early to tackle these problems.

*Naoki Hagiwara, PST Inc., 2 Yamashita-cho, Naka-ku, Yokohama, Kanagawa 231-0023 Japan, +81-45-263-9346 & hagiwara@medical-pst.com

Currently, self-report type psychometric tests such as general health questionnaires (GHQ) [4] and Beck depression inventory (BDI) [5, 6] are used as methods of screening patients with mental health disorders. Besides these tests, there are proposals to use biomarkers such as saliva [7] and blood [8] as a method of evaluating mental health conditions. However, although self-report psychometric tests are effective in determining those conditions at their early stages and in complementing diagnoses, there are issues of reporting biases. A reporting bias refers to the effect of an undervaluation or overvaluation because of the consciousness or unconsciousness of subjects. It has been reported

that the detection rate decreases in organizations that have established hierarchies such as in fire-fighting, police, or military [6-8]. Some causes of undervaluation in self-administered screenings are prejudice, discrimination, resistance to mental health disorder, or anxiety on career disadvantage in the job [6]. Concerning evaluations with biomarkers, they are a burden to those being tested as they are invasive, and the tests cost money. Thus, they cannot be considered as easy or simple solutions. In addition, conventional screening methods such as self-report psychometric tests and biomarkers are often conducted during a routine annual checkup, making it difficult to provide care during middle periods.

On the other hand, prior to this current research, the authors of this paper have been researching and working on a way to estimate the state of a person's mental health as in their state of depression or stress through voice data [14, 15]. Using voice for analysis has the advantage of being not only non-invasive, but also not requiring any dedicated and specialized equipment; it is simple, easy, and can be done remotely. Using this technology to assess the state of one's mental health condition through his or her voice can make it possible to achieve monitoring on a daily basis, which is difficult with conventional methods.

The authors have developed and provided MIMOSYS (mind monitoring system) SDK that estimates the health condition of speakers through their voice. MIMOSYS SDK has a function of recording and analyzing voice from a microphone as input and outputs the index of health condition as a result (see section 2.2). To allow daily monitoring for many people, we have developed a smartphone application using MIMOSYS SDK [16, 17]. The purpose of this study is to compare the vitality score, which is collected and analyzed through MIMOSYS to the BDI and evaluate its validity.

2. Materials and Methods

2.1. The Experiment

The experiment was conducted as part of the University of Tokyo, Center of Innovation Program. With the support of Kanagawa Prefectural government, we provided smartphones with preinstalled application to 50 employees of a company. This company is searching for health care measures and has recruited 50 in-house participants who are interested in the experiment. The 50 participants comprise 39 males and 11 females. The breakdown of age is shown in Figure 1. The model name and Android version of the smartphones used in the experiment are given in Table 1. We required the test participants to record their voice for two months through reading out fixed phrases and through talking using the smartphone provided.

We conducted the BDI test at the beginning of the experiment. The application detects when a call is made or received and automatically records the voice and analyzes it. The fixed phrases that the test participants read out daily are listed in Table 2. The application will erase the recorded voice file after the analysis, but the analysis results and user information are sent over a network to be collected and stored into a database. All the analysis results were obtained via this database.

Figure 1. Age histogram of test participants.

Table 1. Model name of smartphones used in the experiment.

Model	*Number*	*Android version*
SC-01F	31	4.4.2
SC-01F	3	5.0
SC-02E	5	4.3
SO-01E	3	4.1.2
SO-04F	3	4.4.2
SH-04F	3	4.4.2
F-05F	2	4.4.2

Table 2. Manuscript of phrases read by participants.

No.	*Phrase (in Japanese)*	*Purpose*
1	I-ro-ha-ni-ho-he-to.	non-emotional
2	Honjitsu ha seiten nari.	non-emotional
3	Tsurezurenaru mama ni.	non-emotional
4	Wagahai ha neko dearu.	non-emotional
5	Mukashi mukashi aru tokoro ni.	non-emotional
6	Omoeba tooku he kita monda.	non-emotional
7	a-i-u-e-o	check pronunciation
8	Ga-gi-gu-ge-go	check pronunciation
9	Ra-ri-ru-re-ro	check pronunciation
10	Pa-pi-pu-pe-po	check pronunciation
11	Garapagosu shotou	check pronunciation
12	Tsukarete guttari shiteimasu.	emotional
13	Totemo genki desu.	emotional
14	Kinou ha yoku nemuremashita.	emotional
15	Shokuyoku ga arimasu.	emotional
16	Okorippoi desu.	emotional
17	Kokoroga odayaka desu.	emotional

2.2. MIMOSYS

In mental health disorders, a decrease in emotional expression appears and the proportion of sorrow component may increase. Therefore, MIMOSYS estimates the health condition of a speaker focusing on the balance and variation of these emotional indices. MIMOSYS is based on an emotion index obtained by a sensibility technology (ST) [9-11, 18]. It is preferable that natural utterance is inputted, although it is barely affected with utterance by acting. Therefore, in the developed application, we use the voice of calls that can be uttered naturally without consciousness for the analysis. We thought that it is advantageous to continuously use the call voice as an analysis target and accumulate the result.

In emotional recognition using ST, multiple feature quantities are calculated from the input voice, and four emotional components – anger (Ang), sorrow (Srw), joy (Joy), and calmness (Clm) – in the voice are calculated from these features. The degree of excitement (Exc) is also calculated. Thus, five indicators are outputted. MIMOSYS calculates two intermediate feature quantities, namely vivacity and relaxation, from these emotional indexes in internal processing. It calculates vivacity from Joy and Srw, and relaxation from Clm and Ang respectively. From there, a short-term health index, called vitality index, is calculated. Then, a medium-term health index, called mental activity, is calculated from the vitality index accumulated for two weeks. The data calculation flow is shown in Figure 2.

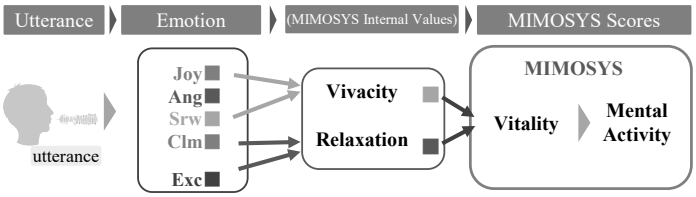


Figure 2. Data calculation flow of MIMOSYS.

Next, we will describe the implementation of the application. When analyzing using ST, it is necessary to input the voice by one utterance, and when analyzing a long voice, it is necessary to divide it into data of one utterance by detecting speech. Here, utterance is a voice uttered in a single breath. The smartphone application, MIMOSYS, implements speech detection to record and analyze a call voice. Specifically, the utterance part is detected by the volume level, and only the extracted utterance part is used for the analysis in ST. Then, an analysis of the health condition estimation is carried out with multiple utterances in one call as one set. The vitality index is outputted as a short-term health condition index from these sets of uttered voices. In addition, in order to suppress variations at individual recording timings and capture long-term trends, we outputted the medium-term health condition index, i.e., the mental activity index, calculated based on the vitality index accumulated for two weeks.

Both indicators for short term (vitality index) and medium term (mental activity) are each outputted as a real number value from 0.0 to 1.0. In addition, having both the recording and processing function on the smartphone eliminates the transfer cost of the recorded voice data that would otherwise be incurred had the recordings been performed on a separate recording device requiring a transfer to another computational device. Furthermore, considering the possibilities of information leaks from the voice recordings and to reduce the storage load, we have decided to delete the voice recordings after analysis. The results of the analysis were stored on the smartphone devices and also sent to a database server via the network for storage. As an index corresponding to the BDI test, the vitality index was calculated from the first recorded voice, and the mental activity was calculated from the voice recorded in the first two weeks.

2.3. Recording and Analyzing the Voice

There are two ways of obtaining the voice over the smartphone: the voice of the user reading some sentences (hereinafter referred to as fixed phrase recordings) and the voice while talking on the phone (hereinafter referred to as call recordings). To monitor by oneself on a daily basis, we speculate that call recordings, in which the caller is not conscious of the recording, would be more suitable. However, as there are differences in the frequency in which individuals use their phone for talking, we included in the analyzing application a function to record a fixed phrase reading on a regular basis for the experiment. In call recordings, the analysis was performed with utterances in one call; however, for fixed phrase recordings, the analysis was performed with utterances of reading the 17 phrases. Additionally, as there may be a difference in the nature of the voice between the fixed phrase recordings and call recordings that were obtained, to make it easier to differentiate them, information on which method was used to obtain the recordings was also included.

The voice to be analyzed was acquired by a microphone of the smartphone provided to the experiment participants (see Table 1). The recording format was a linear pulse-code modulation, the sampling frequency was 11,025 Hz, and the quantization bit rate was set at 16 bits. However, the recorded sound file was deleted immediately after the analysis was completed, and it was not taken out of the smartphone.

2.4. Beck's Depression Inventory

The BDI test is widely used for judgment of depression, and the scores for judgment are listed in Table 3. In this research, a score of 18 or less is considered as low risk of disease, and more than that value is treated as high risk of disease. We conducted a test to evaluate the performance of the application to determine if the voices of the test participants could be used to discriminate between high risk of disease and low risk of disease. The BDI test was incorporated into the application and was executed when the application was first launched at the start of the experimental period.

Table 3. Categories for BDI score.

BDI score	*Categories of depression*	*Judgment*
0–9	None to minimal depression	Low risk of disease (≤ 18)
10–18	Mild to moderate depression	
19–29	Moderate to severe depression	High risk of disease (> 18)
30–63	Severe depression	

2.5. Statistical Analysis of Data

The analysis result of the application is outputted to the csv (comma-separated values) file. We used Microsoft Excel functions and analysis tools to tally and analyze the data. Additionally, we analyzed the data and created diagrams using the R language.

3. Result

The analysis result was obtained from the 50 participants, and 48 valid results were obtained. In total, there were 890 call recordings and 4,357 fixed phrase recordings. The daily total recording count of the test participants in first two weeks is shown in Figure 3.

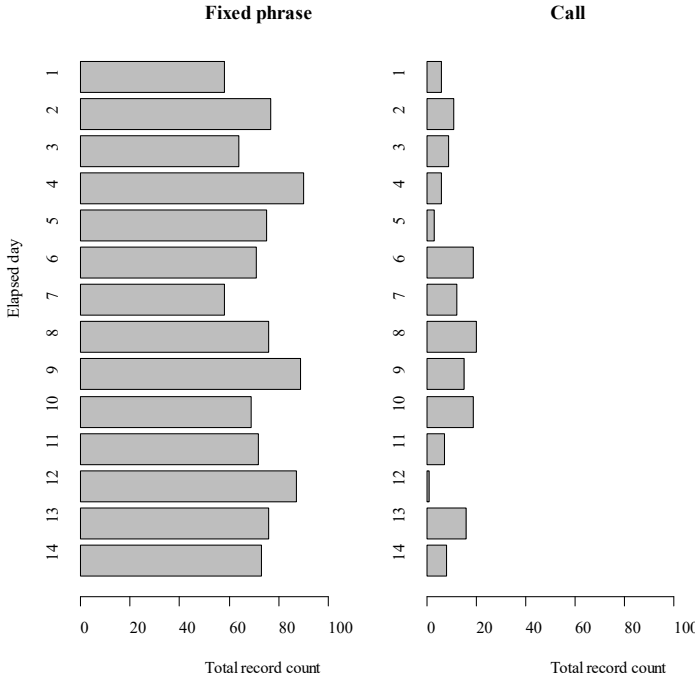


Figure 3. Total recording count of test participants with two methods.

3.1. Effect of the Recording Method

In the experiment, there were two methods in which the voice was recorded and obtained. For each method, we separately tallied the vitality index and the side of the vitality score that shows the short-term mental health condition. The average and standard deviation of the vitality index are given in Table 4. Compared to call recordings, fixed phrase recordings tended to show a lower value and are less consistent. In this research, we evaluated the fixed phrase recordings and phone conversations independently.

Table 4. Average and standard deviation of vitality index by two recording methods.

	Call	***Fixed phrase***
Average	0.554	0.460
Standard deviation	0.0759	0.0981

3.2. Correlation to BDI

The scatter plot of the vitality index by fixed phrase recordings and the BDI score is shown in Figure 4. The correlation coefficient between the vitality index and the BDI score showed a negative correlation value of 0.208. Furthermore, the correlation between the mental activity index by the fixed phrase recordings and the BDI score is shown in Figure 5. The correlation coefficient between the mental activity and the BDI score showed a negative correlation value of 0.285.

The mental activity index, which is the medium-term indicator of the mental health condition, was calculated based on the voice data collected within the first two weeks, which was considered as the mental activity index at the time of the BDI testing. We found that depending on the test participants, there were differences on how often the phone was used. Because some had too few recordings captured or none at all, there were cases when we were unable to calculate the mental activity index corresponding to the BDI. Therefore, we were able to acquire call recordings from only 17 participants, which were used to calculate their mental activity index.

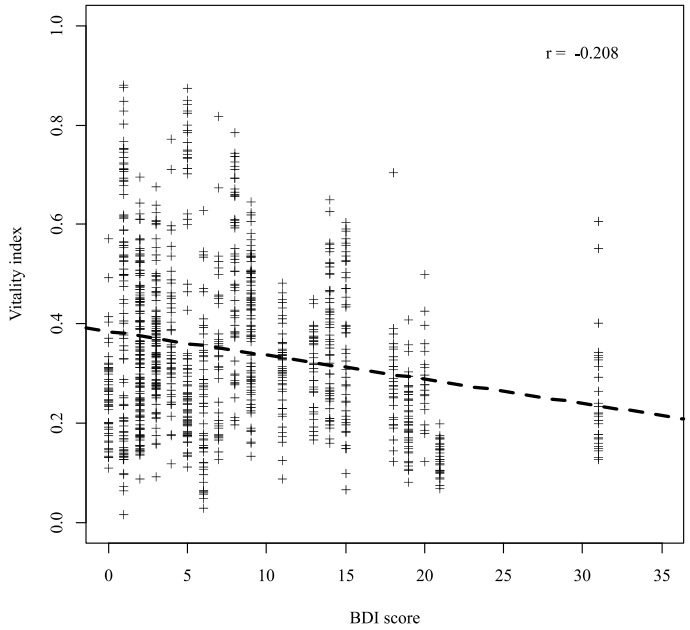


Figure 4. Scatter plot of BDI score and vitality by fixed phrase. The regression line is drawn in dashed line.

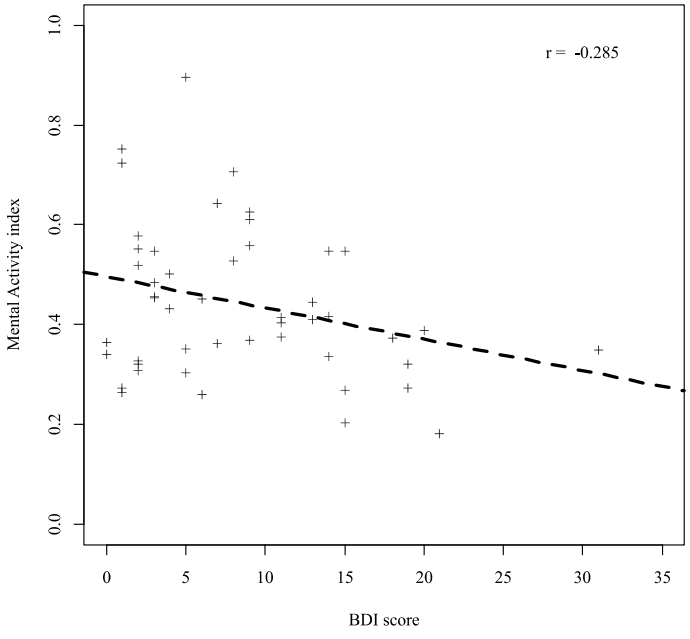
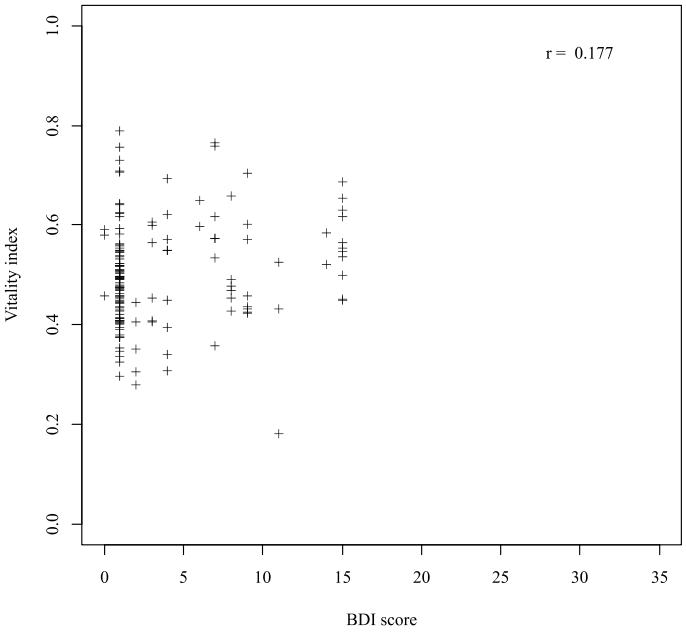


Figure 5. Scatter plot of BDI score and mental activity by fixed phrase. The regression line is drawn in dashed line.



The vitality index by call recording and its correlation to the BDI score is shown in Figure 6. The correlation coefficient between the vitality index and the BDI score was 0.177, which means that there is almost no correlation between the two. The mental activity index by call recordings and its correlation to the BDI score is shown in Figure 7.

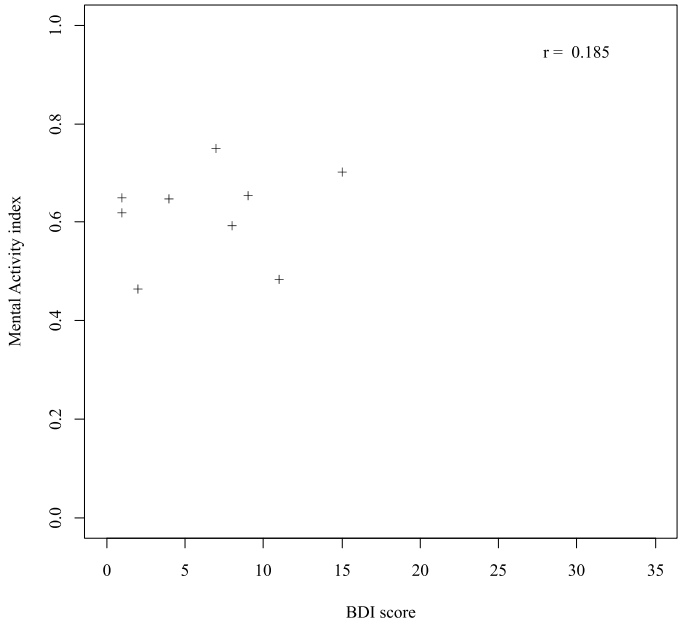


Figure 7. Scatter plot of BDI score and mental activity by call.

The correlation coefficient between the mental activity index by call recording and the BDI score was 0.185, which indicate that there is almost no correlation between the two.

The box and whisker plot is shown in Figure 8 and Figure 9. The results show that both scores showed a significant difference. However, the vitality index overlap of distribution is wider than that of the mental activity index.

3.3. Discrimination Evaluation of High Disease Risk using MIMOSYS

We then changed our perspective and evaluated the performance of discriminating high risk of disease and low risk of disease using the MIMOSYS index. However, the 17 participants with calculated mental activity from the call recordings all had a BDI score of less than 18; therefore, this evaluation was limited to using the fixed phrase recordings. With sensitivity as the vertical axis and 1-specificity as the horizontal axis, the receiver operating characteristic (ROC) curve plotted by changing the threshold is shown in Figure 10 and Figure 11. The ROC curve uses the area under the curve (AUC) as an indicator for performance evaluation. An AUC value in the range of 0.5–0.7 denotes low accuracy, 0.7–0.9 denotes moderate accuracy, and 0.9–1.0 denotes high accuracy.

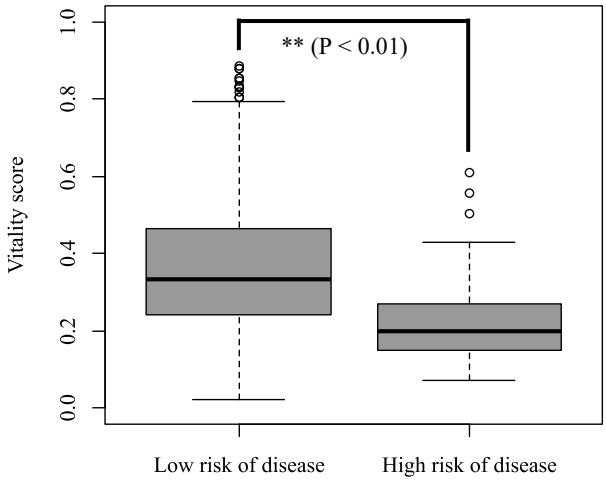


Figure 8. Box and whisker plot of vitality score by fixed phrase with low risk of disease and high risk of disease.

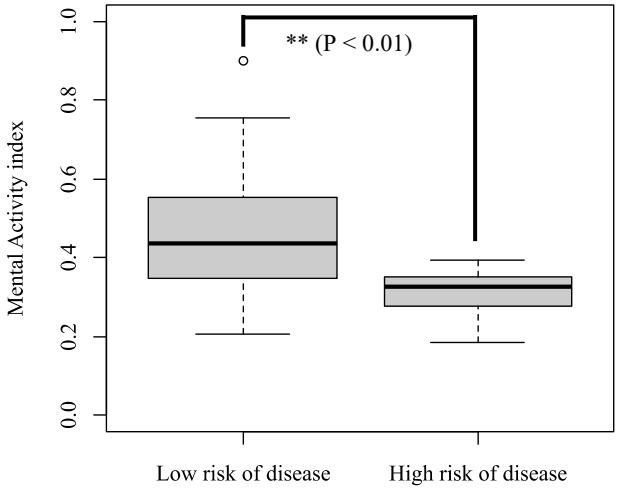


Figure 9. Box and whisker plot of mental activity score by fixed phrase with low risk of disease and high risk of disease.

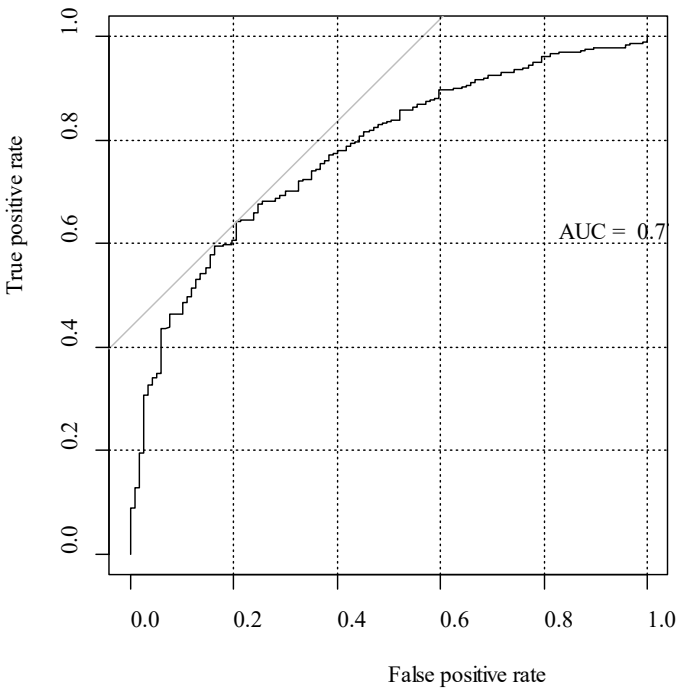


Figure 10. ROC curve used in discriminating between low and high risk of disease using vitality index.

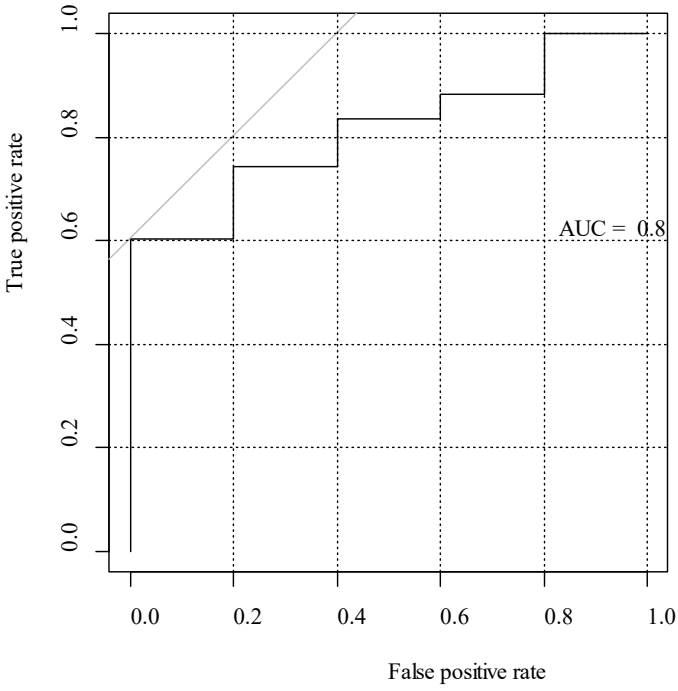


Figure 11. ROC curve used in discriminating between low and high risk of disease using mental activity index.

In the vitality index, the AUC was 0.775, while in the mental activity index, the AUC was 0.814. A well-balanced cutoff can be made at the point closest to the upper left corner in the ROC curve. Figure 12 and Figure 13 show the discriminated counts for the sensitivity, specificity, and accuracy at this point. Sensitivity, specificity, and accuracy are defined as follows:

Sensitivity = True Positive / (True Positive + False Negative)

Specificity = True Negative / (False Positive + True Negative)

Accuracy = (True Positive + True Negative) / Total Count

		BDI	
		High risk of disease > 18	Low risk of disease ≤ 18
Vitality	Positive < 0.2816	93	328
	Negative ≥ 0.2816	24	590

Sensitivity	0.795
Specificity	0.643
Accuracy	0.660

Figure 12. Accuracy of optimal cutoff in vitality.

		BDI	
		High risk of disease > 18	Low risk of disease ≤ 18
Mental Activity	Positive < 0.408	5	17
	Negative ≥ 0.408	0	26

Sensitivity	1.000
Specificity	0.605
Accuracy	0.646

Figure 13. Accuracy of optimal cutoff in mental activity.

4. Discussion

4.1. Effect of the Recording Method

As a result of comparing the vitality index obtained by analyzing the fixed phrase and the call respectively, the fixed phrase tended to have a lower value and the standard deviation was also different. As the calculation of the voice indicator is based on the change in emotion, it is likely that the act of reading a fixed phrase would call for a lesser variation in emotional expression compared to phone conversations. For further consideration, we would like to determine if these differences such as the average and the standard deviation based on the recording methods could be corrected by, for example, changing or converting the values. This is for future research.

4.2. Correlation to BDI

In the fixed phrase recording, the vitality index and mental activity index both had a weak negative correlation to BDI. In addition, a higher correlation was found in the mental activity index. These results indicate that the higher the vitality index or mental activity index, the healthier the individual, and the lower the BDI score. On the other hand, the lower the indices of MIMOSYS, the worse the mental health condition and the higher the BDI. Hence, we consider such negative correlation between the two as actually a valid result.

As for the vitality index of the two classes separated by the BDI score shown by the box and whisker plot, a t-test shows a significant difference when observing the result of the fixed phrase recording, and the reasonable performance as a classifier can be expected. Moreover, the overlapping of boxes is smaller in the mental activity index than in the vitality index, and a higher discrimination ability can be expected.

On the other hand, some participants who made call recordings did not include those belonging to a class with a high disease risk with a BDI score greater than 18in the classification based on the BDI score. Therefore, in call recordings, we were unable to evaluate the discrimination of high disease risk. This is presumably due to the fact that the number of participants was limited, and

those test participants we obtained had low BDI scores, thus skewing the samples.

4.3. Discrimination Evaluation of High Disease Risk using MIMOSYS

As for the discrimination of the high disease risk mentioned in the previous section, the result of evaluation by AUC was 0.775 for the vitality index and 0.814 for the mental activity index. We can say that this application has a moderate ability in discriminating the state of mental health. It shows that the mental activity index has a higher performance than the vitality index. In addition, the accuracy of the mental activity index was also shown to be similar to the vitality index as well.

5. Conclusion

In this research, we conducted an experiment to compare BDI to the vitality score, which was collected and analyzed through MIMOSYS, a smartphone application and a system we developed to use voice as a way of monitoring one's mental health condition. We found that the vitality score had a significant correlation to BDI. Additionally, by its performance in discriminating a person who scored high in the BDI test, we were able to confirm its validity. In addition, the mental activity as a medium-term health index has a higher discrimination ability than the vitality as a short-term health index, indicating that data accumulation is effective. The results thus indicated that this system could be used as a daily monitoring system to supplement the routine checkup. At the same time, as the different recording methods showed a difference in their numerical values, the next challenge would be to eliminate this difference.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This research is (partially) supported by the Center of Innovation Program from Japan Science and Technology Agency. I received generous support in gathering the experiment participants from Kanagawa Prefectural government.

References

[1] R. C. Kessler, H. S. Akiskal, M. Ames, H. Birnbaum, P. Greenberg, R. M. A. Hirschfeld, R. Jin, K. R. Merikangas, G. E. Simon, P. S. Wang, "Prevalence and effects of mood disorders on work performance in a nationally representative sample of U.S. workers" Am. J. Psychiat., **163**, 1561–1568, 2006.

[2] World Health Organization, The Global Burden of Disease: 2004 Update, WHO Press, Geneva, Switzerland, 46–49, 2004.

[3] A. T. Beck, C. H. Ward, M. Mendelson, J. Mock, J. Erbaugh, "An inventory for measuring depression" Arch. Gen. Psychiat., **4**, 561–571, 1961.

[4] A. T. Beck, R. A. Steer, M. G. Carbin, "Psychometric properties of the Beck depression inventory twenty-five years of evaluation" Clin. Psych. Rev., **8**, 77–100, 1988.

[5] S. Izawa, N. Sugaya, K. Shirotsuki, K. C. Yamada, N. Ogawa, Y. Ouchi, Y. Nagano, K. Suzuki, S. Nomura, "Salivary dehydroepiandrosterone secretion in response to acute psychosocial stress and its correlations with biological and psychological changes" Biol. Psychol., **79**(3), 294–298, 2008.

[6] C.W. Hoge, C.A. Castro, S.C. Messer, D. McGurk, D.I. Cotting, R.L. Koffman, "Combat duty in Iraq and Afghanistan, mental health problems, and barriers to care" New Eng. J. Med., **351**(1), 13–22, 2004.

[7] R. N. McLay, W. E. Deal, J. A. Murphy, K. B. Center, T. T. Kolkow, T. A. Grieger, "On-the-record screenings versus anonymous surveys in reporting PTSD" Am. J. Psychiat., **165**(6), 775–776, 2008.

[8] M. Perrin, L. DiGrande, K. Wheeler, L. Thorpe, M. Farfel, R. Brackbill "Differences in PTSD prevalence and associated risk factors among World Trade Center disaster rescue and recovery workers" Am. J. Psychiat., **164**(9), 1385–1394, 2007.

[9] S. Mitsuyoshi, F. Ren, Y. Tanaka, S. Kuroiwa, "Non-verbal voice emotion analysis system" Int. J. Innov. Comput. I., **2**(4), 819–830, 2006.

[10] S. Mitsuyoshi, "Emotion recognizing method, sensibility creating method, device, and software" U.S. Patent 7 340 393, 2003.

[11] S. Shinohara, M. Nakamura, Y. Omiya, N. Hagiwara, S. Mitsuyoshi, S. Tokuno, "A mental health assessment method based on emotional level derived from voice" (Manuscript in preparation)

[12] N. Hagiwara, Y. Omiya, S. Shinohara, M. Nakamura, M. Higuchi, S. Mitsuyoshi, H. Yasunaga, S. Tokuno, "Validity of the mind monitoring system as a mental health indicator" in IEEE BIBE 2016, Taichung, 2016.

[13] A. Sekiyama, "Interleukin-18 is involved in alteration of hipothalamic-pituitary-adrenal axis activity by stress" in Society of Biological Psychiatry Annual Meeting, San Diego, USA, 2007.

[14] S. Tokuno, S. Mitsuyoshi, G. Suzuki, G. Tsumatori, "Stress evaluation by voice: a novel stress evaluation technology" in 9th International Conference on Early Psychosis, Tokyo, 2014.

[15] S. Shinohara, S. Mitsuyoshi, M. Nakamura, Y. Omiya, G. Tsumatori, S. Tokuno, "Validity of a voice-based evaluation method for effectiveness of behavioural therapy" in MINDCARE 2015, Milan, 2015.

[16] Y. Omiya, N. Hagiwara, T. Kirita, S. Shinohara, M. Nakamura, S. Mitsuyoshi, S. Tokuno, "Development of a monitoring system of mental health state using call voice with a smartphone" in HCG symposium, Toyama, 2015.

[17] Y. Omiya, "Development of the mind monitoring system (MIMOSYS) which can be able to monitor mental health status using call voice with a smartphone" in 7th Asia Pacific Regional Conference of IASP, Tokyo, 2016.

[18] S. Mitsuyoshi, Y. Tanaka, F. Ren, K. Shibasaki, M. Kato, T. Murata, T. Minami, H. Yagura, "Emotion voice analysis system connected to the human brain" in IEEE NLP-KE 2007, Beijing, China, 2007.

[19] D. P. Goldberg, Manual of the General Health Questionnaire, NFER Publishing, Windsor, England, 1978.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 345-355 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

The Model of Adaptive Learning Objects for virtual environments instanced by the competencies

Carlos Guevara[1*], Jose Aguilar[2, 3], Alexandra González-Eras[3]

[1]*Centro de Investigaciones en Informática y Tecnologías de la Computación - CITEC, Universidad Nacional Experimental de Guayana, 8050, Ciudad Guayana, Venezuela.*

[2]*Centro de Estudios en Microelectrónica y Sistemas Distribuidos - Universidad de Los Andes, 5101, Mérida, Venezuela. Prometeo Researcher - Universidad Técnica Particular de Loja, Escuela Politécnica Nacional, Ecuador*

[3]*Departamento de Ciencias de la Computación y Electrónica, Universidad Técnica Particular de Loja, 1101608, Loja Ecuador*

ARTICLE INFO

Article history:
Received: 25 March, 2017
Accepted: 04 May, 2017
Online: 17 May, 2017

Keywords:
Adaptive Learning Object
Virtual Environment
Virtual Education Competencies

ABSTRACT

This article presents the instantiation of the Model of Adaptation of Learning Objects (MALO) developed in previous works, using the competencies to be developed in a given educational context. MALO has been developed for virtual environments based on an extension of the LOM standard. The model specifies modularly and independently two categories of rules, of adaptation and conversion, giving it versatility and flexibility to perform different types of adaptation to the learning objects, incorporating or removing rules in each category. In this work, we instance these rules of MALO using the competencies considered in a given educational context.

1. Introduction

The Information and Communication Technologies (ICT) have been introduced in the field of the education, in order to extend and improve the teaching-learning processes. In these environments, the use of learning objects (LO) is becoming more relevant every day to facilitate the teaching-learning process. LO are understood as any digital resource that can be used to support learning [1].

On the other hand, the interest of incorporating adaptive capacities, both to learning objects and virtual learning environments, has increased in recent years. However, there is no uniformity about what should be considered as an Adaptive Learning Object (ALO) [2-12]. In general, there are several works that are being developed on ALO and its application in virtual education. Most of these works are mainly focused on the adaptation of LO to the virtual teaching-learning process considering several aspects, among which we can mention: the learning style of the students; the course offered; the types and means of storages; and the learning trajectories [29], in order to create a favorable and an optimal environment for the successful development of virtual education.

Another aspect to consider is the adaptability of the LO's as support to the formation of competences, assuming that a student has previous competences that will be strengthened during the learning process until reaching an objective competence [26]. In this sense, within the learning trajectories are established states of skills and knowledge to be reached, which determine the level of customization of the LOs in terms of its granularity (knowledge) and skills required in them, for the acquisition of these competence goals [25-29]. This is the approach that we are interested in this work.

Due to the boom of virtual education and the application of emerging paradigms in the teaching-learning processes, combined with the weaknesses found in the various papers reviewed about ALO, we have presented in a previous work a Model of Adaptation of Learning Object (MALO) reconfigurable [2], based on the Learning Object Metadata (LOM) standard [3]. In this paper, we test the capability of reconfiguration of MALO, where its adaptation is guided by the competencies to be reached in a given course, that means, determining the level of adaptation of the LOs using MALO, according to the competencies to be

*Corresponding Author: Carlos Guevara, Centro de Investigaciones en Informática y Tecnologías de la Computación - CITEC, Universidad Nacional Experimental de Guayana, Venezuela | Email: carlos.guevaraf@gmail.com

reached. In this case, the rules are redefined based on an autonomous adaptation of the LO.

This article is organized as follows, the next section presents some theoretical aspects, and then we present in detail to MALO. The next section presents the example of adaptation of MALO using competencies, and finally, we compare MALO with previous work.

2. State of the art

In [4] is proposed an ALO model composed of four components, a knowledge domain model, a user domain model, a context model, and finally a session model. The fourth component is responsible for establishing relationships between the other three components, through navigation rules, content presentation rules, and rules of personalization of the information presented to the user. In that work, the rules of adaptation are specified to each ALO, which makes its implementation difficult. In addition, these are limited to navigation patterns within a LO. Finally, the proposal only presents rule statements, without any implementation. [5] presents an ALO model for t-learning, a term adopted to refer to learning through interactive digital television. The idea that they pursue is to offer LO with different behaviors, depending on the characteristics of the users. To do this, they use an XML file that contains a template with rules of adaptation of parameters, which indicate to the educational object what behavior or appearance to adopt according to the student's preferences. In this work, they do not incorporate autonomous changes in the learning process.

In order to construct Adaptive Learning Systems, in [6] is proposed a model considering the granularity properties of the LO. Its proposal is based on the adaptation of the LO, by adding four levels of functionality: courses, documents, fragments of the document and multimedia pieces. The fragments are made up of multimedia pieces related to some instructional objective and a semantic description, which allows them to assemble a new learning object from the existing ones. The architecture presented in [6] is composed of 3 main components, a Domain Model formed by the concepts that the learner must learn; a Learner Model, which represents the learner's information; and an Adaptation Model, which is responsible for the adaptation of the contents that are presented to the learner, and the way the fragments are assembled. In this model, the authors use a modular and structured treatment of the adaptation process, however, they do not consider the technological limitations, and nor do they describe the rules to be used to adapt to LO.

The work developed by [7], establishes a relationship between adaptation needs and learning materials. It considers that the needs of adaptation are influenced by the user, according to their cognitive and learning styles; the context, through the location and tasks that the user performs; and the learning material, with its own characteristics. They determine 13 adaptation criteria, which can be useful for classifying content and adapting LO. However, they do not present a model guiding how to incorporate these adaptation criteria into a LO. In [29] propose an adaptive framework of LOs, assessment items, and competences. In specific, the competencies are defined by the prerequisite.

3. Theoretical Aspects

3.1. Learning Object

There are several definitions of LO, some of them are: "any digital resource that can be used to support learning" [1], "any entity, digital or non-digital, that maybe be used for learning, education or training" [8], "an element with recyclable digital multimedia content documents, which has a purpose and use in instruction and learning" [9].

According to the definitions cited, [8] considers LO in a rather broad sense, by including in it any resource, both digital and non-digital, that can be used for learning, education or entertainment. On the other hand, [1] delimits it only to digital resources, so this definition is more oriented to digital technology. For its part, [9] makes explicit the use of digital documents with multimedia contents in the learning process.

This shows that there is no consensus on the definition of LO. For the purpose of this research, the interest is focused on LO used in virtual environments, what can be called Virtual Learning Objects. Therefore, LO will be considered a reusable multimedia digital resource, which can be used in learning, and employ metadata for description. This highlights two important aspects in LO: it is a digital resource with multimedia features for educational purposes, and the metadata is used for its description to facilitate its discovery and reuse.

In general, a LO has a set of attributes, such as reusability, portability, modularity, granularity, accessibility, flexibility, interoperability and durability. In specific, the granularity is an important property, because it facilitates its scalability and adaptation, and it simplifies the coupling or combination of different LO to form a new one. Some authors consider granularity from the point of view of the number of concepts or ideas contained in the LO [1,18]. Others, such as the LOM standard of IEEE, consider it as the amount of media or formats that are combined in a LO [6]. For purposes of this work, granularity is assumed from three approaches:

- Concepts or ideas combine in a learning object, which provide the semantic density of the LO
- Media or formats that are combined in a LO. They determine how the content is presented in LO.
- Parts or elements in which the content of the LO is organized. It is like the contents table, where, for example, the document parts, the book chapters, or the course subjects, are indicated.

3.2. Standards for production of Learning Objects

The standards for LO are grouped into metadata standards and standards for the creation of educational platforms. The first seeks to describe and provide LO properties and information to facilitate and optimize their manipulation; while the second helps to define virtual environments that integrate a set of tools that provide support to the online teaching-learning process. Among the most widely used metadata standards are the Dublin Core [10] and the IEEE LOM, while SCORM and IMS are used for the creation of educational platforms.

In our work, we use the LOM standard, which is a data model proposed by the IEEE, through the Learning Technology Standards Committee (LTSC), to describe a learning resource and its components. The purpose of the standard is to facilitate the

search, evaluation, acquisition and use of educational objects, as well as, to promote their exchange, reusability and share use. It consists of 80 descriptors, grouped into nine categories [3]. Each descriptor defines a specific aspect to be considered to describe a LO. Table 1 shows the categories defined in the LOM Standard.

Table 1. LOM Standard Summary **[3]**.

Nº	Category	Description
1	General	This category groups the general information that describes the learning object as a whole
2	Life Cycle	This category describes the history and current state of the learning object and those entities that have affected this learning object during its evolution.
3	Meta-Metadata	This category describes the metadata record itself (rather than the learning object that this record describes).
4	Technical	This category describes the technical requirements and characteristics of this learning object.
5	Educational	This category describes the key educational or pedagogic characteristics of this learning object.
6	Rights	This category describes the intellectual property rights and conditions of use for the learning object
7	Relation	This category defines the relationship between this learning object and other learning objects, if any.
8	Annotation	This category provides comments on the educational use of this learning object, and information on when and by whom the comments were created
9	Classification	This category describes where this learning object falls within a particular classification system.

The LOM standard is considered as an extension of the Dublin Core standard, with additional elements for the description of educational resource properties. For this reason, to represent the LO metadata, in this work we decided to use the standard LOM

3.3. Competencies

The definition of competence is derived from the Latin word "competere", which means "to be adequate", that implies "the combination of skills, attitudes and knowledge necessary to carry out a task effectively" [11]. In this way, the elements of the competence are the skills and knowledge. The knowledge comprises the set of topics and themes that are part of a profession and that the students need to develop, while the skill represents the abilities to use the knowledge in the development of an activity [12]. Thus, the competence defines the person's ability to perform a task [26], as a result of the relations between a specific knowledge, and a know-how [13].

In the learning contexts, the definition of competence refers to the set of skills and knowledge to solve specific problems [26], which relate the previous competences that the student has and the objective competence, i.e. the problem to be solved [29]. Depending on the objective competence, can be set the knowledge and skill levels required, and with the learner's current knowledge and skill state, can be determined the learning trajectory. Based on this learning trajectory, personalized learning paths and adequate LOs can be defined.

The Figure 1 shows the process of constructing of the objective competence: "to establish the difference between a vegetarian dish and others" proposed in Table 11. According to the Figure 1, can be set the learning trajectory knowledge and skills required to fulfill the objective competence, taking as a starting point the previous competence that the student possesses and defining evaluation moments in the learning trajectory.

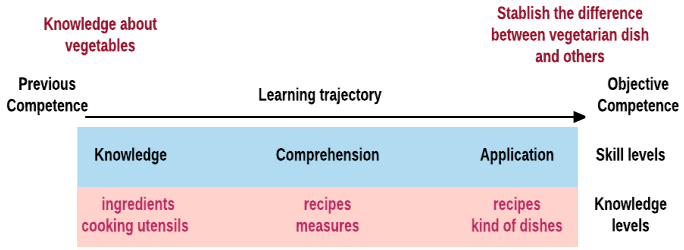


Figure 1. Objective competence construction in a learning trajectory

3.4. MASINA Methodology

MASINA (MultiAgent Systems for INtegrated Automation) is a methodology developed for the modeling of systems of engineering oriented to agents. It uses the Unified Modeling Language (UML), and the Object Systems Development Technique (OSDT), which is a tool for the formal specification of object-oriented models [23, 24]. It covers the phases of conceptualization, analysis, design, coding and testing, and integration. Now, we describe only the phase and model used in our work, for more detail see [23, 24]

- Conceptualization: allows to identify the actors of the system and their use cases. In this phase basically the extraction and acquisition of the knowledge is realized to obtain a first description of the system. The main product of this phase is a conceptualization document, which contains the analysis of the problem, and the use cases with the description of the components of the system. The use case defines the list of actions of each a role in our system, to achieve a goal.

- Analysis: it allows to describe the system, its objectives and tasks, capabilities, interactions, etc. The main product of this phase is a set of models like the task model, that describe each component of the system at the level of its procedure, condition of utilization, etc. of the different component, to be considered.

In our work, we use these two first phases to design MALO, and particularly, the task and user case models of MASINA.

4. Model of Adaptation of Learning Object (MALO)

For the formal description of the model, we used the MASINA methodology proposed in [23-24], and UML diagrams. However, before the presentation of MALO, it is necessary to define an ALO.

4.1. Definition of ALO

An ALO is considered a reusable multimedia digital resource that can be used in a learning process, which contains metadata for its description, and is able to adapt to the context. The ALO concept is presented in Figure 2, through a use case diagram, and in the Figure 3 its activity diagram.

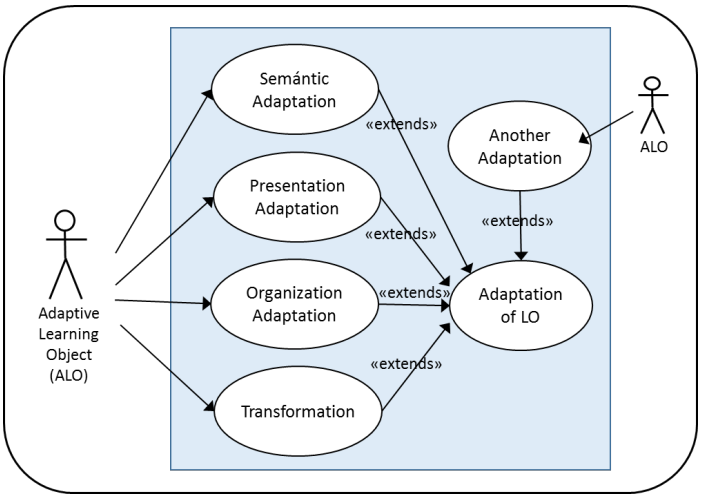


Figure 2. Use Case Diagram of an ALO

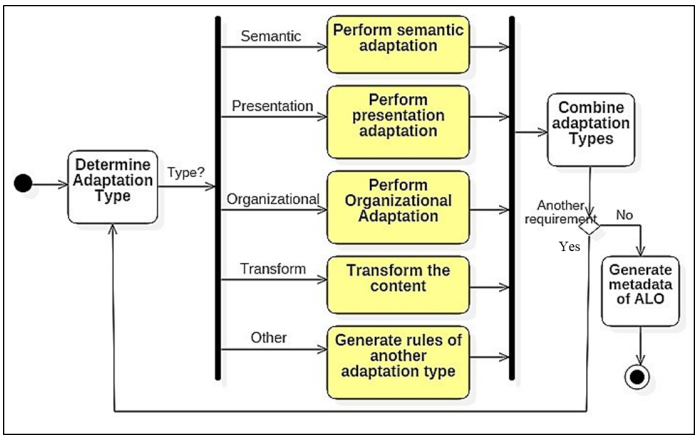


Figure 3. Activity Diagram of an ALO

As can be seen in the ALO activity diagram, its adaptation is a complex, recursive and dynamic process, where several types of adaptation can occur simultaneously, in addition, it has the flexibility to include other types of adaptation in the future. The ALO use case description is presented in Table 2, and the description of the types of adaptation is given in the following subsections. In the use case, we can see the adaptation procedure of the ALO, which can fail if there are not the services for the adaptation process. Also, the precondition conditions are very clear, we need to define the adaptation requirements that can be defined in our case, by the competencies to be reached.

Table 2. Use Case Description of an ALO

Use Case	Adaptive Learning Object
Description	It processes the requirements of adaptation of a LO.
Precondition	Adaptation requirements
Actors	Adaptive Learning Object

Procedure	Begin
	Determine types of adaptation required
	If type = semantic then
	Adapt semantically the LO
	If type = organization then
	Adapt organizationally the LO
	If type = presentation then
	Adapt the presentation of the LO
	If type = transformation then
	Transform the LO content
	If type = other then
	Generate the rules of another type of adaptation of the ALO
	Combine the types of ALO adaptation until to reach the requirements
	Generate Matadata of ALO
	End

Fail condition	Not having the necessary services to process the LO adaptation requirements
Success condition	Build a LO adapted to the required specifications

4.2. MALO Description

MALO has been defined in detail in [21]. In this section, we present a resume of MALO. Figure 4 presents MALO, which is composed of 4 elements:

- LO: are the Learning Objects to adapt, according to the information obtained from the context.
- Units: they are small blocks, such as concepts, media or pieces, that compose a LO. The number of units present in the LO, define their granularity and complexity.
- Rules: two categories of rules are defined, Adaptation and Conversion, which specify, respectively, the structural or contextual adjustments made to a LO.
- Adaptation Metadata: it is an extension of the LOM standard, which describes the data and processes required to facilitate the process of adaptation of the LO.

A LO can be decomposed into smaller units by using rules of conversion, forming a set with simpler LO, increasing its granularity; or conversely, starting from a set of LO, a more complex LO can be formed with less granularity. With the use of the rules of adaptation, the LO is adapted to certain specifications of the context in which it is immersed.

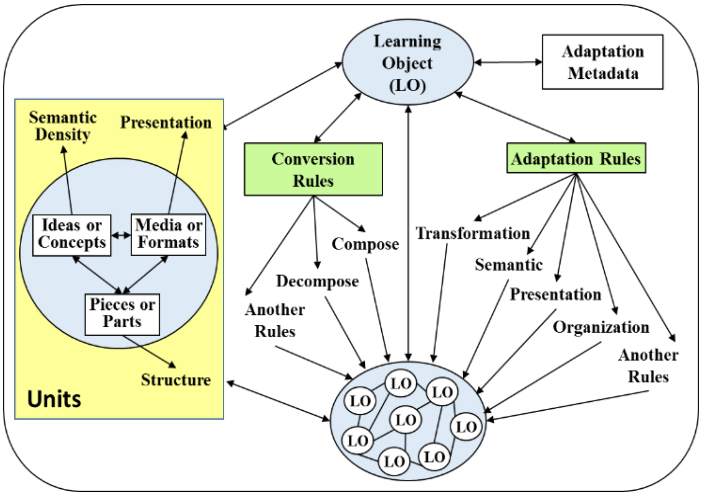


Figure 4. MALO

The result of applying the adaptation and conversion rules to an ALO is a Learning Object with a new semantic, or a new presentation format of the content, or a new organization, among

other things; based on the parameterization of an existing one, or the combination of several of them.

4.2.1. Conversion Rules

The conversion rules indicate that structural adaptations of a LO. In this article, two conversion rules are defined, however, MALO has the flexibility to allow the incorporation of new rules. The basic conversion rules are:

- Composition: based on the principle of reusability, allows to build new LO from other LOs of less granularity. The description of the composition rule is presented in Table 3, which has 5 rules.
- Decomposition: allows to decompose a LO into several LOs of less granularity, or to extract units included in a digital content. The description of the decomposition rule is presented in Table 4, which has 7 rules.

Table 3. Task Model of the Composition Rule

COMPOSITION OF A LEARNING OBJECT	
Name	Composition
Objective	Compose a LO from other LOs.
Description	Composes the content of a LO by total or partial aggregation of units (concepts, pieces, media) contained in other LOs
Precondition	Units required
Subtask	- Define the LO organization to be built - Add concepts or ideas - Add pieces or parts of the structure of a LO - Add media or formats to the contents of a LO - Generate metadata of the new LO
Product	Adaptive Learning Object
Ingredients-Compose Learning Object	
Composition Type	Indicates the type of composition to be made. Values = semantics, organization, format, mixed
Concepts List	Concepts or ideas contained in a LO
Pieces List	Different elements or part in which the LO is organized, similar to a table of contents. For example, document parts, book chapters, course topics, workshop content and activities.
Formats List	Different media or formats that are combined in a LO. Values = Text, Audio, Video, Interactive Animation, Simulator, Image, Table or Chart, Diagram.

Table 4. Task Model of the Decomposition Rule

DECOMPOSITION OF A LEARNING OBJECT	
Name	Decompose
Object	Decompose a LO in order to increase its granularity
Description	Disaggregate the content of a LO by units (ideas or concept, parts or parts, formats or media)
Precondition	Requirement of concept, piece or format
Subtask	- Search for concepts / ideas contained in a LO - Extract concepts or ideas from a LO - Extract organization of a LO - Remove parts of the structure of a LO - Search for formats / media embedded in a LO - Extract media or formats contained in a LO - Generate metadata of the new LOs
Product	Learning objects with greater granularity
Ingredients- Decompose Learning Object	
Decomposition type	- Indicates the type of decomposition required Values = semantics, structure, format, mixed
Concepts List	- Concepts or ideas contained in a LO
Pieces List	- Different elements or part in which the LO is organized, similar to a table of contents. For example, document parts, book chapters, course topics, workshop content and activities.
Formats List	Different media or formats that are combined in a LO. Values = Text, Audio, Video, Interactive Animation, Simulator, Image, Table or Chart, Diagram.

4.2.2. Adaptation Rules

The rules of adaptation define the types of adjustments to be applied to the LOs to contextualize it. In MALO, four types of rules are initially defined:

- Semantics: they allow to define new semantics in a LO, incorporating other ideas, conceptualizations or lexicons, according to its content.
- Presentation: allows to adapt the presentation of the LO, adjusting its formats to the preferences, limitations and use, both technology and user.
- Organization: they allow to reorder the content of a LO according to the pieces that make them up.
- Transformation: they change the units of measurement, of quantification, etc., in the LOs.

Tables 5, 6, 7 and 8 present the use cases of each adaptation rule. The main aspect to remark of these tables is that the *precondition* is one of the main criteria to determine what possible adaptation can be used in a given context, which can be used by a procedure of selection of adaptation mechanisms.

Table 5. Use Case Description of the Semantic Adaptation

Use Case	Semantic Adaptation
Description	It allows defining a new semantics of a LO, incorporating ideas or concepts according to their contents.
Precondition	Concepts or ideas required
Actors	Adaptive Learning Object
Procedure	Begin Determine new concepts or ideas or lexicon Find where to put them into the LO Incorporate them into the LO Generate metadata End
Fail condition	- Do not locate the required concepts in data repositories. - Not having the necessary services for the extraction of concepts, decomposition and composition of the LO
Success condition	Construct a LO with new semantics required.

Table 6. Use Case Description of the Presentation Adaptation

Use Case	Presentation Adaptation
Description	They allow to adapt the presentation of a LO modifying its format
Precondition	- Technological Preferences and limitations - User Preferences and Limitations - Conditions of use and desired functionality of the LO.
Actors	Learning Object, Context, User
Procedure	Begin Determine technological preferences and limitations Determine user preferences and limitations Determine conditions of use and functionality of LO Customize LO to those preferences and limitations Generate metadata End
Fail condition	- Not having the mechanisms to determine preferences, limitations and desired functionality of the context - Not having the necessary services to transform, customize, decompose or compose a LO
Success condition	- Build a LO with a new presentation, transformed and/or customized. - Given some preferences or limitations recommend formats, customization parameters, functionality

Table 7. Use Case Description of the Organizational Adaptation

Use Case	Organizational Adaptation
Description	It allows reorganizing the contents of a LO by rearranging its pieces.
Precondition	- Pieces or parts desired to organize the content of the LO - Concepts or ideas required
Actors	Adaptive Learning Object
Procedure	Begin Determine pieces or part of LO. Determine the reorganization parameters Rearrange LO Generate metadata End
Fail condition	- Do not locate the required pieces in the LO repositories. - Not having the necessary services for the extraction of pieces, decompose or compose a LO
Success condition	Build a LO with the new organization of its content. Recommend an organization for the LO

Table 8. Use Case Description of the Transformation

Use Case	Transformation
Description	Modifies certain properties or contents of a LO
Precondition	- Requirement of concept, piece or format. - Technological or user preferences or limitations - Usage patterns
Actors	Adaptive Learning Object
Procedure	Begin Summarize the content (concepts, organization) of a LO. Transform/Convert values of certain properties of some units to others (for example, of types of images, etc.) Generate metadata End
Fail condition	- Not finding the required units in LO repositories - Not having the necessary services for the transformation of the LO units
Success condition	Build a new LO with new features and/or units

4.3. Adaptation Metadata

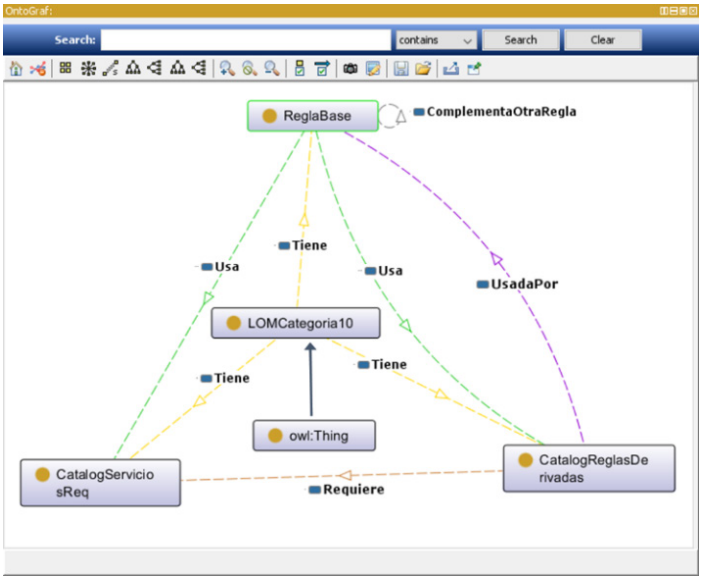
The LOM standard covers quite well the description of a LO. However, it presents limitations for the adaptation of LOs autonomously. Specifically, the adequacy of a LO is guided by a set of mechanisms, which must be defined in the metadata of the LO. The main mechanisms to be defined are for the adaptation, the conversion, according to the information about the context where will be used the LO. This information is not currently included in the LM standard. For this reason, in this paper, it is proposed to incorporate an extension to the standard (defined as category 10 of the standard), based on the data presented in Table 9.

Table 9. Category 10. Extension proposal for the LOM standard

No.	**Category**	**Description**	**Components**
10	Adaptation	This category groups the necessary information, to facilitate the process of autonomous adaptation of a Learning Object	10.1.Adaptation Rules (Name, Type, Description, Associated services, Required inputs, Procedure, Author, Date of creation, Date of modification) 10.2.Conversion Rules (Name, Type, Description, Expected Product, Associated Services, Required Inputs, Procedure, Related Tasks, Author, Date Created, Date Modified) 10.3 Context information (Knowledge Base, Rules of Inference, location of the Knowledge Base) 10.4 Restrictions 10.5 Other additional information

The proposed category specifies the information used to manage an ALO. The conversion and adaptation rules are methods to be applied to an ALO, which are implemented in some LO management platform. Thus, this category defines the knowledge base that allows the interoperability between ALOs, and their adaptation methods.

This category facilitates the search, evaluation, recovery and application of the adaptive methods for an ALO. The ontological model of the Category 10 (OLOMCat10) is presented in Figure 5.



OLOMCat10 relates Category 10 with 3 major components: a base rule, a rules catalog, and a services catalog. The base rule is the initial rule that triggers the ALO adaptation process. The Rules Catalog describes the set of conversion and adaptive rules. The Services Catalog describes and links a set of mechanisms required during the ALO adaptation process. The purpose of catalogs is ubiquity, making ALO independent of both, the rules of adaptation and the services used to implement them.

5. MALO instanced by the competencies

In this section are defined the adequacy of MALO, to consider the competencies during the adaptation process of the LOs.

Considering that the competencies that are applied in the professional field are built on the basis of the training and the professional formation, it is clear to establish a relation between the competencies and the learning results, as expressions of what a person in a learning process knows, understands, and is able to do at the end of an educational process.

As a result, a LO becomes a valid resource within the competence acquisition process, achieving an objective, a result or competence in a specific knowledge [21,25], which is conditioned to the quality of the LO, which involves aspects linked to its structure and granularity, and therefore, its adaptation in all or parts of its components [13].

Adapting a LO based on competencies requires the adjustment of the LO according to the user's previous competencies and the competencies objective (or learning result), in order that the LO has the sufficient characteristics to cover the competence objectives. In addition, should provide facilities to evaluate the competencies acquired. Thus, within the MALO model, the following aspects must be taken into account:

- Through the conversion rules, the LO must adapt its structure, to add units as required to meet the competence objectives, i.e. the composition rule can add a unit from another LO related to the skill levels of the objective competence.
- By contrast, through the adaptation rules, the LO can update its contextualization. For example, by the organization rule, the contents can be reorganized according to the competence requirements; or through the semantic rule; the LO can be enriched with links to external resources; or also through the transformation rule, the LO can incorporate measures establishing the skill level that a user must achieve after learning the specific knowledge that keeps the object.
- And also through adaptation rules, the LO can propose assessment tools to evaluate the competencies acquired in the learning process

Additionally, the enrichment of a LO with competence metadata can be incorporated into MALO, through the category 9 of the LOM standard (see Table 10), which allows the extension of the specification to a semantic model of competencies that uses a combination of pedagogical strategies and competence taxonomies to classify and evaluate the LO's performance in a learning environment [14].

Table 10. Category 9 of the LOM standard extended to include the competencies

Nro	Category	Explanation
9	Classification	This category describes where this learning object falls within a particular classification system.
9,1	Purpose	The purpose of classifying the LO (discipline, idea, prerequisite, educational objective, accessibility, restrictions, educational level, skill level, security level, competence)
9.2	Taxon path	A taxonomic path in a specific classification system.
9.3	Description	Description of the LO relative to the Purpose, such as discipline, idea, skill level, educational objective, competencies, etc.
9.4	Keyword	Keywords and phrases descriptive of the LO relative to the Purpose and competencies.

A hierarchical relationship between the different subcategories of Category 9 of the LOM standard is presented in Figure 6, where the competence (purpose) is defined by a taxonomic classification (taxon path), their description and keywords

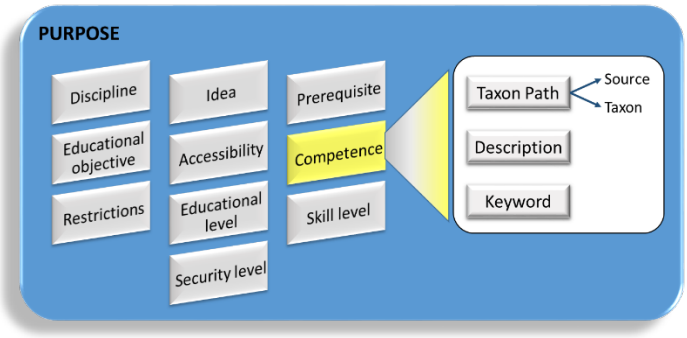
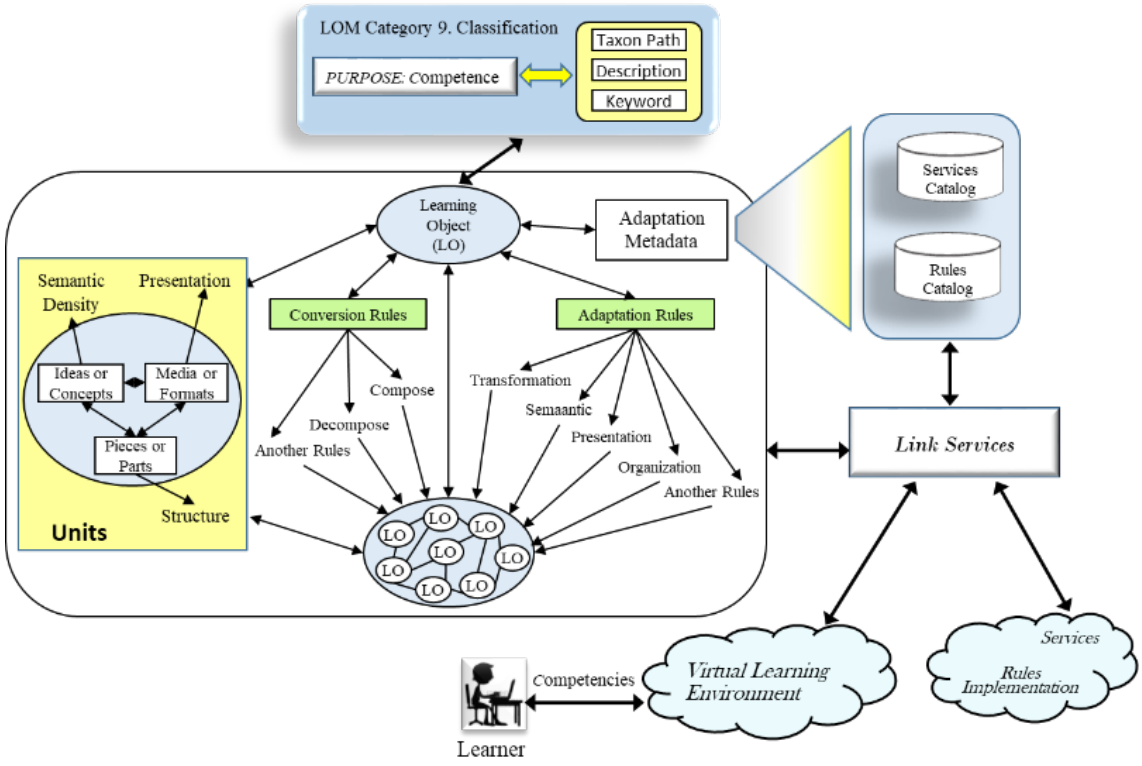


Figure 6. Category 9 of the LOM Standard. Hierarchical description of competence

The use of LO as a resource to generate competencies in an apprentice within a virtual learning environment, requires to adapt the LO according to the required competencies, learning objective

and expected results. For this reason, Figure 7 presents a mechanism to instantiate the ALO in order to facilitate this process. For that, it is required**:**

- Virtual Learning Environment (VLE) provides the necessary environment for the development of the teaching-learning process.
- MALO defines the mechanisms for the LO adaptation
- Adaptation Metadata identifies catalogs of services and rules required to adapt to an ALO.
- Link Services provide ALO connection with VLE, to obtain context information, resources available for learning, as well as access to the Rule and Service catalogs.
- Category 9 of the LOM standard contains the metadata to facilitate the discovery of LO with contents related to the required competencies.



According to Figure 7, the required competencies by the learner are obtained from the VLE. MALO locates resources and LOs, which satisfies the required competencies and LOs. For this, it uses the information of the Category 9 of the LOM standard. MALO provides services with the conversion and adaptation rules defined in the metadata (Category 10), which access the mechanism catalog that allow the adequacy of the ALO according to the requirements. As a result of this process, an ALO is adapted to the competencies.

6. Description of a Case Study

Our case study test the capability of reconfiguration of an ALO using MALO, guided by the competencies to be reached in a given course. An application of the competence approach within MALO can be seen in the Table 11, to which the following considerations are carried out:

- It is established as a previous competence of the user: "knowledge of vegetables".
- It is established as objective competence: "Establish the difference between vegetarian dish and any other".

Table 11. Cooking course: Vegetarian dish preparation

Course activities	- Select vegetables and other ingredients (input) - Prepare mixture (process)
Learning Objects participants	LO1: Interactive virtual course, where the apprentice must follow the instructions and prepare the requested plate LO2: Digital multimedia content with a catalog of vegetables, classified by edible and non-edible vegetables.
Content of the virtual course	The content of the virtual course is composed of two (02) parts: *Ingredients:* in this part the student must demonstrate his/her knowledge, selecting the vegetables and other required ingredients. All ingredients are specified in text form *Preparation:* indicated by a video, step by step, how to prepare the dish.
Scenario	- Previous competence: student with extensive knowledge of vegetables - Limitations: reading problems. - Preference: follow hearing instructions

	- Objective competence: LO does not differentiate between vegetarian and other dishes
Classification	Establish the difference between a vegetarian dish and any other
Previous Competence	Knowledge about vegetables
Objective Competence	Establish the difference between a vegetarian dish and others
Skill Levels	Knowledge→Comprehension→Application

The Table 11 defines the LO to be used, the previous and objective competence, the content of the virtual course to be covered, among other things. Considering that scenario defines that the LO does not meet the objective competence: "Establish the difference between a vegetarian dish and others", then it is necessary to add a LO with knowledge about "Differences between vegetarian and other dishes". For this, in addition to the set of rules explained in section III, should include rules of semantic adaptation, presentation adaptation and transformation, besides to the rule of decomposition and composition. Using MALO, the following procedure will be performed:

1. Application of Presentation Adaptation rules:
   1.1. User preferences are determined.
   1.2. User Competencies are determined
      1.2.1. Previous competence
      1.2.2. Objective competence
      1.2.3. Skill levels of Objective Competence
   1.3. Limitations of the user are determined: reading problems
   1.4. Technological preferences are determined: not specified
   1.5. Technological limitations are determined: not specified
2. Application of rules of Semantic Adaptation,
   2.1. Search for a LO that covers the target competence and user preferences.
   2.2. Search for a LO assessment tool to evaluate the objective competence and its skill levels
   2.3. Using Decomposition rules
      2.3.1. Units of LO are retrieved from the source.
3. Application of Transformation rule
   3.1. According to the limitations and preferences of the user,
      3.1.1. Of the new LO found (LO3), extract audio from video
4. Application of rule of presentation.
   4.1. Adjusts the volume of the extracted audio
5. Application of Composition rules,
   5.1. The content of LO3 is integrated into the adapted LOs.
6. Application of rules of organization,
   6.1. The structure of the adapted LO is reorganized according to objective competence.
7. Generate the new ALO metadata
   7.1. Update the competence metadata in the category 9 of the LO: purpose, taxon path, description and keyword
8. Product: the LO was adapted to add the ALO3, which contains the differentiation between a vegetarian dish and others, in audio format, and assessment tools to evaluate the competencies

7. Comparison of MALO with other Models

For the comparison of MALO with the previous works described in section I, we consider the most similar that are **[4] [6]** and [29]. The comparison criteria were based on the characteristics must fulfill an ALO. We have used the next set of criteria: what adapts each one (because in an ALO we can adapt several things: organization, content, presentation, etc.), the type of rules used (that determine the adaptation of the ALO), the architecture of the framework (that determine if it can be improved very easy), the extensibility (very important for its scalability), the flexibility (in order to include new aspects in the adequacy of an ALO), and if it is based on metadata, and particularly, in a standard. The comparison is presented in Table 12.

Table 12. Comparison of MALO with other proposals

Criteria for Comparison	
Components to adapt	[5]: content and navigation patterns of a particular user, according to user preferences
	[13]: content, navigation and presentation
	[29]: organization of the ALO and assessment items,
	MALO: semantics, presentation and organization of contents. Possibility of adding other types of adaptation
Rules used	[5]: propose the use of content adaptation, navigation and evaluation rules.
	[13]: propose rules of navigation, content and presentation.
	[29]: they propose a set of services, but they do not give a guide about how they must be used
	MALO: two categories of rules, adaptation (semantics, organization, presentation, transformation) and conversion (decomposition, composition) are proposed.
Modularity	[5]: is based on rules of adaptation, knowledge of the domain and the user
	[13]: is based on content granularity, adaptive rules and user domains, context and adaptation
	MALO: modular design based on adaptation components, metadata and internal LO units
Extensibility	[5]: They do not propose it
	[13]: They do not propose it
	[29]: new services can be included
	MALO: It proposes the scalability of the model, allowing the incorporation of new rules of adaptation and conversion
Flexibility	[5]: it is fundamentally based on the preferences and limitations of the user, as well as on their learning needs.
	[13] incorporates preferences and limitations of the user, as well as learning resources, but does not consider technological aspects nor patterns of use
	[29] they use the competencies to guide the adaptation
	MALO: possibility to adapt the preferences and limitations, both user and technologies.
Adaptation Metadata	[5]: They do not propose
	[13]: They do not propose
	[29]: they do not propose an adaptive metadata
	MALO: The model works with the IEEE-LOM standard. It is proposed to incorporate into the standard an adaptation category, which describes the new elements used in the adaptation

When comparing the fourth approaches of adaptation of LO, it is observed that there are different criteria to consider. There is also no uniformity in the description of the adaptation metadata, nor in the modularity, extensibility and flexibility that each approach allows.

With respect to MALO, its modular design, extensibility, and the use of standards provide it with greater robustness and flexibility to adjust its strategies and processes. In addition, it is the most complete in terms of what it adapts, the rules it uses, the descriptions it provides in its metadata, among other things.

8. Conclusions

The main product of this work is to test the capability of adaptation of MALO using the competencies. Among the main features of the MALO are its modular and independent design based on two categories of rules, adaptation and conversion, which are complemented to facilitate the dynamic adaptation of an ALO. The flexibility and extensibility of its design, allow the incorporation and elimination of rules. In addition, MALO considers the preferences and limitations, both user and technology, as well as the Desired functionality of the LO, and the usage patterns detected in the context.

The competencies determine the skills and knowledge to be obtained in a given course, in this way they give useful information to adapt the LO to be used during the course. It determines specific conversion and adaptation rules. For example, according to the knowledge to be reached, it can define conversion rules to include new information in a given ALO. Also, according to the skills to be reached, it can define adaptation rules that allow reaching these skills.

There are several future works, in particular, the proposed extension of the LOM (category 10) and the description of the category 9 how a competence metadata, presented in this article, should be detailed. Finally, it is necessary to implement a middleware that provides services to MALO, inspired by works such as [27,28] to manage repositories of ALOs, which autonomously exploit all the knowledge around the MALO, and can adjust an ALO using different frameworks, such as the competencies.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

Dr. Aguilar has been partially funded by the “Proyecto Prometeo” of the Ministry of Higher Education, Science, Technology and Innovation (SENESCYT) of the Republic of Ecuador.

References

[1] D. Wiley, "Connecting learning objects to instructional design theory: A definition, a metaphor, and a taxonomy," 2000. [Online]. Available: http://reusability.org/read/. [Accessed 15 8 2014].

[2] Q. Jin, Intelligent Learning Systems and Advancements in Computer-Aided Instruction: Emerging Studies, Japan: IGI Global, 2011, p. 378.

[3] M. Rey, R. Díaz, A. Fernández, J. Pazos and M. López, "Objetos adaptativos de aprendizaje para t-learning," *IEEE Latin America Transactions,* vol. VOL. 5 , no. NO. 6, October 2007, pp. pp. 401-408, 2007.

[4] A. García, L. Marcos and E. García, "An empirical study on m-learning adaptation: Learning performance and learning contexts," *Computers & Education,* vol. 82, pp. 450-459, 2015.

[5] J. García, M. Castañón, A. Rodríguez and A. Cristóbal, "Adative Learning Object," in *Proceedings of the International Conference on Artificial Intelligence, IC-AI-04*, Las Vegas, USA, 2004.

[6] N. Matar, "Multi-Adaptive Learning Objects Repository Structure Towards Unified E-learning," *International Arab Journal of e-Technology,* vol. 3, no. 3, pp. 129-137, 2014.

[7] S. Hammami and H. Mathkour, "Adaptive e-learning system based on agents and object petri nets (AELS-A/OPN)," *Computer Applications in Engineering Education,* vol. 23, no. 2, pp. 170-190, 2015.

[8] E. Kurilovasa, S. Kubilinskienea and V. Dagienea, "Web 3.0 – Based personalisation of learning objects in virtual learning environments," *Computers in Human Behavior,* vol. 30, pp. 654-662, 2014.

[9] D. Mérida, R. Fabregat and S. Baldiris, "Sistemas heterogéneos adaptativos basados en el contexto," *RIED. Revista iberoamericana de educación a distancia,* vol. Vol 13, no. Nro 2, 2012.

[10] I. Gutierrez, "Usando Objetos de Aprendizaje en Enseñanza Secundaria Obligatoria," *EDUTEC- Revista Electrónica de Tecnología Educativa,* vol. Nro 27, no. Noviembre 2008, pp. pp. 1-17, 2008.

[11] S. Dietze, S. Sanchez-Alonso and otros, "Interlinking educational resources and the web of data: A survey of challenges and approaches.," *Program 47,* vol. 1, pp. 60-91, 2013.

[12] J. Aguilar, J. Zechinelli and J. Muñoz, "Hacia la creación y administración de repositorios de objetos de aprendizaje," in *Cuarto Encuentro Internacional de Ciencias de la Computación*, 2003.

[13] A. Battou, I. Zohr, A. El Mezouary and D. Mammass, "Towards an Adaptive Learning System Based on a New Learning Object Granularity Approach," in *(IJACSA) International Journal of Advanced Computer Science and Applications*, 2011.

[14] S. Thalmann, "Adaptation criteria for the personalised delivery of learning materials: A multi-stage empirical investigation," *Australasian Journal of Educational Technology,* vol. 30, no. 1, pp. 45-60, 2014.

[15] C. Guevara and J. Aguilar, "Model of Adaptive Learning Objects for Virtual Environments," in *Proceeding of XLII Conferencia Latinoamericana en Informática (CLEI 2016)*, 2016.

[16] IEEE, "Standard for Learning Object Metadata. Final Draft Standard," 2002. [Online]. Available: http://grouper.ieee.org/groups/ltsc/wg12/files/LOM_1484_12_1_v1_Final_Draft.pdf. [Accessed 11 10 2015].

[17] V. Menéndez, M. Castellanos, A. Zapata and M. Emilio Prieto, "Generación de Objetos de Aprendizaje empleando un enfoque asistido," *Pixel-Bit. Revista de Medios y Educación,* vol. Nro 38, no. Julio-Diciembre 2010, pp. pp. 141-153, 2011.

[18] R. Polsani, "Use and abuse of reusable learning objects," in *Journal of Digital Information*, 2003.

[19] DCMI, "Dublin Core Metadata Element Set, Version 1.1," 2012. [Online]. Available: http://www.dublincore.org/documents/dces/. [Accessed 22 02 2016].

[20] A. J. Romiszowski, "Designing instructional systems: Decision making in course planning and curriculum design," *Routledge,* 2016.

[21] A. Gonzalez and J. Aguilar, "Semantic Architecture for the Analysis of the Academic and Occupational Profiles Based on Competencies," *Contemporary Engineering Sciences,* vol. 8, no. 33, p. 1551 – 1563, 2015.

[22] Y. Ossandón Núñez and P. Castillo Ochoa, "Propuesta para el diseño de objetos de aprendizaje," *Revista Facultad de Ingeniería-Universidad de Tarapacá ,* vol. 14, no. 1, pp. 36-48, 2006.

[23] J. Aguilar, I. Bessembel, M. Cerrada, F. Hidrobo and F. Narciso, "Una Metodolog´ıa para el Modelado de Sistemas de Ingeniería Orientado a Agentes," *Inteligencia Artificial,* vol. 12, no. 38, pp. 39-60, 2008.

[24] J. Aguilar, M. Cerrada and H. F., "A methodology to specify multiagent systems," in *Agent and Multi-Agent Systems: Technologies and Applications*, Springer, 2007, pp. 02-101.

[25] D. A. Wiley, Instructional use of learning objects, Agency for instructional technology, 2001.

[26] G. Paquette, "A competency-based ontology for learning design repositories," *International Journal of Advanced Computer Science and Applications,* vol. 5, no. 1, pp. 55-62, 2014.

[27] J. Vizcarrondo, J. Aguilar, E. Exposito and A. Subias, "ARMISCOM: Autonomic Reflective MIddleware for management Service COMposition," in *Proceedings of the 4th Global Information Infrastructure and Networking Symposium*, 2012.

[28] V. Bravo, J. Aguilar, F. Rivas and M. Cerrada, "Diseño de un Medio de gestión de Servicios para Sistemas Multiagentes," in *Proceeding XXX Co*nferencia Latinoamericana de Informática, 2004.

[29] A. Nussbaumer, "A competence-based service for supporting self-regulated learning in virtual environments." Journal of Learning Analytics Vol, 2, no.1, pp. 101-133, 2015.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 356-361 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

An Overview of Traceability: Towards a general multi-domain model

Kamal Souali*, Othmane Rahmaoui, Mohammed Ouzzif

ESTC, RITM Laboratory, Hassan II University ENSEM, Casablanca, Morocco

Email: kml.souali@gmail.com, othmane.rahmaoui@gmail.com, ouzzif@gmail.com

ARTICLE INFO

Article history:
Received: 03 April, 2017
Accepted: 02 May, 2017
Online:17 May, 2017

Keywords:
Traceability
Trace
MDE
Model
Definition

ABSTRACT

Traceability for some people, is merely a tool to keep a history over something important that happened in the past. For others, is has no added value to their actual processes or products. In fact, it is becoming more and more valued. Traceability is still a vast area of research and an undiscovered field that if it is well used and managed, can provide a set of critical information or lead to something bigger. Many researches are still working to enhance its use and its integration by providing solutions to help users better manage and control their different elements (products, source code, documents, requirements, specifications, etc.). Nowadays, it is used in almost all domains as it can provide reliable information and helps improve efficiency and productivity. In this paper, we first present the state of the art on traceability and its use, through several examples. Then we provide a list of major techniques used in this field and propose our own traceability definition models.

1. Introduction

This paper is an extension of work originally presented in the 4th IEEE International Colloquium on Information Science and Technology [1] and was meant to show and explain the important role that traceability plays in different sectors.

Traceability, as defined in ISO (ISO 9001: 2000), is the ability to trace the history, application or location of that which is under consideration. D. Asioli, A. Boecker and M. Canavari say that it is not a new concept but a practice that we need to implement in order to comply with the standards and law rules [2]. Certainly, over the past few years, it has become a necessity in fields where the security or safety of consumers is questioned, especially in medical and food industries.

In software development also, this practice helps in the understanding, capturing, tracking and verification of software artifacts and their relationships and their dependencies during a software life-cycle [3]. As in [4], traceability was initially used to trace requirements from their source to implementation and test, when we talk about software development and now, it plays an increasing role in defect management, change management and project management.

According to the Global Traceability Standard (GS1), Traceability Systems have become an integral part of doing business as they aim to identify and locate unsafe foods and validate the presence or absence of attributes that are important to consumers [5]. Even if they are not yet considered as a catalyst for financial gains, G.G.D. Nishantha, M.K. Wanniarachchige and S.N. Jehan say that they are able to ensure consumer trust, safety, reliability, accuracy and quality [6]. Its importance is reflected through its ability to solve issues and through its power to provide strong proofs or evidences.

These Systems, along with their ability to monitor the composition as well as the position of every lot in a supply chain, are seen as a powerful tool that is capable of defining new management objectives and improve the overall performance [7].

Our personal definition of traceability would be, the ability to keep a detailed history of all activities and changes that a particular object can undergo throughout its entire life cycle, taking into account the different relationships that may appear. This particular object can be a material, a product, a model or even a class in a software development platform.

*Corresponding Author: Kamal Souali ESTC, RITM Laboratory, Hassan II University ENSEM, Casablanca, Morocco | Email: kml.souali@gmail.com

Traceability, if it is used in the right way, can provide a set of critical pieces of information such as the source, the destination, the location, the time, the link, in addition to the actors that were involved in the whole process. As in [8], ubiquitous traceability is achieved automatically, as a result of collecting, analysing, and processing every piece of evidence from which trace data can be inferred and managed.

The remainder of this paper is organized as follows: In Section 2, we will provide a set of definitions extracted from two major sectors, namely the food industry and information technology. Section 3 presents examples of traceability uses in different areas. A list of major techniques that were used to enhance the traceability will be the object of Section 4. In Section 5, we propose a set of definition models related to previously mentioned sectors. We will discuss the proposed traceability definition model in Section 6 and we will give a brief conclusion in Section 7.

2. Definitions

Traceability management is the planning, organization, and coordination of all activities related to traceability, including the creation, maintenance and use of trace links [9], not only in software development, but also in our daily life (e.g. memorizing events, tasks, activities, etc.).

In any area or sector, the definition of "traceability" is based on a number of criteria and limitations according to the used law or standards like the European General Food Law (EGFL) and the GS1 or simply, describes its purpose in a specific context.

Authors in [10] stated that there is no exact, single definition of traceability and that it has a large number of different meanings, which depend on the industry sector, on the supply chain, and on the perspectives of both the suppliers and the users of such information. However, we intend in section 5, to prove that a common definition can be established by means of models.

2.1. Food Industry

The EGFL defines traceability as the ability to trace and follow a food, feed, food-producing animal or substance through all stages of production and distribution.

According to A. F. Bollen and J.P. Emond, traceability is a well-coordinated and a well-documented movement of product and documented activities associated with the product, from producer, through a chain of intermediaries, to the final consumer [10].

M. Gooch and B. Sterling say that it is the ability to follow an item, or a group of items (whether animal, plant, food product, or ingredient) from one point in the value chain to another, either backwards or forwards [11]. Thus, food chain traceability goes from raw materials to consumption. This is almost the same definition given by F. Dabbene, P. Gay and C. Tortia, as they assume that products “moving” along the Food Supply Chain (FSC) are both tracked and traced [7].

Tracking is the process by which a product is followed from upstream to downstream in the Supply Chain. Tracing is the reverse process of tracking. The tracing process tends to reconstruct the history of a product through the information recorded in each step of the Supply Chain, identifying the source of a food or group of ingredients and consequently the real origin of a product [12]. These two primary functions of traceability are known as Trace-Back and Trace-Forward, as the movement can be traced one step backwards and one step forward at any point in the supply chain [6].

As in [13], traceability can either be internal or external. Internal traceability is within one company and relates to data about raw materials. While external traceability focuses on the product information from one link in the chain to the next (tracking a product batch and its history through the entire production chain).

In the food industry, traceability requires that each lot or amount or batch of food material is given a unique identifier which accompanies it and is recorded at all the stages of its progress through its food chain [14].

J. C.C. Martins and R. J. Machado said that a traceability system must record and follow the trail, since products that come from suppliers, are processed and distributed as end products [15]. The traceability presented by these records must contain a set of reliable pieces of information in order to ensure the minimum requirements. In fact, as stated in [11], it has three key essential information components: (1) identification of product attributes, (2) identification of premises and (3) identification of movement.

In the same context, P. Olsen and M. Borit have carried out an insightful comparative study of existing definitions [16]. By combining the best parts of these definitions, they concluded by saying that the simplest yet the most complete definition of traceability is the ability to access any or all information relating to that which is under consideration, throughout its entire life cycle, by means of recorded identifications.

2.2. Information Technology

In the field of software engineering, the IEEE Standard Glossary of Software Engineering Terminology defines traceability as the degree to which a relationship can be established between two or more products of the development process, especially products having a predecessor-successor or master-subordinate relationship to one another [17].

It is the ability to inter-relate any uniquely identifiable software engineering artefact to any other, to maintain the required links over time, and to use the resulting network to answer questions of both the software product and its development process [8]. It is a key element of any rigorous software development process that, provides critical support for many development activities [18].

When we talk about traceability in software development, we often refer to Requirement Traceability, which is an activity that allows creating links between and within software artefacts [19]. The definition of Requirement Traceability (RT), according to O. C. Z. Gotel and A. C. W. Finkelstein, is the ability to describe and follow the life of a requirement, in both a forward and backward direction [20]. Other definitions can be purpose-driven, solution-driven, information-driven or direction-driven.

These authors specify that there are two types of RT: pre-requirements specification traceability (Pre-RST), which is concerned with those aspects of a requirement's life prior to its inclusion in the Requirement Specification (RS), and post-requirements specification traceability (Post-RST), which is concerned with those aspects of a requirement's life that result from its inclusion in the RS.

More details about software traceability were listed in [8], including seven research areas and their associated directions which must be addressed in order to achieve ubiquitous traceability.

3. Uses of Traceability

In the food industry, it is considered as a mechanism used to keep the history of a raw or semi-finished unit during manufacturing and until this unit is delivered. It has a great potential to improve food safety as well as to promote consumer protection, by providing quality information [21].

In the field of Information Technology, traceability is used to list all activities of an entity on a system in execution. An example of such is the use of recovery logs or event logs in some cases. As R. Clayton explained, it is the ability to track down the originator of an action (seen as the flip side idea to "anonymity") and attempts to identify the IP address that caused an action to occur [22]. For instance, Law Enforcement Agencies (LEA) can use traceability to detect "Hi-Tech" crimes through data retention (causing logs to be preserved for a known period) and data preservation (ensuring that logs of special interest are not destroyed).

It is also used to clearly identify the sources behind some statistical analysis. Authors in [23] stated that it is the property which enables the understanding of where the analysis data come from and facilitates transparency. They have proposed a set of traceability pairs (relation criteria and factors) to define all the variables required in an analysis and hence establish the link between the final result and all the sources used. Moreover, traceability can strengthen the link between the requirements put in place, the specifications and the artefacts throughout the phases of a software development, using Requirements Traceability Matrix [19].

As authors in [9] explained in details, it allows creating and using links between software artefacts, which for example allows to connect the origin of a requirement with its specification or development artefacts to each other throughout the software lifecycle. These connections are called trace links, and link a source artefact to a target artefact. These artefacts can be of different types, such as a requirement, a model element, a line of code, or a test case.

In aerospace industry [24], traceability can be used to find the design related causes if a product does not function as expected. It is provided by establishing the relations between the design data and the requirements together with the relations between the components and the identifiers.

In Supply Chain Management (SCM), R.R. Pant, G. Prakash and J. A. Farooquie, traceability is defined in terms of what, how, where, why and when aspects of underlying product along a supply chain [25].

In logistics, traceability may be used to optimize routes and improve planning and management. It may also work with accounting applications to evaluate inventory or with controlling applications to identify process inefficiencies [15].

In electronics, traceability is used to keep track of all information related to changes and transformations which are applied to identified Printed Circuit Board (PCB) or other electronic components. Starting from the original batches and sources, this information is mainly, the Bill of Materials (BOM), the measurements, the list of operations in the process chain and the final destinations to whom or where the boards must be shipped. As stated in [26], it is required for fulfilment of safety standards such as ISO 26262.

Furthermore, traceability is also used in biology. An example of such, is to trace Genetically Modified (GM) animals that may similarly yield improvements in animal breeding, genetics and reproduction [27].

4. Techniques

In order to help users better manage their traceable items, the traceability mechanism has been enhanced by making use of different approaches that vary from using simple information retrieval techniques to the use of ontologies, graphs or even models.

4.1. Information Retrieval

A. De Lucia, A. Marcus, R. Oliveto and D. Poshyvanyk explained that Information Retrieval based methods or techniques like probabilistic, vector space and Latent Semantic Indexing models are used to recover traceability links on the basis of the similarity between the text contained in the software artefacts [28]. The higher the textual similarity between two artefacts is, the higher the likelihood that a link exists between them.

As in [29], this approach focuses on automating the generation of traceability links by similarity comparison between two types of artefacts.

4.2. Ontology

J. C.C. Martins and R. J. Machado proposed the use of software engineering methods and techniques to aggregate, disambiguate and blend existing knowledge [15]. They have used ontologies as a requirements modeling technique and developed specific traceability taxonomy in order to pursue the continuous improvement and answer the requirements of increased efficiency by tracking manufacturing activities information.

S. Bendriss and A. Benabdelhafid used DAML-S which is a generic ontology that can be applied in all areas [30]. They have adapted it by integrating and adding their specific ontology "Product Traceability Service", which describes all the web services of their traceability system. These services are dedicated to supply chain.

4.3. Graphs

As detailed in [19], "TraceMe" is an Eclipse module-based plug-in, that can be used to capture and maintain traceability links between different types of artefacts. According to the authors, this plug-in allows the software engineer to define different artefacts categories, capture traceability links between the defined artefacts categories and manage the traceability information through XML files. Traceability dependencies (trace links) are then displayed as graph.

Other researchers tend to use graph-based techniques in order to create trace links of test case scenarios and therefore, enhance the test coverage measurement and analysis [29].

Additionally, there are other plug-ins in the internet which are capable of tracing issues to both requirements and tests and creating the related traceability matrix.

4.4. Models

According to N. Sannier and B. Baudry, domain-specific modeling, which offers the capability to manipulate business domain concepts and traceability modeling, are Model-Driven Engineering (MDE) techniques that could address various aspects of requirement's formalization [31]. These authors proposed to combine both MDE and Information Retrieval (IR) techniques to improve requirements organization and traceability while handling textual ambiguous requirements documents.

MDE gives the basic principles for the use of models as primary artefacts throughout the software development phases and presents characteristics which simplify the engineering of software in various domains, such as Enterprise Computing Systems. A model is a symbolic system expressed in a language and each kind of model is represented by an appropriate modeling language and can be applied to certain purposes [3].

M. Thakur, B. J. Martens and C. R. Hurburgha defined a data model as a coherent representation of objects from a part of reality [32]. They used the modeling technique to create a database model capable of recording all the transformations related to incoming and outgoing grain lots, as well as the transformations that take place internally in the whole supply chain.

By making use of modeling techniques, C. Szabo and Y. Chen proposed "SeMMA" (Semantic Multi-Modeling Architecture), which is a multi-modeling architecture that permits the semantic integration of models defined in various languages, and ensures multi-model consistency when changes across different models occur and relies on three main modules, namely, the Change Analyzer Module, the Consistency Checker Module and the Warning Module [33].

In the same context, S. Bendriss and A. Benabdelhafid proposed a product data model which takes into account the different elements necessary for traceability, namely, the product in its various states, the various operations on the product, the occurred events, the resources used and the spatiotemporal location of the product [30].

In software development, requirement traceability can be described as a feature model to define a product [29]. It consists of a graph with features as nodes and feature relations as edges. If the number of features is very high, then the representation of features and their relations are displayed by tables.

On the other hand, the authors in [18] presented an approach on how to build a multi-domain traceability framework. It consists of defining first a Traceability Information Model (TIM) which represents the core element of any traceability framework (artefacts/relations) and may refer to artefacts (documents, models, databases, project activities context) from different domains, then deriving traceability information from sources, record the information in a Traceability Model (TM) and finally, performing traceability analyses, based on traceability goals.

Another example of such technique is presented in [26], where authors proposed an Eclipse plugin which uses the Eclipse Modelling Framework (EMF) as its base technology and stores the traceability model as an EMF model. This tool helps both users and project managers to create, customize and maintain traceability links, whose types depend on the company, development context and process used.

For more details about MDE techniques, Galvão and A. Goknil have listed many traceability approaches in MDE and evaluated them using five comparison criteria: representation, mapping, scalability, change impact analysis and tool support [3]. They classified these approaches into three categories: requirements-driven approaches, modeling approaches and transformation approaches.

4.5. Others

F. Furtado and A. Zisman proposed a new traceability approach called "Trace++", a traceability technique that extends traditional traceability relationships to support the transition from traditional to agile software development [34]. This technique extends the use of information sets and consists of six elements: the agile related problem, the trace relations, a set of all source artefacts, a set of all target artefacts, a set of additional information and finally, the type of relations.

Other techniques tend to use XML as the main tool to represent models and trace links.

As stated in [29], these techniques are classified as Hypertext-Based techniques. But there are others which can be either Rule based, Event based, Value-based or Scenario-based.

5. Definition Models

As stated before, there is no single or unique definition for traceability, since the term is described according to both its context and its purpose. Based on this and on the elements extracted from the other definitions, we hereby propose a definition model for the main sectors.

In Food Industry, the purpose of traceability is to trace the initial product with the raw materials from the start, till the very end of the production chain. Figure 1 represents the internal traceability in this field, where ***p*** is the initial item, ***P*** is the final product and $i_0, i_1 \ldots i_n$ are the set of information that describes the movement from one point to another.

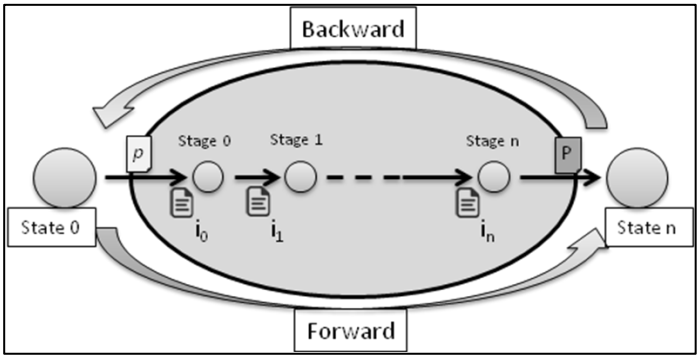


Figure 1: Traceability in Industry

By simply adjusting these elements, this representation can also describe the traceability in the Supply Chain or Logistics, tracing lots from the warehouse to every destination of the distribution chain.

Furthermore, it is only when two or more separate representations of this size are connected, that we basically speak about external traceability. Otherwise, it is still internal.

In the field of Information Technology, one example use of traceability is to create links between customer requirements or

specifications and the supplier software. As shown in Figure 2, ***s*** refers to the initial specification (requirement) or source, while ***O*** refers to the final object. The relation between the different stages of development is represented by $r_0, r_1 \dots r_n$.

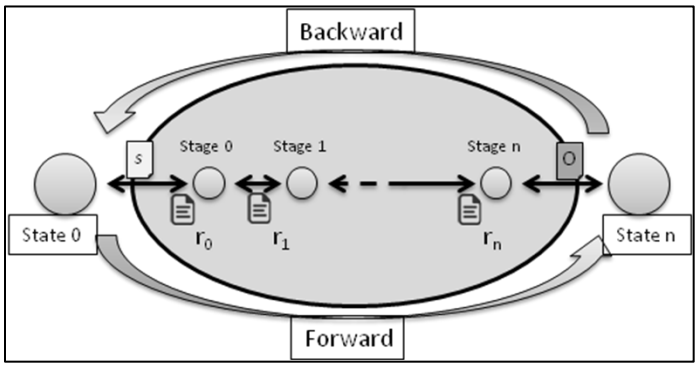


Figure 2: Traceability in Software Development

When this definition is applied to the model driven engineering (Figure 3), relations are replaced by a set of transformations $t_0, t_1 \dots t_n$ between predecessors and successors, representing the same system **S**. Elements ***s*** and ***O*** will be replaced respectively by ***m*** for the initial model and ***M*** the final model. Not to mention that each transformation can be represented likewise.

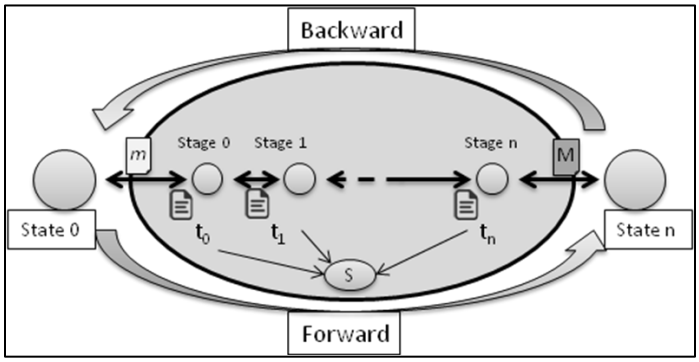


Figure 3: Traceability in Model Transformation

To sum up, we can say that the three proposals have a set of elements in common:

- Items: units that need to be traced and followed.
- Stages: positions where the units are processed.
- Relations: links between predecessors and successors.
- Activities: set of processes that were applied to the units

These elements will lead us to set a common model for traceability, which will be the basis of our future studies.

6. Discussion

Traceability in our point of view, as stated before, is the ability to keep a detailed history of all activities and changes that a particular object can undergo throughout its entire life cycle, taking into account the different relationships that may appear. It can be internal or external and can be used in two different ways either forward or backward.

As presented in the previous section, every traceable item is moving from an initial state to a final state, through numerous stages. In each stage, the output is the result of an activity that takes into account inputs from the previous stage and keeps the link to the origin. At the end, and since inputs and outputs are interrelated, tracing forward and backward is possible.

Thus, we can combine these facts to establish a common definition model that can be used to define "Traceability" everywhere (Figure 4).

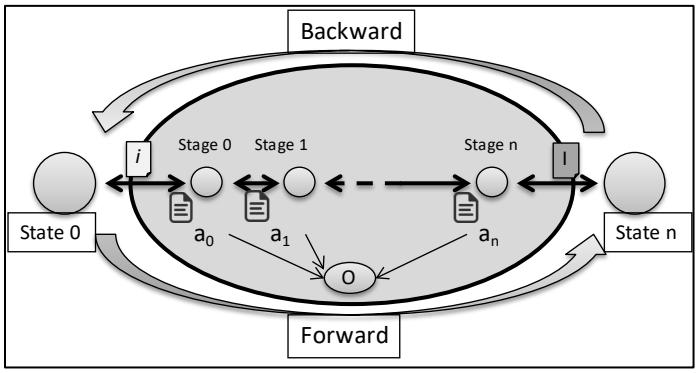


Figure 4: Generic Traceability Model

Here, ***i*** refers to the initial traceable item, where ***I*** is the final traceable item. The activities that the traceable item undergoes are represented by $a_0, a_1 \dots a_n$. ***O*** is the origin or the representation of all original characteristics of the traceable item. These characteristics do not change and are only updated if a new property was discovered when moving from one stage to another.

Certainly, once this generic model is deployed on a particular platform, it will be a subjected to a large amount of data, of different types. Hence, is it mandatory to consider the following challenges:

- How to address the problem of time vs Big Data during information access?
- How can we manage to order the accessed traceability information by degrees of priority or importance?

We intend on enhancing our model by adding a set of rules and other traceability related properties.

For instance, a "weight" or a "priority" measure can be introduced and assigned to each traceability information, after the classification process, or we can improve the representation of trace links by including additional factors. Thus, only the most important set of information is shown when tracing an item, either backward or forward.

7. Conclusion

Traceability can ensure quality, safety, reliability and accuracy. Furthermore, it can help companies improve productivity, reduce costs and gain consumer's trust.

According to GS1 [5], traceability may assess other business systems and tools such as quality management, risk management, information management, logistical flows, commercial advantage and evaluation of management demands.

In this paper, we have listed recent definitions related to traceability from two major sectors. We have presented definition models for these sectors and proposed our generic traceability

model, by combining a set of common elements, which stands as the basis of our research. In the same context, we have listed also, the uses and purposes of traceability as well as the major techniques applied in this research field.

In future work, we intend to refine our model with new elements and then, deploy it and use it in E-learning environments. Furthermore, a study will be initiated to decide whether or not this model can be applied to other fields or needs further enhancements. The final purpose will be to implement a general model, able to satisfy all traceability needs and requirements.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] K. Souali, O. Rahmaoui, M. Ouzzif, "An overview of traceability : definitions and techniques", 4th IEEE International Colloquium on Information Science and Technology (CiSt'16), 2016.

[2] D. Asioli, A. Boecker, M. Canavari, "Perceived traceability costs and benefits in the italian fisheries supply chain", International Journal on Food System Dynamics 2(4), pp 357-375, 2011.

[3] I. Galvão, A. Goknil, "Survey of traceability approaches in Model-Driven Engineering", 11th IEEE International on Enterprise Distributed Object Computing Conference (EDOC), 2007.

[4] G. Regan, F. McCaffery, K. Mc Daid, D. Flood, "The barriers to traceability and their potential solutions: towards a reference framework", 38th Euromicro Conference on Software Engineering and Advanced Applications, 2012.

[5] GS1 Global Traceability Standard Business Process and System Requirements for Full Chain Traceability, http://www.gs1.org/traceability/gts, 2007.

[6] G.G.D. Nishantha, M.K. Wanniarachchige, S.N. Jehan, "A pragmatic approach to traceability in food supply chains", 10th International Conference on Advanced Communication Technology (ICACT 10), 2010.

[7] F. Dabbene, P. Gay, C. Tortia, "Traceability issues in food supply chain management: A review", Biosystems Engineering vol. 120, pp 65–80, 2014.

[8] J. Cleland-Huang, O. Gotel, J H Hayes, P Mäder, A Zisman, "Software traceability: trends and future directions", The 36th International Conference on Software Engineering, 2014.

[9] R. Wohlrab, J. P. Stegh¨ofer, E. Knauss, S. Maro, A. Anjorin, "Collaborative Traceability Management: Challenges and Opportunities", The IEEE 24th International Requirements Engineering Conference, 2016.

[10] A. F. Bollen and J.P. Emond, "Traceability in postharvest systems", Postharvest Handling, pp.485-504, 2014.

[11] M. Gooch, B. Sterling , "Competitive advantage of food traceability to value chain management" , http://vcm-international.com/, 2013.

[12] T.Pizzuti, G. Mirabelli, M. A. Sanz-Bobi, F. Goméz-Gonzaléz, "Food track & trace ontology for helping the food traceability control", Journal of Food Engineering vol. 120, pp 17–30, 2014.

[13] J. T. Mgonja, P. Luning, J. G.A.J. Van der Vorst, "Diagnostic model for assessing traceability system performance in fish processing plants", Journal of Food Engineering vol. 118, pp 188-197, 2013.

[14] B. Welt, J R. Blanchfield, International Union of Food Science And Technology, IUFoST Scientific Information Bulletin (SIB), Food Traceability, 2012.

[15] J. C.C. Martins, R. J. Machado, "Ontologies for product and process traceability at manufacturing organizations: a software requirements approach", Eighth International Conference on the Quality of Information and Communications Technology, 2012.

[16] P. Olsen, M. Borit, "How to define traceability", Trends in food science & technology vol. 29, pp 142-150, 2013.

[17] IEEE Standards Software Engineering, IEEE Standard Glossary of Software Engineering Terminology, IEEE Std. 610-1990.

[18] M. Taromirad, N. Matragkas, R. F. Paige, "Towards a multi-domain model-driven traceability approach", The 7th Workshop on Multi-Paradigm Modeling, 2013.

[19] G. Bavota, L. Colangelo, A. De Lucia, S. Fusco, R. Oliveto, A. Panichella, "TraceME: Traceability Management in Eclipse", 28th IEEE International Conference on Software Maintenance (ICSM), 2012.

[20] O. C. Z. Gotel, A. C. W. Finkelstein, "An analysis of the requirements traceability problem", Proceedings of the First International Conference on Requirements Engineering, 1994.

[21] M. Garcia Martinez, F. M. Brofman Epelbaum, "The role of traceability in restoring consumer trust in food chains", Food Chain Integrity, 2011.

[22] R. Clayton, "The limits of traceability", from http://www.cl.cam.ac.uk/~rnc1/The_Limits_of_Traceability.pdf, 2000.

[23] S. Zhu, L. Yan, "Methods of building traceability for ADaM data", Proceedings of the PharmaSUG2011 Conference - Paper CD05, 2011.

[24] E. Tekin, "A method for traceability and as-built product structure in aerospace industry", Proceedings of the 47th CIRP Conference on Manufacturing Systems, 2014.

[25] R.R. Pant, G. Prakash, J. A. Farooquie, "A framework for traceability and transparency in the dairy supply chain networks", XVIII Annual International Conference of the Society of Operations Management (SOM-14), Procedia - Social and Behavioral Sciences vol. 189, pp 385 - 394, 2015.

[26] S. Maro, J. P. Stegh¨ofer, "Capra: A Configurable and Extendable Traceability Management Tool", IEEE 24th International Requirements Engineering Conference, 2016.

[27] A. Lievens, M. Petrillo, M. Querci, A. Patak, "Genetically modified animals: options and issues for traceability and enforcement", Trends in Food Science & Technology vol. 44, pp 159-176, 2015.

[28] A. De Lucia, A. Marcus, R. Oliveto, D. Poshyvanyk, "Information retrieval methods for automated traceability recovery", Software and Systems Traceability, pp 71-98, 2011.

[29] M. Shahid, S. Ibrahim, H. Selamat, "Test coverage measurement and analysis on the basis of software traceability approaches", International Journal of Information and Electronics Engineering, Vol. 1, No. 2, pp 115-119, 2011.

[30] S. Bendriss, A. Benabdelhafid, "Enabling product traceability through data modeling and semantic web service ontologies", International Conference on Advanced Logistics and Transport (ICALT), 2013.

[31] N. Sannier, B. Baudry, "Toward multilevel textual requirements traceability using model-driven engineering and information retrieval", Model-Driven Requirements Engineering Workshop (MoDRE), IEEE, 2012.

[32] M. Thakur, B. J. Martens, C. R. Hurburgha, "Data modeling to facilitate internal traceability at a grain elevator", Computers and Electronics in Agriculture vol. 75, pp 327-336, 2011.

[33] C. Szabo, Y. Chen, "A model-driven approach for ensuring change traceability and multi-model consistency", The 22nd Australian Conference on Software Engineering, 2013.

[34] F. Furtado, A. Sizman, "Trace++: A Traceability Approach to Support Transitioning to Agile Software Engineering", The IEEE 24th International Requirements Engineering Conference, 2016.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 362-368 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

L-Band SiGe HBT Active Differential Equalizers with Variable, Positive or Negative Gain Slopes Using Dual-Resonant RLC Circuits

Yasushi Itoh* and Hiroaki Takagi

Shonan Institute of Technology , Electrical and Information Engineering, 251-8511, Japan

ARTICLE INFO

Article history:
Received: 01 April, 2017
Accepted: 02 May, 2017
Online: 17 May, 2017

Keywords:
Microwaves
Active Equalizer
Differential Amplifier

ABSTRACT

L-band SiGe HBT active differential equalizers with variable, positive or negative gain slopes have been designed and fabricated for frequency and temperature compensation of microwave and optical systems. The active equalizer employs dual-resonant RLC circuits in the series feedback path of the differential amplifier for positive gain slopes or in the load for negative gain slopes. The implemented active equalizers have achieved positive gain slopes of +54 to +87dB/GHz across 0.2 to 0.6GHz as well as negative gain slopes of -50 to -100dB/GHz over 0.6 to 1.2GHz. The active differential equalizers presented in this paper have an outstanding feature of providing variable, positive or negative gain slopes, which can be easily adjusted to meet with various stringent requirements for frequency and temperature compensation in microwave and optical systems.

1. Introduction

First of all, this paper is an extension of the work originally presented in the 46th European Microwave Conference, 2016 [1]. The research and development of active equalizers have started at the beginning of 1970s as active RC filters with IC operational amplifiers in the audio band because of having many advantages over passive equalizers in terms of inductor-less, cheaper, smaller, easily manufactured and tuned circuits [2]. Since then, various types of active, adjustable amplitude and delay equalizers have been developed, including classical Bode variable equalizers, transfer-function types and switched-capacitor equalizers [3, 4, 5]. With the tremendous growth in semiconductor NMOS, CMOS devices, the R&D activities of active equalizers have moved toward wideband, high speed data transmission systems, a variety of analog adaptive equalizer architectures have been developed to eliminate signal distortions [6, 7, 8, 9]. Meanwhile, wideband microwave and optical receivers have been long tackled with the pass-band negative gain slope due to the falling off of gain of transistors and the insertion loss of passive elements as frequency increases [7, 10]. To address this issue, several types of passive equalizers have been reported, having fixed positive or negative gain slope as a function of frequency [11]. In addition, to realize the adjustability for the required gain slopes, a variety of active equalizers using diodes or transistors have been reported, having variable positive or negative gain slopes [10, 12, 13]. Although these circuits employ active devices, the circuit configuration is basically a passive type where fixed capacitances are replaced by diodes or transistors to provide variable gain slopes. Therefore, the insertion loss is seriously large. To address this problem, the authors have presented various types of the active equalizer with variable, positive or negative gain slopes. The advanced feature is in that active devices are used as an amplifier in place of the conventional switching or variable-capacitance device to achieve high gain. Moreover, to achieve higher frequency operation as well as miniaturized size, LC-resonators in place of the traditional RC filters are employed in the series feedback path or load circuit of the differential amplifier for positive or negative gain slopes. The active differential equalizers are as follows: Capacitance-selectable bridged-T attenuators were incorporated into the series feedback path to achieve variable positive gain slopes, which can be digitally controlled [14]. Series LC resonators were employed in the base or collector bias circuits to realize variable, positive or negative gain slopes [15]. Single-band series or parallel LC resonators were used in the feedback path of the differential amplifier for the same purposes [16]. These active equalizers, however, can vary gain slopes but the frequency positions were fixed. To overcome this problem, a novel active differential equalizer having variable inclination and position of the positive or negative gain slopes have been presented in [1]. It employs dual-resonant RLC circuits in the feedback path of the differential amplifier for positive gain slopes or in the load for negative gain

*Corresponding Author: Yasushi Itoh, 1-1-25 Tsujido-Nishikaigan, Fujisawa, Kanagawa, 251-8511 Japan, +81-466-30-0185 | Email: itoh@elec.shonan-it.ac.jp

slopes. To extend the work, the theory and design for the active differential equalizer using the dual-resonant RLC circuits are described in more detail. Then the circuit design, simulation and performance are newly added for the case that dual capacitors (actually two varactor diodes) consisting of the dual-resonant RLC circuit are independently varied. These active differential equalizers are relatively narrowband as well as provide abrupt gain slopes but can be easily extended to wideband design with the proper choice of circuit parameters.

2. Active Differential Equalizer Using Dual-Resonant RLC Circuits

Two types of the active differential equalizers with variable inclination and position of the positive or negative gain slopes are graphically shown in Figure 1. f_{PP} and f_{PN} are peak frequencies. f_{BP} and f_{BN} are bandstop frequencies. The conventional active equalizers [14, 15, 16] can vary either a peak or bandstop frequency. Thus only gain slope can be varied. On the other hand, the novel active equalizers can vary both peak and bandstop frequencies. Thus both the inclination and position of gain slopes can be varied.

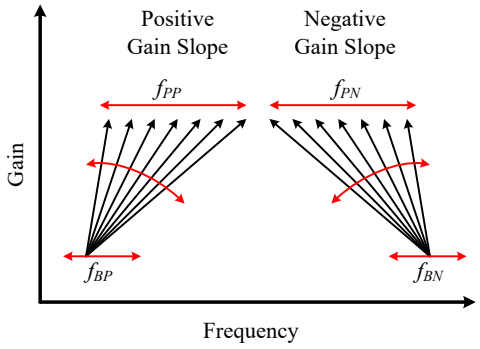
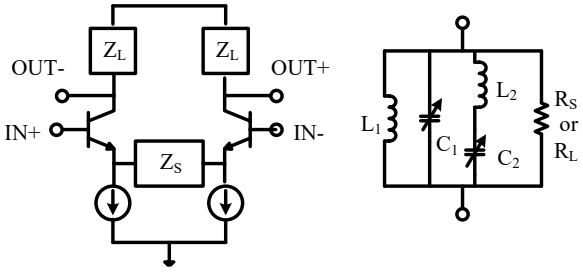


Figure 1 Two types of the active differential equalizers with variable inclination and position of the positive or negative gain slopes



(a) Differential amplifier (b) Dual-resonant RLC circuit

Figure 2 Schematic diagram of the differential amplifier and dual-resonant RLC circuit

Schematic diagrams of the differential amplifier and dual-resonant RLC circuit are shown in Figure 2. The dual-resonant circuit consists of series LC (L_2, C_2) and parallel LC (L_1, C_1) circuits combined with a resistor (R_S or R_L). The dual-resonant RLC circuit becomes purely resistive at the frequency f_1 and f_2, which are given as follows:

$$f_1 = \frac{1}{2\pi}\sqrt{\frac{b-\sqrt{b^2-4a}}{2a}} \tag{1}$$

$$f_2 = \frac{1}{2\pi}\sqrt{\frac{b+\sqrt{b^2-4a}}{2a}} \tag{2}$$

$$a = L_1L_2C_1C_2 \tag{3}$$

$$b = L_1C_1 + L_2C_2 + L_1C_2 \tag{4}$$

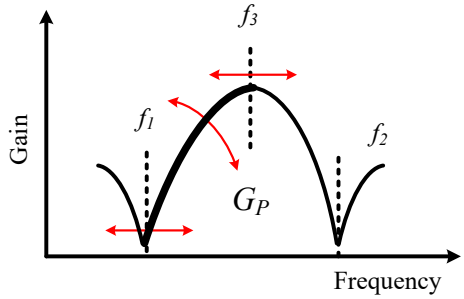
The circuit also becomes short-circuited at the frequency f_3.

$$f_3 = \frac{1}{2\pi\sqrt{L_2C_2}} \tag{5}$$

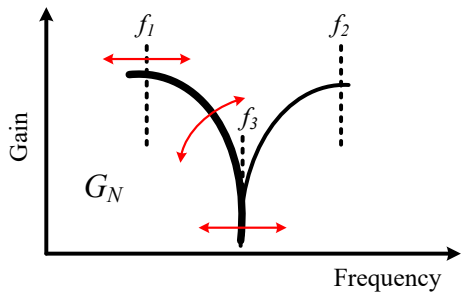
The voltage gain of the differential amplifier (G_V) in Figure 2(a) can be approximated as the following equation, where Z_L and Z_S are an impedance of the load and feedback circuits.

$$G_V = \frac{Z_L}{Z_S} \tag{6}$$

When the dual-resonant RLC circuit is applied to Z_S, G_V becomes minimum at f_1 and f_2 as well as maximum at f_3, which is graphically shown in Figure 3(a). Whereas, when the dual-resonant RLC circuit is applied to Z_L, G_V becomes maximum at f_1 and f_2 as well as minimum at f_3, which is graphically shown in Figure 3(b). The differential amplifier provides variable, positive gain slopes (G_P) between f_1 and f_3 by adjusting f_1 or f_3. Actually, f_1 or f_3 can be varied with C_1 or C_2. It must be noted here that both f_1 and f_3 can be varied with C_1. Meanwhile, the differential amplifier produces variable, negative gain slopes (G_N) between f_1 and f_3 by adjusting f_1 or f_3. In the similar way, f_1 or f_3 can be varied with C_1 or C_2. However, it must be also noted that both f_1 and f_3 can be varied with C_1.

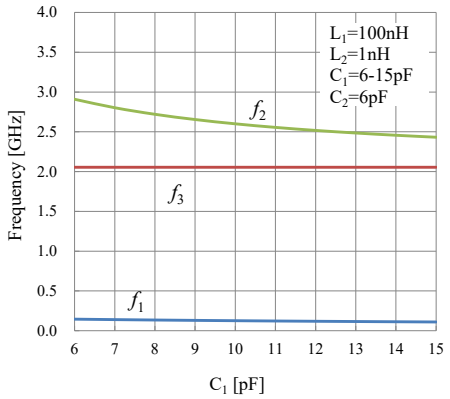


(a) Positive Gain Slope

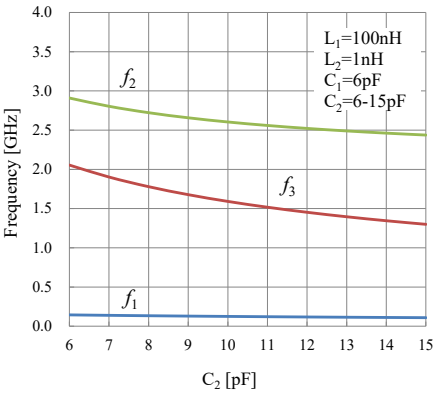


(b) Negative Gain Slope

Figure 3 Voltage gains of the differential amplifier when the dual-resonant RLC circuit is applied to Z_S or Z_L.



(a) Variable C_1 and fixed C_2



(b) Variable C_2 and fixed C_1

Figure 4 Calculated resonant frequencies f_1, f_2 and f_3

Table 1 Circuit element values

Element	Value
R_L [Ω]	56
R_S [Ω]	10
L_1 [nH]	100
C_1 [pF]	6 to 15
L_2 [nH]	1
C_2 [pF]	6 to 15

The resonant frequencies f_1, f_2 and f_3 are calculated by using (1) to (5) and are shown in Figure 4. The circuit element values are listed in Table 1. It is clearly shown that the inclination and position of positive or negative gain slopes can be tuned by varying C_1, C_2 or both. As described above, f_1 and f_3 cannot be varied independently with C_2 changes in Figure 4(b).

3. Circuit Design

Schematic diagrams of the active differential equalizer with variable, positive or negative gain slopes are displayed in Figures 5 and 6, respectively. Dual-resonant RLC circuits are employed in the feedback path of the differential amplifier for positive gain slopes or in the load for negative gain slopes.

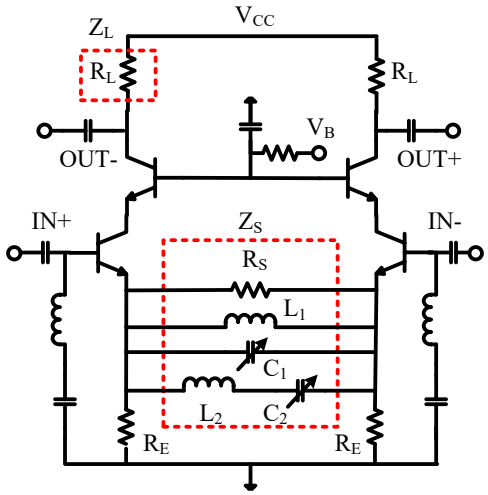


Figure 5 Schematic diagram of the active differential equalizer with variable, positive gain slopes

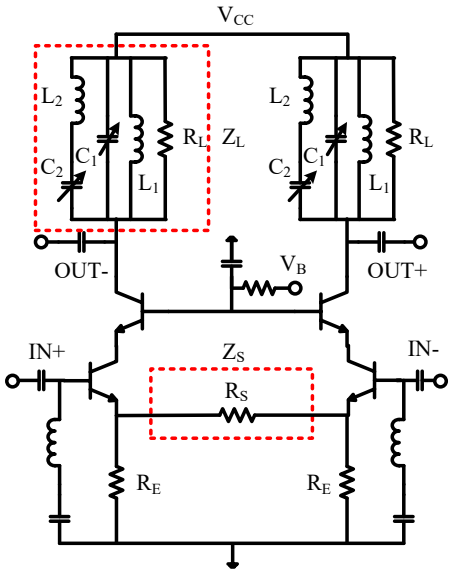


Figure 6 Schematic diagram of the active differential equalizer with variable, negative gain slopes

As shown in Figure 2(b), the dual-resonant RLC circuit consists of a series LC circuit (L_2, C_2), a parallel LC circuit (L_1, C_1) and a resistor (R_S or R_L). R_S or R_L plays an important role of circuit stability and impedance matching. R_L is basically chosen as 50 ohms for impedance matching. But R_S has to be carefully determined to achieve both high stability and high gain. When the dual-resonant RLC circuit is employed in Z_S, R_L is used in Z_L. Meanwhile, the dual-resonant RLC circuit is employed in Z_L, R_S is utilized in Z_S.

Now the voltage gain with positive gain slopes (G_P) and that with negative gain slopes (G_N) can be given from (6) as follows:

$$G_P = R_L\sqrt{\mathrm{Re}[Y_S]^2 + \mathrm{Im}[Y_S]^2} \tag{7}$$

$$\mathrm{Re}[Y_S] = \frac{1}{R_S} \tag{8}$$

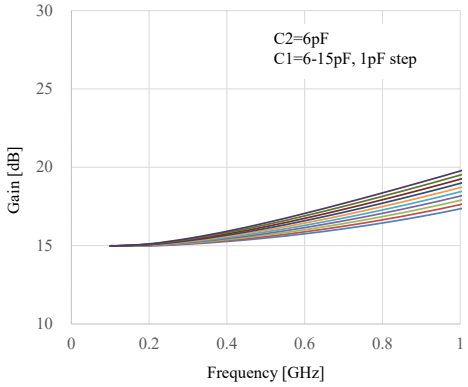
$$\mathrm{Im}[Y_S] = \mathrm{Im}[Y_L] \tag{9}$$

$$G_N = \frac{1}{R_S\sqrt{\mathrm{Re}[Y_L]^2 + \mathrm{Im}[Y_L]^2}} \tag{10}$$

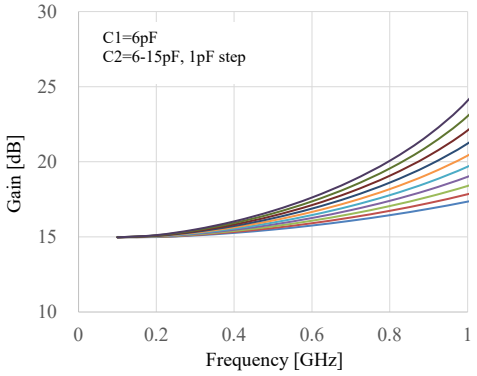
$$\mathrm{Re}[Y_L] = \frac{1}{R_L} \tag{11}$$

$$\mathrm{Im}[Y_L] = \frac{\omega^4 L_1 L_2 C_1 C_2 - \omega^2 (L_1 C_1 + L_2 C_2 + L_1 C_2) + 1}{\omega L_1 (1 - \omega^2 L_2 C_2)} \tag{12}$$

G_P and G_N are calculated by using (7) to (12) and the circuit element values in Tables 1. The calculation was done for variable C_1 and fixed C_2 or variable C_2 and fixed C_1. The calculated G_P and G_N are plotted in Figures 7 and 8, respectively. It is clearly demonstrated that variable, positive or negative gain slopes can be obtained. Due to a large value of L_1, a variation of the bandstop frequency f_1 is small in Figure 7. Thus the positive gain slopes greatly change with f_2 (C_2) in Figure 7(b). In the similar way, due to a large value of L_1, a variation of the bandstop frequency f_1 is also small in Figure 8. The negative gain slopes largely vary with f_2 (C_2) in Figure 8(b).

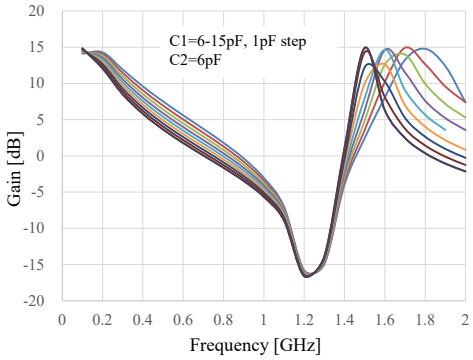


(a) Variable C_1 and fixed C_2

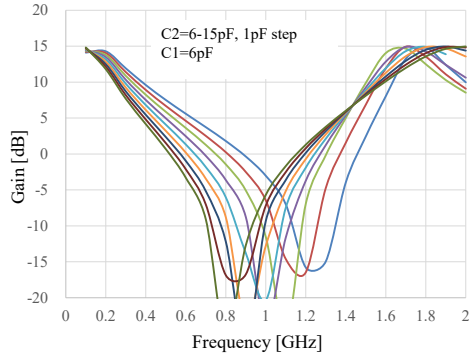


(b) Variable C_2 and fixed C_1

Figure 7 Calculated G_P for variable C_1 and fixed C_2 or variable C_2 and fixed C_1

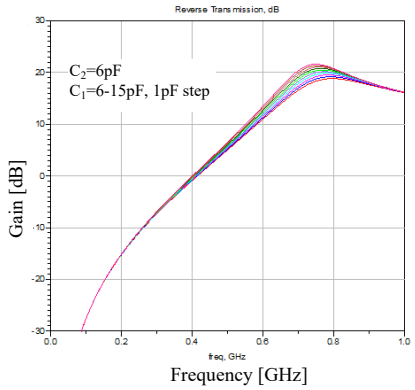


(a) Variable C_1 and fixed C_2

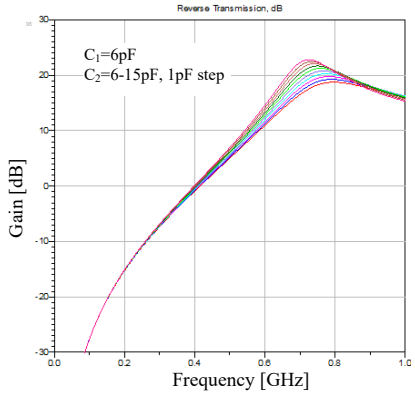


(b) Variable C_2 and fixed C_1

Figure 8 Calculated G_N for variable C_1 and fixed C_2 or variable C_2 and fixed C_1

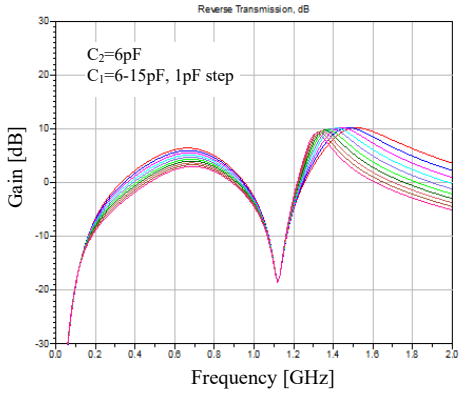


(a) Variable C_1 and fixed C_2

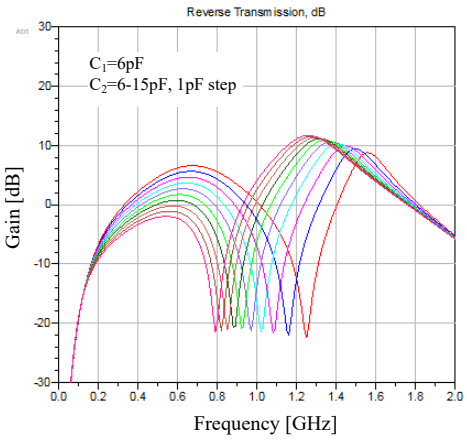


(b) Variable C_2 and fixed C_1

Figure 9 Simulated gains of the active differential equalizer with variable, positive gain slopes



(a) Variable C_1 and fixed C_2



(b) Variable C_2 and fixed C_1

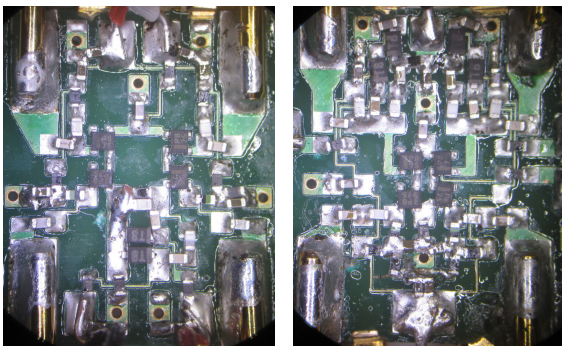
Figure 10 Simulated gains of the active differential equalizer with variable, negative gain slopes

4. Circuit Simulation

Based on the schematic diagrams of Figures 5 and 6, the circuit simulation was accomplished by using ADS. A cascode connection of SiGe HBTs is employed to control gain. In order to improve input and output matching, reactive elements of L and C are employed in the input matching circuit as well as the value of R_L was determined as 56 ohms. The value of R_S was finally chosen as 10 ohms to keep high gain and stabilize the circuit. The simulation was done for variable C_1 and fixed C_2 or variable C_2 and fixed C_1. The simulated gains of the active differential equalizer with variable, positive or negative gain slopes are plotted in Figures 9 and 10, respectively. The variation of positive gain slopes is a little bit larger for f_3 (C_2) changes, which is in good agreement with Figures 4 and 7. In the similar way, the variation of negative gain slopes is also larger for f_3 (C_2) changes. In Figure 10(b), both the inclination and position of negative gain slopes have drastically changed.

5. Circuit Fabrication

Photographs of the active differential equalizer with variable, positive or negative gain slopes are shown in Figure 11. The active differential equalizers were fabricated on the FR-4 substrate with a dielectric constant of 4.5. 0.35μm SiGe HBTs with an f_t of 25GHz (Toshiba MT4S102T), 1005-type resistors, inductors and capacitors are mounted on the substrate by soldering. A surface mount type of the varactor diode with a capacitance ratio of 2.5 (Toshiba 1SV279) was used. The circuit size is 14 x 16 x 1.2 mm^3.



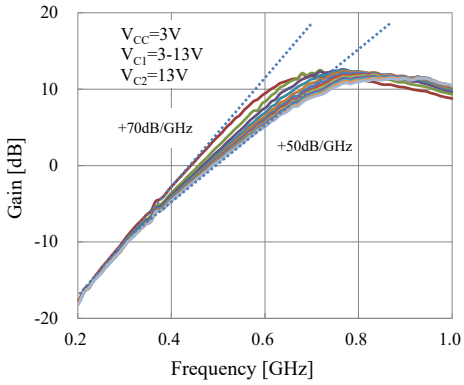
(a) Positive Slope (b) Negative Slope

14 x 16 x 1.2 mm^3

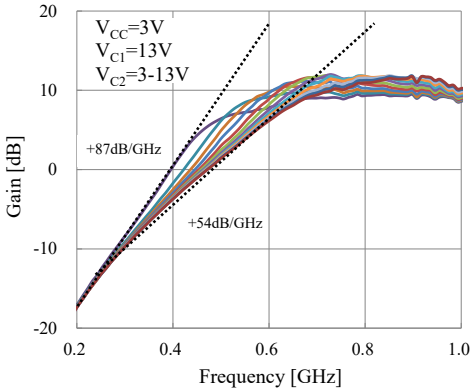
Figure 11 Photographs of the active differential equalizer with variable, positive or negative gain slopes

6. Circuit Performance

The measured gains of the active differential equalizers are plotted in Figure 12 for the positive gain slopes and in Figure 13 for the negative gain slopes. The measurement was done for variable C_1 and fixed C_2 or variable C_2 and fixed C_1. Actually C_1 or C_2 was varied with a control voltage V_{C1} or V_{C2} of the varactor diode. The measurement conditions are shown in Figures 12 and 13, respectively.

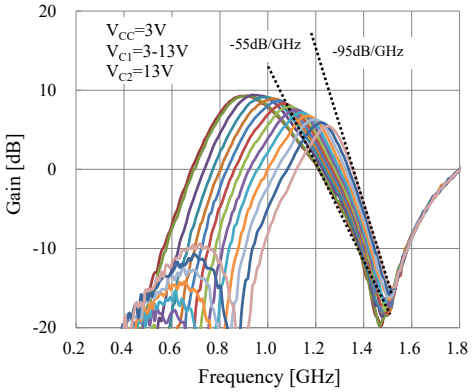


(a) Variable V_{C1} and fixed V_{C2}

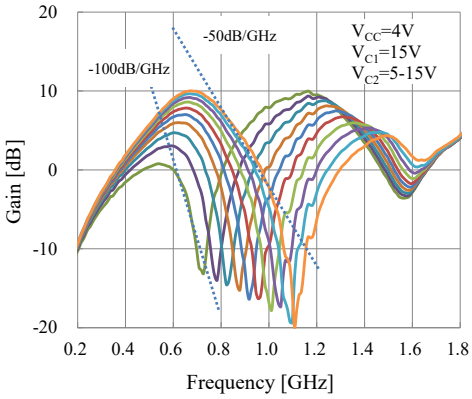


(b) Variable V_{C2} and fixed V_{C1}

Figure 12 Measured gains of the active differential equalizer with variable, positive gain slopes



(a) Variable V_{C1} and fixed V_{C2}



(b) Variable V_{C2} and fixed V_{C1}

Figure 13 Measured gains of the active differential equalizer with variable, negative gain slopes

The active differential equalizer with variable, positive gain slopes has achieved a gain slope of +50 to +70dB/GHz over 0.2 to 0.6GHz in Figure 12(a) as well as +54 to +87dB/GHz across 0.2 to 0.6GHz in Figure 12(b). The input and output return losses were better than 2dB and 15dB, respectively, which can be easily improved by employing a lossy match circuit in the input matching circuit.

The active differential equalizer with variable, negative gain slopes has achieved a gain slope of -55 to -95dB/GHz over 1.1 to 1.5GHz in Figure 13(a) as well as -50 to -100dB/GHz across 0.6 to 1.2GHz in Figure 13(b). The input and output return losses were better than 2dB. In a similar way as the positive gain slope, the input and output matches can be further improved with the use of a lossy match configuration.

These performances were compared with the conventional active equalizers and summarized in Table 2. It is clearly shown that the active differential equalizers presented in this paper can provide narrow-band but abrupt gain slopes as well as higher gain. With the proper choice of the circuit parameters, it is very easy to meet with the moderate gain slopes over a wide bandwidth.

Table 2 Summary of performances of active equalizers

Ref.	Active Device	Role	Equalizer Type	Circuit Topology	Frequency Band	Gain Slope	Gain
[10]	GaAs MESFET	Swiching	Negative Slope	Bridged-T Attenuator	DC - 18GHz	-0.67dB/GHz	<-2.7dB
[11]	unkown	Filter	Positive, Negative Slope	Non-reflective Transmission Filter	2 - 4GHz	plusmin 30dB/GHz	unkown
[12]	PIN Diode	Variable R	Positive, Negative Slope	Tee Attenuator	3 - 5GHz	plusmin 16dB/GHz	<-2dB
This work	0.35 micron SiGe HBT	Amplifier	Positive, Negative Slope	Differential Amplifier	0.2 - 0.6GHz, 1.1 - 1.5GHz	+54 to +87dB/GHz, -55 to -95dB/GHz	< +10dB

7. Conclusions

Two types of the active differential equalizers with variable, positive or negative gain slopes have been presented. With the use of the dual-resonant RLC circuit in the feedback or load circuit, the active differential equalizers have achieved positive gain slopes of +54 to +87dB/GHz across 0.2 to 0.6GHz as well as negative gain slopes of -50 to -100dB/GHz over 0.6 to 1.2GHz. The active differential equalizers presented in this paper show narrow bandwidth but abrupt gain slopes, which can be easily extended to wide bandwidth and moderate gain slopes with the proper choice of the circuit parameters. Moreover, the input and output matches can be improved with the use of a lossy match structure. It can be concluded from these results that the active differential equalizers would be one candidate for frequency and temperature compensations in microwave and optical systems.

References

[1] Y. Itoh and H. Takagi, "L-Band SiGe HBT Active Differential Equalizers with Variable Inclination and Position of the Positive and Negative Gain Slopes", Proceeding of the 46th European Microwave Conference, pp. 1015-1018, October 2016

[2] Paul E. Fleischer, "Active Adjustable Loss and Delay Equalizers", IEEE Trans. Circuits and Systems, vol.CAS-21, No. 4, pp. 527-531, July 1974.

[3] Franc Brglez, "Minimally Active RC Variable Equalizers", IEEE Trans. Circuits and Systems, vol.CAS-22, No. 8, pp. 688-691, August 1975.

[4] Robert R. Cordell, "A New Family of Active Variable Equalizers", IEEE Trans. Circuits and Systems, vol.CAS-29, No. 5, pp. 316-322, May 1982.

[5] Antonio Petraglia and Sanjit K. Mitra, "Switched Capacitor Equalizers with Digitally Programmable Tuning Characteristics", IEEE Trans. Circuits and Systems, vol.CAS-38, No. 11, pp. 1322-1331, November 1991.

[6] Tadayoshi Enomoto, Kasaaki Yasumoto, Tsutomu Ishihara and Kohjiro Watanabe, "Monolithic Analog Adaptive Equalizer Integrated Circuit for Wide-Band Digital Communication Networks", IEEE Journal of Solid-State Circuits, vol.SC-17, No. 6, pp. 1045-1054, December 1982.

[7] K. W. Goossen, J. A. Walker, D. T. Neilson, J. E. Ford and W. H. Knox, "Micromechanical Gain Slope Compensator for Spectrally Linear Optical Power Equalization", IEEE Photonics Technology Letters, Vol. 12, No. 7, pp. 831-833, 2000.

[8] H. Kim, F. Bien, S. Chandramouli, J. Cha, E. Gebara and J. Laskar, "A 0.25-um BiCMOS Feed Forward Equalizer Using Active Delay Line for Backplane Communication", IEEE Symposium on Circuits and Systems, pp. 193-196, May 2007.

[9] Melikyan Vazgen Sh.,, Sahakyan Arthur S., Safaryan Karo H., Dingchyan Hayk H.,, "High Accuracy Equalization Method for Receiver Active Equalizer", Proceeding of the East-West Design & Test Symposium, 2013

[10] H. J. Sun and B. C. Morley, "A DC-18GHz GaAs MESFET Monolithic Variable Slope Gain-Equalizer IC", IEEE 1989 Microwave and MMW Monolithic Circuit Symposium Digest, pp. 79-82.

[11] H. Wang, B. Yan, Z. Wang and R. Xu, "A Broadband Microwave Gain Equalizer", Progress in Electromagnetic Research Letters, Vol. 33, pp. 63-72, 2012.

[12] S. C. Bera, "Amplitude Tilt Active Equalizer for Frequency and Temperature Compensation", IEEE Microwave and Wireless Component Letters, Vol. 21, No. 7, pp. 344-346, 2011.

[13] M Morgan, T. Newton, T. B. Hayward, and T. Boyd, "Non-Reflective Transmission-Line Filters for Gain Slope Equalization", 2007 IEEE/MTT-S International Microwave Symposium Digest, pp. 545 – 548.

[14] Y. Itoh, N. Mizuo and A. Ohta," An L-Band 4-bit SiGe HBT Active Differential Equalizer Using Capacitance-Selectable Bridged-T Attenuators", Proceeding of the ISMOT 2015, pp.164 -166, June 2015.

[15] Y. Itoh and H. Takagi, "L-Band SiGe HBT Active Differential Equalizers with Varactor-Tuned Bias Circuits for Variable Positive or Negative Gain Slopes", Contemporary Sciences, Vol. 9, No. 15, pp. 727-733, 2016.

[16] Y. Itoh and H. Takagi, "L-Band SiGe HBT Active Differential Equalizers Providing Variable Positive or Negative Gain Slopes", Proceeding of the 21th International Conference on Microwaves, Radar and Wireless Communications, May 2016.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 369-375 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Moving Towards Reliability-Centred Management of Energy, Power and Transportation Assets

Kang Seng Seow, Loc K. Nguyen*, Kelvin Tan, Kees-Jan Van Oeveren

Clean Technology Centre, DNV GL - Energy, 118227, Singapore

A R T I C L E I N F O

Article history:
Received: 12 April, 2017
Accepted: 02 May, 2017
Online: 17 May, 2017

Keywords:
Physical Asset Management
Risk-based Asset Management
Reliability Centered Maintenance

A B S T R A C T

Infrastructure Asset Management was first adopted in 1980s for challenges faced by the utilities, road and rail transport systems, mining, process and manufacturing industries in asset life cycle value optimization. International standards ISO 55000 and PAS 55 were subsequently developed to provide a basis for a structured methodology in the development of Asset Management System. This methodology with the basic framework is considered inadequate in addressing the changing and challenging environment faced by the Energy, Power and Transportation sectors. This paper introduces the concept of Reliability-Centred Management System that could leverage on the advancement in Information and Communication Technologies and the entry of the 'Big Data' and 'Smarter Grid' era.

1. Introduction

Incidents in the late 1980's, such as the North Sea Piper Alpha disaster that caused 167 fatalities, and the ensuing oil price crash and the falling levels of service, escalating costs and poor planning in the public sector of Australia and New Zealand has necessitate a radical change in managing the assets, such as the oil platforms and the public infrastructures. The lessons learnt from these widely different environments have been remarkable similar, with significant convergence of conclusions. This has led to the approach of asset life cycle value optimization for the utilities, road and rail transport systems, mining, process and manufacturing industries at an accelerating rate. [1]

On the other hand, international standards, such as PAS 55 and ISO 55000 developed in the last decade provide a basis for a structured methodology in the development of Asset Management System. However, this methodology with the basic framework is not adequate in addressing the changing and challenging environment faced by the power utilities. How utilities manage their assets is undergoing profound changes. Asset Management approaches and analytical tools have finally matured to the point where decisions can be based on data and facts – not just "gut feel" instincts and intuition.

2. Current Asset Management Concepts

Facing with the changing and challenging environment brought about in the global economic and local political pressures, the Asset Management mind-set for the energy, power and transportation sectors needs to focus on:

- Prudent performance targets, not maximum performance;
- Risk management not risk avoidance; and
- Spending justified by performance improvement and risk mitigation not blanket budgets.

The quality of assets managed by the energy, power and transportation sectors cannot therefore be assured by simple compliance with the general ISO 9001 management system standard [2]. This is because ISO 9001 provides little guidance on what constitutes good Asset Management practice with regard to capital-intensive infrastructures with relatively long asset lives.

The crux of Asset Management in these sectors is to have an integrated, whole-life and risk based approach. The issues faced by the energy, power and transportation sectors are many and varied. These sectors are required by their stakeholders to procure, operate, maintain and subsequently replace their assets at acceptable risk levels with satisfactory quality of service and reasonable return of capital investments. These assets comprise complex assets which typically have operational lives of several decades. At the same time, the sectors have to avoid unnecessary investments planning while ensuring the reliability, security and quality of services provided by these assets.

Cost-effective operation and maintenance of the assets are particularly important in the management of the assets throughout

*Corresponding Author: Loc K. Nguyen, Clean Technology Centre, DNV GL - Energy, 118227, Singapore | Email: khanh.loc.nguyen@dnvgl.com

the useful life of these assets. Figure 1 shows the possible strategies that can be adopted in the maintenance of the assets.

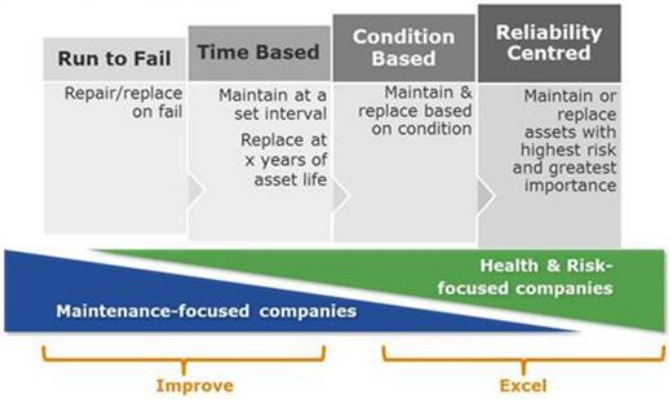


Fig. 1. Maintenance Strategy Evolution

3. Technical Asset Management System

Figure 2 shows a typical step-wise implementation approach that has been developed for Asset Management Systems of energy, power and transportation sectors under ISO 55000. [3,4]

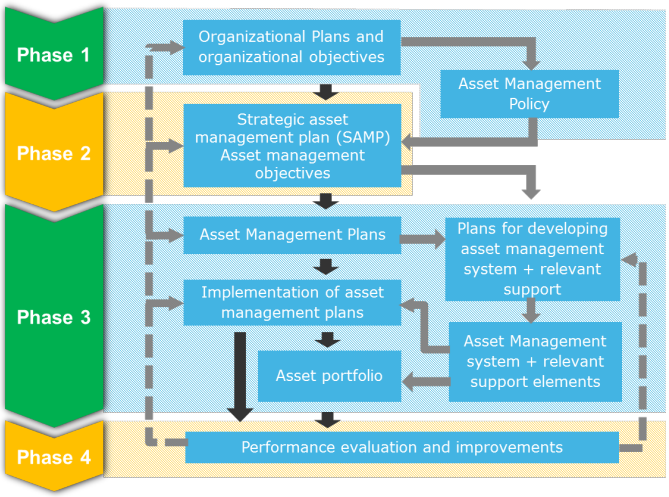


Fig. 2. Typical Four-Phase Approach for Asset Management System Implementation

Phase 1 is to have an in-depth understanding of the organizational objectives and goals for meeting both the internal and external requirements. Demand growth, failure behaviour and other major trends and drivers impacting the power utility with regard to reliability, financial, safety, social, political, legal, and regulatory perspective are the most common external considerations.

Phase 2 is the scoping of the Asset Management System. In this phase, the power utility needs to develop Asset Management objectives and a Strategic Asset Management Plan (SAMP) which are in alignment with organizational objectives. The SAMP is developed in consideration of expectations and needs from all stakeholders, and includes business values and Key Performance Indicators (KPIs); decision making processes and criteria; boundaries of the Asset Management System with scope; and plans and programs to achieve the Asset Management System objectives.

Phase 3 is the resource planning and implementation phase, which is dedicated to the actual implementation of the Asset Management System including the required resources. This involves the development of processes and procedures, development of the right technical and IT tools and systems, and competence building, and awareness & communication on a regular basis.

The technical scope of Asset Management needs to be aligned with the asset life cycle. In particular, different Asset Management practices should address the following four different stages of life span of the assets: - specification and design, commissioning, maintenance and operation, and retiring.

According to Kristian Steenstrup, VP and Gartner Fellow, Energy and Utilities, "There are many strategies to improve reliability. Rather than thinking of them as conflicting or contradictory, develop a road map that matches the best tool to the characteristics of specific assets."

Technical Asset Management is therefore focusing on complex assets, where maintenance and analytics are used to support the reliable operation of the asset and to extend its useful life.

Technical Asset Management methodologies in each stage of the life span of the asset are shown in the Figure 3. The methodologies have to be selected to manage the asset in each stage properly in order to reduce the risk, achieving the real advantages and benefits of the Asset Management System.

Fig. 3. Technical Asset Management Methodologies in Different Stages of Asset Life

Phase 4 is the evaluation and improvement phase as the Asset Management System should be continuously evaluated for possible improvements and corrections to be in alignment with inherent objectives. In this respect, systematic measurement of performance with quantitative and qualitative indices can contribute to the Asset Management System improvement with SMART targets for a specific process or the system.

However, technical and financial constraints of the energy, power and transportation sectors usually hinder Asset Management System realization as planned. In order to overcome such constraints and to plan the Asset Management System in a tailor-made manner for short/long-term goals, a technical gap analysis can be carried out to understand the gap between current practices with technical competence and required scenario's, prioritizing required technologies, tools or competences in an appropriate order.

This technical gap analysis not only enables the energy, power and transportation sectors in assessing their current practice, but can also be used to list and prioritize the technologies and methodologies to which the energy, power and transportation sectors have to pay attention to within their objectives and resource boundary in order to plan for the resourcing of required tools on a short or long-term basis. Asset Management supporting Information Technology (IT) tools therefore play an important role to support the technical performance of the Asset Management System in order to maintain, control, and analyse the information and its trend over the time, understanding of assets and their risks throughout their lifecycle. [5]

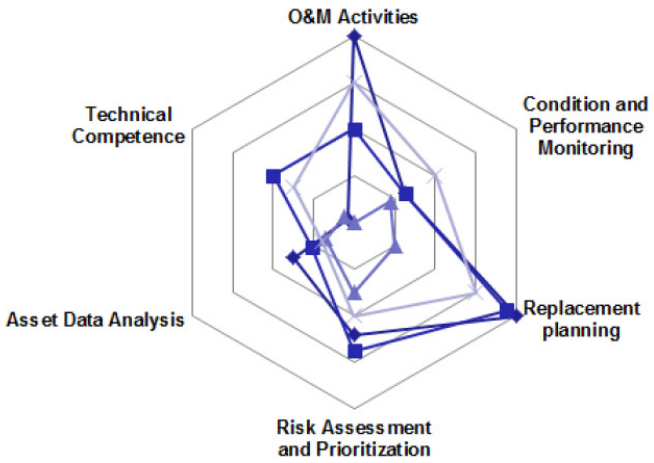


Fig. 4. Example of a Technical Gap analysis

Since sound decisions can be highly dependent on having a high quality of information, support by efficient IT tools for data acquisition and analysis is therefore critical. In this sense, some considerations of IT tools are

1. *Scalability*: IT tools should accommodate all assets the power utility likes to include for monitoring and maintenance;

2. *Lifecycle Coverage*: IT tools should cover all available information and its trend over the lifecycle of the asset;

3. *Compatibility*: When energy, power and transportation sectors decide to adapt new condition assessment or monitoring software solution, IT tools should communicate core condition information on assets to one another; and

4. *Flexibility*: IT tools should support any specific methods and information as a template in which the energy, power and transportation sectors like to add or remove at any time.

In recent years, more and more energy, power and transportation sectors are considering to utilize on-line monitoring devices to extract meaningful dynamic condition parameters such as Partial Discharge, Tangent Delta, or Dissolved Gas Analysis for their power assets and other parameters such as vibration, temperature, corrosion, etc. for other critical assets.

4. Asset Management Decision Support Modeling Using Health Indexing

With the rapid growth of demand and the maturing electricity sectors in Asia, energy, power and transportation sectors are facing many challenges in managing their power assets, such as increasing demands from stakeholders with regard to safety, reliability, environmental impact and financial results, aging assets with increased risks of failure, and, as a result, a decreased level of reliability and (possibly) safety, etc. In planning maintenance and replacement of the vast number of growing assets, asset managers of these energy, power and transportation sectors need to identify which assets require attention, and what is the best action to take at what time. The choice becomes all the more complex as each asset type has different failure modes, and each failure may have different ways to influence the network. An ability to estimate the expected time to failure of assets will help the energy, power and transportation sectors make the optimal maintenance and replacement programs.

We will describe a Health Index model, which has been adopted by several energy, power and transportation sectors and enables their asset managers to oversee the health of all installed assets, recognizing the required additional maintenance or replacement need per asset, and representing the results as a function of time. [5, 6,7, 8]

The model combines Health Indexing and Risk Analysis, thereby providing full decision support on intervention actions to manage the power asset. The Health Index is a single indicator that represents the condition of an asset in relation to its specified performance and lifetime. The Health Index Tool is a dashboard for the asset manager and provides condensed information on the status of all relevant assets.

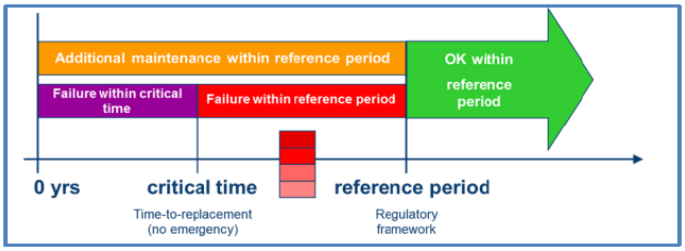


Fig. 5. Health Index Classification Scheme

The Health Indexing Methodology developed by DNV GL assesses the Health Index as accurately as the available data allow, and presents the results in a clear way, enabling the user to interpret the results in a blink of an eye. A simple colour scheme has been selected to represent asset health categories, each colour directly representing the required replacement or required additional maintenance need and urgency. A reference period is defined by the user, as an analysis horizon. Further to this, a critical time period is used to distinguish critical replacements from regular replacements. The reference period may be tuned to the regulator review period, or to any other relevant timeframe, and is typically 10 to 15 years. The critical time is set to the time required for a "standard" replacement plan, typically 1 to 3 years. These timeframes are defined as health classes as indicated in Figure 5. These classes are characterized by the colours green, orange, red and purple, which represent "good condition", "additional maintenance required", "replacement within reference period" and "immediate replacement" respectively.

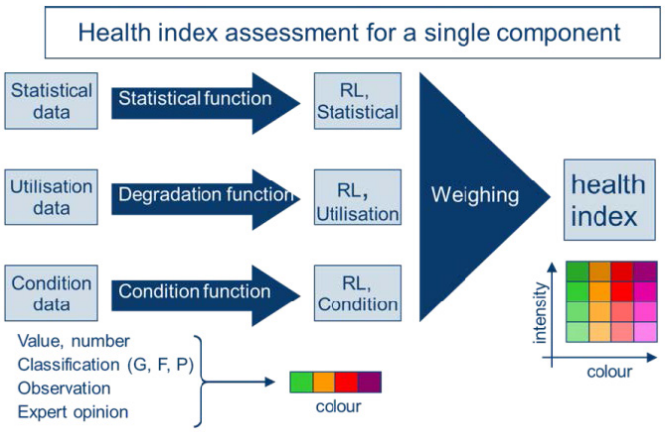


Fig. 6. Health Index Concept Using Three Types of Assessment Functions

The core of the Health Index Model is formed by the assessment functions that estimate the remaining life of the asset from available data. This estimation is based on whatever information is available, be it asset type, failure data, age, utilization data, maintenance and condition data. The most direct way of assessing the remaining life is direct condition data. However, often sufficient data are not available. Under such a situation, alternatives are built in, based on statistical information or operational (utilisation) information. In other words, if the remaining life cannot be assessed directly, the model produces an estimation based on statistical expectations, or on expected degradation due to stress.

As a result, three different types of assessment functions are used as shown graphically in Figure 6, i.e.:

- the statistical remaining life function;

- the degradation (utilisation) remaining life function; and

- the condition remaining life function.

Each of the above assessment functions estimates the most likely remaining life of the asset under study.

Energy, power and transportation sectors may also use a risk-based Asset Management System to support decision making in managing the power assets. To run a risk-based Asset Management model, there is a need to estimate the risks the assets are facing. Ideally, the risk of all assets shall be estimated individually, enabling a prioritization of all assets requiring attention.

In order to apply the Health Index model to risk-based decision support, the model is extended to include a Risk Index. The Health Index, estimating the remaining life of assets based on failure probability, is combined with the assets criticality or failure impact, thereby opening the option to plot all assets in a risk matrix, tuned to the specific power utility's core business values and circumstances.

To match best practice Asset Management processes, this risk matrix can be defined for the typical most relevant business values and KPIs such as reliability (frequency and duration of service disruptions), safety (number of accidents / injured persons) and financial (revenue lost/money spent). For example, the reliability impact of failure may be estimated by the load and number of customers connected to the individual assets and the system topography, like redundancy applications. The safety impact may be analysed using the probability of fire and explosion per asset type, in combination with the asset's housing, type of substation and distance to the public areas.

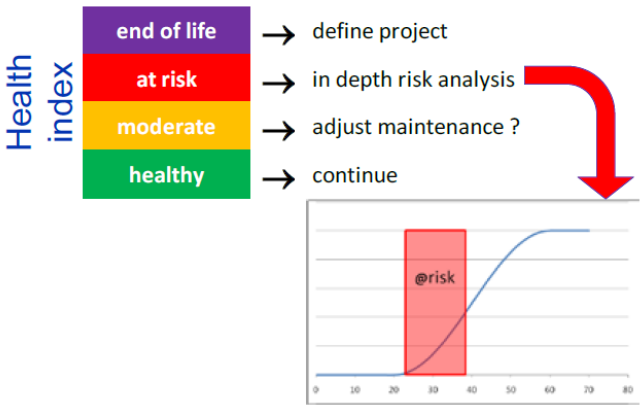


Fig. 7. Focused In-Depth Risk Analysis

With the risk matrix in place, and the health and criticality indices available, it is possible to plot all assets into the risk matrix and identify the assets that require urgent and critical attention. Assets that require attention may then be prioritized in terms of both on safety and reliability.

Once the assets with the highest risks are identified, individual in-depth analysis can be performed on selected assets, if required. For assets in the purple category, replacement is required as soon as possible. In case of a high level of uncertainty, the asset may require further detailed analysis before a replacement decision is made; whereas, assets with a low level of uncertainty may be replaced with minor or without risk analysis.

For assets in the red category, a longer timeframe is available for more detailed studies and condition assessment. This enables focused in depth analysis if needed to better substantiate a decision. Figure 7 gives a graphical representation of this principle.

The Health and Risk Indexing approach is part of the Asset Management decision cycle, an example of which is shown in Figure 8.

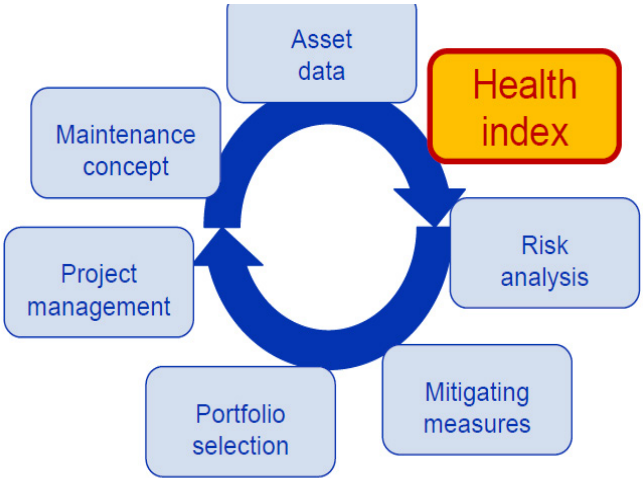


Fig. 8. Health Indexing as part of the Asset Management Cycle

Health and Risk Indexing processes data to produce information to decide what measures are required with what urgency and priority. The way in which it is embedded in the Asset Management process differs from user to user, but often it consists

of a two-step approach. The first step, as indicated in Figure 7, divides the assets in one of four categories based on the remaining life and the additional activities required to survive the reference time period. The assets that score excellent or very poor need no detailed further analysis as they will be either continued, or replaced on short notice. For the other assets, before deciding on the basis of lifetime (equivalent to failure probability), the impact of failure is assessed, yielding a risk ranking.

Starting with the highest risk, solutions can then be generated. Each solution is again ranked on the basis of Solution Added Value. Finally, the asset manager will decide on a prioritization in the portfolio of solutions. The highest risk does not necessarily lead to the highest priority as this also depends on the cost and risk reduction potential of the solution.

After prioritization of the solutions, the asset manager will be able to select and plan the actions. For this he will use the ranking obtained. However, the decision will of course be influenced by the budget available and by the (un)acceptability of the risks assessed. Some actions are mandatory for meeting the power utility's directives and policies, and others may be highly relevant but not mandatory for the short term.

This will ultimately develop asset management strategy and policy leading towards reliability-centred maintenance for critical assets.

5. Asset Management System Based on Reliability–Centered Maintenance

Reliability-Centered Maintenance (RCM) is a structured process to determine the maintenance strategies. Various maintenance strategies have been evolved over time, i.e. Reactive maintenance (RM), Preventive maintenance (PM), Predictive maintenance (PdM). RM and PM approaches have been implemented in the earlier stage of maintenance strategies. However, these strategies were recognized as ineffective methodologies in 1960s because they do not capture the asset condition and the impacts and risks of the failures. Soon after, RCM has been first introduced by Nowlan in 1978 and applied to the Department of Defense of the US. Thanks to its effectiveness, RCM has been applied widely to other industries.

An example of RCM hierarchy is shown below:

Table 1: An example of RCM Approach

Reactive Element Applications	Preventive Element Applications	Predictive Element Applications
Small parts and equipment	Equipment subject to wear	Equipment with random failure patterns
Non-critical equipment	Consumable equipment	Critical equipment
Equipment unlikely to fail	Equipment with known failure patterns	Equipment not subject to wear
Redundant systems	Manufacture recommendations	Systems which failure may be included by incorrect preventive maintenance

In terms of methodologies, the three maintenance strategies, i.e. RM, PM and PdM, are broken down into four maintenance methodologies, i.e. Run-to-Fail maintenance, Time-based maintenance, Condition-based maintenance, and Risk-based maintenance. Figure 9 below shows the relationship between the maintenance strategies and methodologies.

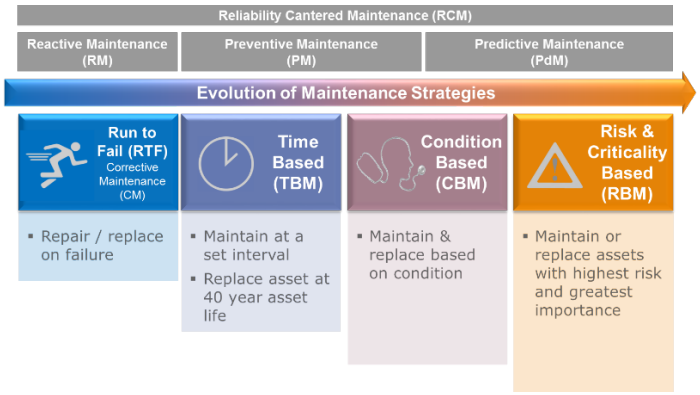


Fig. 9. Consistent and Technically Sound Process towards a Reliability-Centred Asset Management System

The condition-based approach:

The condition based approach bases on the understanding of asset's condition i.e. health index of the asset, which indicate how good or bad the condition of asset is. It also indicates the remaining life time of the asset, the likelihood of the asset's failure. The health index system enables asset managers to oversee the health of all installed assets, recognize the required additional maintenance or replacements per location and plotted in time. An example of asset health index result is shown in Figure 10 below.

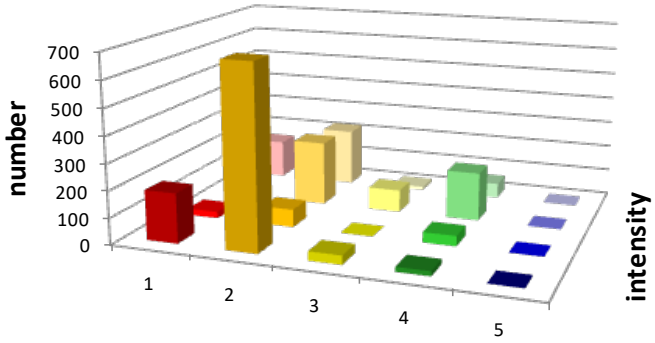


Fig. 10. Example of health index results of a fleet of assets

A simple color scheme presents the results indicating the required replacement or required additional maintenance attention. A reference period is agreed with the asset managers and usually a critical planning time period is defined to distinguish critical replacements from regular replacements. Based on these timeframes we can now define health classes as indicated in **Error! Reference source not found.**. The classes are characterized by the colors blue, green, yellow, orange, and red.

Risk-based approach:

Once the Health Index has been developed and determined for an asset, that asset can also be graded according to its Criticality of failing. Its Criticality can be determined with reference to all relevant asset management business values, like safety, reliability etc. The level of Criticality will be determined by the asset type, location, local grid situation etc. Once the Criticality and the Health Index are ready, an asset management risk matrix can be developed with on one axis the Criticality and on the other axis the Health Index. The Health Index gives an indication of remaining

life, but it can also be seen as an indicator of probability of failure. Criticality indicates the impact of the failure. Combining both will show the risk level of each asset in the risk matrix. An example of a risk matrix is shown in Figure 11.

Remaining Life	Catastrophic	Severe	Moderate	Small	Negligible
≤ 1 year	Unacceptable	Unacceptable	High	High	Medium
2-3 years	Unacceptable	High	High	Medium	Low
3-7 years	High	High	Medium	Low	Low
7-15 years	High	Medium	Low	Low	Negligible
> 15 years	Medium	Low	Low	Negligible	Negligible

Fig. 11. Example of a risk matrix specifically for the health index

As a general observation, the following trend has started to emerge [9, 10, 11, 12]:

- Moving from time-based to condition-based maintenance and further from condition-based to risk-based maintenance for most critical assets - Using Fukushima Nuclear Power Station as an example, an incident in a nuclear power station can have dire consequences to the environment, human life, properties and the livelihood of the population in vicinity. The reliability-centred design and maintenance based on multiple KPIs will be the most appropriate choice in maintaining such an asset;
- Moving from risk-based Asset Management to Reliability-Centred Asset-Management, whereby
  - Near-real-time information is captured and analysed; and
  - Real-time failure prediction function plays a key role in decision making support.
- Moving from Big Data to Smart Data
  - With a more comprehensive data considered in the Smarter-Asset-Management, the Big Data will be the next challenge in Asset Management;
  - Smart Data is the next level of dealing with Big Data in a smarter and more effective way; and
  - More advanced data analytics techniques will be adopted such as machine learning/artificial intelligence approach.
  - Integrated Energy Asset Management System - merging Asset Management System (AMS) with Energy Management System (EMS): - The merger between the EMS and AMS will help address the Big Data more effectively.

Energy, power and transportation sectors are moving in the data analytics era. This is the result of several global forces - one being the proliferation of less expensive electronic monitoring technologies and the speed and availability of communications systems. Furthermore, more and more energy, power and transportation sectors want to embark on developing the 'Smartest' asset systems possible. As a result, an unprecedented amount of raw data is being collected by energy, power and transportation sectors each day.

On the other hand, criticality and health of each of the assets are influenced by the asset operational environment. A systematic approach to assess the criticality and health will deliver a practical approach to risk assessment across the entire asset base.

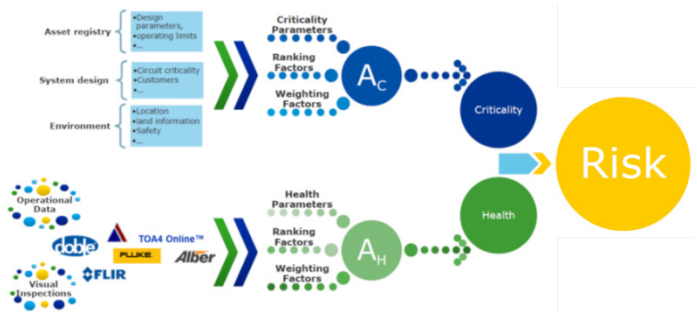


Fig. 12. Consistent and Technically Sound Process towards a Reliability-Centred Asset Management System

Using the ranking factors and weighting factors, we are able to tailor the assessment to specific business priorities and objectives in achieving reliability-centred Asset Management strategy and policies.

As criticality and health are two essential factors in assessing asset risks, the trend mentioned in the start of this section enables the integration of Asset Management Systems with other IT systems, such as Geospatial Systems, distribution management systems, energy management systems, signalling and control systems, etc. will help these systems in:

- improving the accuracy in identifying critical contingencies by making use of more accurate component's criticality and utilization information;
- improving the performance by filtering out low failure probability;
- better big-data management by using a single database;
- improving the effectiveness by avoiding redundancy of assessments and processes; and
- optimising decision making processes based on a more comprehensive assessment considering both global and local aspects

6. Conclusion

The energy, power and transportation sectors need to move from time-based to Predictive Maintenance / Condition Based Maintenance of their asset and ultimately towards Reliability-Centred Asset Management strategy and policies. This will involve:

- Apply Reliability-Centred Maintenance philosophy in selection of effective maintenance tasks;
- Apply Reliability-Centred Maintenance philosophy to determine what operational and conditional data is required;
- Reliability-centred maintenance provides a platform for developing an effective Reliability-Centred Asset Management strategy; and
- Lastly, a platform with appropriate software applications to ensure a sustainable Reliability-Centred Asset Management strategy.

To achieve this state of managing the assets, particularly the critical asset, cost-effectively with minimum risks, energy, power and transportation sectors need to embark on a journey to develop resources and capability within the organisation to manage such changes. External helps from experts with in-depth knowledge, vast experience and up-to-date competency will definitely be necessary to leverage on the evolving trends in the ICT infrastructures.

References

[1] Institute of Asset Management: Asset Management – an anatomy (December 2011 version 1)

[2] "ISO 9001:2015 Quality management systems" standard

[3] "ISO 55001:2014 Asset Management" standard. BSI Group.

[4] Sungin Cho, Khanh-Loc Nguyen and Jos Wetzer, "Conceptual Design for Asset Management System under the Framework of ISO 55000", proceedings CIRED 23rd International Conference on Electricity Distribution Lyon, 15-18 June 2015, Paper 1025

[5] Mischa Vermeer, Ashutosh Sharma, Jos Wetzer, Peter Zonneveld and Dirk Boender, "Asset Management Decision Support Modelling, using a Health Index, for Maintenance and Replacement Planning", proceedings The 9th GCC – CIGRE International Conference, 2013

[6] Mischa Vermeer and Jos Wetzer, "Transformer health and risk indexing, including data quality management", Proceedings 45th CIGRÉ Session, 2014, Paper A2-116

[7] Fred Steennis and et al, "The Intriguing Behavior over Time of PD's from Defects in MV Cables and Accessories; Lessons Learned with SCG, an On-Line Monitoring System", proceedings CIRED 21st International Conference on Electricity Distribution, 2011

[8] Mischa Vermeer and et al, "Asset Management Decision Support Modeling, using a Health Index, for Maintenance and Replacement Planning", PowerTech, 2015 IEEE Eindhoven, 2015

[9] Khanh-Loc Nguyen, Liang Hongzhu Daniel, Kelvin Tan, "Energy Asset Management System (EAMS) – Preparation for future power system with high penetration of DGs and RE", ACEF Manila Philippines, 2015 June. Presentation online: http://d335hnnegk3szv.cloudfront.net/wp-content/uploads/sites/837/2015/ 06/KEMCO-DDWPresentation_Session-II-Khanh-Loc-Nguyen.pdf

[10] Khanh-Loc Nguyen, Liang Hongzhu Daniel, Kelvin Tan, "Energy Asset Management System – A Hybrid Solution for Future Distribution System with High Penetration of DGs and RE", WES (World Engineers Summit), Track 3, 2015 July, Singapore. Online link: http://www.wes-ies.org/summit/tracks/

[11] Khanh-Loc Nguyen, Liang Hongzhu Daniel, Kelvin Tan, "Energy Asset Management System for offshore rigs", SIEW ACES Singapore, 2015 October

[12] Khanh-Loc Nguyen, Daniel Liang, Terence Goh, Kelvin Tan, "A smart asset management method for offshore rigs", EASEC-14 Conference, Vietnam, 2016 January. Online link: http://easec14.hcmut.edu.vn/wp-content/uploads/2015/06/S4-Summary-and-list-of-papers.pdf

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 2, No. 3, 376-383 (2017)
www.astesj.com
Special issue on Recent Advances in Engineering Systems

ASTES Journal
ISSN: 2415-6698

Secure Path Selection under Random Fading

Furqan Jameel*, Faisal, M Asif Ali Haider, Amir Aziz Butt

Department of Electrical Engineering, COMSATS Institute of Information Technology, Islamabad, Pakistan

ARTICLEINFO

Article history:
Received: 30 March, 2017
Accepted: 02 May, 2017
Online: 22 may, 2017

Keywords :
Wireless Sensor Networks (WSNs)
Secrecy Capacity
Nakagami-m Fading
Optimal Scheme
Dissimilar fading
Rician-K Factor

ABSTRACT

Application-oriented Wireless Sensor Networks (WSNs) promises to be one of the most useful technologies of this century. However, secure communication between nodes in WSNs is still an unresolved issue. In this context, we propose two protocols (i.e. Optimal Secure Path (OSP) and Sub-optimal Secure Path (SSP)) to minimize the outage probability of secrecy capacity in the presence of multiple eavesdroppers. We consider dissimilar fading at the main and wiretap link and provide detailed evaluation of the impact of Nakagami-m and Rician-K factors on the secrecy performance of WSNs. Extensive simulations are performed to validate our findings. Although the optimal scheme ensures more security, yet the sub-optimal scheme proves to be a more practical approach to secure wireless links.

1 Introduction

Wireless sensor network (WSN) consist of a collection of sensors nodes which monitor the specific elements from the surrounding and eventually transmit this information to other nodes or sink. In addition to different types of sensors, these nodes are also equipped with a small battery and a low power transceiver. Sensors periodically sense the environment and collects measurements. Depending on the application, these measurements are transmitted to one or multiple nodes for processing. Therefore, WSNs are found to be useful in multiple areas of life. In military, these networks are found to be useful for surveillance and monitoring of hostile territories [1]. Similarly, WSNs are also used to monitor the growth of forests and vegetation. They also evaluate temperature, humidity and pressure of these areas to examine the gradual variations in environment. In health care, these sensor networks monitor the condition of patient and in case of any abnormalities, these networks immediately alert the concerned authorities [1, 2].

Despite the advancements in WSNs over the past decade, security at Physical Layer of WSNs has not been sufficiently explored. Since sensor nodes are low powered and energy limited devices, therefore, security techniques that consume large amount of power are not feasible for WSNs [3, 4]. In addition to this, provisioning of security under minimal hardware complexity is also a daunting task. In this backdrop, research community is being gravitated towards Physical Layer Security (PLS) techniques. PLS exploits the physical characteristics of the wireless channel such as random fading and thermal noise, for securing the wireless link between two legitimate users [5]. In this regard, many existing works suggest that WSNs experience multiple type of fading (i.e. Rayleigh, Rician and Nakagami-*m*). Rayleigh fading channel has been used in literature to model the fading characteristics of the wireless channel. In [6], Cheng et.al concluded that in mobile to mobile communication, the line of sight (LOS) component gradually diminishes and the received signal amplitudes follow Rayleigh distribution. The empirical studies conducted in [6] also suggest that in 5.9 GHz range when the distance between transmitted and received node is less than 5 m the fading characteristics of channel follow Rician distribution. The authors in [7] proposed as a result of extensive simulations that Rayleigh fading is most suitable in congested city roads. In [8], the authors describe that Nakagami-*m* being a versatile model is suitable for various channel conditions. In case when nodes are close to each other it converges to Rician fading. Similarly, when nodes are far from each other it approaches Rayleigh fading channel.

Recently, a few studies have considered PLS in wireless networks. Saad et.al in [9] utilized tree based formulation of network to provide secure communication in uplink wireless networks. A distributed tree formation algorithm was proposed which converged

*Furqan Jameel, COMSATS Institute of Information Technology, 44000, Islamabad, Pakistan, Email: furqanjamil01@yahoo.com

https://dx.doi.org/10.25046/aj020348

to equilibrium state. The authors in [9] also evaluated secrecy per node and path qualification probability based on tree formation algorithm. In [10], Pinto et.al studied the PLS of a Poisson distributed network. The authors quantified the isolation probability and secrecy capacity of all the neighbors of a node. It was found that there exist innate connection between spatial position of eavesdropper and secrecy capacity. Similarly, the authors in [11] derived secrecy outage expression for wireless sensor networks under correlated Weibull fading. Using graph theory, Zhang et.al in [12] provided in-depth analysis of secrecy, rate and network connectivity. The authors illustrated that there exist a trade-off between secrecy capacity and network connectivity. Finally, in [13], authors provided a resource allocation scheme under the constraints of PLS. The paper applied Kuhn-Munkres (KM) algorithm on weighted bipartite graphs to gain significant improvements in secrecy capacity.

It may be highlighted that above mentioned works considered similar fading at the main and wiretap link. However, due to random attributes of fading channel (considering mobility of objects such as vehicles and human beings in the environment), two sensor nodes may experience completely different kind of fading. Therefore, the assumption of similar channel fading is not true for all practical scenarios. Moreover, these studies assumed that the Channel State Information (CSI) of eavesdropper is available at the transmitter. Since a passive eavesdropper does not frequently communicate, therefore, it is very difficult to obtain the CSI of all the eavesdroppers in the network. In view of foregoing arguments, this paper provides a comprehensive analysis of outage probability of secrecy capacity under dissimilar fading at the main and eavesdropper link. Moreover, intermediate trusted relays have been used to provide Optimal Secure Path (when CSI of receiver and eavesdropper is available) and Sub-optimal Secure Path (when CSI of receiver is available only). Random Selection (RS) of nodes has been used as a benchmark scheme to prove the significance of our proposed schemes.

The remainder of this paper is organized as follows. Sec. II provides details of system model; Sec. III derives the outage probability when links are subjected to Rician and Nakagami-m fading. In Sec. IV detailed discussion on simulation results is presented. Finally, Sec. V, provides concluding remarks.

2 System Model

Let us consider a group of $N \times M$ sensor nodes deployed in a two dimensional square area as shown in Figure 1. It is assumed that all the sensor nodes are equipped with communication and sensing capability. Also, all network entities are assumed to have single antenna which experience statistically independent flat fading. The nodes share information from one end of the network to another end using the intermediate sensor nodes. This scenario can occur quite often in area where sensor nodes are deployed far away from each other and direct communication link is not available between source and destination nodes. Some common examples are forest fire detection system and industrial WSNs. Let us now consider that a source node wants to send a message to destination node. In the absence of a direct link between source and destination, the source node delivers the confidential message to nodes present in the very next hop i.e. receiving node. The received signal with power P at the next hop from can be expressed as

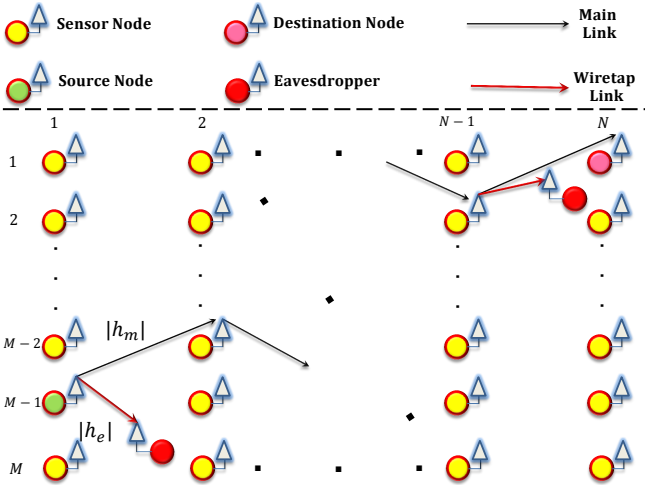


Figure 1. System Model.

$$y_m(t) = \sqrt{P} h_m s(t) + n_s, \tag{1}$$

where s is the transmitted signal, h_m represents the main channel between source node and the receiving node. Additionally, n_s represents the zero mean additive white Gaussian noise (AWGN) with variance N_0 due to the electronics at receiver. Then the instantaneous signal-to-noise ratio (SNR) for the received signal is written as $x_m = \frac{|h_m|^2 P}{N_0}$.

Due to the broadcast nature of wireless signal, the eavesdropper also receives the transmitted signal. It is assumed that all the eavesdroppers in the network are passive and do not participate in the network to jam or interfere the transmission of messages. Moreover, these eavesdroppers do not collude and work independently. In view of above, the received signal at a single eavesdropper (Eve) is written as

$$y_e(t) = \sqrt{P} h_e s(t) + n_e, \tag{2}$$

where n_e is the AWGN at eavesdropper, h_e represents the wiretap channel between Tx and Eve. The instantaneous signal-to-noise ratio (SNR) at eavesdropper for the received signal is given as $x_e = \frac{|h_e|^2 P}{N_0}$.

The channel capacity for both the main link and wiretap link can be written as $C_m = \log_2(1 + x_m)$ and $C_e = \log_2(1 + x_e)$, respectively. According to Wyner, the positive difference between main link capacity and

wiretap link capacity is called the secrecy capacity and expressed as [14]

$$C_{sec} = [C_m - C_e]^+. \tag{3}$$

3 Outage Probability Analysis

Following section derives a closed-form outage probability expression under the consideration of dissimilar fading at main and wiretap link. To be specific, we consider two different scenarios: (1) When Eavesdropper link experiences arbitrary fading and main link experiences Rayleigh fading; and (2) when main link experiences arbitrary fading and wiretap link experiences Rayleigh fading. In this regard, an outage event occurs when C_{sec} falls below some target rate $R_s > 0$. Hence, the probability of outage is given as

$$P_{out} = \Pr\{C_{sec} < R_s\}\,. \tag{4}$$

The outage probability can also be re-written as

$$P_{out} = 1 - P_{cov}, \tag{5}$$

where $P_{cov} = \Pr\{C_{sec} > R_s\}$ is the coverage probability, which can be expressed as

$$P_{cov} = \Pr\left[\log_2\left(\frac{1+x_m}{1+x_e}\right) > R_s\right]. \tag{6}$$

$$P_{cov} = \Pr\left(x_m > 2^{R_s}(1+x_e) - 1\right). \tag{7}$$

Using (7) we can define probability limits as

$$P_{cov} = \int_0^\infty \int_{2^{R_s}(1+x_e)-1}^\infty f_{X_m,X_e}(x_m, x_e)\, dx_m dx_e, \tag{8}$$

where $f_{X_m,X_e}(x_m, x_e)$ is the joint PDF of x_m and x_e. By exploiting the independence of x_m,x_e, we get

$$P_{cov} = \int_0^\infty \left[1 - F_{X_m}\left(2^{R_s}(1+x_e) - 1\right)\right] . f_{X_e}(x_e)\, dx_e, \tag{9}$$

where F_{X_m} is the Cumulative Distribution Function (CDF) of x_m.

3.1 Random Fading at Eavesdropper

If the main link is subjected to Rayleigh fading then F_{X_m} is written as

$$F_{X_m}(x_m) = 1 - \exp\left(-\frac{x_m}{\bar{x}_m}\right), \tag{10}$$

where $\bar{x}_m$ is the average SNR of the main link. After simplification we obtain above equation as

$$P_{cov} = \exp\left(-\frac{2^{R_s}-1}{\bar{x}_m}\right) \int_0^\infty f_{X_e}(x_e) \exp\left(-\frac{2^{R_s} x_e}{\bar{x}_m}\right) dx_e\,. \tag{11}$$

According to the definition of Moment Generating Function (MGF), the MGF of a positive random variable β is written as

$$\mathcal{M}(r) = \int_0^\infty f_\beta(t) \exp(rt)\, dt, \tag{12}$$

where f_β is the PDF of β.

3.1.1 Rician Fading

Let us now consider the case when eavesdropper experiences Rician fading then the CDF of instantaneous SNR can be written as

$$F_{X_e}(x_e) = 1 - Q\left(\sqrt{2K}, \sqrt{2\frac{1+K}{\bar{x}_e} x_e}\right), \tag{13}$$

where K is the ratio of power of LOS component and the power of multipath components and $Q(.,.)$ is the Marcum - Q function. The MGF is given by [15]

$$\mathcal{M}_{Rician}(r) = \left(\frac{1+K}{1+K-r\bar{x}_e}\right) \exp\left(\frac{rK\,\bar{x}_e}{1+K-r\bar{x}_e}\right), \tag{14}$$

then the outage probability is given as

$$P_{out} = 1 - \exp\left(-\frac{2^{R_s}-1}{\bar{x}_m}\right) \mathcal{M}_{Rician}(r). \tag{15}$$

3.1.2 Nakagami-m Fading

When the wiretap link undergoes Nakagami-m fading, then the CDF of instantaneous SNR is given as

$$F_{X_e}(x_e) = \frac{\gamma\left(m, \frac{m x_e}{\bar{x}_e}\right)}{\Gamma(m)}, \tag{16}$$

where m is the Nakagami Shape factor which represents the number of multipath clusters in the environment and $\gamma(.,.)$ and $\Gamma(.)$ are the incomplete and complete Gamma function respectively.

$$\mathcal{M}_{Nakagami}(r) = \frac{1}{\left(1 - \frac{r\bar{x}_e}{m}\right)^m}. \tag{17}$$

Hence, the expression of outage probability, when eavesdropper experiences Nakagami-m, is obtained as

$$P_{out} = 1 - \exp\left(-\frac{2^{R_s}-1}{\bar{x}_m}\right) \mathcal{M}_{Nakagami}(r). \tag{18}$$

Also note that for $K = 1$ or $m = 1$, (18) resolves to well known case of Rayleigh fading.

3.2 Random Fading at Transmitter

Now we consider the case when main link experiences random fading and the wiretap link undergoes Rayleigh fading. To evaluate this condition the above mentioned expression of outage probability can be re-written as

$$P_{cov} = \Pr\left(x_e < \frac{1+x_m}{2^{R_s}} - 1\right). \tag{19}$$

After some mathematical manipulations, above expression can be written as

$$P_{cov} = \int_{2^{R_s}-1}^{\infty} F_{X_e}\left(\frac{(1+x_m)}{2^{R_s}} - 1\right) . f_{X_m}(x_m)\, dx_m$$

$$P_{cov} = 1 - F_{X_m}\left(2^{R_s}-1\right) - \exp\left(-\frac{1-2^{R_s}}{\overline{x}_e}\right)$$

$$\times \int_{2^{R_s}-1}^{\infty} f_{X_m}(x_m)\exp\left(-\frac{x_m}{2^{R_s}\overline{x}_e}\right) dx_m. \qquad (20)$$

It is to be noted that the incomplete MGF of a random variable β is given as

$$\mathcal{M}(r,\omega) = \int_{\omega}^{\infty} f_{\beta}(t)\exp(rt)\, dt \qquad (21)$$

where f_β is the PDF of β.

3.2.1 Rician Fading

Using the above expression of CDF of Rician fading, we can obtain the incomplete MGF of x_m as [15]

$$\mathcal{M}_{Rician}(r,\omega) = \left(\frac{1+K}{1+K-r\overline{x}_m}\right)\exp\left(\frac{sK\overline{x}_m}{1+K-r\overline{x}_m}\right)$$

$$\times Q\left(\sqrt{2\frac{(1+K)K}{1+K-r\overline{x}_m}}, \sqrt{2\left(\frac{1+K}{\overline{x}_m} - r\right)\omega}\right). \qquad (22)$$

Now P_{out} is given as

$$P_{out} = F_{X_m}\left(2^{R_s}-1\right) + \exp\left(-\frac{1-2^{R_s}}{\overline{x}_e}\right)\mathcal{M}_{Rician}(r,\omega)\,. \qquad (23)$$

3.2.2 Nakagami-m Fading

Similarly, Using the CDF of Nakagami-m fading,the incomplete MGF of x_m is given as [15]

$$\mathcal{M}_{Nakagami}(r,\omega) = \frac{\Gamma\left(m, \left(\frac{m}{\overline{x}_m} - r\right)\omega\right)}{\Gamma(m)\left(1-\frac{r\overline{x}_m}{m}\right)^m}. \qquad (24)$$

where $\Gamma(.,.)$ is the complementary incomplete Gamma function. Outage probability now yields

$$P_{out} = F_{X_m}\left(2^{R_s}-1\right) + \exp\left(-\frac{1-2^{R_s}}{\overline{x}_e}\right)\mathcal{M}_{Nakagami}(r,\omega)\,. \qquad (25)$$

4 Secure Path Algorithms

This section proposes two algorithms namely, Optimal Secure Path (OSP) and Sub-optimal Secure Path (SSP) as shown in Figure 2 and Figure 3, respectively. The objective of both algorithms is to find the most secure path from source to destination under specified constraints. However, before discussing these algorithms individually, it is pertinent to highlight that following assumptions are made regarding the characteristics of sensor devices:

(1) Each sensor node is given a unique ID.
(2) Each sensor node is able to locate its position using techniques like trilateration and multilateration or with the help of GPS.
(3) Each sensor node has the ability to acquire the position of nodes that fall under its communication range.
(4) Eavesdroppers are uniformly distributed in the network.

4.1 OSP

In the literature of PLS, a common assumption is that the CSI of eavesdropper is available at the transmitter [16, 17]. In this case, transmitting node can decide by calculating the secrecy capacity of each node in the next hop. After performing calculations, the transmitter can select a node which ensures maximum secrecy capacity. The objective function for OSP can be written as

$$\arg\max_{i\in M} C_{sec} \qquad (26)$$
$$\textbf{s.t.}\ R_s > 0$$

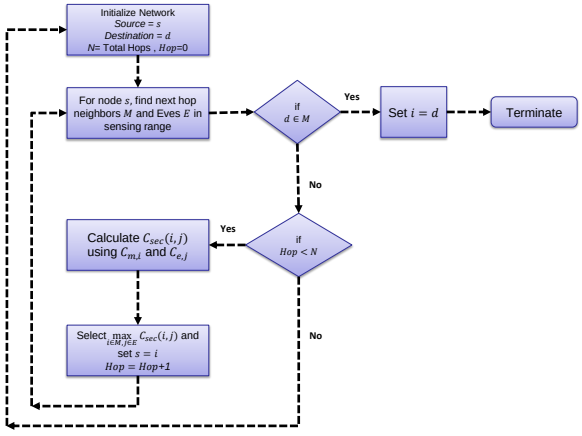


Figure 2. Optimal Secure Path (OSP).

4.2 SSP

In contrast, it is more realistic approach to assume non-availability of eavesdroppers' CSI. It is due to the fact that CSI of passive eavesdroppers' is difficult to obtain. Moreover, the CSI obtained during one time slot may become outdated after a while, which may lead to imperfect estimation. In this context, it is feasible to select a path which maximizes the capacity of main link. Hence, the objective function for SSP becomes

$$\arg\max_{i\in M} C_s \qquad (27)$$
$$\textbf{s.t.}\ R_s > 0$$

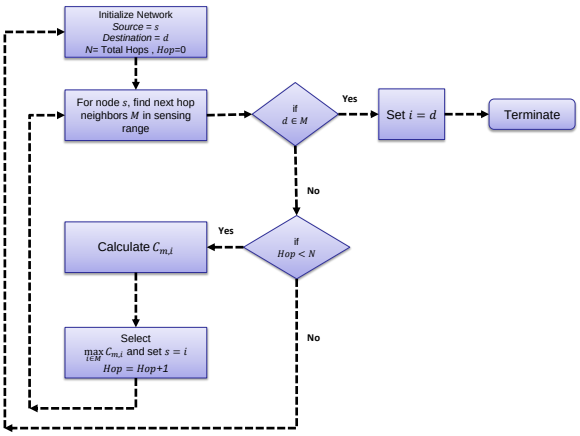


Figure 3. Sub-optimal Secure Path (SSP).

It is worthwhile to note that our proposed schemes can easily be reduced to the case when partial information of eavesdroppers is available, i.e. the CSI of some of the eavesdroppers is available. This scenario can occur in WSNs where some of the eavesdroppers are legitimate nodes of the network, while others are not part of the networks.

5 Results and Discussion

This section provides comparison of results obtained from the simulation of OSP and SSP algorithms when main/ wiretap link undergo random fading as discussed in Section 3. For generation of simulation results, we have considered sparse network and dense network as shown in Figure 4 and Figure 5, respectively. Simulation parameters are provided in Table 1.

Parameter	Value
Simulation Area	1000×1000
Sensor Nodes (Sparse Network)	25
Sensor Nodes (Dense Network)	361
No. of Eavesdroppers	10
Rayleigh Channel Variance	1
Rician Channel Variance	1
Nakagami-m Channel Variance	1
Rician K factor	10
Nakagami-m shape parameter (m)	3
Channel Realizations	10^3
Noise	-50 dB
Target Secrecy Rate (R_s)	1 bit/s/Hz

Table 1. Simulation Parameters

Figure 6 plots the end to end outage probability against the increasing values of transmit power for OSP, SSP and RS schemes for sparse network shown in Figure 4. It can be easily observed from the graphs (a) and (b) that the outage probability reduces with the increase in transmit power. Moreover, it can be seen that OSP out performs SSP and RS in case of both Rician and Nakagami-m fading.

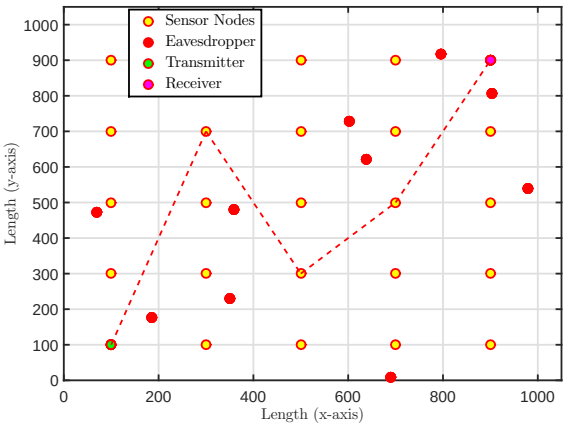


Figure 4. Sparse Network.

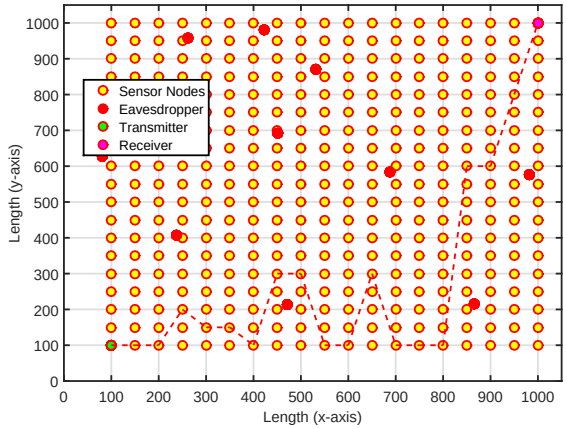
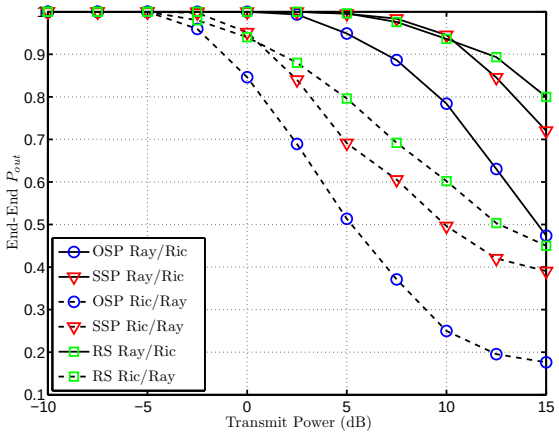


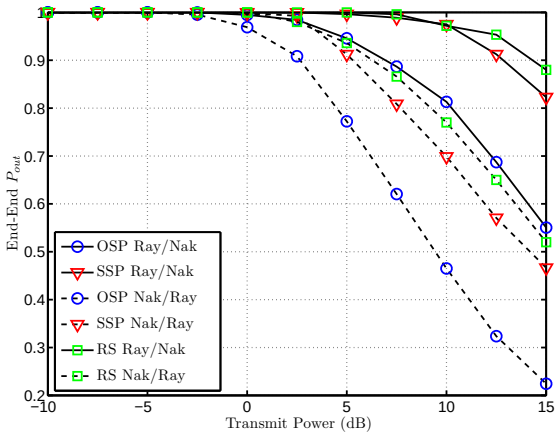
Figure 5. Dense Network.

In Figure 6(a) it can be observed that outage performance of Rician/Rayleigh link is better than Rayleigh/Rician link for all schemes. It is due to the fact that in case of Rician/Rayleigh link the main link undergoes Rician fading with $K = 10$. Since $K > 1$ refers to improvement in LOS power over the power of multipath components, therefore the main link experiences less fading in comparison to wiretap link and the outage probability is reduced. Similarly, in Figure 6(b), it is evident from the plot that for a particular value of transmit power, the outage probability of Nakagami-m/Rayleigh link is less than Rayleigh/Nakagami-m link where $m = 3$. It is due to the fact that $m > 1$ corresponds to reduction in fading and therefore reduction in outage probability.

Figure 7 reiterates the observations highlighted in Figure 6 by plotting end to end outage probability as a function of (a) Rician-K factor and (b) Nakagami-m shape parameter for sparse network. In Figure 7(a) it can be seen that the end to end outage probability rapidly decreases with the increase in Rician-K parameter when main/wiretap link undergoes Rician/Rayleigh fading. Conversely, the outage performance deteriorates with the increase in Rician-K parameter when main/wiretap link undergoes Rayleigh/Rician fading.



(a)

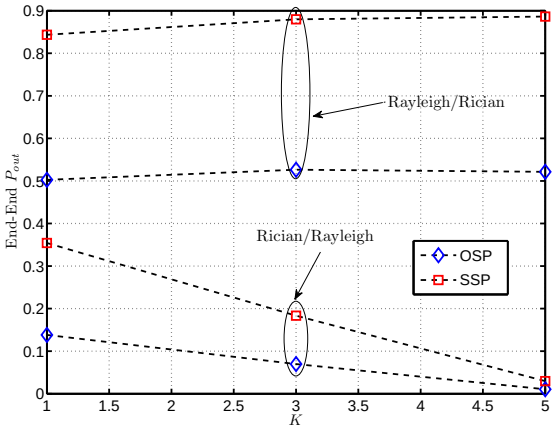


(b)

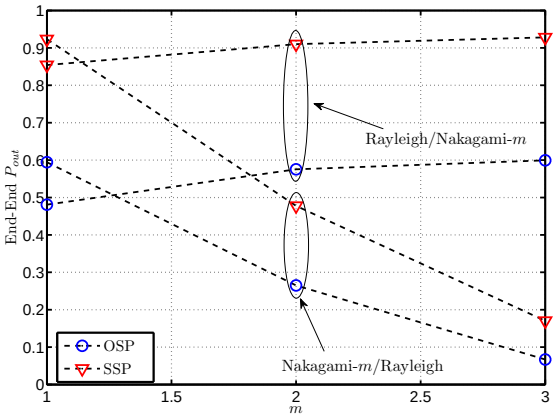
Figure 6. End-End outage probability versus transmit power OSP, SSP and RS (a) Rician (b) Nakagami-m.

The Figure 7(b) shows similar response for Nakagami-m channel. It can easily be observed that for the increasing values of shape parameter m, the Nakagami-m/Rayleigh channel performs better than Rayleigh/Nakagami-m. It is worth noting that for both Rician and Nakagami-m fading, OSP performs better than the SSP. However, if the main link experiences less fading in comparison to the wiretap link then the sub-optimal scheme (i.e. SSP) performs very close to OSP. This fact signifies the importance of using SSP in practical scenarios where legitimate channel is better than eavesdropper's channel.

Figure 8 plots the end to end outage probability against the target secrecy rate R_s for (a) Rician fading (b) Nakagami-m fading for sparse network. It can be seen from the graphs that with the increase in R_s, the outage probability also increases for both OSP and SSP. It is due to the fact that as the target threshold increases, it becomes difficult to maintain secure communication under given channel conditions. This leads to increase in the occurrence of outage events which eventually increases the outage probability. It is can also be observed that in case of Rayleigh/Rician and Rayleigh/Nakagami-m, OSP and SSP converge to 1 with increase in R_s. On the contrary, the curves of OSP and SSP diverge with the increase in R_s for Rician/Rayleigh and Nakagami-m/Rayleigh.



(a)



(b)

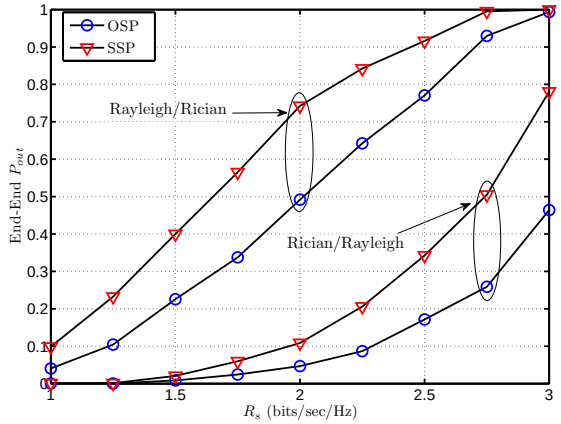
Figure 7. End-End outage probability versus (a) Rician-K factor (b) Nakagami-m shape parameter.

Figure 9 shows the end to end outage probability against the increasing values of transmit power for sparse and dense network. It can be seen that the end to end outage probability in case of dense network is better than that of sparse network for same number of eavesdroppers. The advantage obtained by using multiple relays is also evident from the plot. We can see that using multiple relays offers spatial diversity, due to which the end to end outage probability decreases as the network become dense. It can be observed that effect of spatial diversity is more pronounced where main link is better than wiretap link.

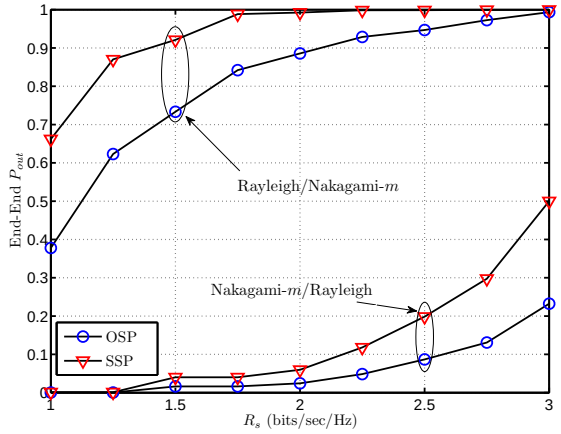
6 Conclusion

In this article, we have discussed implications of dissimilar fading on the secrecy performance of WSNs. We derived closed form expression of outage probability for four scenarios i.e. Rayleigh/Rician,

Rician/Rayleigh, Rayleigh/Nakagami-m, Nakagami-m/Rayleigh. We then provided two algorithms i.e. OSP and SSP to enhance link security. A detailed comparison of OSP, SSP and RS was presented with OSP performing better than SSP and RS. However, when CSI of eavesdropper is unavailable to the transmitter, the SSP algorithm is a better and more practical choice. Finally, it was revealed that OSP performs better in dense network due to increased spatial diversity.



(a)



(b)

Figure 8. End-End outage probability as a function of R_s (a) Rician (b) Nakagami-m.

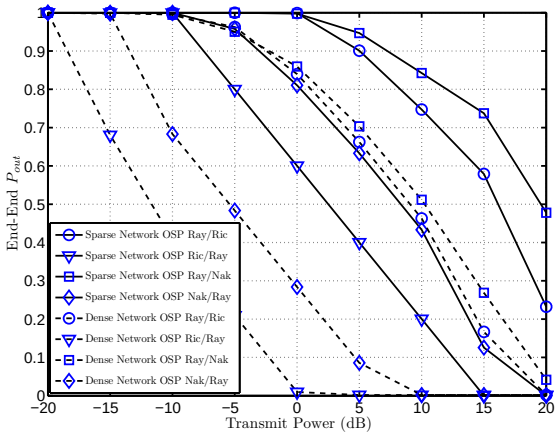


Figure 9. End-End outage probability versus transmit power for Sparse and Dense Network.

Conflict of Interest The authors declare no conflict of interest.

References

[1] M. M. Warrier and A. Kumar, "Energy efficient routing in Wireless Sensor Networks: A survey," 2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), Chennai, 2016, pp. 1987-1992.

[2] G. Han, J. Jiang, C. Zhang, T. Q. Duong, M. Guizani and G. K. Karagiannidis, "A Survey on Mobile Anchor Node Assisted Localization in Wireless Sensor Networks," in IEEE Communications Surveys and Tutorials, vol. 18, no. 3, pp. 2220-2243, thirdquarter 2016.

[3] R. C. Luo and O. Chen, "'Mobile sensor node deployment and asynchronous power management for wireless sensor networks,"' IEEE Transactions on Industrial Electronics, vol. 59, no. 5, pp. 2377-2385, 2012.

[4] J.-C. Wang, C.-H. Lin, E. Siahaan, B.-W. Chen, and H.-L. Chuang, "'Mixed sound event verification on wireless sensor network for home automation,"' IEEE Transactions on Industrial Informatics, vol. 10, no. 1, pp. 803-812, 2014.

[5] Y. Zou, J. Zhu, X. Wang and V. C. M. Leung, "Improving physical-layer security in wireless communications using diversity techniques," in IEEE Network, vol. 29, no. 1, pp. 42-48, Jan.-Feb. 2015.

[6] L. Cheng, B. E. Henty, D. D. Stancil, F. Bai and P. Mudalige, "Mobile Vehicle-to-Vehicle Narrow-Band Channel Measurement and Characterization of the 5.9 GHz Dedicated Short Range Communication (DSRC) Frequency Band," in IEEE Journal on Selected Areas in Communications, vol. 25, no. 8, pp. 1501-1516, 2007.

[7] Y. Ibdah and Y. Ding "Mobile-to-Mobile Channel Measurements at 1.85 GHz in Suburban Environments," IEEE Trans. Commun., vol. 63, no.2,pp. 466-475, 2015.

[8] A. Goldsmith, Wireless Communications. Stanford, CA, USA: Stanford Univ., pp. 23-48, 2004.

[9] W. Saad, X. Zhou, B. Maham, T. Basar and H. V. Poor, "Tree Formation with Physical Layer Security Considerations in Wireless Multi-Hop Networks," in IEEE Transactions on Wireless Communications, vol. 11, no. 11, pp. 3980-3991, November 2012.

[10] P. C. Pinto, J. Barros and M. Z. Win, "Physical-layer security in stochastic wireless networks," Communication Systems, 2008. ICCS 2008. 11th IEEE Singapore International Conference on, Guangzhou, 2008, pp. 974-979.

[11] F. Jameel; S. Wyne; I. Krikidis, "Secrecy Outage for Wireless Sensor Networks," in IEEE Communications Letters , vol.PP, no.99, pp.1-1

[12] R. Zhang, C. Comaniciu and H. V. Poor, "On Rate, Secrecy, and Network Connectivity Tradeoffs for Wireless Networks," in IEEE Communications Letters, vol. 20, no. 8, pp. 1559-1562, Aug. 2016.

[13] H. Zhang, T. Wang, L. Song and Z. Han, "Radio resource allocation for physical-layer security in D2D underlay communications," 2014 IEEE International Conference on Communications (ICC), Sydney, NSW, 2014, pp. 2319-2324.

[14] A. D. Wyner, The wire-tap channel, Bell System Technical Journal, vol. 54, no. 8, pp. 1355-1387, 1975.

[15] Lopez-Martinez, F. Javier, Juan M. Romero-Jerez, and Jose Francisco Paris. "On the Calculation of the Incomplete MGF with Applications to Wireless Communications." IEEE Transactions on Communications (2016).

[16] Y. Zou, X. Wang, and W. Shen, Optimal relay selection for physical layer security in cooperative wireless networks, IEEE Journal on Selected Areas in Communications, vol. 31, no. 10, pp. 2099-2111, 2013.

[17] Y. Zou, X. Wang, W. Shen, and L. Hanzo, Security versus reliability analysis of opportunistic relaying, IEEE Transactions on Vehicular Technology, vol. 63, no. 6, pp. 2653-2661, 2014.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 2, No. 3, 384-388 (2017)
www.astesj.com
Special issue on Recent Advances in Engineering Systems

ASTES Journal
ISSN: 2415-6698

Security in SWIPT with Power Splitting Eavesdropper

Furqan Jameel*, Faisal, M Asif Ali Haider, Amir Aziz Butt

Department of Electrical Engineering, COMSATS Institute of Information Technology, Islamabad, Pakistan

ARTICLEINFO

Article history:
Received: 31 March, 2017
Accepted: 04 May, 2017
Online: 22 May, 2017

Keywords :
SWIPT
Secrecy Capacity
Weibull Fading
Power Splitting
Secrecy Performance
Weibull Shape Parameter

ABSTRACT

Simultaneous wireless information and power transfer (SWIPT) has drawn significant research interest in recent years. In this paper, we investigate the information theoretic secrecy of a SWIPT system under Weibull fading channel. To be specific, we analyze the information security in the presence of an energy harvesting and information decoding eavesdropper. All links are subjected to Weibull fading which is more complex than traditional Rayleigh fading. Moreover, we derive closed-form expressions for the probability of strictly positive secrecy capacity and the ergodic secrecy capacity. We evaluate the effect of Weibull shape parameter and power splitting factor on the secrecy performance of SWIPT system. Numerical and simulation results are provided to demonstrate that our findings are instantly applicable on the design of wireless power networks.

1 Introduction

Most of the conventional electrical energy transfer takes palace with the help of conductors between power source and load. However, in case of wireless power transfer (WPT), the transmission of energy from source to load takes place without interconnecting conductors. In other words, this simply means that WPT allows transfer of energy without wires [1]. In WPT, a power source connected to a wireless transmitter conveys the energy wirelessly across an intervening space to once or more wireless receivers. At the receiving end, electromagnetic energy is again converted to electrical current which can be stored and used by the device. Since most wireless devices are powered either through power cables or battery replacement, which limits the scalability, sustainability, and mobility of wireless communications [2]. In practice, wireline charging and battery replacement may be infeasible under some conditions; for instance, it is very difficult to replace the battery of implanted medical devices in human bodies. Besides, wireline charging and battery renewal shortens working period of wireless mobile devices. The WPT technique becomes appealing since it can be used to wirelessly charge the devices in hazardous conditions.

In this context, simultaneous transfer of power and information is emerging as one of the most promising technique to increase the lifetime of wireless devices [2]. However, provisioning of security over a Simultaneous wireless information and power transfer (SWIPT) system has remained a daunting task due to the broadcast nature of wireless networks. Traditional cryptographic techniques are not suitable for SWIPT because they require complex hardware and consume large amounts of energy that are typically not affordable by wireless devices. Moreover, an eavesdropper with unlimited computing power may still decipher these techniques using brute-force attack. In this context, Physical Layer Security (PLS) has emerged as an attractive solution for securing wireless transmissions by exploiting the wireless channel characteristics [3]. The PLS techniques such as artificial noise generation and cooperative relaying are not suitable for day to day wireless devices due to the latters limited energy resources.

Weibull distribution was studied by authors in [4] to understand the effect of small scale fading (SSF) in outdoor environment for 800/900 MHz frequency band. The authors in [5] performed channel measurements for indoor channels and found Weibull distribution to conform closely with measurements for less than 10 dB of fading variations. In [6] the authors found the Weibull model to be a good fit for the SSF in narrow-band channel measurements for wireless body area networks. Authors in [7] and [8]investigated the SSF for 5 and 10-GHz band and proposed to use the Weibull distribution with its parameter β taking values between 1.77 to 3.9 to perfectly match the measurement curves. Authors in [9] measured the wireless channel based on pedestrians movement in a forest environment for 5-GHz band. These measure-

*Furqan Jameel, COMSATS Institute of Information Technology,Islamabad, 44000, Pakistan, Email: furqanjamil01@yahoo.com

https://dx.doi.org/10.25046/aj020349

ments suggested that Weibull distribution can be used to specify amplitude fading in a forest environment.

Despite this importance, a handful of studies have investigated Weibull fading from security point of view. Moreover, most of the existing work in SWIPT security is limited to analysis of Rayleigh fading which is a special case of Weibull fading. In this regard, it is pertinent to note that this paper is an extension of our previous work [10], where expressions of optimal power splitting and time switching were derived under Weibull fading. However, this paper evaluates the secrecy performance of SWIPT system under the influence of Weibull fading. Here, we provide closed-form expressions of probability of strictly positive secrecy capacity and ergodic secrecy capacity in the presence of power splitting eavesdropper. Additionally, we evaluate the impact of power splitting factor and Weibull parameter β, on determining the overall secrecy of information in SWIPT system.

The rest of this paper is organized as follows. System model is provided in Sec. 2; Sec. 3 contains derivation of the SPSC probability and Secrecy capacity under Weibull fading. In Sec. 4 numerical results are discussed. Finally, in Sec. 5, conclusions are highlighted.

2 System Model

A downlink SWIPT system is assumed which consists of a Base Station (BS) and an information receiving node S in the presence of an energy harvesting and information decoding node represented as E, as shown in Figure 1. BS conveys information to S and transfers energy to E during a single scheduling slot. Energy harvesting node is assumed to store energy from received radio signal. The BS, node S and eavesdropper E are equipped with single antenna. Antennas in all the network entities are assumed to be experiencing statistically independent flat Weibull fading. Moreover, we assume quasi-static fading to incorporate the effect of fading during each block of time. The BS is assumed to have channel state information (CSI) for the channels to node S. Being part of the coverage range of the BS, the energy harvesting node can act as a potential eavesdropper. It is assumed that eavesdropper uses power splitting (PS) scheme [1] to incorporate the process of energy harvesting and information decoding. According to PS, signal is divided into two streams for with ratio ρ and $(1-\rho)$for information decoding and energy harvesting, respectively; where $0 \leq \rho \leq 1$.

When BS transmits its signal s to S with power P; the received signal at S can be expressed as

$$y_m = \sqrt{\frac{P}{P_{Loss,s}}} h_m s + n_m, \tag{1}$$

where h_m represents the main channel between BS and S with $|h_m|$ being Weibull-distributed. Furthermore, n_m represents the zero mean additive white Gaussian noise (AWGN) with variance N_0 due to the receiver electronics at node S. $P_{Loss,s} = \frac{(4\pi)^2 d_s^{\gamma}}{G_t G_r \lambda^2}$ is the path loss at main link, d_s is the distance between BS and S. γ is the path loss exponent, G_t and G_r are transmitting and receiving antenna gains. Since the transmission from BS is also picked up by the eavesdropper, the signal received at the eavesdropper is given as

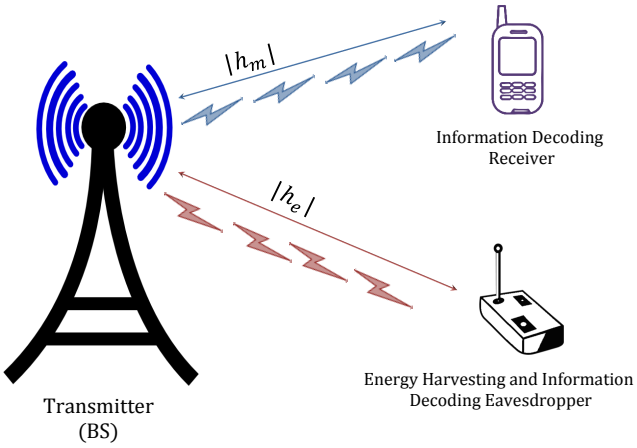


Figure 1. System Model.

$$y_e = \sqrt{\rho}\left(\sqrt{\frac{P}{P_{Loss,e}}} h_e s + n_e \right) + z, \tag{2}$$

where n_e is the AWGN at eavesdropper, h_e represents the channel between BS and eavesdropper with $|h_e|$ being Weibull-distributed. Also, $P_{Loss,e} = \frac{(4\pi)^2 d_e^{\gamma}}{G_t G_r \lambda^2}$ is the path loss at wiretap link, d_e is the distance between BS and E and n_e represents the zero mean additive white Gaussian noise (AWGN) with same variance as N_0 . z is the signal processing noise which is modeled as AWGN with zero mean and variance σ.

Then the instantaneous signal-to-noise ratio (SNR) at S for the received signal is written as

$$x_m = \frac{|h_m|^2 P}{P_{Loss,s} N_0}. \tag{3}$$

The instantaneous signal-to-noise ratio (SNR) at eavesdropper for the received signal is

$$x_e = \frac{\rho |h_e|^2 P}{P_{Loss,e} N_0 \left(\rho + \frac{\sigma^2}{N_0}\right)}. \tag{4}$$

Since main and wiretap link are subjected to Weibull fading, therefore, the Probability Density Function (PDF) and Cumulative Distribution Function (CDF) of the SNR of received signal at S and E is given by

$$f_{X_a}(x_a) = \beta_a \left(\frac{\Gamma\left(1+\frac{1}{\beta_a}\right)}{\bar{x}_a} \right)^{\beta_a} x_a^{\beta_a - 1} \times e^{-\left(\frac{\Gamma\left(1+\frac{1}{\beta_a}\right) x_a}{\bar{x}_a} \right)^{\beta_a}}. \tag{5}$$

$$F_{X_a}(x_a) = 1 - e^{-\left(\frac{\Gamma\left(1+\frac{1}{\beta_a}\right)x_a}{\bar{x}_a}\right)^{\beta_a}}, \quad (6)$$

where $a \in (m,e)$ for main and wiretap link, β_a represents the Weibull shape parameter and $\Gamma(.)$ is the well-known Gamma function. The channel capacity for both main and wiretap link can be written as $C_m = \log_2(1+x_m)$ and $C_e = \log_2(1+x_e)$, respectively [11]. Now, the positive difference between main link capacity and wiretap link capacity is called the secrecy capacity which is expressed as [12]

$$C_{\text{sec}} = [C_m - C_e]^+. \quad (7)$$

3 Analysis of Secrecy Performance

In this section, we will derive closed form expression of probability of SPSC and Ergodic Secrecy Capacity.

3.1 Probability of Strictly Positive Secrecy Capacity (SPSC)

Let us now derive an expression for the probability of strictly positive secrecy capacity (SPSC) which is the probability that the secrecy capacity is greater than zero. Mathematically it is written as

$$\Pr(C_{\text{sec}} > 0) = \Pr\left[\log_2\left(\frac{1+x_m}{1+x_e}\right) > 0\right]. \quad (8)$$

$$\Pr(C_{\text{sec}} > 0) = \Pr(x_m > x_e). \quad (9)$$

The above expression can be written as

$$\Pr(C_{\text{sec}} > 0) = \int_0^\infty \int_{x_e}^\infty f_{X_m,X_e}(x_m,x_e)dx_m dx_e, \quad (10)$$

where $f_{X_m,X_e}(x_m,x_e)$ is the joint PDF, which can be decomposed by using independence of x_m, x_m. Consequently (10) can be written as

$$\begin{aligned}\Pr(C_{\text{sec}} > 0) &= \int_0^\infty \int_{x_e}^\infty f_{X_m}(x_m) f_{X_e}(x_e)dx_m dx_e \\ &= \int_0^\infty f_{X_e}(x_e)[1 - F_{X_m}(x_e)]dx_e. \end{aligned} \quad (11)$$

Using the definition of CDF of a random variable Y we can write

$$F_Y(y) = \Pr(Y < y) = \int_{-\infty}^{y} f(t)dt. \quad (12)$$

Replacing values from (5) and (6) in (11) we can write above equation as

$$\begin{aligned}Pr(C_{\text{sec}} > 0) = \beta_e \left(\frac{\Gamma(1+\frac{1}{\beta_e})}{\bar{x}_e}\right)^{\beta_e} \times \int_0^\infty x_e^{\beta_e - 1} e^{-\left(\frac{\Gamma(1+\frac{1}{\beta_e})x_e}{\bar{x}_e}\right)^{\beta_e}} \\ \times e^{-\left(\frac{\Gamma\left(1+\frac{1}{\beta_m}\right)x_e}{\bar{x}_m}\right)^{\beta_m}} dx_e, \end{aligned} \quad (13)$$

where $\bar{x}_m = \frac{\Omega_s}{P_{Loss,s}}$ and $\bar{x}_e = \frac{\rho\Omega_e}{(P_{Loss,e}(\rho+\sigma^2/N_0))}$; also Ω_s and Ω_e is the average SNR of main and wiretap link, respectively.

Assuming $\beta_e = \beta_m = \beta$ and substituting $u = \left(\frac{\Gamma\left(1+\frac{1}{\beta_e}\right)x_e}{x_e}\right)^{\beta_e}$ in (10), we can solve integral as

$$\Pr(C_{\text{sec}} > 0) = \frac{(\bar{x}_m)^\beta}{(\bar{x}_m)^\beta + (\bar{x}_e)^\beta}, \quad (14)$$

3.2 Ergodic Secrecy Capacity (C)

Another measure of interest is the ergodic secrecy capacity. It is defined as the average of secrecy rate over x_m and x_e. It can be mathematically written as

$$\begin{aligned} C &= \mathbf{E}\{[\log_2(1+x_m) - \log_2(1+x_e)]^+\} \\ &= \int_0^\infty \underbrace{\int_0^{x_m} [\log_2(1+u) - \log(1+v)] f_{x_e}(v)dv}_{\mathcal{G}} f_{x_m}(u)du \end{aligned} \quad (15)$$

Using integration by parts, the above expression can be simplified as

$$\begin{aligned}\mathcal{G} &= \log_2(1+u)F_{x_e}(u) - \left[\log_2(1+u)F_{x_e}(u) - \frac{1}{\ln 2} \right. \\ &\quad \left. \times \int_0^u \frac{1}{1+v} F_{x_e}(v)dv\right] \\ &= \frac{1}{\ln 2}\int_0^u \frac{F_{x_e}(v)}{1+v}dv. \end{aligned} \quad (16)$$

Replacing (16) in (15) and by changing the order of integration, we get

$$\begin{aligned} C &= \frac{1}{\ln 2}\int_0^\infty \frac{F_{x_e}(v)}{1+v}\left[\int_v^\infty f_{x_m}(v)du\right]dv \\ &= \frac{1}{\ln 2}\int_0^\infty \frac{F_{x_e}(v)}{1+v}[1 - F_{x_m}(v)]dv. \end{aligned} \quad (17)$$

After replacing (6) for $\beta_e = \beta_m = \beta$ in (17) and performing some algebraic manipulations, we obtain

$$\begin{aligned} C = \frac{1}{\ln 2}\int_0^\infty \frac{1}{1+x_e} e^{-\left(\frac{\Gamma\left(1+\frac{1}{\beta}\right)x_e}{\bar{x}_m}\right)^\beta} - \frac{e^{-\frac{(\bar{x}_m)^\beta + (\bar{x}_e)^\beta}{(\bar{x}_m\bar{x}_e)^\beta}}}{1+x_e} \\ \times e^{-2\left(\Gamma(1+\frac{1}{\beta})x_e\right)^\beta} dx_e. \end{aligned} \quad (18)$$

Above equation contains single integral which can be readily solved using computational software packages such as MATHEMATICA and MATLAB.

4 Numerical Results

In the following section we provide some numerical results along with discussion on the above analysis. Unless mentioned otherwise, the simulation parameters along with their respective values are provided in Table 1.

S No.	Simulation Parameter	Value
1.	Path loss at Main Link $P_{Loss,s}$	0 dB
2.	Path loss at Wiretap Link $P_{Loss,e}$	0 dB
3.	Power splitting factor ρ	0.8
4.	Weibull shaper parameter β	2
5.	Noise at main/ wiretap link	- 60 dB
6.	Channel realizations	10^6

Table 1. Simulation Parameters

Figure 2 shows the SPSC probability plotted against $\bar{x}_m$ for different values of $\bar{x}_e$. It is vividly clear from the figure that probability of SPSC increases with the increase in $\bar{x}_m$. It is because with the increase in $\bar{x}_m$ the channel capacity of the main link increases which resultantly increases the SPSC probability. Moreover, it is also evident from the graph that increase in x_e, decreases the probability of SPSC. In addition to this, we can observe that for the same values of $\bar{x}_m$, $\bar{x}_e$ and $P_{Loss,e}$, the probability of SPSC significantly decreases with the increase in $P_{Loss,s}$. Furthermore, we observe that the impact of decrease in $P_{Loss,s}$ is more significant for larger values of $\bar{x}_e$ as compared to smaller $\bar{x}_e$. It can also be seen that the simulation results closely coincide with the analytical results which proves the accuracy of our analysis.

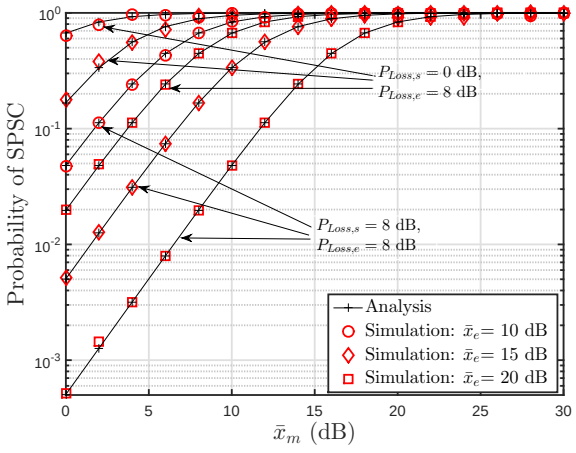


Figure 2. SPSC probability plotted as a function of $\bar{x}_m$.

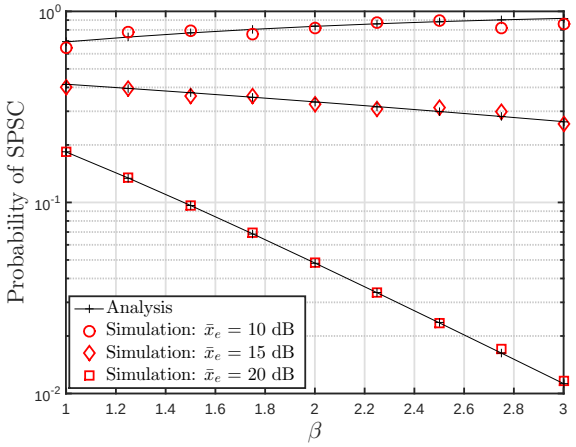


Figure 3. SPSC probability against Weibull shape parameter β.

Figure 3 plots SPSC probability as a function of Weibull shape parameter β. It is observed from the figure that the SPSC probability increases with the increase in β until main links SNR is equal or greater than eavesdroppers SNR. However, when $\bar{x}_m < \bar{x}_e$ then the probability of SPSC decreases with increase in β. Additionally, this effect is more significant for large values of $\bar{x}_e$ as compared to small values.

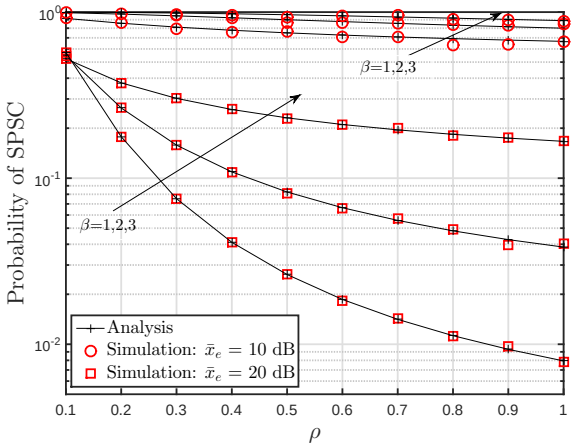


Figure 4. Probability of SPSC vs power splitting factor ρ.

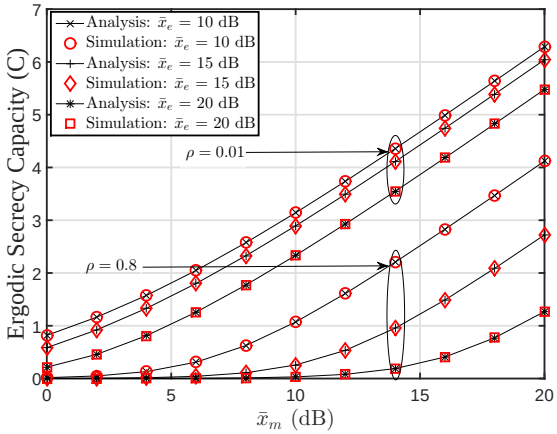


Figure 5. Ergodic secrecy capacity as a function of $\bar{x}_m$.

Figure 4 gives the SPSC probability as a function of ρ for different values of $\bar{x}_e$. It can be observed from the figure that for the increasing values of ρ the probability of SPSC decreases. It is due to the fact that larger values of ρ indicate that more power at the eavesdropper is being used for information decoding which results into reduction in the secrecy capacity. Furthermore, the reduction in probability of SPSC is more significant at higher values of eavesdroppers SNR. Additionally, it can be seen from the plot that influence of β increases with the increase in ρ and $\bar{x}_e$.

Finally, Figure 5 illustrates the achievable secrecy capacity as a function of increasing values of main links average SNR. It is well established by now that the increase in $\bar{x}_m$ increases the secrecy capacity. It is noteworthy that for a specific value of $\bar{x}_m = 12$ dB and $\bar{x}_e = 10 \rightarrow 20$ dB, the secrecy capacity increases from 3 to 3.8 for $\rho = 0.01$, whereas, secrecy capacity increases from 0.1 to 1.8 for $\rho = 0.8$. This result is shows that secrecy capacity varies rapidly for higher values of ρ as compared to lower ones.

5 Conclusion

In this article we have analyzed the performance of SWIPT system from the perspective of physical layer security. We have derived the closed-form expression for probability of SPSC. We have characterized the effect of Weibull shape parameter in SPSC. Our results also demonstrate the significance of power splitting factor on the confidentiality of information. Finally, we evaluated the impact of power splitting factor on ergodic secrecy capacity while varying the Weibull shape parameter. These findings can be of utility for ensuring data security for wireless communication.

Conflict of Interest The authors declare no conflict of interest.

References

[1] X. Zhou, R. Zhang and C. K. Ho, "Wireless information and power transfer: Architecture design and rate-energy tradeoff", IEEE Trans. Commun., vol. 61, no. 11, pp. 4754-4767, 2013.

[2] L. R. Varshney, Transporting information and energy simultaneously, in Proc. I EEE Int. Symp. Inform. Theory (ISIT), Toronto, pp. 16121616, 2008.

[3] Y. Zou, J. Zhu, X. Wang and V. C. M. Leung, "Improving physical-layer security in wireless communications using diversity techniques," in IEEE Network, vol. 29, no. 1, pp. 42-48, Jan.-Feb. 2015.

[4] N. S. Adawi et al., Coverage prediction for mobile radio systems operating in the 800/900 MHz frequency range, IEEE Transactions on Vehicular Technology, vol. 37, pp. 372, 1988.

[5] H. Hashemi, The indoor radio propagation channel, Proceedings of IEEE, vol. 81, pp. 943968, 1993.

[6] D. Smith, J. Zhang, L. Hanlen, D. Miniutti, D. Rodda, and B. Gilbert, A Simulator for the Dynamic On-body Area Propagation Channel, in IEEE International Symposium on Antennas and Propagation Society, pp. 14, 2009.

[7] I. Sen, D. W. Matolak, Vehicle-vehicle channel models for the 5-GHz band, IEEE Transactions on Intelligent Transportation Systems, vol. 9, no. 2, p. 235 - 245, 2008.

[8] Q. Wu, D. W. Matolak, and I. Sen, 5-GHz-band vehicle-to-vehicle chan-nels: Models for multiple values of channel bandwidth, IEEE Transactions on Vehicular Technology, vol. 59, no. 5, pp. 26202625, 2010.

[9] D. W. Matolak, F.-C. Yang, and H. B. Riley, Short range forest channel modeling in the 5 GHz band, in Proceedings of 6th European Conference on Antennas Propagation, Prague, Czech Republic, pp. 33373341, 2012.

[10] F. Jameel, A. Ali and R. Khan, "Optimal time switching and power splitting in SWIPT," 2016 19th International Multi-Topic Conference (INMIC), Islamabad, 2016, pp. 1-5.

[11] A. D. Wyner, The wire-tap channel, Bell System Technical Journal, vol. 54, no. 8, pp. 1355-1387, 1975.

[12] M. Bloch, J. Barros, M. R. D. Rodrigues, and S. W. McLaughlin, Wireless information theoretic security, IEEE Transactions on Information Theory, vol.54, no.6, pp.25152534, 2008.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 2, No. 3, 389-394 (2017)
www.astesj.com
Special issue on Recent Advances in Engineering Systems

ASTES Journal
ISSN: 2415-6698

Performance Analysis of Phased Array and Frequency Diverse Array Radar Ambiguity Functions

Shaddrack Yaw Nusenu*

Koforidua Technical University, Electrical / Electronic Department, Faculty of Engineering, Ghana.

ARTICLEINFO

Article history:
Received: 04 April, 2017
Accepted: 12 May, 2017
Online: 22 May, 2017

Keywords :
Frequency diverse array
Phased array radar
Radar ambiguity functions
ECDF

ABSTRACT

Radar ambiguity function is widely employed to analyze a radar system's time delay (range) resolution, Doppler frequency (range rate) resolution, mainlobe width and sidelobes in both range and Doppler. In this paper, ambiguity functions of frequency diverse array (FDA) which offers range-angle dependent transmit beampattern and Phased array which yields angle transmit beampattern is investigated. The FDA and Phased array radar ambiguity function is derived and evaluate its performance characteristics in terms of time-delay and Doppler frequency resolutions. Performance analysis has also been done in terms of empirical cumulative distribution function (ECDF). Simulation and results have verified the effectiveness of both radar systems by comparing their performance characteristics.

1 Introduction

Phased array radar system has been widely utilized in many modern communications, radar and navigation systems [1, 2, 3]. In a standard phased-array antenna, the elements are fed coherently and phase shifters or time-delayers are used to scan the beam to the desired direction in space [4]. The offered directional gain is beneficial for target detection and tracking weak targets in the antenna looking direction and at the same time suppressing strong sidelobe interferences from other directions. The phased-array antenna beam steering is fixed in an angle for all ranges and thus, the range and angle of targets cannot be directly estimated from its beamforming output due to an inherent range ambiguity [5].

Recently, a flexible array, namely, frequency diverse array (FDA) has been proposed [6, 7, 8]. The most significant FDA difference from its counterpart phased-array is that a small frequency increment, compared to the carrier frequency, is applied between the FDA elements. This frequency increment yields a range-angle-dependent beampattern for which the beam focus direction will be a function of the angle, range and time [9, 10, 11]. FDA has sparked much investigations because of its promising applications. In [12, 13, 14] FDA have been investigated in terms of range-dependent beam with applications in suppressing range ambiguous clutter. FDA for synthetic aperture radar (SAR) high-resolution imaging was reported in [15]. Furthermore, [16, 17] exploited FDA capabilities in range and angle localization of targets.

It is well established that the radar ambiguity function is a useful tool for evaluating performance metrics of radar system, such as time delay (range) and Doppler frequency resolutions and the probabilities of detection and false alarm. Moreover, it helps in selecting appropriate waveforms depending on the required performance parameters.

Therefore, in this perspective, this paper derived and analyzed the performance characteristics of phased array and FDA radar ambiguity functions in time delay and Doppler frequency resolutions. Ambiguity function performance metric is also provided in terms of empirical cumulative distribution function (ECDF). The performance characteristics are compared and FDA radar exhibited better performance results than the phased array radar.

The remaining sections are organized as follows: In Section II, conventional ambiguity function is reviewed. Section III, ambiguity functions for FDA and phased array is derived. Section IV, provides simulation results. Finally, Section V, concludes the paper.

2 Review Of The Conventional Ambiguity Function

In a conventional radar system, the radar ambiguity function (AF) is a common tool for investigating radar signals [18, 19, 20, 21]. The AF is a two dimensional function that describes the output of a matched filter when the input signal of time delay t_d and Doppler frequency f_d relative to

*Shaddrack Yaw Nusenu, Koforidua Technical University, Ghana, Email: nusenu2012gh@yahoo.com

https://dx.doi.org/10.25046/aj020350

nominal values for which the matched filter has been designed. The AF can be expressed as

$$|\chi(t_d, f_d)| \triangleq \left| \int_{-\infty}^{\infty} u(t)\, u^*(t+t_d) \exp\{j2\pi f_d t\}\, dt \right| \quad (1)$$

where $(\cdot)^*$ denotes the conjugate operator, $u(t)$ is the radar waveform, t_d and f_d are the mismatch in time delay and Doppler frequency, respectively. The ambiguity function plays a great role in determining the t_d and f_d resolutions of radar. The value of this function is maximum at $(t_d, f_d) = (0, 0)$, and it represents the matched filter output without any mismatch. Hence, the narrower the AF around origin, the better the t_d and f_d resolutions. Therefore, the ideal AF would be an impulse at the origin, however, there is no waveform promising such kind of an AF.

3 Frequency Diverse Array And Phased Array Radar Ambiguity Function

First, we will analyzed FDA signal model and its ambiguity function derivation and then followed by phased array radar. FDA elements can be either excited by the same waveform or different waveforms. Consider N elements of FDA radar signal model. Assume the nth element transmitted signal is given by

$$\psi_n(t) = \exp(j2\pi z_{n,1} t)\, s(t) \quad (2)$$

where $z_{n,1}$ denotes the nth transmit signal pulse carrier frequency and $s(t)$ can be expressed as

$$s(t) \triangleq \begin{cases} 1, & t \in [0, T_\psi] \\ 0, & otherwise \end{cases} \quad (3)$$

with T_ψ denotes the pulse duration. The nth FDA element emits the signal $\psi_n(t)$ with a frequency increment Δf_n. The signals radiated by the nth element can be expressed as

$$\begin{aligned} s_n(t) &= \sqrt{N}\psi_n(t) \exp(-j2\pi f_n t)\, w_n^* \\ &= \sqrt{N} u_n(t)\, w_n^* \end{aligned} \quad (4)$$

where $f_n = f_0 + \Delta f_n$, with f_0 denotes the FDA radar carrier frequency, w_n being weights and $u_n(t)$ defined in (5)

$$u_n(t) = \exp(-j2\pi(f_0 - z_{n,1} + \Delta f_n)t)\, s(t) \quad (5)$$

By following [16], the transmitted signal propagating along (θ, r) in the far-field is given by

$$S_n(t, \theta, r) = \sqrt{N} \sum_{n=1}^{N} w_n^* a_n(t, \theta, r)\, u_n(t) \quad (6)$$

and $a_n(t, \theta, r)$ can be written as [22]

$$a_n(t, \theta, r) = e^{\left\{ j2\pi \left(\Delta f_n t - \frac{\Delta f_n r}{c} + n \frac{d f_0 \sin\theta}{c} \right) \right\}} \quad (7)$$

Without loss of validity, a single-antenna is employed in the receiver. Assumed the round-trip time delay to a target in the angle θ and range r is t_d and $w_n = 1$ is used. The received signal is written as

$$\hat{r}_s(t, t_d, f_d, \theta, r) = \sqrt{N} \sum_{n=1}^{N} a_n(t, \theta, r)\, u_n(t - t_d) \exp(j2\pi f_d t) \quad (8)$$

where f_d denotes the Doppler frequency. When the received signal is matched with the assumed parameters $(t_d', f_d', \theta', r')$, the output of the matched filter is used to define FDA radar ambiguity function as expressed in (9), where the cross-ambiguity function is

$$\chi_{n,n'}(t_d, f_d) \triangleq \int_0^{T_\psi} u_n(t)\, u_{n'}{}^*(t + t_d) \exp(j2\pi f_d t)\, dt \quad (10)$$

The following summarized FDA radar characteristics:

(1) If the frequency increment Δf_n is fixed, the beam direction is a function of range r dependent.

(2) If the range r is fixed, the beam direction is a function of Δf_n dependent.

(3) If the frequency increment Δf_n is not employed $(\Delta f_n = 0)$, the corresponding FDA simplified to phased array radar.

(4) As the frequency increment Δf_n influences the beamwidth, a higher resolution may be achieved for the FDA as compared to the phased array radar.

By using the expression of the FDA radar ambiguity function derived in (9), the phased array ambiguity function can also be derived in a similar manner as in (11)

4 Simulation Results And Discussions

For the simulations, we consider a uniform linear FDA and phased array antenna with 16 elements spaced by d. The radar operational frequency is $10GHz$, frequency increment adopted is $30kHz$, $r = 10kHz$ and $\theta = \frac{\pi}{3}$.

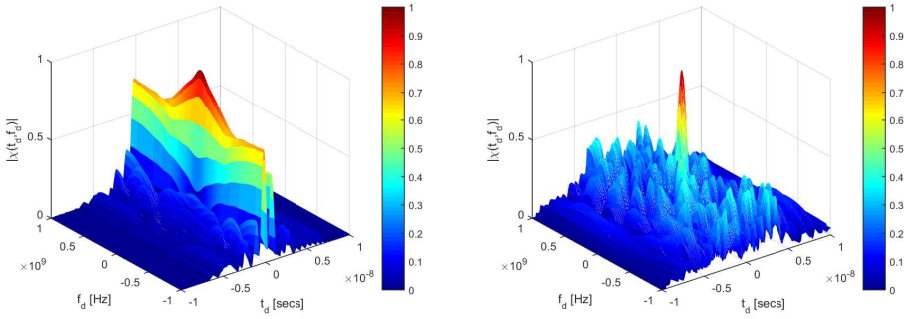
Figure 1(a) shows phased array radar ambiguity function when $\theta = \frac{\pi}{3}$ is adopted and Figure 1(b) depicts FDA radar ambiguity function when $r = 10kHz$ and $\theta = \frac{\pi}{3}$ is also adopted. It can be seen in Figure 1(a) that the phased-array radar has range-Doppler coupled and defocused ambiguity function. This implies that the phased-array radar cannot effectively distinguish range-angle dependent targets, while in Figure 1(b), FDA radar ambiguity function accumulates into a spike with relatively low sidelobe peak. It is evident from Figure 1(b) that, FDA radar can effectively distinguish range-angle dependent targets.

Figure (2a) and 2(b) illustrates the ambiguity functions in time delay and Doppler frequency dimensions. It can be observed in Figure 2(b) that by employment of frequency increment in FDA radar system produces a more focused peak which is useful in target localization than the phased array radar in Figure 2(a).

Figure 3(a) and 3(b) shows the comparisons of ambiguity function versus time delay of phased-array and FDA

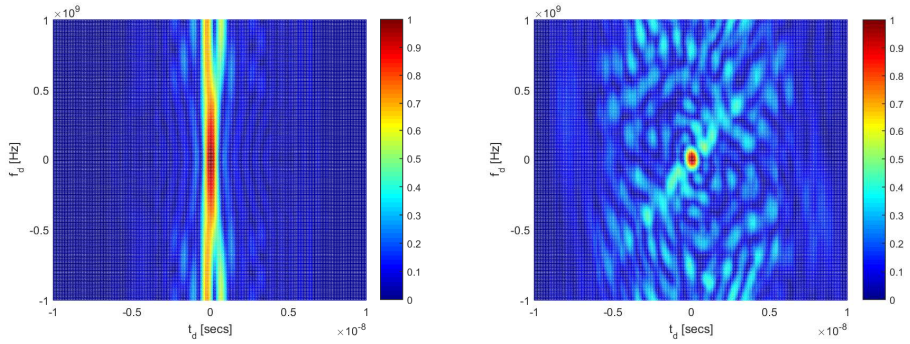
$$\chi(t_d, f_d, \theta, r, t_d', f_d', \theta', r') = \int_{-\infty}^{\infty} \hat{r}_s^H(t, t_d', f_d', \theta', r')\, \hat{r}_s(t, t_d, f_d, \theta, r)\, dt$$
$$= N \cdot \sum_{n=1}^{N} \sum_{n'=1}^{N} \left\{ \exp\left\{ j2\pi \left(\begin{array}{l} T_0(\Delta f_n - \Delta f_{n'}) - \frac{(\Delta f_n - \Delta f_{n'})}{c} \\ + \frac{f_0 d((n-1)\sin\theta - (n'-1)\sin\theta')}{c} \end{array} \right) \right\} \chi_{n,n'}(t_d, f_d) \right\} \tag{9}$$

$$\chi(t_d, f_d, \theta, t_d', f_d', \theta') = \int_{-\infty}^{\infty} \hat{r}_s^H(t_d', f_d', \theta')\, \hat{r}_s(t_d, f_d, \theta,)\, dt$$
$$= N \cdot \sum_{n=1}^{N} \sum_{n'=1}^{N} \left\{ \exp\left\{ j2\pi \left(\frac{f_0 d((n-1)\sin\theta - (n'-1)\sin\theta')}{c} \right) \right\} \chi_{n,n'}(t_d, f_d) \right\} \tag{11}$$



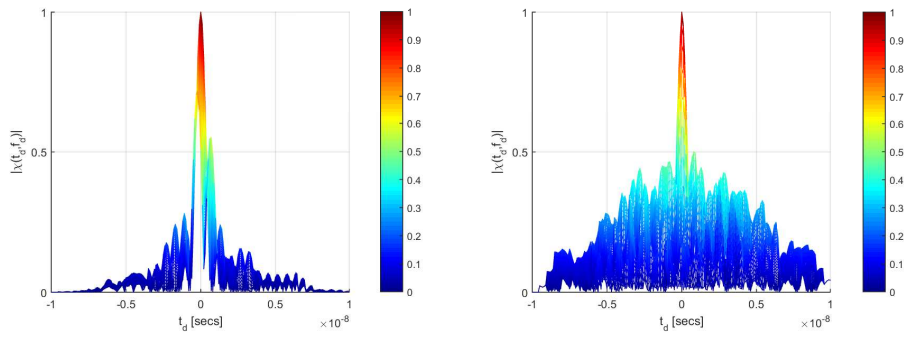
(a) (b)

Figure 1: Comparisons of 3-D ambiguity functions: a) Phased-array radar, b) FDA radar.



(a) (b)

Figure 2: Comparisons of ambiguity functions: a) Phased-array radar, b) FDA radar.



(a) (b)

Figure 3: Comparisons of ambiguity function versus time delay: a) Phased-array radar, b) FDA radar.

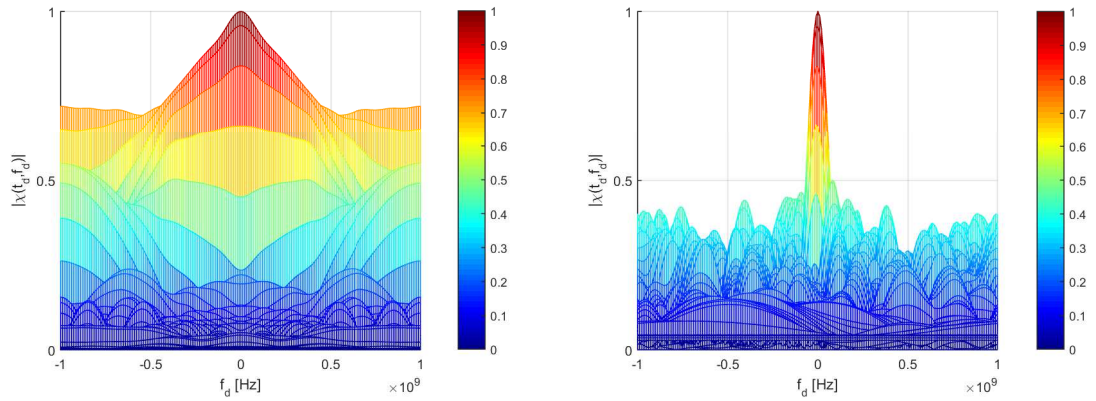


Figure 4: Comparisons of ambiguity function versus Doppler frequency: a) Phased-array radar, b) FDA radar.

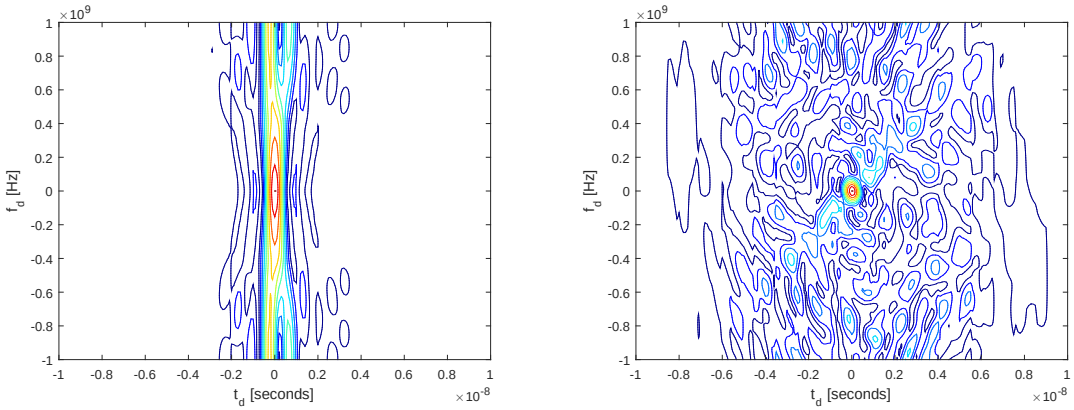


Figure 5: Comparisons of contour plot: a) Phased-array radar, b) FDA radar.

radar, respectively. It can be noticed in Figure 3(a) that the phased-array radar produces narrow beamwidth with maximum sidelobe peak ratio of 0.5510 dB. On the other hand, the FDA radar in Figure 3(b) also achieves a narrow beamwidth with maximum sidelobe peak ratio of 0.5003 dB. Hence, FDA radar outperforms phased array in terms of sidelobe peak ratio.

Figure 4(a) and 4(b) compares the ambiguity function versus Doppler frequency of phased-array and FDA radar, respectively. It can be observed in Figure 4(a) that the phased-array radar produces poor Doppler frequency resolution which is coupled. In contrast, the FDA radar in Figure 4(b) produces a narrow beamwidth Doppler frequency resolution with maximum sidelobe peak ratio of 0.5003 dB. Therefore, FDA radar can estimates Doppler frequency effectively than phased array radar.

Figure 5(a) and 5(b) examines the contour plots of phased-array and FDA radar systems, respectively. It can be seen in Figure 5(b) that the FDA radar contour plot produces a narrow resolution in time delay and Doppler frequency dimensions than the phased array radar in Figure 5(a). It is also evident from Figure 5(b) that the fuzzy area will reduce with increase in N, which would also finally affect the time delay and Doppler frequency resolutions of the system.

Finally, an effective metric used to evaluate the ambiguity function is the empirical cumulative distribution function (ECDF), which represents the percentage of samples of $|\chi(t_d, f_d)|$ lower than a given magnitude. Figure 6 illustrates the ECDF curve to compare both radar systems. The ECDF performance means that the FDA radar yields smaller sidelobe energy than the phased array radar.

5 Conclusion

In this paper, FDA radar and Phased-array radar ambiguity functions has been formulated. Simulation results shows that the phased-array radar has range-Doppler coupled and defocused ambiguity functions. This implies that the phased-array radar cannot effectively distinguish range-angle dependent targets. In contrast, the FDA radar ambiguity obtained lower maximum sidelobe level without degrading the mainlobe performance and can effectively distinguish range-angle dependent targets. In addition, the performance analysis has also been done in terms of empirical cumulative distribution function (ECDF) which proved that FDA radar ambiguity function outperforms the phased array radar ambiguity function. Furthermore, FDA radar ambiguity function produces narrow resolution in time delay and

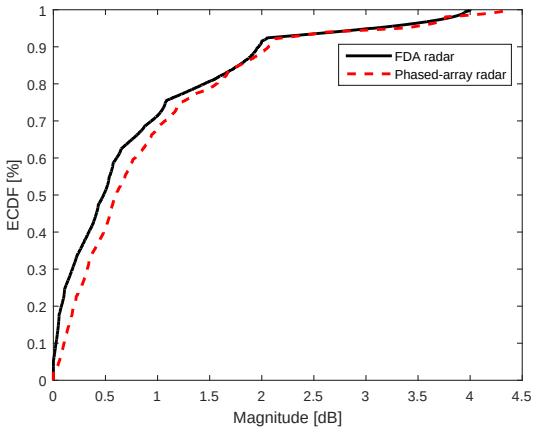


Figure 6: Comparisons of ECDF performances.

Doppler frequency dimensions than its counterpart phased-array radar.

Conflict of Interest The author declare no conflict of interest.

Acknowledgment The author would like to thank Respected Professor Wen-Qin Wang (Senior Member - IEEE) of UESTC for his great contributions, support and encouragement.

References

[1] K. T. Wong, Y. I. Wu, Y. S. Hsu, and Y. Song. "A lower bound of DOA estimates by an array randomly subject to sensor-breakdown,"IEEE Sensors J., vol. 12, no. 5, pp. 911-913, May 2012.

[2] N. I. Giannoccaro and L. Spedicato. "A new strategy for spatial reconstruction of orthogonal planes using a rotating array of ultrasonic sensors,"IEEE Sensors J., vol. 12, no. 5, pp. 1307-1316, May 2012.

[3] S. Yong and J. T. Bernhard. "A pattern reconfigurable null scanning antenna,"IEEE Trans. Antennas Propag., vol. 60, no. 10, pp. 4538-4544, Oct. 2012.

[4] J. J. Adams and J. T. Bernhard. "A modal approach to tuning and bandwidth enhancement of an electrically small antenna,"IEEE Trans. Antennas Propag., vol. 59, no. 4, pp. 1085-1092, Apr. 2011.

[5] W.-Q. Wang. "Mitigating range ambiguities in high PRF SAR with OFDM waveform diversity,"IEEE Geosci. Remote Sens. Letter, vol. 10, no. 1, pp. 101-105, Jan. 2013.

[6] P. Antonik, M. C. Wicks, H. D. Griffiths, and C. J. Baker. "Frequency diverse array radars,"in Proc. IEEE Radar Conf., Verona, Italy, Apr. 2006, pp. 215-217.

[7] P. Antonik, M. C. Wicks, H. D. Griffiths, and C. J. Baker. "Multi-mission multi-mode waveform diversity,"in Proc. IEEE Radar Conf., Verona, Italy, Apr. 2006, pp. 580-582.

[8] P. Antonik, M. C. Wicks. "Method and apparatus for simultaneous synthetic aperture and moving target indication,"U.S. Patent 0129584, Jun. 5, 2008.

[9] P. Antonik. "An investigation of a frequency diverse array,"Ph.D. dissertation, Dept. Electron. Electr. Eng., Univ. College London, London, U.K., 2009.

[10] A. Aytun. "Frequency diverse array radar,"M.S. thesis, Dept. Comput. Sci., Naval Postgraduate School, Monterey, CA, USA, 2010.

[11] S. Brady. "Frequency diverse array radar: Signal characterization and measurement accuracy,"M.S. thesis, Dept. Math., Air Force Inst. Technol., Patterson AFB, OH, USA, 2010.

[12] P. Antonik, M. C. Wicks, H. D. Griffiths, and C. J. Baker. "Range dependent beamforming using element level waveform diversity,"in Proceedings of the International Waveform Diversity and Design Conference,LasVegas, USA, January 2006, pp. 1-4.

[13] S. Mustafa, D. Simsek, and H. A. E. Taylan. "Frequency diverse array antenna with periodic time modulated pattern in range and angle,"in Proceedings of the IEEE Radar Conference, Boston, April 2007, pp. 427-430.

[14] S. Huang, K. F. Tong, and C. J. Baker. "Frequency diverse array: Simulation and design,"in Proceedings of the LAPS Antennas and Propagation Conference, Loughborough, UK, May 2009, pp. 253-256.

[15] J. Farooq. "Frequency diversity for improving synthetic aperture radar imaging,"Ph.D. dissertation, Air Force Institute of Technology, 2009.

[16] W.-Q. Wang and H. C. So. "Transmit subaperturing for range and angle estimation in frequency diverse array radar,"IEEE Transactions on Signal Processing, vol. 62, no. 8, pp. 2000-2011, April 2014.

[17] Y. B. Wang, W.-Q. Wang, and H. Chen. "Linear frequency diverse array manifold geometry and ambiguity analysis,"IEEE Sensors Journal, vol. 15, no. 2, pp. 984-993, February 2015.

[18] N. Levanon and E. Mozeson. "Radar Signals,"New York, NY, USA: Wiley-IEEE Press, 2004.

[19] N. Levanon and E. Mozeson. "Removing autocorrelation sidelobes by overlaying orthogonal coding on any train of identical pulses,"IEEE Trans., Aerospace and Electronics Systems, vol. 39, no. 2, April 2003.

[20] Dennis W. Ricker. "Echo signal processing,"The Springer International Series in Engineering and Computer Science, vol.725, 2003.

[21] Lu Zhang, Bo Yang, and Mao Kang Luo. "Joint delay and Doppler shift estimation for multiple targets using exponential ambiguity function,"IEEE Transactions on Signal Processing, DOI 10.1109/TSP.2017.2655494, pp. 1–13, 2017.

[22] W.-Q. Wang, H. C. So, and H. Z. Shao. "Nonuniform frequency diverse array for range-angle imaging of targets,"IEEE Sensors Journal, vol. 14, no. 8, pp. 2469-2476, August 2014.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 2, No. 3, 395-400 (2017)
www.astesj.com
Special issue on Recent Advances in Engineering Systems

ASTES Journal
ISSN:2415-6698

Adaptive Discrete-time Fuzzy Sliding Mode Control For a Class of Chaotic Systems

Hanene Medhaffar[*,1], Moez Feki[2], Nabil Derbel[1]

[1]*Research Unit ICOS, CEMLab, Sfax Engineering School, University of Sfax, Tunisia*

[2]*Laboratoire de Mathématiques: Modélisation Déterministe et Aléatoire (LAMMDA), ESSTHS, University of Sousse, Tunisia*

ARTICLEINFO

Article history:
Received: 04 April, 2017
Accepted: 01 May, 2017
Online: 22 May, 2017

Keywords:
Sliding mode control
Adaptive fuzzy system
Discrete time chaotic system

ABSTRACT

In this paper, we propose an adaptive discrete-time fuzzy sliding mode control for a class of chaotic systems. For this aim, a discrete sliding mode controller and a fuzzy system are combined to achieve an adequate control. The Laypunov stability theorem is used to testify the asymptotic stability of the whole system and the consequence parameters of the adaptive fuzzy system are tuned on-line by adaptive laws. The simulation example of the 3D Henon chaotic model is giving to confirm the effectiveness and the robustness of the proposed method.

1 Introduction

In recent years, the control of chaotic systems has increasingly interested researchers. The first chaos control method has been proposed by Ott et al. [1], nowadays known as the OGY (Ott-Grebogi-Yorke) method. This is a discrete technique that considers small perturbations applied in one system parameter when the trajectory approaches the vicinity of the desired orbit when crossing a specific surface. Since then, numerous control techniques have been proposed for controlling chaos in different chaotic systems such as backstepping [2–4], adaptive control algorithms [5–7] and sliding-mode control [8–10].

The sliding mode control (SMC) has undergone extensive and detailed studies in the last three decades. It is noted that SMC is a powerful robust control strategy treating the model uncertainties and external disturbances. The design and the implementation of discrete time sliding mode control have later been considered, and still in progress because a large classes of continuous systems are controlled by digital signal processors and microprocessors. Indeed, discrete sliding mode control is well studied in the literature [11–14]. The aim of this work is to give a further contribution in this field. The main objective is to design a discrete time sliding mode control strategy and enhanced by a stable adaptive fuzzy inference system to cope with modeling inaccuracies and external disturbances that can arise. Many researches on introducing the concept of fuzzy logic and especially fuzzy approximators into sliding mode control have been developed in the past years [15–18]. Historically, fuzzy logic systems have been proposed in an attempt to control nonlinear systems whose parameters are inaccuracy, and presenting neglected dynamics as well as time varying systems [19]. Since then, fuzzy logic control becomes an active research area and it has been implemented in several industrial applications. Accordingly, fuzzy logic with discrete sliding mode control are proposed to improve system control performances. Thus, the overall control system drives the tracking error to zero even in the presence of external disturbances. This paper is organized as follows: Section 2 and 3 deal with discrete sliding mode and fuzzy system respectively. Moreover, the detailed design procedure of fuzzy sliding mode controller is explained in section 4. Numerical simulations are carried out in section 5 for illustration and verification of the proposed controller. Finally some concluding remarks are given in section 6.

2 Discrete-time sliding mode controller

Consider the following nonlinear discrete time system

$$\begin{cases} x_1(k+1) = x_2(k) \\ x_2(k+1) = x_3(k) \\ \vdots \\ x_{n-1}(k+1) = x_n(k) \\ x_n(k+1) = f(x(k)) + u(k) + d(k) \end{cases} \quad (1)$$

[*]Corresponding Author: Hanene Medhaffar, Email:medhaffar_hanene@yahoo.fr, hanene.medhaffar@isgis.rnu.tn

https://dx.doi.org/10.25046/aj020351

where $x(k) = [x_1(k), x_2(k), \ldots, x_n(k)]^T$ is the state vector and $u(k)$ is the input signal.
$f(x(k))$ is unknown but it is a bounded function and $d(k)$ is a bounded external disturbance such that $|d(k)| < D$.
When $u(k) = 0$, system (1) behaves chaotically. Therefore, the aim of this work is to apply a discrete-time sliding mode controller $u(k)$ in order to track a desired trajectory.
Let $x_d(k) = [x_{d1}(k), x_{d2}(k), \ldots, x_{dn}(k)]^T$ the well known desired trajectory. Then, the tracking error can be expressed as

$$e(k) = x(k) - x_d(k) \tag{2}$$

where $e(k) = [e_1(k), e_2(k), \ldots, e_n(k)]^T$.
The sliding surface can be defined as

$$\begin{aligned} s(k) &= c_1 e_1(k) + c_2 e_2(k) + \ldots + c_{n-1} e_{n-1}(k) + e_n(k) \\ &= Ce(k) = 0 \end{aligned} \tag{3}$$

where $C = [c_1, c_2, \ldots, c_{n-1}, 1]$ can be selected as $h(z) = z^{n-1} + c_{n-1} z^{n-2} + \ldots + c_2 z + c_1$ is stable. The sliding mode controller is designed by adopting the reaching law defined by [12]

$$\Delta s(k+1) = s(k+1) - s(k) = -Qs(k) - Ksign(s(k)) \tag{4}$$

where $0 < Q < 1$ and $K > 0$.
If $f(x(k))$ is supposed known and $d(k) = 0$ then

$$s(k+1) = \sum_{i=1}^{n-1} c_i e_i(k+1) + f(x(k)) + u(k) - x_{dn}(k+1) \tag{5}$$

therefore $\Delta s(k+1)$ can be expressed as

$$\begin{aligned} \Delta s(k+1) &= \sum_{i=1}^{n-1} c_i e_i(k+1) + f(x(k)) + u(k) \\ &- x_{dn}(k+1) - \sum_{i=1}^{n-1} c_i e_i(k) - e_n(k) \end{aligned} \tag{6}$$

The equivalent control $u_{eq}(k)$ can be derived by using $\Delta s(k+1) = 0$ such that

$$\begin{aligned} u_{eq}(k) &= -\sum_{i=1}^{n-1} c_i e_i(k+1) - f(x(k)) + x_{dn}(k+1) \\ &+ \sum_{i=1}^{n-1} c_i e_i(k) + e_n(k) \end{aligned} \tag{7}$$

Since $d(k) \neq 0$ then a switching type control must be added such that

$$\begin{aligned} u(k) &= u_{eq}(k) + u_s(k) \\ &= -\sum_{i=1}^{n-1} c_i e_i(k+1) - f(x(k)) + x_{dn}(k+1) \\ &+ \sum_{i=1}^{n-1} c_i e_i(k) + e_n(k) + u_s(k) \end{aligned} \tag{8}$$

where $u_s(k)$ is defined by

$$u_s(k) = -Qs(k) - Ksign(s(k)) \tag{9}$$

where the switching gain K will be determined afterwards .
Furthermore, if $f(x(k))$ is unknown then a fuzzy system $\hat{f}(x(k))$ will be used to approximate $f(x(k))$ in order to obtain the sliding mode control law. Moreover, an adaptive adjusting law will be designed.

3 Fuzzy system

The knowledge base for the fuzzy logic system comprises a collection of fuzzy IF-THEN rules of the form:

$$\begin{aligned} \mathbf{R_1^{(j)}}: \quad & \text{If} \quad \left[(x_1(k) \text{ is } A_1^j) \ldots \text{and } (x_n(k) \text{ is } A_n^j)\right] \\ & \text{then} \quad \left[(y(x(k)) = b^j)\right] \end{aligned} \tag{10}$$

pour $j = 1, \ldots, H$. H is the rule number of the fuzzy logic system.
$x_i(k)$, $i = 1, \ldots, n$ and $y(x(k))$ denote the linguistic variables associated with the inputs and the output of the fuzzy logic system.
By the use of the singleton fuzzification strategy, product inference and center-average defuzzification, the output of the fuzzy system is expressed as:

$$y(x(k)) = \frac{\sum_{j=1}^{H} y^j(x(k)) \left(\prod_{i=1}^{n} \mu_{A_i^j}(x_i(k)) \right)}{\sum_{r=1}^{H} \left(\prod_{p=1}^{n} \mu_{A_p^r}(x_p(k)) \right)} \tag{11}$$

where $\mu_{A_i^j}$ is the membership function of the linguistic variable A_i^j.
By introducing the concept of fuzzy basis function vector $\xi(x(k))$, (11) can be rewritten as:

$$y(x(k)) = \theta^T \xi(x(k)) \tag{12}$$

where

$$\theta = \begin{bmatrix} y^1(x(k)) \\ : \\ y^H(x(k)) \end{bmatrix}, \quad \xi(x(k)) = \begin{bmatrix} \xi^1(x(k)) \\ : \\ \xi^H(x(k)) \end{bmatrix} \tag{13}$$

and

$$\xi^j(\mathbf{x}) = \frac{\prod_{i=1}^{n} \mu_{A_i^j}(x_i(k))}{\sum_{r=1}^{H} \left(\prod_{p=1}^{n} \mu_{A_p^r}(x_p(k)) \right)} \tag{14}$$

$\xi^j(x(k))$ which are called fuzzy basis functions (FBF's). It has been proved that these FBF's are universal approximators [20].

4 Fuzzy Sliding mode control

In order to derive the sliding mode control, the fuzzy system $\hat{f}(x(k)/\theta_f)$ is used to approximate $f(x(k))$ [21]. The fuzzy logic system $\hat{f}(x(k)/\theta_f)$ is expressed by:

$$\hat{f}(x(k)/\theta_f) = \theta_f^T \xi_f(x(k)) \tag{15}$$

where $\xi_f(x(k))$ is the vector of fuzzy basis supposed to be fixed, while parameters θ_f are variables which will be designed by adaptive laws.
Let θ_f^* the optimal parameter vectors of the fuzzy logic system. Minimum approximation error can be defined as follows:

$$w_f(k) = f(x(k)) - \hat{f}(x(k)/\theta_f^*) + d(k) \tag{16}$$

Then, we can select the control law as:

$$\begin{aligned} u(k) &= u_{eq}(k) + u_s(k) \\ &= -\sum_{i=1}^{n-1} c_i e_i(k+1) - \hat{f}(x(k)/\theta_f) + x_{dn}(k+1) \\ &+ \sum_{i=1}^{n-1} c_i e_i(k) + e_n(k) - Qs(k) - Ksign(s(k)) \end{aligned} \tag{17}$$

where $0 < Q < 1$ and K will be determined afterwards. Therefore, $\Delta s(k+1)$ can be rewritten as

$$\begin{aligned} \Delta s(k+1) &= f(x(k)) - \hat{f}(x(k)/\theta_f) - Qs(k) - Ksign(s(k)) \\ &= \hat{f}(x(k)/\theta_f^*) - \hat{f}(x(k)/\theta_f) + w_f(k) \\ &- Qs(k) - Ksign(s(k)) \\ &= \Phi_f^T(k)\xi_f(x(k)) + w_f(k) \\ &- Qs(k) - Ksign(s(k)) \end{aligned} \tag{18}$$

where $\Phi_f(k)$ represent the fuzzy parameter errors such that:

$$\Phi_f(k) = \theta_f^* - \theta_f(k) \tag{19}$$

Theorem 1 *The following adaptive law for adjusting the parameter vector* θ_f

$$\Delta\theta_f(k) = -\alpha\xi_f(x(k))s(k) \tag{20}$$

asymptotically stabilizes system (1) *controlled by* (17), *where* α *is a positive constant which determines the rate of adaptation.*

Proof: The Lyapunov function candidate is chosen as

$$V(k) = \frac{1}{2}s^2(k) + \frac{1}{2\alpha}\Phi_f^T(k-1)\Phi_f(k-1) \tag{21}$$

Then, $\Delta V(k)$ can be obtained as

$$\begin{aligned} \Delta V(k+1) &= V(k+1) - V(k) \\ &= \frac{1}{2}s^2(k+1) - \frac{1}{2}s^2(k) + \frac{1}{2\alpha}\Phi_f^T(k)\Phi_f(k) \\ &- \frac{1}{2\alpha}\Phi_f^T(k-1)\Phi_f(k-1) \end{aligned} \tag{22}$$

Let

$$\Delta\tilde{\theta}_k = \frac{1}{2\alpha}\Phi_f^T(k)\Phi_f(k) - \frac{1}{2\alpha}\Phi_f^T(k-1)\Phi_f(k-1) \tag{23}$$

then

$$\begin{aligned} \Delta V(k+1) &= \frac{1}{2}s^2(k+1) - \frac{1}{2}s^2(k) + \Delta\tilde{\theta}_k \\ &= \frac{1}{2}(\Delta s(k+1) + s(k))^2 - \frac{1}{2}s^2(k) + \Delta\tilde{\theta}_k \\ &= \frac{1}{2}\Delta s(k+1)^2 + s(k)\Delta s(k+1) + \Delta\tilde{\theta}_k \\ &= \frac{1}{2}\Delta s(k+1)^2 + \Delta\tilde{\theta}_k \\ &+ s(k)\Big[\Phi_f^T(k)\xi_f(x(k)) + w_f(k) - Qs(k) - Ksign(s(k))\Big] \\ &= \frac{1}{2}\Delta s(k+1)^2 + s(k)\Phi_f^T(k)\xi_f(x(k)) + s(k)w_f(k) \\ &- Qs^2(k) - Ks(k)sign(s(k)) + \Delta\tilde{\theta}_k \end{aligned} \tag{24}$$

Moroever, $\Delta\tilde{\theta}_k$ can be transformed as

$$\begin{aligned} \Delta\tilde{\theta}_k &= \frac{1}{2\alpha}\Phi_f^T(k)\Phi_f(k) - \frac{1}{2\alpha}\Phi_f^T(k-1)\Phi_f(k-1) \\ &= \frac{1}{2\alpha}\Phi_f^T(k)\Phi_f(k) \\ &- \frac{1}{2\alpha}(\Phi_f(k) - \Delta\Phi_f(k))^T(\Phi_f(k) - \Delta\Phi_f(k)) \\ &= \frac{1}{\alpha}\Phi_f^T(k)\Delta\Phi_f(k) - \frac{1}{2\alpha}\Delta\Phi_f^T(k)\Delta\Phi_f(k) \end{aligned} \tag{25}$$

Substituting (25) into (24), we obtain

$$\begin{aligned} \Delta V(k+1) &= \frac{1}{2}\Delta s(k+1)^2 + s(k)\Phi_f^T(k)\xi_f(x(k)) \\ &+ s(k)w_F(k) - Qs^2(k) - Ks(k)sign(s(k)) \\ &+ \frac{1}{\alpha}\Phi_f^T(k)\Delta\Phi_f(k) - \frac{1}{2\alpha}\Delta\Phi_f^T(k)\Delta\Phi_f(k) \\ &= \frac{1}{2}\Delta s(k+1)^2 + s(k)w_f(k) - Qs^2(k) - Ks(k)sign(s(k)) \\ &+ \Phi_f^T(k)(s(k)\xi_f(x(k)) + \frac{1}{\alpha}\Delta\Phi_f(k)) \\ &- \frac{1}{2\alpha}\Delta\Phi_f^T(k)\Delta\Phi_f(k) \end{aligned} \tag{26}$$

By applying the adaptive law (20), equation (26) can be rewritten as

$$\begin{aligned} \Delta V(k+1) &= \frac{1}{2}\Delta s(k+1)^2 + s(k)w_F(k) - Qs^2(k) \\ &- Ks(k)sign(s(k)) - \frac{1}{2\alpha}\Delta\Phi_f^T(k)\Delta\Phi_f(k) \end{aligned} \tag{27}$$

According to (18), we have

$$\begin{aligned} |\Delta s(k+1)| &\leq |\Phi_f^T(k)\xi_f(x(k))| + |w_f(k)| \\ &+ |Qs(k)| + |Ksign(s(k))| \end{aligned} \tag{28}$$

Furthermore, it's obvious that $\xi_f(x(k))$ remains bounded such that

$$\|\xi_f(x(k))\| \leq \psi_\xi \quad \forall k \geq 0 \tag{29}$$

where ψ_ξ is a positive constant.
According to (20) and (29) and from $k > N_0$, we have

$$|\Delta\theta_f(k)| \leq \psi_\theta|s(k)| \tag{30}$$

where ψ_θ is positive constant.
Consequently, we can deduce that

$$|\Phi_f(k)| \leq \psi_\Phi|s(k)| \tag{31}$$

It is obvious that the term $|w_f(k)| \leq \psi_w$ where ψ_w is a positive constant.
If we define $u_{sgn} = -Ksign(s(k))$ then

$$\begin{aligned} |\Delta s(k+1)| &\leq \psi_\xi \psi_\Phi |s(k)| + \psi_w + Q|s(k)| + |u_{sgn}| \\ &\leq (\psi_\xi \psi_\Phi + Q)|s(k)| + \psi_w + |u_{sgn}| \\ &\leq \beta|s(k)| + (\psi_w + K) \end{aligned} \quad (32)$$

where $\beta = (\psi_\xi \psi_\Phi + Q)$. By taking square from both sides of (32), we can get

$$\begin{aligned} |\Delta s(k+1)|^2 &\leq \beta^2|s(k)|^2 + (\psi_w + K)^2 \\ &+ 2\beta|s(k)|(\psi_w + K) \end{aligned} \quad (33)$$

Then,

$$\begin{aligned} \frac{1}{2}|\Delta s(k+1)|^2 &\leq \frac{1}{2}\beta^2|s(k)|^2 + \frac{1}{2}(\psi_w + K)^2 \\ &+ \beta|s(k)|(\psi_w + K) \end{aligned} \quad (34)$$

Thus, we can obtain

$$\begin{aligned} &\frac{1}{2}\Delta s(k+1)^2 - Qs^2(k) + s(k)u_{sgn} \\ \leq\ & \frac{1}{2}(\psi_w + K)^2 + \frac{\beta^2}{2}|s(k)|^2 + \beta|s(k)|(\psi_w + K) \\ -\ & Q|s(k)|^2 + K|s(k)| + \psi_w|s(k)| - \psi_w|s(k)| \\ \leq\ & \frac{1}{2}(\psi_w + K)^2 + (\beta+1)|s(k)|(\psi_w + K) - \psi_w|s(k)| \\ +\ & (\frac{1}{2}\beta^2 - Q)|s(k)|^2 \end{aligned} \quad (35)$$

If we choose

$$Q > \frac{1}{2}\beta^2 \quad (36)$$

and

$$\begin{aligned} K &= \left(-(\beta+1) + \sqrt{[(\beta+1)^2 + 2(Q - \frac{1}{2}\beta^2)]}\right)|s(k)| \\ &- \psi_w \end{aligned} \quad (37)$$

then

$$\begin{aligned} \frac{1}{2}(\psi_w + K)^2 &+ (\beta+1)|s(k)|(\psi_w + K) \\ &+ (\frac{1}{2}\beta^2 - Q)|s(k)|^2 = 0 \end{aligned} \quad (38)$$

Therefore, (27) can be expressed as

$$\begin{aligned} \Delta V(k+1) &\leq s(k)w_f(k) - \psi_w|s(k)| \\ &- \frac{1}{2\alpha}\Delta\Phi_f^T(k)\Delta\Phi_f(k) \end{aligned} \quad (39)$$

It's clear that $s(k)w_f(k) - \psi_w|s(k)| \leq 0$ since $|w_f(k)| \leq \psi_w$, furthermore $\Delta\Phi_f^T(k)\Delta\Phi_f(k) \geq 0$, then

$$\Delta V(k+1) \leq 0 \quad (40)$$

By using the Barbalat's lemma [23], we can readily prove that $s(k) \to 0$ as $k \to +\infty$ thus $\lim_{k\to+\infty} e(k) = 0$.

5 Simulation results

To illustrate the above design approach, a 3D Henon chaotic model is considered.
The model for the discrete-time 3D Henon map is given as [22]

$$\begin{cases} x_1(k+1) &= x_2(k) \\ x_2(k+1) &= x_3(k) \\ x_3(k+1) &= -0.2x_1(k) - 0.3x_2(k) - 1.65x_3(k) - x_3^2(k) \\ &\quad +u(k) \end{cases} \quad (41)$$

The non controlled ($u(k) = 0$) Henon map has a chaotic strange attractor. The result of computation is shown in figure 1 where $[x_1(0), x_2(0), x_3(0)]^T = [0.1, 0, 0.1]^T$.

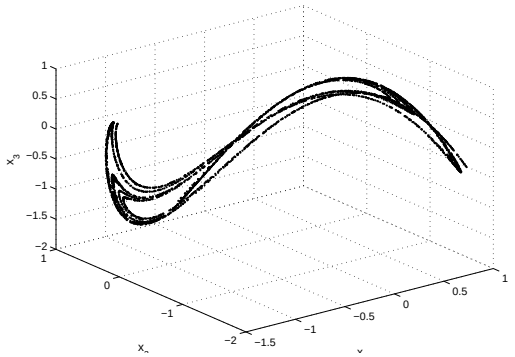


Figure 1: Hénon map attractor.

In order to control this autonomous chaotic system, controller $u(k)$, given by (8), is applied at $k = 60$ to reach the desired trajectory $x_d(k) = [0, 0, 0]^T$.
For the estimation of $f(x(k))$, we consider five fuzzy levels, i.e. NB, NS, EZ, PS and PB on x_1, x_2 and x_3. We use fuzzy logic systems with center-average defuzzifier, product inference, singleton fuzzifier and Gaussian membership functions to approximate $f(x(k))$. Slopes c_i are chosen as $c_1 = 0.1$ and $c_2 = 0.9$.
In figure 2, we represent the evolution of system states which converge rapidly to the desired values. Figure 3 represent the sliding surface and the control signal whose amplitude is always within an acceptable range compared to the system states.

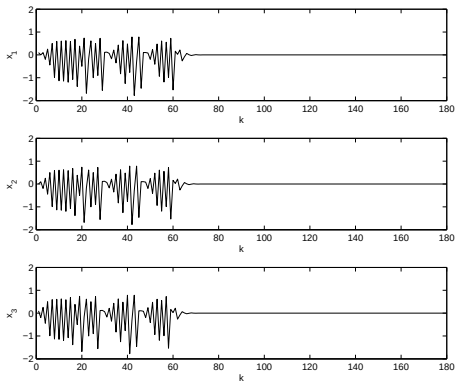


Figure 2: Evolution of system states

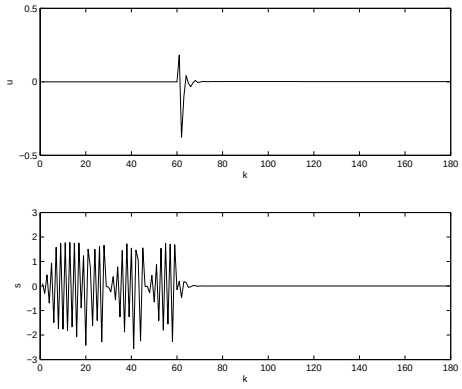


Figure 3: Representation of the sliding surface and the fuzzy sliding mode controller.

In order to illustrate the robustness of our design, a white gaussian noise with variance equal to 0.01 has been considered as an external disturbance. The control objective has been achieved as is illustrated in figures 4 and 5. Thus, the robustness of the discrete fuzzy sliding mode controller is proved.

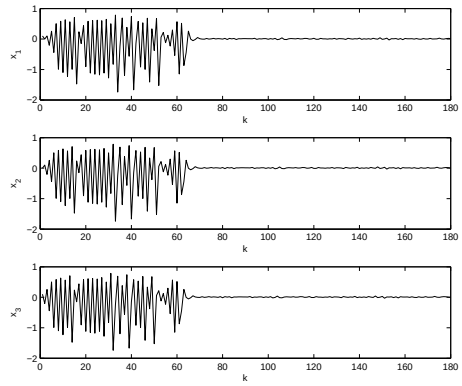


Figure 4: Evolution of system states in presence of a white gaussian noise.

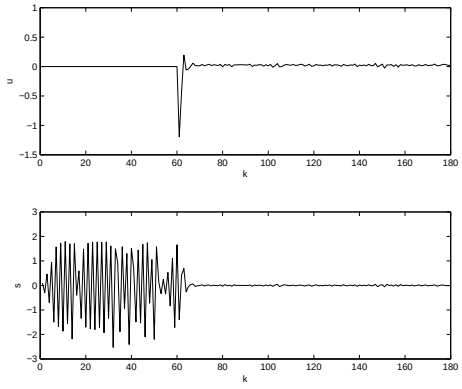


Figure 5: Representation of the sliding surface and the fuzzy sliding mode controller.

6 Conclusion

In this paper, an adaptive discrete-time fuzzy sliding mode controller is proposed. This controller, inherits the advantages of both sliding mode control and fuzzy systems. It has been proposed for stable control of a class of chaotic systems. The sliding mode control is proposed as a robust method to control nonlinear and uncertain systems. Consequence parameters of fuzzy control rules have been adjusted on-line in order to guarantee the reaching condition. Simulation results of a 3D Henon chaotic model show the applicability and the effectiveness of the proposed approach.

Conflict of Interest The authors declare no conflict of interest.

References

1. E. Ott, C. Grebogi, J. A. Yorke, "Controlling chaos", Physical Review Letters, **64**, 1196–1199, 1990.

2. X. Tan , J. Zhang, Y. Yang, "Synchronizing chaotic systems using backstepping design", Chaos, Solitons and Fractals, **16**, 37–45, 2003.

3. J. H. Park, "Synchronization of Genesio chaotic system via backstepping approach", Chaos, Solitons and Fractals, **27**(5), 1369–1375, 2006.

4. H. Adelipoor, O.A. Babaei, "Stability and Tracking in the New Chaotic System Using Backstepping Method", J. Basic. Appl. Sci. Res., **3**, 446–452, 2013.

5. H.Zhang, H. Quin, G. Chen, "Adaptive control of chaotic systems with uncertainties", Int. J. Bifurcation and Chaos, **8**(10), 2041–2046, 1998.

6. M. Feki, "Model-independent adaptive control of chua's system with cubic nonlinearity", Int. J. Bifurcation and Chaos, **14**(12), 4249–4263, 2004.

7. M. P. Aghababa, B. Hashtarkhani, "Synchronization of Unknown Uncertain Chaotic Systems Via Adaptive Control Method", Journal of Computational and Nonlinear Dynamics, **10**(5), 051004, 2015.

8. H.T. Yau, C.K. Chen, C. Li Chen , "Sliding mode control of chaotic systems with uncertainties", Int. J. Bifurcation and Chaos, **10**(5), 1139–1147, 2000.

9. J. J. Yan, Y. S. Yang, T. Y. Chiang, C. Y. Chen, "Robust synchronization of unified chaotic systems via sliding mode control", Chaos, Solitons and Fractals, **34**(3), 947–954, 2007.

10. M. Feki, "Sliding mode control and synchronization of chaotic systems with parametric uncertainties", Chaos Solitons and Fractals, **41**(3), 1390–1400, 2009.

11. K. Furuta, "Sliding mode control of a discrete systems", Systems and Control Letters, **14**(2), 145–152, 1990.

12. W. Gao, Y. Wang, A. Homaifa, "Discrete-time variable structure control systems", IEEE Trans. Ind. Electronics, **42**(2), 117–122, 1995.

13. H. Sira-Ramirez, "Non-linear discrete variable structure systems in quasi-sliding mode", International Journal of Control, **54**(5), 1171–1187, 1991.

14. S.Z. Sarpturk, Y. Istefanopulos, O. Kaynak, "On the stability of discrete-time sliding mode systems", IEEE Transaction Automatic Control, **AC-32**(10), 930–932, 1987.

15. J.Y. Chen, "Rule regulation of fuzzy sliding mode controller design: direct adaptive approach", Fuzzy sets and systems, **120**, 159–168, 2001.

16. O. Kaynak, "Fuzzy adaptive sliding mode control of a direct drive robot, Robotics and Autonomous Systems", **19**, 215–227, 1996.

17. H. Medhaffar, T. Damak, N. Derbel, "A decoupled fuzzy indirect adaptive sliding mode controller with application to robot manipulators", IJMIC, **1**(1), 23–29, 2006.

18. H. Medhaffar, T. Damak, N. Derbel, "Direct adaptive fuzzy moving sliding mode controller design for robotic manipulators", IJCIA, **5**(1), 1–20, 2005.

19. L.A. Zadeh, "Outline of a new approach to the analysis of complex systems and decision processes", IEEE Trans. Systems, Man. cybernetics, **3**(1), 28–44, 1973.

20. L.X. Wang, Adaptive fuzzy systems and control, Englewood Cliffs, NJ: Prentice-Hall, 1994.

21. B. Yoo, W. Ham, "Adaptive fuzzy sliding mode control of nonlinear systems", IEEE Trans. Fuzzy systems, **6**(2), 315–321, 1998.

22. S. V. Gonchenko, I. I. Ovsyannikov, C. Simo, D. Turaev, "Three Dimensional Hénon-like Maps and Wild Lorenz-like Attractors", International Journal of Bifurcation and Chaos, **15** (11), 3493–3508, 2005.

23. J. J. E. Slotine, W. Li, Applied Nonlinear Control, Englwood Cliffs, NJ:Prentice Hall, 1991.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 2, No. 3, 401-411 (2017)
www.astesj.com
Special issue on Recent Advances in Engineering Systems

ASTES Journal
ISSN: 2415-6698

Fault Tolerant Inverter Topology for the Sustainable Drive of an Electrical Helicopter

Igor Bolvashenkov*, Jörg Kammermann, Taha Lahlou, Hans-Georg Herzog

Technical University of Munich (TUM), Department of Electrical and Computer Engineering, Institute of Energy Conversion Technology, Arcisstr. 21, 80333 Munich, Germany

ARTICLEINFO

Article history:
Received: 05 April, 2017
Accepted: 12 May, 2017
Online: 22 May, 2017

Keywords :
Electrical Helicopter
Traction Drive
Multilevel Inverter
Thermal Behavior
Reliability
Fault Tolerance

ABSTRACT

The paper describes the results of the system analysis of various options of electric inverter topologies for the use in the traction drive of an electrical helicopter. The paper explains the optimal choice, in terms of the design requirements and future operational conditions, and the type and topology of the electric inverter to supply the traction permanent magnet synchronous motor with the batteries. The main focus has been set on the study of the overload capacity, thermal behavior, and the fault tolerance of the multilevel electric inverter. In order to estimate these parameters, the Markov models of the multi-state system's reliability evaluation were used.

1 Introduction

Nowadays, according to the long-term plans for the electrification of various types of vehicles, the task of the electrification of air traffic has now become very important. First of all, the application of electric drives is a promising one for small helicopters designed for search and rescue operations with a nominal power of 510 kW, a rotational speed of 400 rpm, and a voltage of 800 V. Regarding the choice of an optimal electric inverter topology for the traction drive the main attention was paid to the flight safety.

The simplest solution would be to maximize the number of redundant equipment, but strict limitations on weight and dimensions do not allow this. The papers [1, 2] show that only multi-phase electric motors are capable to satisfy the requirements to the fault tolerance ($\lambda \leq 10^{-9}$/h), set up for the electric traction drive of the helicopter.

Taking into account the safety and sustainability of helicopter's operations, the relevant options for inverter topologies were evaluated. Thus according to [2] for the comparative analysis the semiconductor inverters for 6-, 7-, and 9-phase motors were taken.

Due to the great relevance of the creation of electric inverters for special applications, there are many publications describing the topologies and the analysis of inverters with different levels of fault tolerance for various vehicle types. The results of these studies were taken into account by the authors at carrying out of their investigations. Hence, the main task of this research is to find the decision, how it becomes possible to achieve the required values of the inverter's fault tolerance.

2 Approach and Parameters for Evaluation

For each component of the helicopter's traction drive, based on the project requirements and conditions of future operations, the most important parameters are:

- reliability and fault tolerance;
- weight, volume, and dimensions.

2.1 Reliability and Fault Tolerance

For the safety-critical aircraft applications accurate assessment of reliability and fault tolerance indices of the electric inverter at the design stage of the helicopter is crucial for further operation.

The operating experience of electric inverters indicates that the most vulnerable elements of inverter are the capacitors, semiconductors, and printed circuit board (PCB) [3, 4]. Statistical data of failure in various parts of inverter are shown in Figure 1. The main reason of the failures is the overheating of the semiconductors caused by current overloads, as can be seen in Figure 2.

*Phone number: +49 89 28928428, e-mail: igor.bolvashenkov@tum.de

https://dx.doi.org/10.25046/aj020352

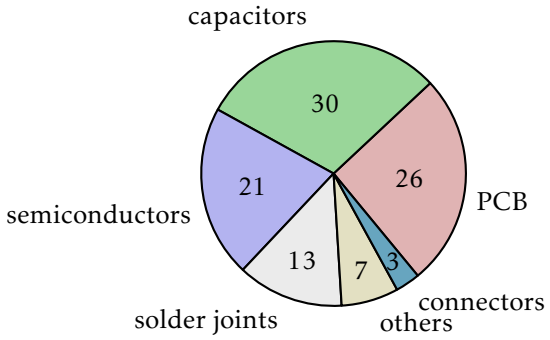


Figure 1: Failure statistics of the components [3].

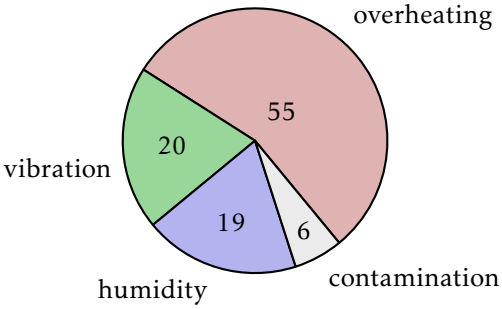


Figure 2: Causes for the failures [3].

Based on the above data charts, reliability level of an electric inverter λ_{EI} can generally be estimated considering the reliability values of its components by the equation:

$$\lambda_{\mathrm{EI}}(t) = \Sigma\lambda_{\mathrm{T}i}(t) + \Sigma\lambda_{\mathrm{D}j}(t) + \Sigma\lambda_{\mathrm{C}k}(t) + \Sigma\lambda_{\mathrm{B}n}(t) \qquad (1)$$

where $\lambda_{\mathrm{T}i}(t)$, $\lambda_{\mathrm{D}j}(t)$, $\lambda_{\mathrm{C}k}(t)$, and $\lambda_{\mathrm{B}n}(t)$ are the failure rates of the all components of electric inverter, respectively of transistor, diode, capacitor and printed circuit board.

In accordance with [3–6] the following failure cases were considered, shown in Figure 3.

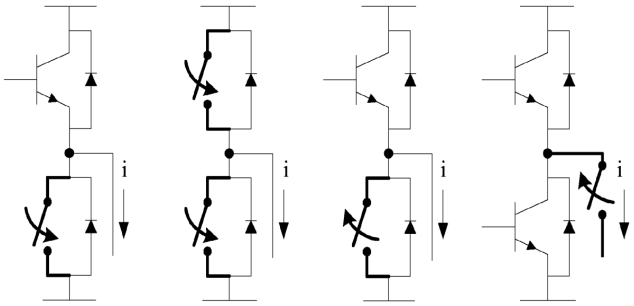


Figure 3: Considered inverter failures [6].

Thus, the considered failures are (Figure 3):

1. single inverter switch short-circuit;
2. phase-leg short-circuit;
3. single inverter switch open-circuit;
4. single phase open-circuit (internal or external).

In these cases for the fault-tolerant operation either the inverter module must be shut down or must be switched to the possibly existing redundant inverter leg or module. Each of these modes is connected with a rapid increase in the current at the emergency site and an overload of semiconductors. On the basis of data on the standard overload capacity of power electronics has been plotted the graph presented in Figure 4.

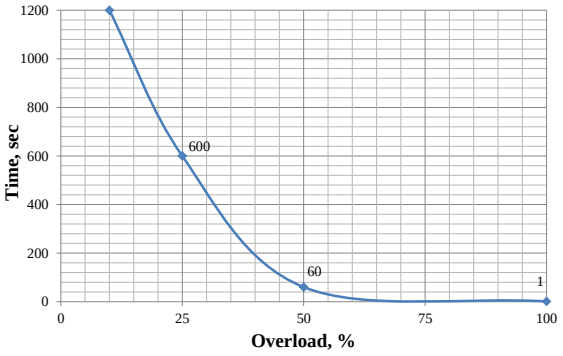


Figure 4: Overload capacity of the inverter.

The most of the fail operational modes caused by technological overloads. At the overload power losses occur in the p-n-transition and its temperature due to the low heat capacity increases dramatically. In the case of exceeding a certain critical temperature of p-n-transition, the semiconductor device fails. So overheating temperature is the main parameter characterizing the overload capacity of semiconductor devices as can be seen in Figure 5.

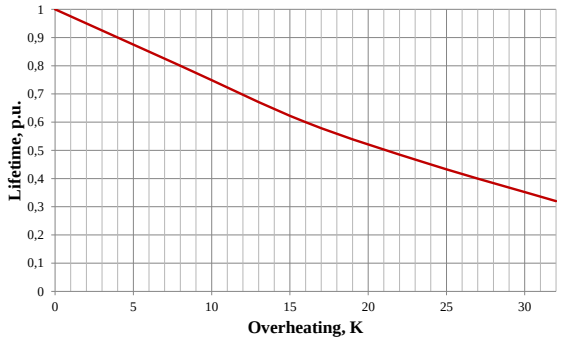


Figure 5: Dependency of the inverter's lifetime on overheating.

So, the consequences of overload in failure operational modes are overcurrent and overheating of an inverter, which lead to a reduction of the reliability indices of the motor and to a decrease in lifetime [7].

2.2 Weight, Volume, and Dimension

The modern options of the fault tolerant topologies of electrical inverters [6, 8–10] have a different number of components designed for different values of operating parameters. Therefore, the number of components, the flexibility of their placement, and the inverter parameters are largely determining such important characteristics of the helicopter, such as weight, volume, and dimensions. The comparative analysis of these values for different inverter topologies are presented in the next sections of this paper.

3 Topologies for the Comparative Analysis

The first part of this section explains the conventional topologies, such as 6-pulse-bridge (B6-bridge) and the H-bridge. However, the second part analyzes the use of multilevel inverters, i.e. cascaded H-bridge (CHB) inverters.

3.1 Conventional Full- and Half-Bridge Topologies

One of the previous papers [2] compared multi-phase permanent magnet synchronous motors (PSM) for safety-critical drive trains. As a result, 6-, 7-, and 9-phase motors are preferable in order to reach the requirements regarding fault tolerance. Hence, possible 9-phase machine topologies are presented in Figure 6, as well as in Figure 7. Those are commonly used and can be adapted to all different numbers of phases.

The already mentioned inverter topologies are compared for a failure case (open circuit) in one phase. This is also valid for a short circuit failure, if the fault is instantaneously detected and isolated, as described in [11, 12]. Hence, the advantage of H-bridges is the separation of every single phase in order to monitor failures. For this inverter topology comparison, an existing electrical machine with three phases serves as an exemplary machine for the failure simulations. At first sight, the results can relatively be adapted to other power requirements.

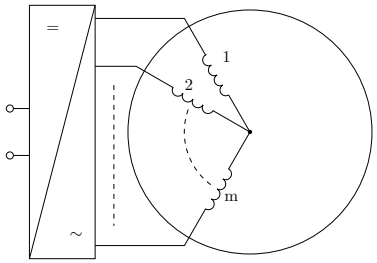


Figure 6: General visualization of a multi-phase topology with windings connected in one star.

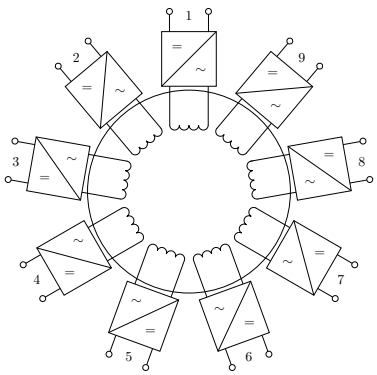


Figure 7: General visualization of a multi-phase modular topology with galvancally insulated windings.

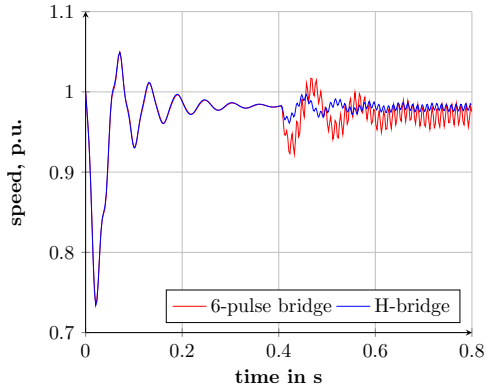


Figure 8: Simulation results for a failure case after 0.4 s for both the 6-pulse-bridge and the H-bridge considering speed over time.

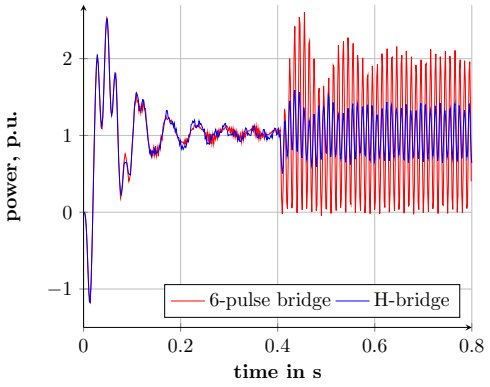


Figure 9: Simulation results for a failure case after 0.4 s for both the 6-pulse-bridge and the H-bridge considering power over time.

After the transient response of the simulation model, the failure takes place at 0.4 seconds and the results are shown in Figure 8 and Figure 9. The speed fluctuations are clearly visible for both machines in Figure 8, whereas the amplitude of the 6-pulse bridge is higher, as well as the speed ripple. Considering Figure 9, the results show, that the power or torque ripple get a lot smoother for the H-bridge, since the peaks for the B6-bridge even drop down to 0. The peak to peak value equals up to ±100% for the 6-pulse-bridge, compared to approximately ±40% for the H-bridge.

In summary, considering a one phase failure and hence fault tolerance, the H-bridge is recommendable for machines with multiple phases, because of the lower power degradation. Due to the high requirements, the machine is already highly saturated in nominal operation; hence, overload in a failure case is not applicable. Thus, in case of a 6-pulse-bridge, three phases fail, since the inverter is unable to drive two out of three phases. In contrast, each H-bridge drives one phase and consequently in case one phase fails, eight phases still operate. Assuming an equally distributed power per phase, the H-bridge topology provides 89% of the nominal power and the B6-bridge only 67% respectively for a 9-phase machine. For the 6-phase machine the situation gets even worse.

However, after first analytical approximations and

considering the application case, weight is 20%, volume 60% higher, and the power losses are slightly higher (10%) for the H-bridge compared to the 6-pulse-bridge topology. This results out of the increased number of semiconductors being necessary for the H-bridge topology. Therefore, the next section includes the analysis of a multilevel inverter trying to solve the problems regarding weight, volume, and power losses, since there is no definitely preferred option. [11–13]

3.2 Multilevel Inverter – Cascaded H-Bridge Inverter

Multilevel inverters offer several advantages compared to their two-level counterparts [14]: smaller power filters, smaller voltage ratings for semiconductors, lower switching frequencies, and less power losses. They offer also more modularity and are more reliable. The inverter with three H-bridges and the B6-bridge inverter need higher voltage range semiconductors (1000V range).

Considering [15–17], CHB inverters with the lower number of components than other classic multilevel counterparts, such as the Neutral Point Clamped (NPC) or Flying Capacitor (FC), are the best candidate for the use in electric drives. They need more components than CHB and have less modularity, due to the central storage unit. The number of components (power switches, diodes and capacitors) for one phase with 17-level output voltage steps is: 288 for the NPC, 168 for the FC and 40 for the CHB (see Figure 10).

Figure 10 shows, that the cascaded H-bridge topology has the most advantages with more than 5-level. CHB inverters offer higher modularity and have the capability for embedding energy storage units in a split manner, given the existence of several submodules operating at low voltages (see Figure 13).

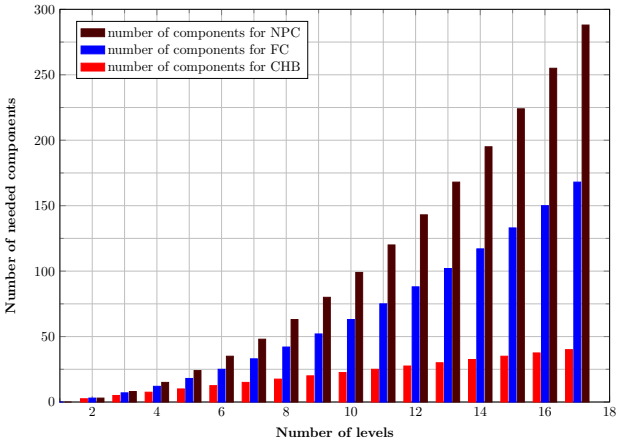


Figure 10: Number of components using three multilevel inverter topologies.

Table 1 shows the parameters of the system. It also shows the structure of the batteries used for the B6-bridge inverter with centralized energy source (battery rack) and CHB inverter with decentralized energy sources (battery modules). V_L is the phase voltage of the machine, V_{cell} the battery cell voltage, V_{mod} the battery module voltage, and V_{rack} the battery rack voltage. The capacity/energy content of the storage units is neglected.

A multilevel inverter with 17 level output voltage steps needs 8 battery modules in each phase. The number of required submodules depends on the value of the minimal phase voltage of the machine and on the minimal battery module voltage $V_{battmin}$ and is calculated as follows:

$$n_{submodules} = \frac{v_{L1}}{V_{battmin}} = \frac{353\,V}{44.8\,V} = 7.88 \approx 8 \qquad (2)$$

Using a higher number of submodules the output voltage is near to a sinusoidal wave form and the harmonics on the electrical machine are lower.

The CHB inverter uses battery modules as shown in Figure 13. Each module is connected to an H-bridge with 4 MOSFETs. The use of MOSFETs enhances the efficiency of the CHB inverter, because of the low conduction losses. Usually, IGBTs are used for a higher voltage range, such as for B6-bridge inverters.

The prototype of the three phase CHB inverter is compared to a commercial B6-bridge inverter (Siemens S120) with the same nominal power 36 kW, taking into account weight, volume, and efficiency. The power-to-volume ratio of the B6-bridge inverter including the power filter is 36 kW/30.8 l = 1.17 kW/l.

Finally, Table 2 shows the results of the comparison between B6-bridge and CHB inverter and it proofs the advantages of CHB over the B6-bridge inverter in volume and weight. The next section discusses results of reliability and fault tolerance analysis.

In conclusion and based on the given project parameters, a 17-level version with MOSFETs was selected, thus, 8 submodules per phase. [14–17]

4 Preliminary Assessment

Based on initial data and the technique for reliability evaluation described in [4,9,18–23], reliability and fault tolerance indices were analyzed. Statistical data for calculating the reliability function of the multilevel inverter in 6-, 7-, and 9-phase implementation are shown in Table 3 and Table 4.

Figure 11 shows the reliability function of the multilevel inverter. It can be seen that the probability of failure-free operation of 6-phase inverter is higher than 7- and 9-phase, due to a smaller number of components. However, the difference between the 6- and 9-phase options does not exceed 20% for 10,000 operating hours.

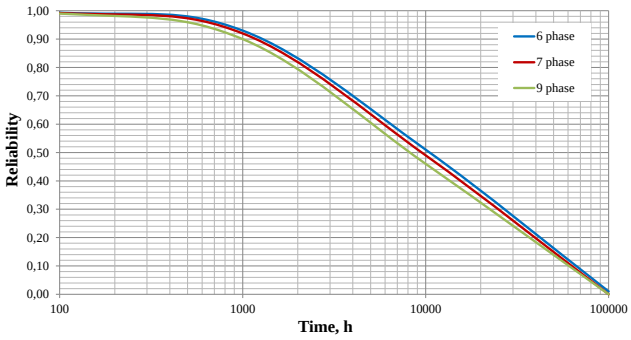


Figure 11: Reliability function of a CHB inverter.

On the basis of the schedule of acceptable inverter's overload (see Figure 4) and the method described in [24], the qualitative analysis of the inverter's ability to function was carried out in case of a failure of 1, 2, and 3 submodules in a single phase. The analysis results are summarized in Table 5 and Table 6.

Considering the requirements of the project, the most critical failure mode of the traction drive, as can be seen from Table 5, is when it is necessary to maintain stably during 2 minutes 113% of nominal power. Further, this mode will be more carefully simulated and estimated for the failure cases using a Markov model.

5 Multi-State System Reliability Markov Model

For the maximum accurate estimation of the fault tolerance value of each comparable inverter option and its compliance (or non-compliance) with design requirements, a Multi-State System Reliability Markov Model (MSS MM) was constructed, which theoretical base is described in [25, 26].

5.1 State Diagram of Markov Model

For the present study the phase open-circuit failure has been considered as the main dangerous failure for electric inverter, i.e. it is the most severe kind of failure, to which less dangerous failures can be summarized. The impact of this failure on the safety operation of the power drive is very significant. In this way, the phase open-circuit failure of a multilevel inverter's submodule can be considered as one of the main causes of failures in traction electric drives, since practically all other electrical failures of the inverter can be attributed to it [7].

Thus, the multilevel inverter can be considered as a multi-state system with a loaded functional redundancy and consequently, with an appropriate reserve of the fault tolerance. A more accurate approach to build the Markov model is described in [2, 27].

The number of the degraded states of the MSS MM has to be determined in accordance with the requirements of the project on the fault tolerance and on the technical capabilities of the object to continue functioning with reduced performance as a result of the critical failure.

Regarding the above requirements on the fault tolerance of safety-critical drives, as well as statistical data on the reliability of electric inverters, it was determined that the optimal model for the analysis of fault tolerance in such conditions is a multi-state Markov model with a minimum of five states, as shown in Figure 12. With one degradative state of inverter, it is not feasible to realize the required values of fault tolerance.

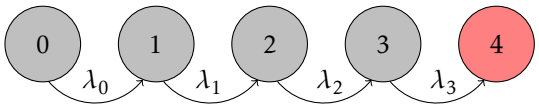


Figure 12: State diagram of MSS MM for a not repairable system.

The first state "0" corresponds to a completely failure-free operation of the inverter. The second, third, and fourth state are the states of degradation and correspond to failure cases. The last red state of the model corresponds to the failure of one phase considering the CHB (see Table 6) or respectively, a completely failed conventional inverter (see Table 5) when the traction motor does not receive the necessary power for a safe flight. Thus, every following state of the degradation corresponds to the worst case – a loss of an inverter's submodule in the same phase.

Table 1: Parameter of the system.

Parameter	Value	Unit
Machine nominal power P	510	kW
$V_{\text{L1min}} \dots V_{\text{L1nom}} \dots V_{\text{L1max}}$	$353 \dots 565 \dots 636$	V
$V_{\text{cellmin}} \dots V_{\text{cellnom}} \dots V_{\text{cellmax}}$	$1.4 \dots 2.2 \dots 2.5$	V
Battery rack structure	360s	
$V_{\text{rackmin}} \dots V_{\text{racknom}} \dots V_{\text{rackmax}}$	$500 \dots 800 \dots 900$	V
Battery module structure (1 of 24)	32s	
$V_{\text{modmin}} \dots V_{\text{modnom}} \dots V_{\text{modmax}}$	$44.8 \dots 70.4 \dots 80$	V

Table 2: Comparison of the data of the B6-bridge and CHB inverter.

Topology	B6-bridge	CHB – three phase
Power	36 kW	36 kW
Weight	31 kg	14.3 kg
Volume	30.8 l	8.08 l
Power-to-weight ratio	1.16 kW/kg	2.52 kW/kg
Power-to-volume ratio	1.17 kW/l	4.45 kW/l
efficiency	meas. 95 %	meas. 99 %

Table 3: Failure rates of inverter components.

Components	MOSFET	Diode	Capacitor	PCB	CU
Failure rate, FIT	250	100	80	200	1000

Table 4: Multilevel inverter and number of components.

Components Phase number	MOSFET	Diode	Capacitor	PCB	CU
6	192	192	48	48	1
7	224	224	56	56	1
9	288	288	72	72	1

Table 5: Comparison of overload value in failure cases for the conventional B6-bridge.

Phase number	6			7			9		
Fault number Load level	1	2	3	1	2	3	1	2	3
113 %	+	MF	MF	++	+	MF	+++	+++	++
85 %	+++	+	MF	+++	+++	+	+++	+++	+++
65 %	+++	+++	MF	+++	+++	+++	+++	+++	+++

+++ 0–10 %; ++ 10–25 %; + 25–40 % overload; MF – motor failure

Table 6: Comparison of overload value in failure cases for the CHB.

Phase number	6			7			9		
Fault number Load level	1	2	3	1	2	3	1	2	3
113 %	+	PhF	PhF	++	+	PhF	+++	+++	++
85 %	+++	++	+	+++	+++	++	+++	+++	+++
65 %	+++	+++	+++	+++	+++	+++	+++	+++	+++

+++ 0–10 %; ++ 10–25 %; + 25–40 % overload; PhF – phase failure

5.2 Transition Probabilities

The most important and difficult fact for a simulation by Markov models is to determine the transition probabilities and determination of the number of states with a reduced level of functionality.

The values of the Markov Model transition probabilities $\lambda_1, \lambda_2, \ldots, \lambda_k$ are derived from the results of calculations of the Degree of Fault Tolerance ($DOFT$) for the states $1, 2, \ldots, k$, respectively, according to the method [28] based on the analysis of an overload capacity and thermal stability of an inverter:

$$\lambda_{\mathrm{R}i} = 1 - DOFT_{\mathrm{R}i} \tag{3}$$

Here R is the value of reduced level of performance according to the project requirements and i is the number of critical failures. A large number of different factors affect the magnitude of the transition probabilities, from the environmental parameters to the used maintenance strategies, monitoring, and diagnostics.

As can be seen from (3), for the calculation of the transition probabilities the main challenge is the correct calculation of $DOFT$ for given project required performance levels. The method of determining its value in accordance with [24] is presented below.

Considering the definition of fault tolerance of a technical system as an ability to maintain the required functional level of the system, in case of one or more failures of its components, the $DOFT$ can be defined as the amount of time, in which the system may remain in a degraded state without irreversible changes in its functionality. Mathematically, in general form this can be written as in (4):

$$DOFT_{\mathrm{R}i} = \frac{W_{\mathrm{R}}}{W_{\mathrm{N}}} \cdot \frac{\Delta t_i}{\Delta t_{\mathrm{N}}} \tag{4}$$

where W_{R} and W_{N} are the reduced (R) and nominal (N) values of the performance of technical system; Δt_i and Δt_{N} are the duration of functioning after i failures and without failure, respectively.

For quantitative assessment of the overload capability and thermal stability of the inverter's submodule, their thermal behavior for various load modes has been experimentally investigated and is discussed below.

5.3 Thermal Behavior of the Inverter

The thermal stability and overload capacity of the power electronic inverter is an important part for a save and sustainable operation of an electrical helicopter. Therefore, the definition of these parameters must be performed with the great accuracy.

On the basis of an experimental study, the compliance of the graph of Figure 4 with the real values of the inverter's overload capacity was verified, as well as the legitimacy (or illegitimacy) of its use in determining the transition probabilities for MSS reliability Markov Models.

As mentioned above the optimal option, considering the project requirements is a cascaded H-bridge 17-level inverter as shown in Figure 13. It consists of 8 submodules in each phase and 9 similar phases. Each submodule consists of a battery storage unit and a MOSFET full bridge and is connected in series with other submodules.

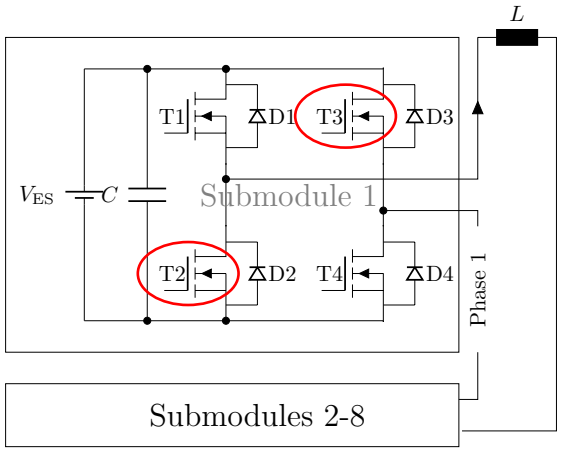


Figure 13: Multilevel CHB inverter submodule.

A full bridge used in a cascaded H-bridge multilevel inverter with lithium batteries was taken into account for an experimental verification of the normative thermal stability of the MOSFET's power inverter.

For this verification of EN61000-2-4 [29] was the full bridge tested for 20 min with 10% more current than the nominal current, 10 min with 25%, 60s with 50% more current and 1 s with 100% (double of the nominal current). Furthermore the thermal capacity of the inverter has been estimated and the thermal behavior of the full bridge was observed.

Table 7 shows the characteristics of the MOSFET type IPP023NE7N3, which has been investigated.

Figure 14 shows an experimental test bench for investigating the thermal behavior of the inverter's submodule. For the construction of the test bench, the results of the research, discussed in [30–32], were taken into account.

Figure 14: Test bench setup for the thermal tests of the full bridge.

In general, the test bench consists out of a power

Table 7: Description of the used MOSFET.

Type	IPP023NE7N3
Maximum voltage	75 V
On-state resistance	2.3 mΩ
Maximum continuous drain current	120 A

supply, the submodule, load resistors, and the measurement devices. The submodule is directly connected to the power supply on the DC-side and the resistors on the AC-side. The DC current at the power supply was set by adjusting the resistance of the load resistors. Therefore the power supply was working in control voltage mode, excepted from the last test (double nominal current), which would exceed the capabilities of the power supply.

The output voltage of the power supply had to be reduced in order not to exceed the maximum output power of 3 kW. The full bridge was working with a pulse width modulated (PWM) signal. The constant pulse width of 500 s was generated by a PWM frequency of 1 kHz and a duty cycle of 50%. The submodule had to operate covered under a box with just a few holes near the top and bottom side. This box ensures that only natural convection can take place and a possible airflow in the measurement environment would not affect the measurement results.

The measurement data consists out of the timestamps and two MOSFET temperatures, namely MOSFET 2 and MOSFET 3 of the submodule from Figure 13. The temperature measurements of the test item, as can be seen in Figure 15, were taken by a thermocouple sensor. Two bimetal temperature sensors, which were evaluated by multi-meters, were used as measuring means. Thermal imaging was also used to validate the temperature measurement by the thermocouple. The measurement was evaluated as representative.

Figure 16 shows the picture of the thermal imaging camera. It illustrates that the heat transfer to the cooling elements work properly and that the chosen measurement points offer a representative result. A STM32F4 controller generates the PWM signal for the control of the full-bridge. The standard PWM frequency within the inverter is 1 kHz. It should be noted that each experiment with different load levels continued until the failure of the tested MOSFETs or when the normative values of the duration of fail-safe operation were significantly exceeded.

The Figures 17–21 show the thermal characteristics of the submodule for various load modes, i.e. nominal load, 10%, 25%, 50%, and 100% overload of power electronic devices. The experiments were stopped when the time range significantly exceeded the normative values of Figure 4 for every load mode. The experimental targets were clearly exceeded. The tested submodule has shown large reserves in the overload capacity. During the tests no unexpected defects have been occurred. Thus, the normative overload data (Figure 4) can be used to determine the transition probabilities of MSS MM without any restrictions.

Figure 15: Test item for thermal measurement with measurement points on MOSFET 2 and 3.

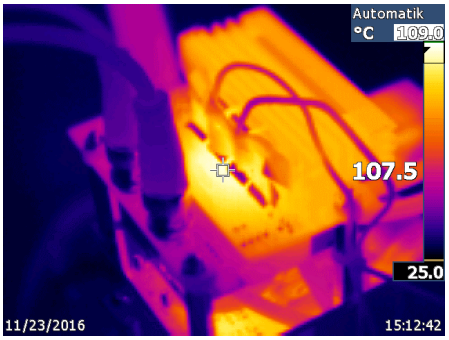


Figure 16: Resulting picture of thermal imaging camera.

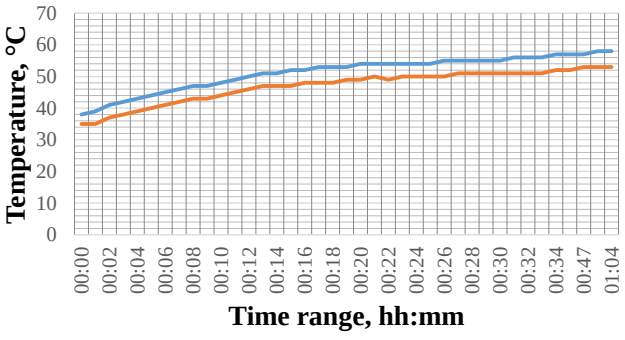


Figure 17: Thermal characteristic for nominal mode without overload.

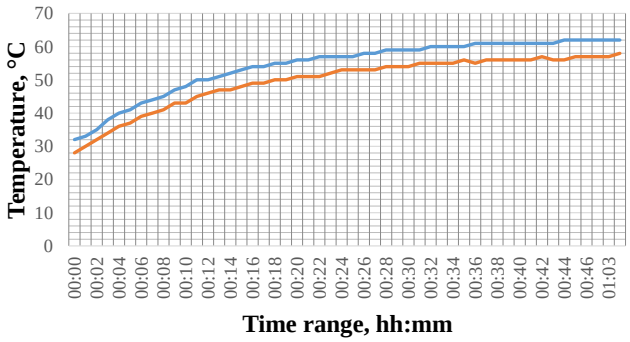


Figure 18: Thermal characteristic for 10% overload.

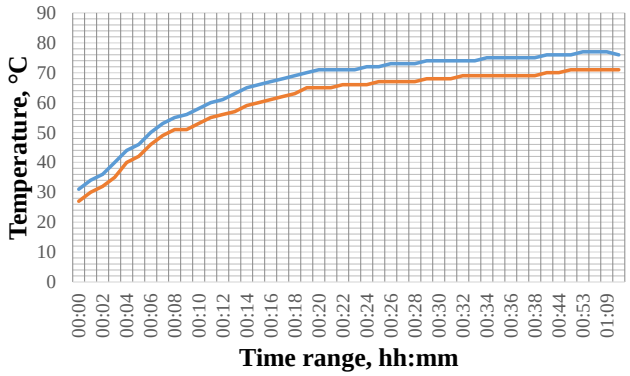


Figure 19: Thermal characteristic for 25% overload.

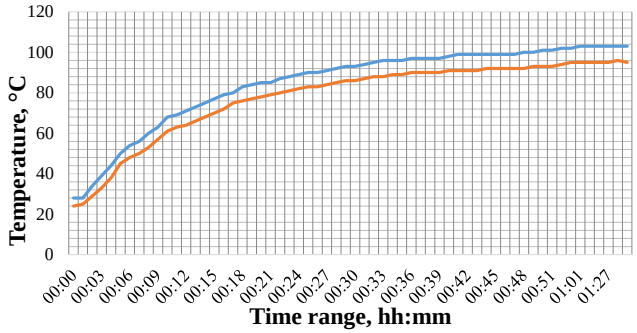


Figure 20: Thermal characteristic for 50% overload.

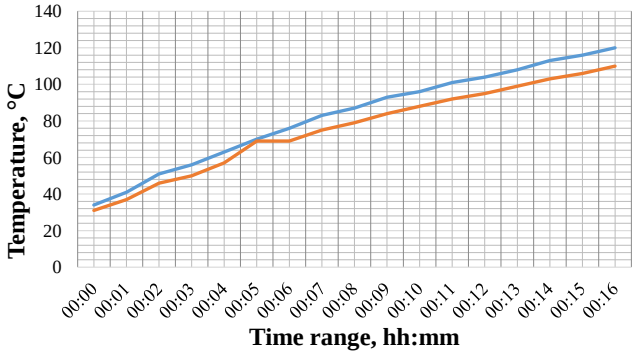


Figure 21: Thermal characteristic for 100% overload.

6 Results of Simulation

By the simulation has been considered the worst case, critically dangerous variant of failure – the submodule failures occur consistently in the same phase. In case of the simulation of a not safety-critical failure, as well as the possibility of partial recovery of power drive's operating ability in the degraded state, the value of the fault tolerance will be significantly higher. Considering the design requirements on the fault tolerance, the sustainable functionality of the inverter in the failure cases was analyzed using the above MSS MM for a power demand of 113% of the nominal value. The corresponding graphs for a different number of phases are presented in Figure 17.

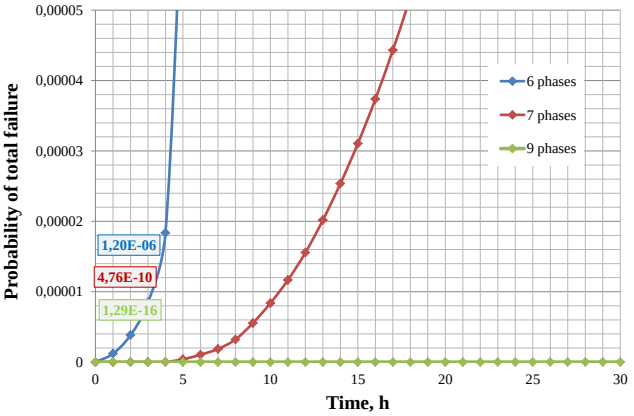


Figure 22: Probability of total failure one inverter phase.

The results of simulation of three consecutive critically dangerous failures allow quantifying the degree of fault tolerance of a multilevel inverter, which is a one important part of the traction drive of helicopters. The 7- and 9-phase options have shown the maximum compliance with the requirements relating to the safety-critical drives. For further studies of reliability and fault tolerance of safety-critical electrical drives, it is advisable to evaluate the integrated performance of electric drive as a whole, considering the reliability characteristics of electric power source, power electronics, and electric motor.

It is also of significant interest to quantitatively assess the impact of the maintenance strategy on the reliability and fault tolerance of electrical drive of the helicopter.

7 Conclusion

The paper presents the results of complex comparative analysis of an important part of the safety-critical drive train – the electric inverter, including an analysis of the reliability and fault tolerance of inverter in emergency operational modes.

On the basis of the developed techniques and models, taking into account the strict requirements on the safety-critical drive of an electrical helicopter, the authors carried out a comparative analysis of the fault tolerance degree of considered inverter topologies, which, in turn, led to the conclusion about the optimal range of applicability and feasibility of each compared variant of topology considering the specified safety requirements.

References

[1] I. Bolvashenkov, J. Kammermann, T. Lahlou, and H.-G. Herzog, "Comparison and Choice of a Fault Tolerant Inverter Topology for the Traction Drive of an Electrical Helicopter", 3rd Int. Conf. on Electrical Systems for Aircraft, Railway, Ship Propulsion, and Road Vehicles (ESARS16), 02–04 November 2016, Toulouse, France, pp.1–6.

[2] I. Bolvashenkov, J. Kammermann, S. Willerich, and H.-G. Herzog, "Comparative Study of Reliability and Fault Tolerance of Multi-Phase Permanent Magnet Synchronous Motors for Safety-Critical Drive Trains", in Proc. of International Conference on Renewable Energies and Power Quality (ICREPQ16), 04–06 May 2016, Madrid, Spain, pp.1–6.

[3] H. Wang, K. Ma, F. Blaabjerg, "Design for Reliability of Power Electronic Systems", in Proc. of the 38th Annual Conference of the IEEE Industrial Electronics Society (IECON12), 25–28 October 2012, Montreal, Canada, pp.33–44.

[4] E. Lauger, "Reliability in electrical and electronic components and systems", North-Holland Publishing Co., Amsterdam, 1982, p.1171.

[5] D. Kastha, B. K. Bose, "Investigation of fault modes of voltage-fed inverter system for induction motor drive", IEEE Transactions on Industry Applications, Vol.30, No.4, Jul/Aug 1994, pp.1028–1038.

[6] B. A. Welchko, T. A. Lipo, T. M. Jahns and S. E. Schulz, "Fault Tolerant Three-Phase AC Motor Drive Topologies: A Comparison of Features, Cost, and Limitations", IEEE Transactions on Power Electronics, Vol.19, No.4, July 2004, pp.1108-1116.

[7] I. Bolvashenkov and H.-G. Herzog, "Approach to Predictive Evaluation of the Reliability of Electric Drive Train Based on a Stochastic Model", 5th International Conference on Clean Electric Power (ICCEP15), 16–18 June 2015, Taormina, Italy, pp.1–7.

[8] R. Bojoi, M.G. Neacsu and A. Tenconi, "Analysis and Survey of Multi-Phase Power Electronic Converter Topologies for the More Electric Aircraft Applications", In Proc. of International Symposium on Power Electronics, Electrical Drives, Automation and Motion (SPEEDAM), 20–22 June 2012, Sorrento, Italy, pp.440–445.

[9] A. L. Julian and G. Oriti, "A Comparison of Redundant Inverter Topologies to Improve Voltage Source Inverter Reliability", IEEE Transactions on Industry Applications, Vol.43, No.5, Sept./Oct. 2007, pp.1371–1378.

[10] Y. Ding, P. C. Loh, K. K. Tan, P. Wang, and F. Gao, "Reliability Evaluation of Three-Level Inverters", In Proc. of IEEE Twenty-Fifth Applied Power Electronics Conference (APEC), 21–25 Feb. 2010, Palm Springs, CA, USA, pp.1555–1560.

[11] D. Fodorean, M. Ruba, L. Szabo, A. Miraoui, "Comparison of the main types of fault-tolerant electrical drives used in vehicle applications", In Proc. of International Symposium on Power Electronics, Electrical Drives, Automation and Motion, (SPEEDAM), 11–13 June 2008, Ischia, Italy, pp.895–900.

[12] M. Villani, M. Tursini, G. Fabri and L. Castellini, "Multi-Phase Permanent Magnet Motor Drives for Fault-Tolerant Applications", In Proc. of IEEE International Electric Machines & Drives Conference (IEMDC), 05–18 May 2011, Niagara Falls, Canada, pp.1351–1356.

[13] J. Kammermann, I. Bolvashenkov and H.-G. Herzog, "Approach for Comparative Analysis of Electric Traction Machines", In Proc. of 3rd International Conference on Electrical Systems for Aircraft, Railway, Ship Propulsion, and Road Vehicles (ESARS) Aachen, Germany, March 2015, pp.1–5.

[14] M. Malinowski, K. Gopakumar, J. Rodriguez, and M. Perez, "A Survey on Cascaded Multilevel Inverters", IEEE Transaction on Industrial Electronics, Vol. 57, No. 7, July 2010, pp.2197–2206.

[15] B. Sarrazin, N. Rouger, J. P Ferrieux, J. C Crébier, "Cascaded Inverters for electric vehicles: Towards a better management of traction chain from the battery to the motor?", In Proc. of IEEE International Symposium on Industrial Electronics, 27-30 June 2011, Gdansk, Poland, pp.153–158.

[16] S. Fazel, S. Bernet, D. Krug, K. Jalili, "Design and Comparison of 4-kV Neutral-Point-Clamped, Flying-Capacitor, and Series-Connected H-Bridge Multilevel Converters", IEEE Transactions on Industry Applications, Vol.43, No.4, July/Aug. 2007, pp.1032–1040.

[17] O. Josefsson, T. Thiringer, S. Lundmark, H. Zelaya, "Evaluation and comparison of a two-level and a multilevel inverter for an EV using a modulized battery topology", In Proc.of IEEE 38th Annual Conference on Industrial Electronics Society (IECON), 25–28 Oct. 2012, Montreal, Canada, pp.2949–2956.

[18] N. P. Ermolin and I. P. Zerichin, "Zuverlssigkeit elektrischer Maschinen", Berlin, Verlag Technik, 1981, p.227. (in German)

[19] M. Molaei, H. Oraee and M. Fotuhi-Firuzabad, "Markov Model of Drive-Motor Systems for Reliability Calculation", In Proc. of IEEE International Symposium on Industrial Electronics, 09–13 July 2006, Montreal, Canada, pp.2286–2291.

[20] T. Geyer and S. Schroder, "Reliability Considerations and Fault-Handling Strategies for Multi-MW Modular Drive Systems", IEEE Transactions on Industry Applications, Vol.46, No.6, Nov./Dec. 2010, pp.2442–2451.

[21] P. Wikström, L. A. Terens and H. Kobi, "Reliability, Availability, and Maintainability of High-Power Variable-Speed Drive Systems", IEEE Transactions on Industry Applications, Vol.36, No.1, Jan./Feb. 2000, pp.231–241.

[22] A. H. Ranjbar, M. Kiani and B. Fahimi, "Dynamic Markov Model for Reliability Evaluation of Power Electronic Systems", In Proc. of IEEE International Conference on Power Engineering, Energy and Electrical Drives (POWERENG), May 2011, Malaga, Spain, pp.1–6.

[23] Y. Song and B. Wang, "Survey on Reliability of Power Electronic Systems", IEEE Transactions on Power Electronics, Vol.28, No.1, January 2013, pp.591–604.

[24] I. Bolvashenkov and H.-G. Herzog, "Degree of fault tolerance of the multi-phase traction electric motors: methodology and application", In Proc. of 16th IEEE International Conference on Environment and Electrical Engineering (EEEIC), 07–10 June 2016, Florence, Italy, pp.1–6.

[25] B. Natvig, "Multi-state systems reliability theory with applications", John Wiley & Sons, New York, 2011, p.232.

[26] A. Lisnianski, I. Frenkel and Y. Ding, "Multi-state System Reliability Analysis and Optimization for Engineers and Industrial Managers", Berlin, New York, Springer, 2010, p.393 p.

[27] I. Bolvashenkov, J. Kammermann, and H.-G. Herzog, "Reliability Assessment of a Fault Tolerant Propulsion System for an Electrical Helicopter", Proceedings of 12th International Conference and Exhibition on Ecological Vehicles and Renewable Energies (EVER), 11–13 April 2017, Monaco, pp. 1–6.

[28] I. Bolvashenkov, J. Kammermann, and H.-G. Herzog, "Methodology for determining the transition probabilities for multi-state system Markov models of fault tolerant electric vehicles", In Proc. of Asian Conference on Energy, Power and Transportation Electrification (ACEPT), 25–27 Oct. 2016, Singapore, pp.1–6.

[29] EN61000-2-4, https://www.vde-verlag.de/normen/0839015/din-en-61000-2-4-vde-0839-2-4-2003-05.html, visited on 01.09.2016.

[30] C. D. Tudryn, B. Blalock, G. Burke et al., "Low Temperature Thermal Cycle Survivability and Reliability Study for Brushless Motor Drive Electronics", In Proc. of IEEE Aerospace Conference, 04–11 March 2006, Big Sky, MT, USA, pp.1–37.

[31] Z. Zhao, K. Li, Y. Jiang, S. Lu, and L. Yuan, "Overview on Reliability of Modular Multilevel Cascade Converters", In Chinese Journal of Electrical Engineering, Vol.1, No.1, December 2015, pp.37–49.

[32] I. Vernica, K. Ma and F.Blaabjerg, "Modelling and Improvement of Thermal Cycling in Power Electronics for Motor Drive Applications", In Proc. of IEEE Energy Conversion Congress and Exposition (ECCE), 18–22 Sept. 2016, Milwaukee, WI, USA, 2016, pp.1–8.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 412-419 (2017)

www.astesj.com

Special issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Computational Intelligence Methods for Identifying Voltage Sag in Smart Grid

Turgay Yalcin*, Muammer Ozdemir

Electrical & Electronic Engineering Faculty, Ondokuz Mayis University, 55139, Samsun, TURKEY

ARTICLE INFO

Article history:
Received: 16 March, 2017
Accepted: 11 May, 2017
Online: 23 May, 2017

Keywords:
C4.5 decision trees,
Classification methods
Feature extraction
K-Means Clustering
Support vector machines
Voltage sag

ABSTRACT

In recent years pattern recognition of power quality (PQ) disturbances in smart grids has developed into crucial topic for system equipments and end-users. Undoubtedly analyzing the PQ disturbances develop and maintain smart grids effectiveness. Voltage sags are the most common events that affect power quality. These faults are also the most costly. This paper represents performance comparisons of different computer intelligence methods for voltage sag identification. PQube Analyzer which is installed in Ondokuz Mayis University Computer Laboratory for collecting real time disturbances data for each three phases in order to test for proposed algorithms. Firstly, we used Hilbert Huang Transform to genarate Instantaneous Amplitude (IA) feature signal. Then Characteristic features are attained from IA. The 4 features, mean, standard deviation, skewness, kurtosis of IA are calculated. Support Vector Machines (SVMs) and C4.5 Decision Tree methods are conducted for classification of the disturbance. Secondly we used Fishers Discriminant Ratio for selecting statistical features such as mean, standard deviation, skewness and kurtosis of the normal and voltage sag signals for this part K Means Clustering Method were performed for classification of the disturbance. Consecuently, SVMs, C4.5 Decision Tree and K Means Clustering Methods were performed also their achievements were matched for error rates and CPU timing.

1. Introduction

PQ problems occur due to frequency, voltage swells, voltage sags, light flicker, voltage transients, harmonics, unbalanced 3-phase signal. All equipments influence the grid by the reproduction harmonics, voltage variations. Also the performance of transformers and generators are affected by the power quality on the distribution network. Unanticipated alteration of the voltage and current signal from healthy conditions can devastate or stoppage the critical electrical machineries created for various aim [1-5].

1.1. Voltage sag

Voltage sags are short-duration (less than 1 second) reductions in voltage magnitude. The sag is mostly related with the short circuit faults single-line to ground, line to line, and for three-phase to ground faults. This type of disturbance is nowadays one of the main PQ disturbances. Momentary increase of current has many origins in power systems such as energizing of transformers, short circuits, earth faults and starting of induction motors. Voltage sags are the most typical disturbances that affect power quality in smart grids. Moreover, they are the most costly. When voltage sag happens, the power supply inside electronic devices uses some of its stored energy to make up for the failure of input voltage. If enough energy is lost owing to the sag, then the power supply may fail its task to maintain adequate

DC voltage to components. Even for nanoseconds such integrated circuits, inside the device are affected by. This is very short moment to destroy data in microprocessor based electronics and to bring about malfunctions of digital equipment especially PCs. Malfunction of PCs integrated. in a real-time system is linked with conceivably bigger results, because there are losses correlate with the controlled process. The sensitivity of PCs to voltage sags is mostly expressed only in terms of the magnitude and duration of the voltage sag [25, 26, 30, 31, 40]. Considering previous studies about PQ disturbances analysis, Short-Time Fourier Transform (STFT) have been applied for stationary and periodic signals in frequency domains. But when handling non-stationary signals, it is problematic to analyze with STFT [47, 48, 49]. It was experimentally verified and demonstrated the WT is better than STFT [47]. Wavelets are employed in PQ when the time information is crucial however, it is not crucial to perceive the specific frequency of a fault in waveforms. Using WT, high time resolution is arranged for high- frequency components and low time resolution is attained for low-frequency components of the disturbance. Another method, ST is acquired by multiplying WT

*Turgay Yalcin, Ondokuz Mayis University, Electrical&Electronic Engineering Faculty, 55139, Samsun, TURKEY, +903623121919,turgay.yalcin@omu.edu.tr

with a phase correction factor. This method's disadvantage is having large computation time. Since ST is a Fourier-based technique, it also deteriorates from the handicaps of Fast Fourier Transform (FFT) [18, 20, 47, 48, 49]. On the other hand, the Hilbert Huang Transform (HHT), signals are extracted for different oscillation modes called intrinsic mode functions (IMFs). These functions are related to different time scales. Application of Hilbert transform to these IMFs can indicate the instantaneous amplitudes (IA) and the instantaneous frequencies (IF). Therefore, HHT can give a reasonably better time-frequency pattern representations for non-stationary signals [6, 11, 12, 13, 18, 19, 21]. In the light of this information, we used the HHT for generating IA signals from the voltage sag signal Firstly, for 3 phases (L1-N, L2-N, L3-N) real time processing the first intrinsic mode function is removed with the addition (superposition) of remain components to reconstruct the analyzed signal. Secondly, we generate features with the help of statistical functions such as mean, Standard deviation, skewness and kurtosis. For Diagnosis part we utilized Support Vector Machines, Decision Tree, and K Means Clustering Methods. These methods have performed also their achievements were matched for error rates and CPU timing.

2. Material and Methods

2.1. Signal Processing Methods for Feature Generation

Unlike the traditional methods such as STFT (Short time fourier Transform), Wavelet Transform (WT) and Stockwell Transform (ST), Hilbert Huang Transform (HHT), especially for noise assisted PQ disturbances signals, provides better and useful representation in order to detect the disturbance. Disadvantages and advantages of signal processing methods for feature generation procedure summarized in Table I.

Table I. Comparison of feature extraction methods for PQ disturbances [11, 18, 21]

	Disadvantages	Advantages
STFT	**Basis: Non-adaptive Completely Theoretical** Not acceptable for non-stationary signal as it does not follow signal oscillation accurately because of limitation of fixed window width	Profitably handle for stationary signals where properties of signals do not derive in time. Quite simple in application.
WT	**Basis: Non-adaptive Completely Theoretical** Fully affected by noise present in the signal, adversity from spectral leakage effects	Gives local illustration in both time and frequency. Thus, suitable good time– frequency resolution is appropriate.
ST	**Basis: Non-adaptive Completely Theoretical** Does not accomplish real-time requirement based on block processing, false harmonics measurement owing to dependency of frequency window width	Maintain time and frequency representation. Good time-frequency resolution.
HHT	Limited for narrow band conditions Mode mixing and end effect problems	**Basis: Adaptive Empirical** Conveniently applicable to non-linear and non-stationary processes. Generates perpendicular IMFs whereby instantaneous amplitude and phase can be easily assessed be easily evaluated

2.2. Stockwell transform (S- transform)

The S-transform is a time–frequency plane gadget produced by the mixture of wavelet and short time fourier transform It generates a time– frequency plane of a time series. It individually merges a frequency dependent resolution that contains the real and imaginary part of spectrum. Gaussian modulated co-sinusoids are the primary function for the S-transform [6, 7]. The Short Time Fourier Transform (STFT) of signal $x(t)$ is explained in equation 1.

$$STFT(\tau, f) = \int_{-\infty}^{\infty} x(t) g(\tau - t) e^{-j2\pi ft} dt \quad (1)$$

τ and f stand for the time of spectral localization and Fourier frequency, $g(t)$ shows a window function. The S transform could stem from (1) by putting into $g(t)$ with the Gaussian function, explained in equation 2.

$$g(t) = \frac{|f|}{\sqrt{2\pi}} e^{-\frac{t^2 f^2}{2}} \quad (2)$$

Ultimately the S transform is achieved in equation 3.

$$S(\tau, f) = \int_{-\infty}^{\infty} x(t) \frac{|f|}{\sqrt{2\pi}} e^{\frac{(\tau - t)^2 f^2}{2}} e^{-j2\pi ft} dt \quad (3)$$

If the window of S transform is wider in time, it gives better frequency resolution for lower frequency. Whereas the window is narrower, it gives superior time resolution for higher frequency, namely trade-off between time and frequency resolution [18, 20].

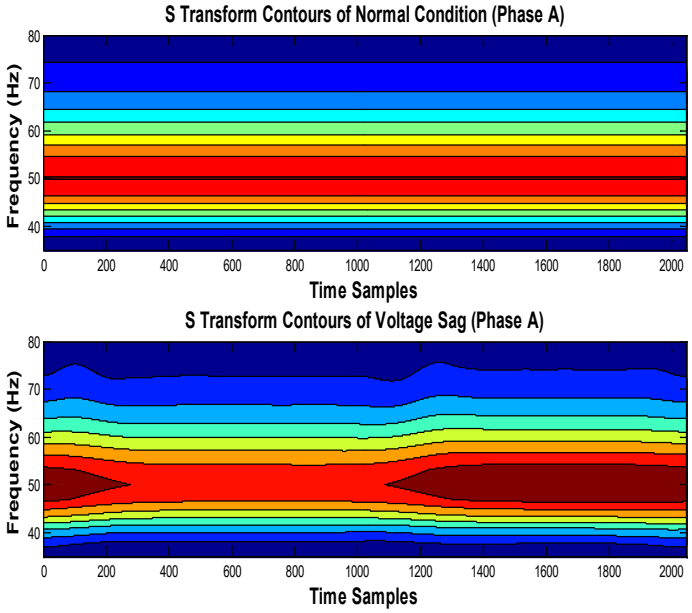


Figure 1. STransform Contours of Normal – Voltage Sag Cases

S-Transform is not capable of detecting noise as much as HHT in real time measurements [8, 32] (Figure 1). However, it is clearly shown that ST can produce valid features for detecting voltage sag.

2.3. Proposed Method: Hilbert Huang transform

HHT, a self-adaptive signal-processing method, which has been applied in non-stationary signal-processing successfully, was developed by Huang. EMD method is a ***sifting*** iteration, to decompose one Intrinsic Mode Function (IMF) component needs many times sifting till obtaining last IMF, residual monotonic function [9, 10]. Main steps of HHT shown in figure 2.

The EMD [11, 12, 13] algorithm the steps:

Beginning of Sifting Process

1. Find all maximas and minimas of the signal $x(t)$.
2. Attain the upper and lower envelope by the cubic spline interpolation.
3. Find the mean function of the upper and lower envelope $mean(t)$.
4. Calculate the difference signal $f(t) = x(t) - mean(t)$.
5. When $f(t)$ becomes a zero-mean process, then the iteration stops and $f(t)$ is an Intrinsic Mode Function IMF1, $c_1(t)$; otherwise, go first step and replace $x(t)$ with $f(t)$.
6. Calculate the residue signal $res(t) = x(t) - c_1(t)$. Iterate the procedure from steps (i) to (vi) to obtain IMF2, named $c_2(t)$. To attain $c_n(t)$, continue steps (1) till step (6) after n iterations.

End of the Sifting Process

7. The algorithm stops when the final residual signal $res(t)$ becomes a monotonic function.

2.3.1. Ensemble EMD (EEMD) Algorithm

The major challenge of the EMD algorithm is the appearance of mode mixing. Therefore the EEMD algorithm was performed for overcoming the problem in this study [12, 13].

Analyzed signal is decomposed using the ensemble empirical mode decomposition (EEMD) into intrinsic mode function (IMF) that includes instantaneous frequencies (IF) and instantaneous amplitudes (IA). When the signal is turn into IMFs, the Hilbert transform can formerly be carried out to each IMF giving the IA and IF versus time plot. This merging of EEMD process and Hilbert transform is known as the HHT [6, 11, 12, 13, 18, 19, 21].

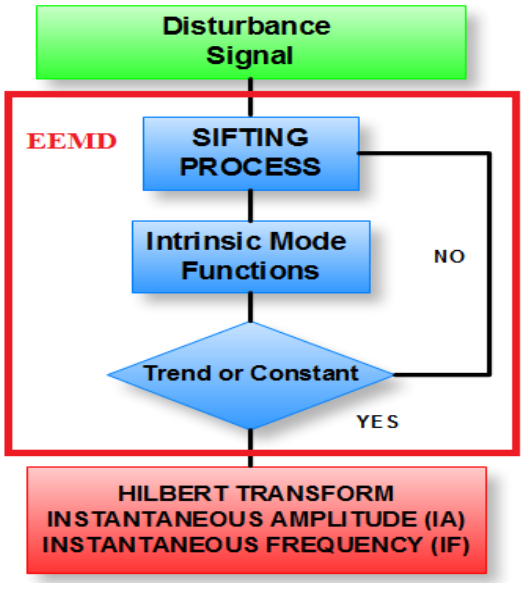


Figure 2. Main steps of the feature generation routine with HHT

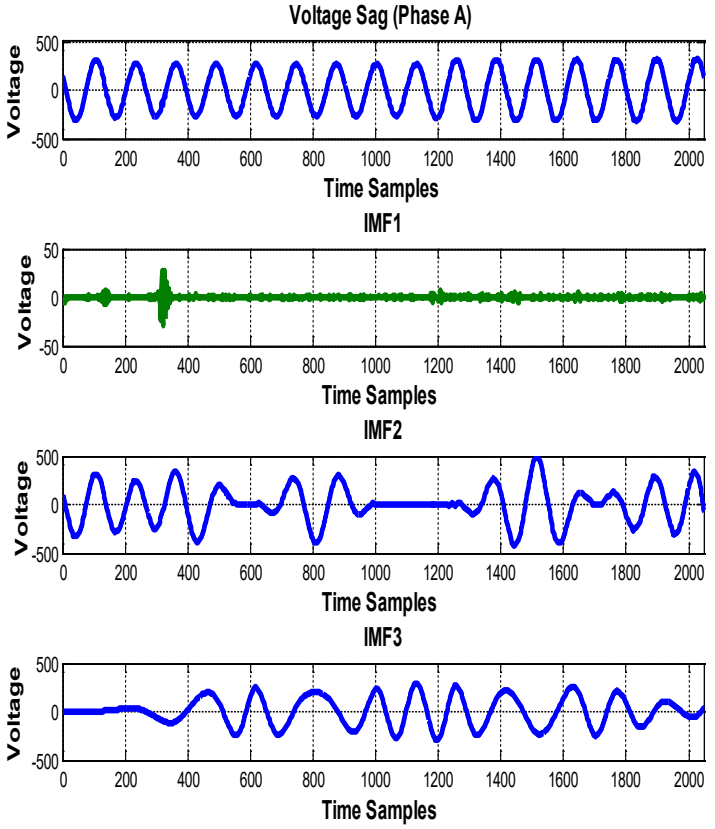


Figure 3a. Feature vectors (IMF1 (noise) to IMF3) for a voltage sag signal with EEMD

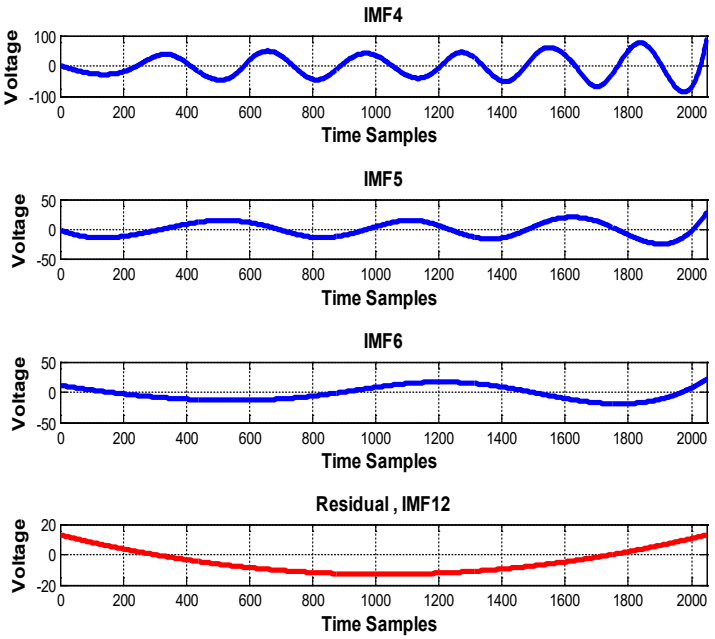


Figure 3b. Feature vectors (IMF4 to Residual (IMF12)) for a voltage sag signal with EEMD

For 3 phases (L1-N, L2-N, L3-N) real time processing the first intrinsic mode function is removed with the addition (superposition) of remain components to reconstruct the analyzed signal. Fig 3 illustrates first component IMF1 the noise (lowest

magnitude highest frequency signal) on the line (L1-N) to the lowest frequency component residual, last IMF12.

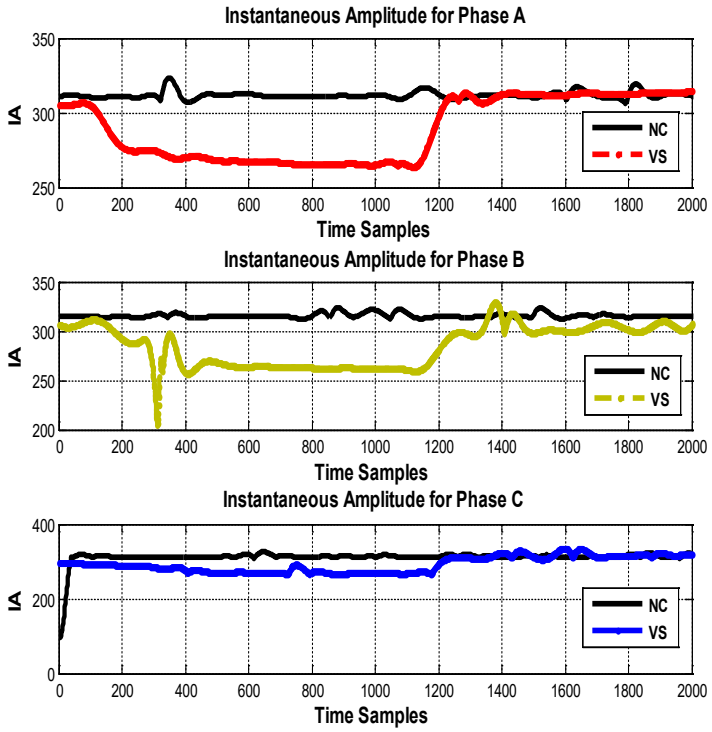


Figure 4. IA corresponding to remove first IMF voltage sag condition of signal

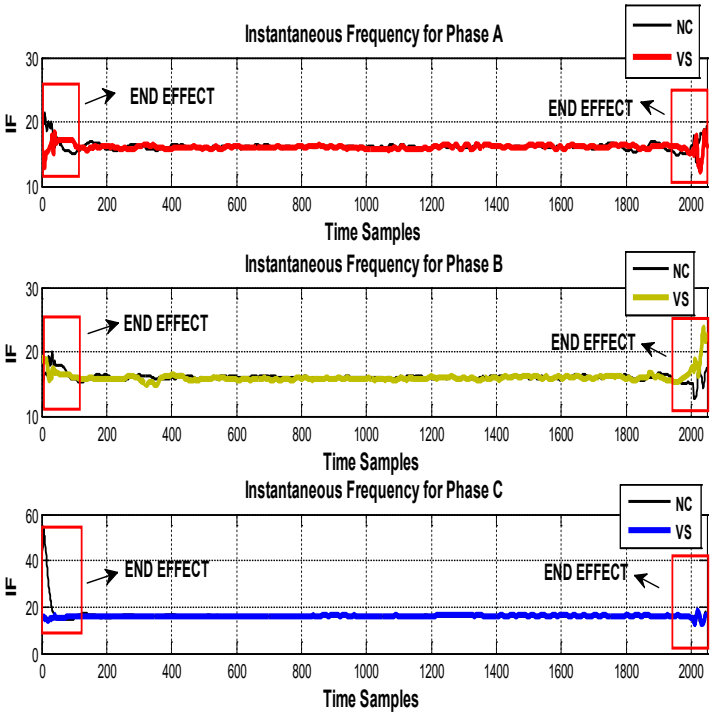


Figure 5. IF corresponding to remove first IMFs voltage sag condition of signal

IF signal can be used for separation for two cases but there is end effect problem that has to be solved. When cubic spline fitting is computationally demanding, it generates distortions near the end points. This is a technical problem that causes data failures and peaks at the beginning and at the end of the signal. This fault will be investigated on Hilbert-Huang transform (Figure 5).

3. PQ Disturbances Computational Intelligence Methods

3.1. Support vector machine (SVM)

SVM methods, is developed by Vapnik, whereby statistical learning technique being the basis contributes a novel machine learning method. SVMs are linked supervised learning methods used for classification and regression. SVMs target obtaining optimal hyper-planes different classes of input data in a high dimensional feature space, thus new test data can be classified by acquired the hyper-planes Support vectors lie closest to the optimal hyper-plane [7, 14, 18, 21, 28, 34].

Let $\{x_i,y_i\}$, $i = 1,2,\ldots,M$ be M training data vectors x_i with class label y_i. Given an input vector x, an SVM constructs a classifier of the form:

$$f(x) = sign(\sum_{i=1}^{M} a_i y_i K(x_i, x) + b) \quad (4)$$

a_i denotes non-negative Lagrange multipliers each of which corresponds to a training data, b denotes bias, $K(.\ ,\ .)$ is a kernel fullfilling the conditions of *Mercer's theorem* [11, 34]. In this study we performed two kernel functions are the polynomial kernel $K(x_i, x_j) = (x_i . x_j + 1)^d$ we tested for d=2, 3. Gaussian Radial Basis Function (RBF) $K(x_i, x_j) = e^{-\frac{|x_i - x_j|^2}{2\sigma^2}}$ we tested sigma 0.01 and 1 in this study.

3.2. Decision Trees

Decision trees are methods that utilize divide-and-conquer approaches as structure learning by induction [15, 16]. The C4.5 algorithm was developed by Qinlan, contains the generation of a tree whereby a training set, finding the information gain criterion with the help of Shannon Entropy in order to find the finest attribute/feature to be used at each node.

3.2.1. Shannon Entropy

It is given a probability distribution $D = (d_1, d_2, \ldots, d_m)$ the Information carried by this distribution entropy of D is explained in equation 5.

$$Entropy(D) = -\sum_{i=1}^{m} d_i \log(d_i) \quad (4)$$

3.2.2. The Information Gain

The Information defines the gain for a test t and a position d

$$Gain(D,t) = Entropy(D) - \sum_{j=1}^{m} d_j Entropy(d_j) \quad (5)$$

that values (*dj*) is the set of all possible values for attribute *t*. Equation 5 can calculate highest ranking attributes and build the decision tree where at each node is located the attribute with the highest information gain. Furthermore, the algorithm applies the post pruning approach to diminish the size of the tree and prohibit over fitting. C4.5 is a technique for approximating discrete-valued functions that is powerful tool to noisy data and suitable for learning distinctive statements [16, 17, 18, 19, 20, 21, 28, 29].

3.3. K-Means Clustering

K-means clustering algorithm is a significant unsupervised classification technique to separation of set of dissimilar patterns. Same featured patterns are arranged into same clusters, different features are replaced in different sets [22, 23].The steps of the algorithm:

1. Apply K points into the space illustrated by objects being clustered. These points symbolize first group centroids. Determine each object to the nearest centroid group.
2. As objects have been completed, recalculate the localizations of the K centroids.
3. Iterate Steps 1 and 2 until the centroids do not change anymore. This creates a separation of the objects into groups from which the metric with squared Euclidean distance objective function which is defined in equation 6, can be calculated.

The objective of this algorithm is to make an objective function, minimize the squared error function [35, 36, 37].

$$Obj(x) = \sum_{i=1}^{K} \sum_{x_j \in S_i} \left\| x_i - \mu_i \right\|^2 \qquad (6)$$

K denotes clusters S_i, $i = 1, 2, \ldots, k$, and μ_i is the mean points of centroids.

3.3.1. Fisher's discriminant ratio (FDR)

FDR or SNR (in machine learning) is commonly performed to compute the inequitable way of specific features between two clusters. Linked with the values of a feature in two classes, μ_1 and μ_2 be the respective mean values and σ_1^2 and σ_2^2 the respective variances defined in equation 7 [38, 39].

$$FDR = \frac{(\mu_1 - \mu_2)^2}{(\sigma_1^2 + \sigma_2^2)} \qquad (7)$$

4. Experimental Set up and Discussions

For experimental studies, we installed PQube Analyzer in Ondokuz Mayis University Computer Laboratory for collecting real time disturbances data for each three phases in order to test for proposed algorithms. 3 phases connected with totally 28 numbers of PC loads. PQube power meter installation properties explained in appendix of the paper. This paper includes computational intelligence methods in order to classify voltage sag disturbance. Figure 6 shows 3 phase voltage sag condition PQube sample measurements from Computer Laboratory in Ondokuz Mayis University including July 2016 to December 2016.

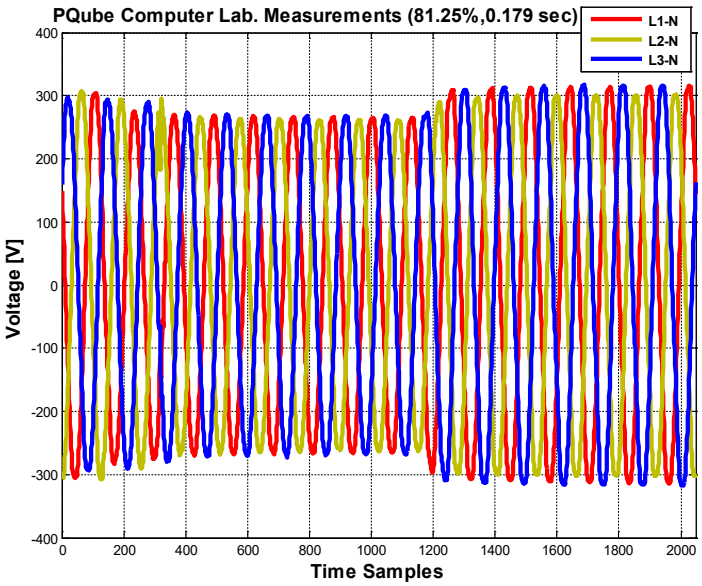


Figure 6. Computer laboratory power quality monitor with PQube voltage sag

In this study, we apply two methods for identification of voltage sag. Selecting appropriate features of voltage sag events are highly crucial for diagnosing of the fault properly. Classification process of real-time voltage sag disturbances are shown in Figure 7.

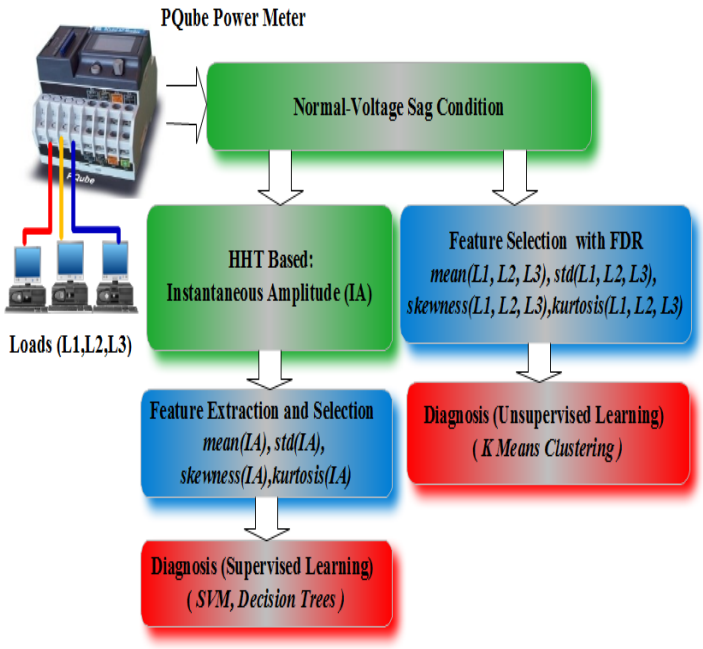


Figure 7. Schematic model of Voltage Sag Classification

First Method, firstly extracted features are produced from ensemble empirical mode decomposition (EEMD) method so as to classify the voltage sags in grid. After reconstruction signal without noisy part, removing first intrinsic mode function (IMF). Then instantaneous amplitude (IA) is generated by means of HHT. The following features were extracted: mean, standard deviation, skewness function results of IA.

Second Method, we generated the statistical features from PQube 3 phase line voltages. In order to select the best feature from the functions, we performed Fisher Discriminant Ratio criteria. Selected features which give high ratio for Phase A and B are standart deviation and kurtosis, for phase C are mean and skewness. Then we used generated features for K Means Clustering Methods. Figure 8 shows selected features space and cluster assignments and centroids of normal and voltage sag condition sets.

Table II. Performances of disturbance diagnose algorithms

	Phase A (L1-N)		Phase B (L2-N)		Phase C (L3-N)	
	Error rate (ER)(%)	Execution time(ET) (sec)	ER	ET	ER	ET
SVM-poly d=2	10	4.088	0	3.321	20	5.883
SVM-poly d=3	10	3.929	0	3.636	20	5.489
SVM-RBF sigma =0.01	50	0.738	**0**	**3.064**	20	5.805
SVM-RBF sigma =1	**0**	**5.254**	0	5.368	50	0.674
DecisionTree C4.5	20	6.20	20	6.835	20	3.514
K-Means (Euclidean distance)	6.66	0.693	16.6	0.510	**13.3**	**0.480**

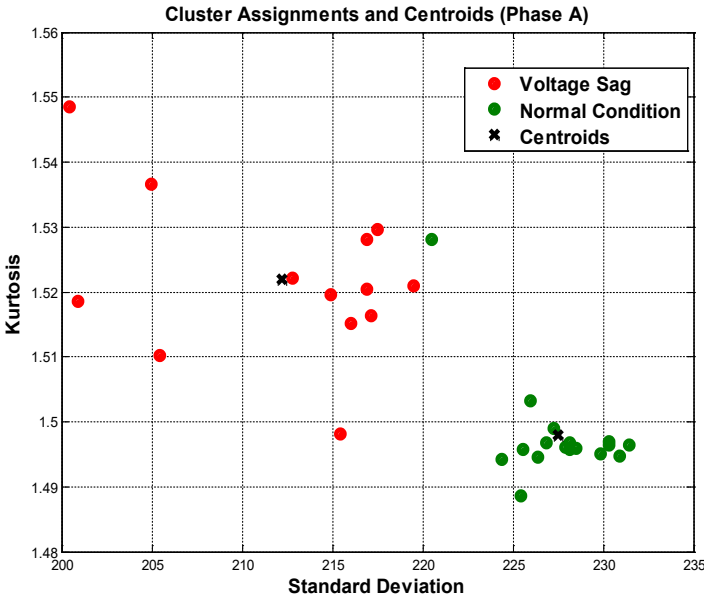


Figure 8. Feature Space and Clusters for Phase A

To figure out the performance of the proposed power quality classification algorithm, a total number of 30 PQube Analyzer real time disturbances data were used 15 for normal conditions 15 for voltage sag conditions. The PQ signals are divided into two categories; 20 of them were used for training and 10 of them were used for testing the proposed algorithm with shuffling the data.

We analyzed for each three phases separately owing to main source connected with different loads (number of computers). This study is the first step of our work, second step will be searching relations with number of computers for each phases and fault percentage.

- In the light of Table II., for phase A, it is concluded that for radial basis kernel function sigma 1, gives more proper result precision of 100% and CPU time (5.254 sec) in non linear SVM. Kernel function for SVM is $K(x_i, x_j) = e^{-\frac{|x_i - x_j|^2}{2(1)^2}}$

- For phase B, it is deduced that for radial basis kernel function sigma 0.01, gives more accurate result precision of 100% and CPU time (3.064 sec) in non linear SVM.

  Kernel function for SVM is $K(x_i, x_j) = e^{-\frac{|x_i - x_j|^2}{2(0.01)^2}}$

- For phase C, it is deduced that for K-Means clustering (Euclidean distance function ($d_{i,j}$)), gives better result precision of 86.7% and CPU time (0.480 sec). $Euclidean(d_{i,j} = \sqrt{\sum_{k=1}^{n}(x_{ik} - x_{jk})^2})$

- It is clearly shown in Table II when examining each phases K- means clustering algorithm is the fastest algorithm.

- For phase A and B, SVM with RBF sigma=1 gives better results for detection voltage sag.

- For three phases, C4.5 decision tree gives same 80% recognition rate.

(Note: the most proper and robust classifiers for each phase are showed by Red font in Table II).

5. Conclusion

In this study, we aimed to generate non-time consuming features for detection voltage sag in smart grid. We also process and classify real time 3 phase voltage sag signals. In contrast to literature studies we do not normalize the features. Thus we can relate load types and fault magnitude and duration. Also it is processed real-time noise- assisted data. For this reason this study gives effective ideas about real time disturbance detection. As a result, SVMs, C4.5 Decision Tree and K Means Clustering Methods were performed also their achievements were matched for error rates and CPU timing. Simulations indicate that K-means clustering algorithm is the fastest algorithm. For phase A and B, SVM with RBF sigma=1 gives better result for detection voltage sag. Simulation results will be using for relation with loads types and the magnitude and duration of the voltage sag.

Table III provides discussions about related signal processing methods such as WT, ST HT and for decision part SVM, DT, k-nearest neighbor (k-nn), Probability Neural Network (PNN). It also gives percentages error rates literature studies and the proposed methods. It is clearly shown in Table III most of the studies reported the results on synthetic normalized an done phase data which are absolutely far from real PQ disturbances. Apart from the previous studies, detection capability validates the potential of the proposed algorithm. To analyze the real time power signals and classification techniques were employed with Matlab ™ Toolboxes.

Table III. Performances of disturbance diagnose algorithms

Reference	Methods for PQ Detection	Data types with Noise or Not ?	Error rate (ER)(%)
Eristri and Demir [41]	WT + SVM	Noisy data+Practical data	4,19%
Xiao et al. [42]	ST + SVM	Without noise+Synthetic data	7,70%
Samantaray [43]	ST + DT	Noisy data+Practical data	2,44%
Hu et al. [44]	WT + SVM	Without noise+Synthetic data	1,50%
Gargoom et al. [45]	HT + k-NN	Without noise+Synthetic data	19,40%
Shukla et al. [46]	HT + PNN	Without noise+Synthetic data	2,50%
Proposed Methods	***Table II**	**Noisy data+Practical data (Not normalized data for each 3 phase measurements)**	***Table II**

Acknowledgment

This scientific study is supported by TUBITAK.

(Project Number: 114E919)

APPENDIX

Table IV. PQUBE MAINS VOLTAGE MEASURING CHANNELS [33]

Connection	L1, L2, L3, N PQube screw terminals
Frequency Range	40 Hz ~ 70 Hz and 320 Hz ~ 560 Hz. Nominal 50 Hz, 60 Hz, or 400 Hz auto, 320-560 Hz manually selected.
Specifications below apply at 50/60 Hz.	
Mains Configuration	Single-phase, split-phase, delta, wye or star. User selected or auto-selected.
Range of Nominal Input Voltage	100 VAC ~ 690 VAC L-L (69 VAC ~ 400 VAC L-N). User selected or auto-selected.
Measurement Channels	Line-to-Neutral, Line-to-Line, Neutral-to-Earth.
Sampling Rate	256 samples per cycle, phase-locked to input frequency.
Measurement Range	0 VAC ~ 900 VAC L-L (520 VAC L-N)
Accuracy	±0.05% rdg ±0.05% FS typical (10%~150% of nominal). Factory tested at better than ±0.04% rdg ±0.04% FS. Note: FS = 345 VAC or 520 VAC, selected based on nominal line-to-earth voltage.

References

[1] T. Yalcin, M. Ozdemir, "Pattern Recognition Method for Identifying Smart Grid Power Quality Disturbance", IEEE PES ICHQP 2016, Belo Horizonte, Brazil,2016.

[2] RC Dugan, MF McGranaghan, HW Beaty. Electrical power systems quality. New York: McGraw-Hill, 1–38,1996.

[3] RA Flore. State of the art in the classification of power quality events, an overview. In: 10th international conference on harmonics and quality of power, vol. 1; p. 17–20,2002.

[4] M. H. J. Bollen and I. Y. H. Guo, Signal Processing of Power Quality Disturbances,New York: Wiley, 2006.

[5] P. F. Ribeiro, Duque. C.A, P. M. Silveria, and A. S. Cerqueira, eds., Power Systems Signal Processing for Smart Grids. Chichester, UK: John Wiley & Sons, Inc., 2013.

[6] J. C. Alfonso-Gil, S. Orts-Grau, N. Munoz-Galeano, F. J. Gimeno- Sales, S. Segui-Chilet, "Measurement system for a power quality improvement structure based on IEEE Std. 1459.", IEEE Trans Instrum Meas;62(August (12)):3177–88, 2013.

[7] B. Biswal, PK. Dash, S. Mishra "A hybrid ant colony optimization technique for power signal pattern classification", Expert Syst Appl, 38: 6368–75, 2011.

[8] M. Uyar, S. Yildirim, M.T.Gencoglu,, "An effective wavelet-based feature extraction method for classification of power quality disturbance signals", Electr. Power Syst. Res, 78, (10), pp. 1747–1755, 2008.

[9] M. Inigo, L. Carlos, R. Jorge, G. Antonio, M. Jose, C. M. Juan, "Classification of electrical disturbances in real time using neural Networks", IEEE Trans Power Deliv;22(7):1288–96, 2007.

[10] Standard definitions for the measurement of electric power quantities under sinusoidal, non-sinusoidal, balanced, or unbalanced conditions. Revision of IEEE Std. 1459-2000, IEEE Standard 1459-2010; March 2010.

[11] O. Ozgonenel, T. Yalcin, I. Guney, U. Kurt, "A New Classification for Power Quality Events in Distribution System", Electric Power System Research (EPSR), 95, 192-199, 2013.

[12] Z. Wu, N.E. Huang,"A study of the characteristics of white noise using the Empirical Mode Decomposition method" ,Proc. Roy. Soc. London A. 2002.

[13] Z. Wu, N.E. Huang, "Ensemble empirical mode decomposition: a noise-assisted data analysis method. *Adv. Adapt. Data. Anal.* 1, 1–41, 2009.

[14] K. K. Hoong, S. P. Lam, C. Y. Chung, "An output regulation based unified power quality conditioner with Kalman filters", IEEE Trans Ind Electron;59(November (11)):4248–62, 2012.

[15] I. H. Witten and E. Frank, *Data Mining: PraticalMachine Learning Tools and Techniques*. San Mateo, CA, USA: Morgan Kaufmann, 2005.

[16] M. T. Hagan and M. B. Menhaj, "Training feedforward networks with the Marquardt algorithm," *IEEE Trans. Neural Netw.*, vol. 5, no. 6, pp. 989– 993, Nov. 1994.

[17] R. J. Quinlan, *C4.5: Programs for Machine Learning*, vol. 1. San Mateo, CA, USA: Morgan Kaufmann, 1993.

[18] S. Mishra, T. Nagwani, "A Review on Detection and Classification Methods for Power Quality Disturbances" Volume 6, Issue No. 3, International Journal of Engineering Science and Computing, March 2016.

[19] A. S. F. Borges, A. S. R. Fernandes, N. I. Silva, B. S. C. Silva, "Feature Extraction and Power QualityDisturbances Classification Using Smart Meters Signals", IEEE Transactions on Industrial Informatics, Vol. 12, No. 2, April 2016.

[20] O. Ozgonenel, D. W. P. Thomas, and T. Yalcin, "Superiority of decision tree classifier on complicated cases for power system protection," in 11th International Conference on Developments in Power Systems Protection, pp. 134–134, Birmingham, UK, 2012.

[21] O. P.Mahela, A. G. Shaik, N. Gupta , "A critical review of detection and classification of power quality events", Renewable and Sustainable Energy Reviews Volume 41, Pages 495–505, January 2015.

[22] J.M. Pena, J.A. Lozano,P. Larranaga, "An empirical comparison of four initialization methods for the k-means algorithm", Pattern Recognition Lett. 20 (10), 1027–1040,1999.

[23] St. Douglas, J. B. Michael, "Initializing K-means Batch Clustering: A Critical Evaluation of Several Techniques", Journal of Classification 24 (1), 99-121, 2007.

[24] T. Sergios, K. Konstantinos, Introduction to Pattern Recognition, 4th edition, Academic Press, 2009.

[25] M. Kezunovic, Y. Liao, "A new method for classification and characterization of voltage sags"Electric Power Systems Research;58(1):27–35, 2001.

[26] V. Ignatova, P. Granjon, S. Bacha, "Space vector method for voltage dips and swells analysis", IEEE Transactions on Power Delivery;24(4):2054–61, 2009.

[27] PK. Dash, BK. Panigrahi, G. Panda, "Power quality analysis using s-transform". IEEE Transactions on Power Delivery;18(2):406–11, 2003.

[28] SR. Samantaray, "Decision tree-initialised fuzzy rule-based approach for power quality events classification". IET Generation Transmission & Distribution;4(4):538–51, 2010.

[29] IWC Lee, PK. Dash, "S-transform-based intelligent system for classification of power quality disturbance signals". IEEE Transactions on Industrial Electronics ;50(4):800–5, 2003.

[30] MHJ. Bollen, What is power quality? Electric Power Systems Research ;66(1):5–14, 2003.

[31] MHJ. Bollen, E. Styvaktakis, IYH. Gu, "Categorization and analysis of power system transients", IEEE Transactions on Power Delivery;20(3):2298–306, 2005.

[32] S. Mishra, CN. Bhende, BK. Panigrahi, "Detection and classification of power quality disturbances using s-transform and probabilistic neural network", IEEE Transactions on Power Delivery;23(1),pp.280–287, 2008.

[33] PQube® Installation & User's Manual, Firmware Rev 2.1, July 2012.

[34] V. N. Vapnik, Statistical Learning theory, Wiley, New York, 1998.

[35] C. Zhang, S. Xia, " K-means Clustering Algorithm with Improved Initial center," in Second International Workshop on Knowledge Discovery and Data Mining (WKDD), pp. 790-792, 2009.

[36] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, A. Y. Wu, "An Efficient k-Means Clustering Algorithm: Analysis and Implementation", IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 24, NO. 7, JULY, pp.881-892, 2002.

[37] J. Dong, M. Qi. "K-means Optimization Algorithm for Solving Clustering Problem. Knowledge Discovery and Data Mining", pp.52-55, 2009.

[38] R. O. Duda, P. E. Hart, D. G. Stork, Pattern Classification, 2nd Edition, Wiley, Newyork, 2001.

[39] T. Sergios, K. Konstantinos, Introduction to Pattern Recognition, 4th edition, Academic Press, 2009.

[40] S. Ž. Djokic´, J. Desmet, , G. Vanalme, J. V. Milanovic´, K. Stockman, "Sensitivity of Personal Computers to Voltage Sags and Short Interruptions", IEEE TRANSACTIONS ON POWER DELIVERY, VOL. 20, NO. 1, JANUARY,pp 375-383, 2005.

[41] H. Eristi, Y. Demir, "A new algorithm for automatic classification of power quality events based on wavelet transform and Svm", Expert Systems With Applications, 37(6):4094–102, 2010.

[42] X. Xiao, F. Xu, H. Yang, "Short duration disturbance classifying based on stransform maximum similarity", International Journal of Electrical Power and Energy Systems,31(7-8):374–8, 2009.

[43] SR. Samantaray, "Decision tree-initialised fuzzy rule-based approach for power quality events classification", IET Generation Transmission & Distribution, 4(4):538–51, 2010.

[44] G-S. Hu, F-F. Zhu, Z. Ren, "Power quality disturbance identification using wavelet packet energy entropy and weighted support vector machines", Expert Systems With Applications 2008;35(1-2):143–9, 2008.

[45] AM. Gargoom, N. Ertugrul, WL. Soong, "Investigation of effective automatic recognition systems of power quality events", IEEE Transactions on Power Delivery, 22(4):2319–26, 2007.

[46] S. Shukla, S. Mishra, B. Singh, "Empirical-mode decomposition with Hilbert transform for power-quality assessment", IEEE Transactions on Power Delivery, 24(4):2159–65, 2009.

[47] IYH. Gu, MHJ. Bollen, Time-frequency and time-scale domain analysis of voltage disturbances. IEEE Transactions on Power Delivery,15(4):1279–84, 2000.

[48] PS. Addison, The Illustrated Wavelet Transform Handbook. Institute of Physics, 2002.

[49] F. Jurado, N. Acero, B. Ogayar. "Application of signal processing tools for powerquality analysis", IEEE Canadian Conference on Electrical & Computer Engineering, 1:82–7, 2002.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 420-426 (2017)

www.astesj.com

Special issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

A Highly-Secured Arithmetic Hiding cum Look-Up Table (AHLUT) based S-Box for AES-128 Implementation

Ali Akbar Pammu*, Kwen-Siong Chong, Bah-Hwee Gwee

School of Electrical and Electronic Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798

ARTICLE INFO

Article history:
Received: 22 March, 2017
Accepted: 12 May, 2017
Online: 24 May, 2017

Keywords:
Arithmetic Hiding
Side-Channel Attack
AES-128
Substitution-Box
CPA
CEMA

ABSTRACT

Side-Channel Attack (SCA) is an effective method in extracting the secret key of cryptographic algorithms by correlating the physical leakage information with the processed data. In this paper, we propose an arithmetic hiding cum Look-up Table (AHLUT) based Substitution-Box (S-Box) in AES-128 cryptographic algorithm implementation to countermeasure against SCA. There are three key features in our proposed AHLUT S-Box. First, the arithmetic hiding performs four types of arithmetic operations such that their total physical leakage information sufficiently overshadows the correlated physical leakage information of the S-Box operation. This is to reduce the correlation of the AES-128 physical leakage information with the processed data. Second, the AHLUT S-Box pre-stores all the 256 bytes of possible output values based on the conventional S-Box and selects a corresponding output value with respect to the input accordingly. In this context, it dissipates significantly lower power when compared to the conventional S-Box which performs multiplication inversion and affine transformation. Third, we propose a methodology to determine a minimum number of the arithmetic operations to sufficiently overshadow the physical leakage information of the S-Box operation. Based on the measurement results of performing AES-128 algorithm on Sakura-X FPGA encryption-board and in term of power dissipation, our proposed AHLUT S-Box dissipates 1.6mW and features 11.56× lower power dissipation than the conventional S-Box. In term of security which is based on Correlation Power Analysis attack, it requires 73× more power traces to reveal the secret key for our proposed AHLUT S-Box than the conventional S-Box. As for the non-invasive Correlation Electromagnetic Analysis attack, it requires 25× more electromagnetic traces for our proposed AHLUT S-Box than the conventional S-Box.

1. Introduction

Side-Channel Attack (SCA) is an effective method to extract the secret key of cryptographic algorithms, such as Advanced Encryption Standard-128 (AES-128) algorithm, by correlating physical leakage information, generated during the encryption process, with processed data. The physical leakage information such as power dissipation [1], Electromagnetic (EM) emanation [2], temperature [3] and timing [4] information, which are measured during the encryption process, are dependent on the processed data, determined based on plaintext/ciphertext. Due to the simplicity, in term of measurement, the power dissipation and EM emanation are the most preferred by adversary to be employed in the SCA compared with other physical leakage information [1].

Fig. 1 depicts the attacking scenario of the SCA, by intercepting wireless communication, transmitting the encrypted plaintext, ciphertext, to receiver and at the same time, measuring the physical leakage information. In this scenario, the correlation based SCA, such as Correlation Power Analysis (CPA), is employed to compute and analyze the correlation between the processed data and power dissipation measurements (power traces). To protect the secret key against SCA, countermeasure techniques are employed to reduce (break) the correlation between the physical leakage information with processed data.

*Corresponding Author: Ali Akbar Pammu, School of Electrical and Electronic Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798 | Email: ali1@e.ntu.edu.sg

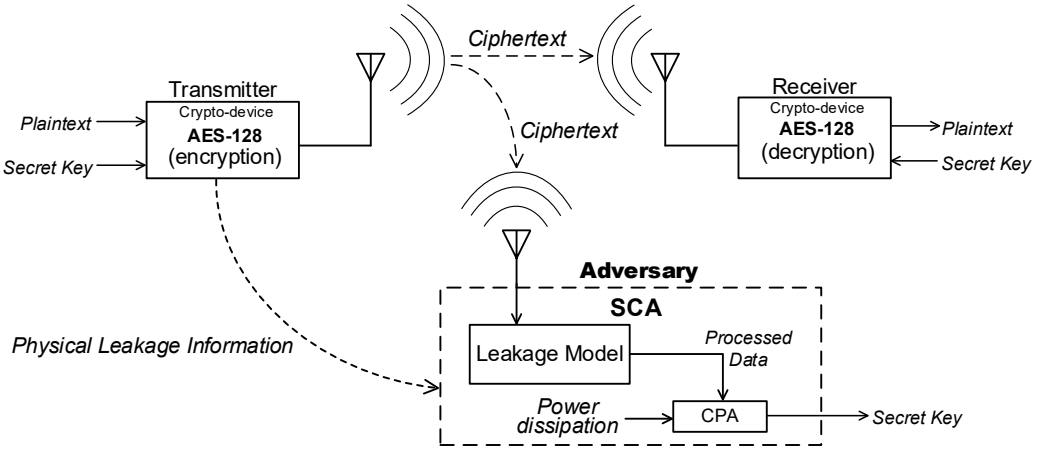


Fig. 1: Attacking scenario of SCA of wireless communication based on AES-128 implementation

The countermeasure technique is classified into two main classifications, hiding and masking, which are based on hardware and software approaches respectively [1]. The hiding technique breaks the correlation between power dissipation and processed data by balancing the power dissipation for different processed data whereas the masking technique employs masking variable (*m*) to mask the processed data against SCA. The main drawback of the masking technique is to mask and unmask of the *m* which can degrade the performance of the cryptographic algorithm implementations, such as low throughput, speed reduction and high power dissipation [2].

There are two main approaches of hiding technique, cell and block level approaches. In the cell level, several techniques have been reported such as Sense Amplifier Based Logic (SABL) [5], Wave Dynamic Differential Logic (WDDL) [6], Three-phase Dual-rail Pre-charge Logic (TDPL) [7] and Pre-Charge Static Logic (PCSL) [8]. The concept of SABL is to balance internal charges by fully charging and discharging all internal node for different processed data (i.e. bit-0 or bit-1). However, during the implementation in crypto-device, the internal charges is not fully discharged at high frequency (>100MHz) due to small variation on the internal parasitic capacitance [6]. The WDDL and PCSL implement Pre-charge and Evaluation cycle with differential logic to make a constant power dissipation for different logic transition. In the AES-128 implementation, the WDDL occupies over 3.1× area, dissipate 3.7× dynamic power and 3.8× reduction in throughput compared with standard cell implementation [5]. For the PCSL implementation, the power dissipations tend to leak information during the pre-charge cycle [6] and hence vulnerable against CPA attack. The TDPL employs dual-rail dynamic logic with three-phase clocking system (Pre-charge, Evaluation and Discharge). The three-phase clock is to ensure that the remaining internal charge is fully discharge to make a constant amount of charge for each cycle. However, the TDPL features 4.6× slower speed compared with conventional CMOS implementation [1].

For the hiding approach at the block level, the power dissipation is balanced directly at the main power supply, V_{DD} point, of the crypto-device. The hiding techniques based on block level approach are a Switching Capacitor Current Equalizer (SCCE) [9], an intermittent Supply-Current Equalizer (iSCE) [10] and A Dynamic Voltage and Frequency Switching (DVFS) [11]. The SCCE is the same principal as in the TDPL. The current equalizer implemented with integrated switching capacitors, which isolates the encryption circuits activity by equalizing the current. However, it is 33% power overhead 2× slower than conventional differential logic [8]. The iSCE is an improvement of the SCCE performance, which is only at the vulnerable round of AES-128 (i.e. 1st and 10th rounds) implement the equalizer. The current equalizer techniques (i.e. SCCE and iSCE) are both vulnerable against EM based attack by measuring the EM emanation generated after the equalizer module. The DVFS hides the correlated power dissipation against SCA by dynamically changing the scale of the voltage and frequency during the encryption. The noise generated during the operation can be filtered by Finite Impulse Response (FIR) filter with optimized parameters to increase the Signal-to-Noise Ratio (SNR). Therefore, the correlated power dissipation still can be detected and the secret key can be revealed with required low number of power traces.

In this paper, we propose an arithmetic hiding cum Look-up Table (AHLUT) based Substitution-Box (S-Box) in AES-128 cryptographic algorithm implementation to countermeasure against SCA. There are three key features in our proposed AHLUT S-Box. First, the arithmetic hiding performs four types of arithmetic operations such that their total physical leakage information sufficiently overshadows the correlated physical leakage information of the S-Box operation. This is to reduce the correlation of the AES-128 physical leakage information with the processed data. Second, the AHLUT S-Box pre-stores all the 256 bytes of possible output values based on the conventional S-Box and selects a corresponding output value with respect to the input accordingly. In this context, it dissipates significantly lower power when compared to the conventional S-Box which performs multiplication inversion and affine transformation. Third, we propose a methodology to determine a minimum number of the arithmetic operations to sufficiently overshadow the physical leakage information of the S-Box operation. Based on the measurement results of performing AES-128 algorithm on Sakura-X FPGA encryption-board and in term of power dissipation, our proposed AHLUT S-Box dissipates 1.6mW and features 11.56× lower power dissipation than the conventional S-Box. In term of security which is based on the CPA attack, it requires 73× more power traces to reveal the secret key for our proposed AHLUT S-Box than the conventional S-Box. As for the non-invasive Correlation Electromagnetic Analysis (CEMA) attack, it requires 25× more electromagnetic traces for our proposed AHLUT S-Box than the conventional S-Box.

This paper is organized as follows. Section II reviews the AES algorithm, various S-Box implementations, CPA and CEMA. Section III presents the proposed AHLUT S-Box. Section IV describes the measurement results on CPA and CEMA attack and finally, conclusions are drawn in Section V.

2. Advanced Encryption Standard, Substitution-Box operation, CPA and CEMA

In this section, an overview of the AES algorithm is briefly described followed by a description of the various topologies of the S-Box operation and the correlation based attack, the CPA and CEMA.

2.1. Advanced Encryption Standard

The AES algorithm has been employed in a variety of security systems including the defense and banking applications since 2001 [6]. It is categorized as a symmetric-key encryption algorithm, in which the transmitter and receiver employ the same secret key for encryption and decryption respectively. The AES algorithm transforms a plaintext into a ciphertext using the secret key by several iterative processes. The processed data block length is fixed at 128 bits, while the key length can be 128, 192, or 256 bits [1]. For the 128, 192 and 256 secret key length, there are 10, 12 and 14 round of iterations are required respectively.

Fig. 2 depicts the flow chart of the encryption process in the AES algorithm. Each round of iteration consists of four operations, namely S-Box, ShiftRow, MixColumn and AddRoundKey, except for the last round which does not have MixColumn operation. The decryption is a reverse operation of the encryption process, i.e. transforming the ciphertext into plaintext (original message) using the same secret key. The decryption structure can be derived by inverting the encryption structure directly [12]. The equivalent decryption structure has the same sequence of operation as in the encryption structure, thus, the resources sharing is allowed for the encryption and decryption process.

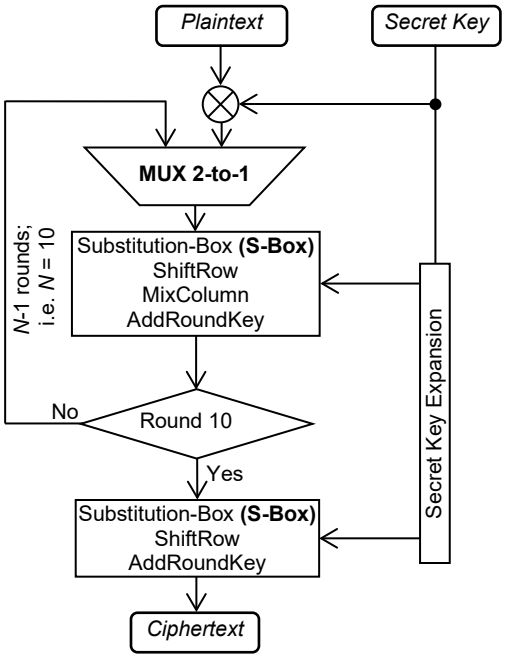
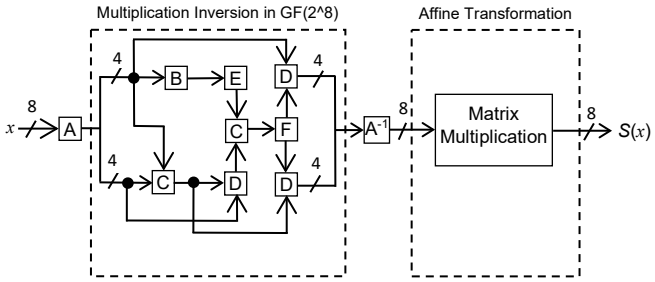


Fig. 2: Flow chart of the encryption process of the AES-128 algorithm with 10 round iterations

2.2. Substitution-Box

The S-Box is one of the critical operations in AES algorithm and it consists of two sub-modules [1], namely the multiplicative inversion sub-module in GF(2^8) and the Affine transformation sub-module as depicted in Fig. 2. Each input to the S-Box is a 1-byte of intermediate data, x, and the S-Box will generate 1-byte of output $S(x)$. In term of power, it dissipates 65% - 80% of the total power dissipation of the AES implementation [1]. Based on these two sub-modules, the S-Box features a non-self-inverse function, which effectively protects the data against the brute force attacks.



A = isomorphic mapping
A^{-1} = inverse isomorphic mappings
B = square operation in GF(2^4)
C = sum operation in GF(2^4)
D = multiplication operation in GF(2^4)
E = multiplication with constant operation
F = inverse operation in GF(2^4)

Fig. 3: The two sub-modules of a conventional S-Box of AES algorithm

The S-Box operation can be implemented in the form of the LUT, in which all the possible output (2^8 = 256) are pre-stored in the LUT memory, as depicted in Fig. 4. The analysis of the LUT S-Box [13] shows that the power dissipation is reduced significantly by 5.5× lower than conventional S-Box (reduces from 10.5mW to 1.9mW). However, the security features can only protect the key against CPA attack up to 13×10^3 power traces. This is due to the selection function (in the multiplexer) dissipates relatively small differences of the dynamic power for different input and output values of the S-Box.

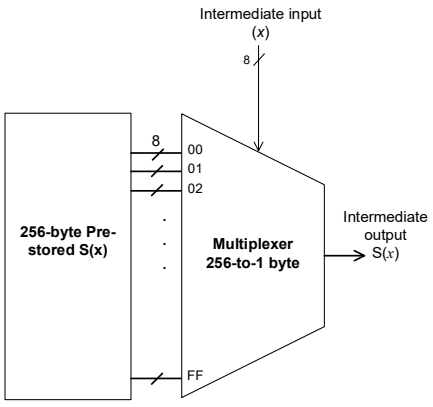


Fig. 4: LUT S-Box implementation

The data dependency with physical parameters is relatively high in the S-Box although the power dissipation can be significantly reduced by LUT technique. The variance of power traces for different input values is still detectable by CPA attack [13]. However, the low power dissipation at LUT architecture is possible to apply dummy operations, which hide the correlated power dissipation against CPA attack without sacrificing the power overhead.

2.3. CPA and CEMA

The CPA attack is a byte-based power analysis attack. Each byte of key (sub-key) is estimated by means of 256 possible values ($2^8 = 256$). The CPA attack is performed by analyzing the correlation coefficient ($r_{i,j,t}$) of two variables, power model ($X_{i,j,m}$), and the power traces ($Y_{t,m}$), for i = 1, …, 16 sub-keys, j = 1, …,256 sub-key candidates, t = 1, …, N sampling points, as follows:

$$r_{i,j,t} = \frac{\sum_{m=1}^{n}\left(X_{i,j,m} - \overline{X}_{i,j}\right)\left(Y_{t,m} - \overline{Y}_{t}\right)}{\sqrt{\sum_{m=1}^{n}\left(X_{i,j,m} - \overline{X}_{i,j}\right)^2 \left(Y_{t,m} - \overline{Y}_{t}\right)^2}} \quad (1)$$

The correct sub-key, i, corresponds to the highest $r_{i,j,t}$ at the particular sub-key candidate, j, and sampling point of power traces, t. The common power model used is either Hamming Distance (HD) or Hamming Weight (HW). The higher number of power traces required to reveal the correct sub-key, the higher CPA-resistant to the hardware, hence more secured.

The CEMA attack applies the same procedure as the CPA attack. However, the process of acquiring the physical leakage information is less invasive compare with the CPA. The adversary can simply measure the EM emanation by placing the EM probe on the crypto-device and apply the Eq. (1) to reveal the secret key.

3. Proposed AHLUT S-Box

The power dissipation generated from the crypto-device is resulted from current (I_{DD}) and voltage (V_{DD}) consumed during the processing one plaintext (consists of 16 bytes). The collection of power dissipation ($P = I_{DD} \cdot V_{DD}$) measurements form power traces [1] and the power traces can be decomposed as the total sum of power dissipations of an operation, P_{OP}, processed data, P_{DATA}, noise, P_{NOISE} and constant power (static power dissipation), P_{CONST} as described in Equation (2).

$$P_{TRACES} = P_{OP} + P_{DATA} + P_{NOISE} + P_{CONST} \quad (2)$$

The P_{OP} and P_{DATA} are generated when performing the operations of the AES-128 algorithm with different value of input data, which are the main physical leakage information employed for the SCA. The P_{NOISE} can be generated from the crypto-device and measurement tools (i.e. oscilloscope) which exhibit different noise level for different application (i.e. ASIC or FPGA) and specification respectively. The P_{NOISE} can be filtered out by means of FIR filter with optimized parameters and hence increase the SNR value of the power traces. The P_{CONST} is relatively irrelevant to the SCA, since the value is generated constantly for different operation and processed data. In this context, the SCA only consider two power dissipation components, P_{OP} and P_{DATA}, to leak out the information of the secret key.

The proposed arithmetic hiding based LUT S-Box technique, in this paper, is focused on the P_{OP} and P_{DATA}, to decorrelate between the physical leakage information and the processed data, based on the CPA attack. The main idea is to generate dummy power dissipation by performing dummy operation (P_{D_OP}) with dummy input data (P_{D_DATA}). Therefore, the total power traces measurement is the sum of main power dissipation and dummy power dissipation as expressed in Equation (3).

$$TP_{TRACES} = P_{OP} + P_{D\_OP} + P_{DATA} + P_{D\_DATA} \quad (3)$$

To break the correlation between power dissipation measurement and processed data, based on the CPA attack, the dummy power dissipation must be able to dominate the total power traces. In other words, the dummy power dissipation is generated in such a way that overshadow the main power dissipation of the AES-128 algorithm, as expressed in Equation (4). In this context, the changes of P_{OP} and P_{DATA}, are negligible respect to the total power traces which are used in the CPA attack as expressed in Equation (5). Eventually, the broken correlation, between power dissipation and processed data, is achieved due to the total power traces measured during the encryption is always referring to dummy operation, which is performing irrelevant operation and data with the operations in the AES-128 algorithm.

$$TP_{TRACES} = P_{OP}\downarrow + P_{D\_OP}\uparrow + P_{DATA}\downarrow + P_{D\_DATA}\uparrow \quad (4)$$

When P_{OP} and P_{DATA} are negligible respect to P_{D_OP} and P_{D_DATA}, the Eq. (4) can be rewrite as follows.

$$TP_{TRACES} \simeq P_{D\_OP} + P_{D\_DATA} \quad (5)$$

The total power traces as expressed in the Eq. (5) can be realized by performing the arithmetic operations in parallel with the operation (i.e. S-Box) of the AES-128. The S-Box operation dissipates 80% of the total power dissipation [13] and leak more information about the secret key, whereas other three operations (AddRoundKey, ShifRow and MixColumn) insignificantly leak the information of the secret key [1] due to non-data dependent operations. To overcome the power overhead and yet secure S-Box, we adopt LUT S-Box which is performed in parallel with arithmetic operation to overshadow the P_{OP} and P_{DATA} of the S-Box. The LUT S-box is adopted due to the power dissipation is 5.5× lower than the conventional S-Box (1.9mW) [13]. Consequently, it is worthwhile to implement the arithmetic operation with sufficient power overhead to overshadow the power dissipation of the S-Box operation. Fig. 5 depicts our proposed AHLUT S-Box implementation, performed in parallel with LUT S-Box. Four types of arithmetic operations, implemented, Addition (ADD), Subtraction (SUB), Division (DIV) and Multiplication (MULT) are implemented which generate unused output, Dummy output, D(x).

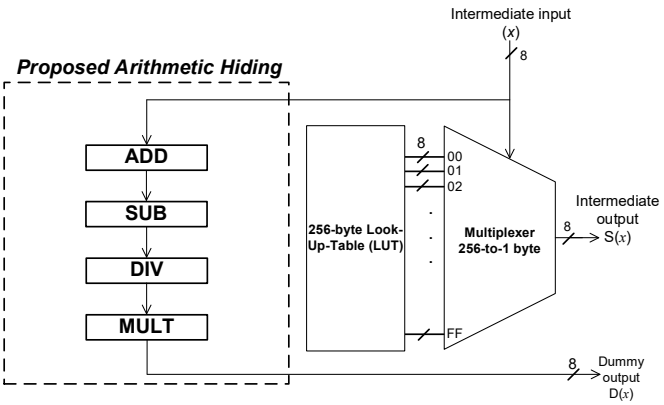


Fig. 5: Proposed arithmetic hiding is performed in parallel with LUT S-Box

In term of circuit implementation, each type of arithmetic operation requires different number of gates and dissipate different power. In addition to our proposed arithmetic hiding based LUT S-Box, we propose a methodology to select the number of arithmetic operations such that the power dissipation of these arithmetic operations sufficiently overshadows the S-Box power dissipation. Table I tabulates the number of gates and power dissipation for each arithmetic operation based on frequency of 16MHz.

TABLE. I. GATE COUNTS AND POWER DISSIPATION IN AHLUT S-BOX

Arithmetic Operations	Number of Gates			Power dissipation (mW*)
	XOR	AND	OR	
ADD	2	2	1	0.094
SUB	2	2	2	0.102
DIV	2	6	2	0.123
MULT	2	7	2	0.160

*dynamic power dissipation @16MHz

The methodology of selecting the number of arithmetic operations of our proposed AHLUT S-Box is explained as follow:

1. The power dissipation of the LUT S-Box is measured at the first stage to estimate the minimum power dissipation which will be generated by arithmetic hiding (overshadow power dissipation).
2. The power dissipation for each arithmetic operation is measured and sorted from lowest to highest, which are denoted as *a*, *b*, *c* and *d* in ascending order respectively.
3. The rule of thumb for estimating the overshadowed power dissipation is $P_{Arithmetic\_operations} > P_{LUT\_S\text{-}Box}$, in which $P_{Arithmetic\_operations} = a{\cdot}P_{ADD} + b{\cdot}P_{SUB} + c{\cdot}P_{DIV} + d{\cdot}P_{MULT}$; $a = b = c = d \geq 1$, as expressed in Equation (6).

$$a \cdot P_{ADD} + b \cdot P_{SUB} + c \cdot P_{DIV} + d \cdot P_{MULT} > P_{LUT\_S-Box} \quad (6)$$

4. The number of arithmetic operations are increased starting from *d* and evaluating the power dissipation in the Eq. (6) every increment.
5. If power dissipation of arithmetic operation is overshoot as in Eq. (6), the selection is gradually descended from *c* to *a*.
6. The overshoot power dissipation is only allowed with incremental number of *a*, which is the lowest power dissipation in the arithmetic operation.
7. The selection process for the number of arithmetic operation is terminated when the $P_{Arithmetic\_operations}$ is slightly higher than $P_{LUT\_S\text{-}Box}$

The output of the arithmetic operation is unconnected to the AES-128 encryption process and therefore, the output ciphertext will not be affected. The flow chart of selecting the arithmetic operation is depicted in the Fig. 6.

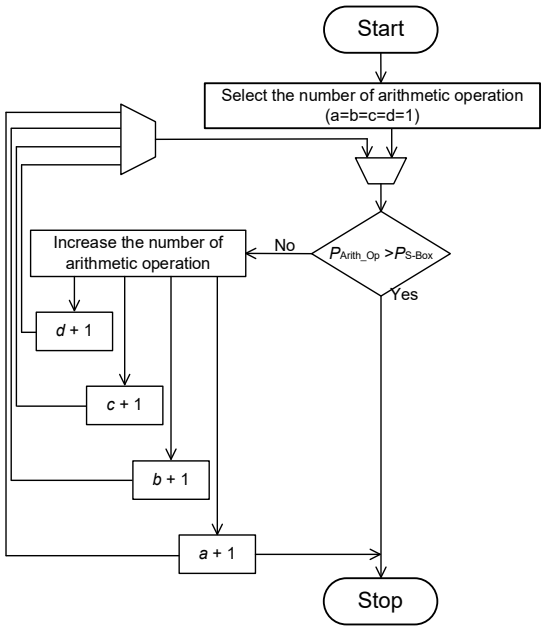


Fig. 6: Flow chart of selection the Arithmetic operations

4. Measurement Results

The experiment is performed based on Sakura-X, FPGA board [2], incorporating our proposed arithmetic hiding LUT S-Box in the AES-128 with operating frequency of 16MHz. The experimental setup comprises two parts, power dissipation and EM emanation measurements as depicted in Fig. 7. A 10-bit ADC 2.5Giga samples/second oscilloscope is used to record the power dissipation and EM emanation of the AES-128 implementation. We attack the last round of the AES-128 designs with the HD leakage model (power and EM model).

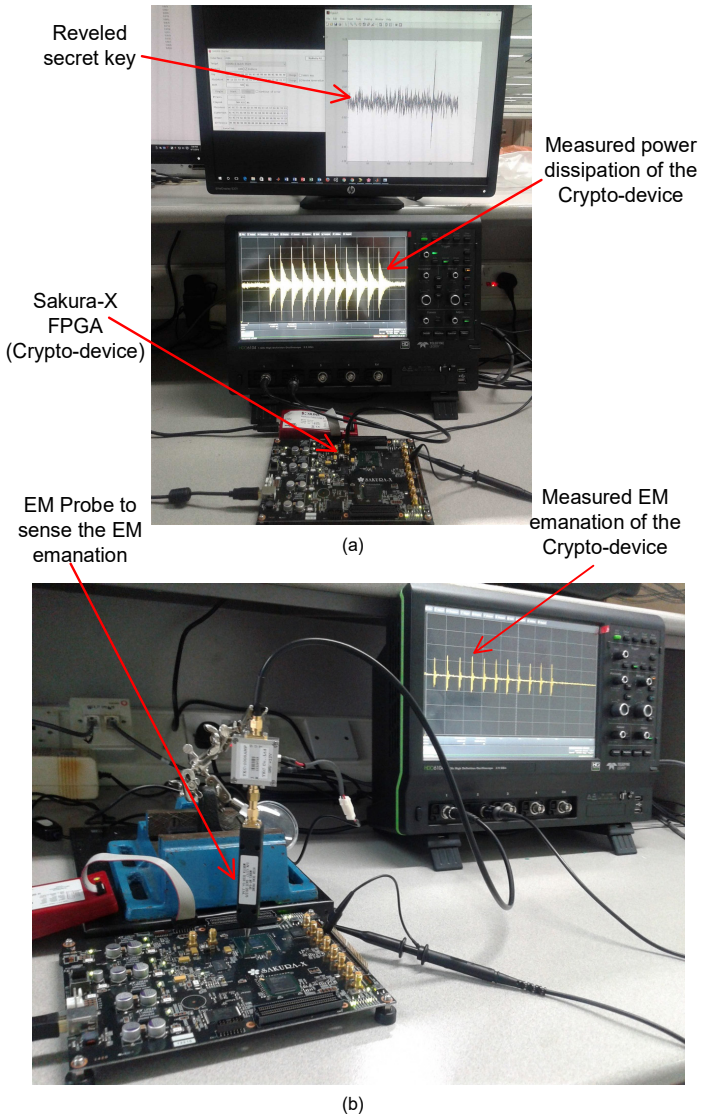


Fig. 7: The experimental setup (a) power dissipation measurement for CPA and (b) EM emanation measurement for CEMA attacks

Based on the power dissipation measurement of the AES-128 with LUT S-Box, we can estimate the number of the arithmetic operations can be employed to overshadow the power dissipation LUT S-Box as depicted in the Fig. 5. Table II tabulates the measurement result of LUT S-Box and the number of arithmetic operation required to overshadow the LUT S-Box.

TABLE. II. CIRCUIT MODULE IMPLEMENTATION OF AHLUT S-BOX

Circuit Module	Power dissipation (mW)
LUT S-Box	0.785
Arithmetic Operations*	0.813

*2 ADD; 1 DIV; 1 SUB; 4 MULT

The power dissipation of the arithmetic operation in the Table II is slightly higher than LUT S-Box with 0.028mW power overhead. The implementation of the arithmetic operation performed in parallel with LUT S-Box is depicted in Fig. 8. The AND logic gate is embedded to activate the arithmetic operations only when the LUT S-Box is performing the operation to prevent leakage current and dissipate additional power at the arithmetic operation (static power dissipation). The input data for arithmetic operation is 16-bit which comes from input LUT S-Box (x). The 8-bit input is inverted to make the input value opposite against the x such that uncorrelated with processed data.

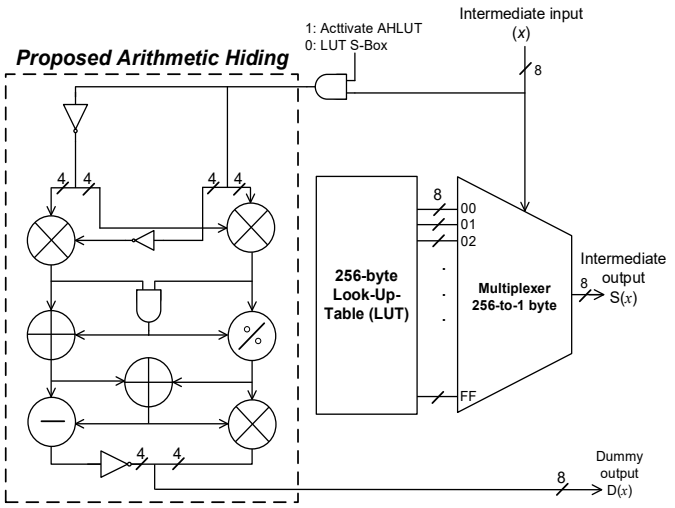


Fig. 8: Arithmetic operation is performed in parallel with LUT based S-Box

During the implementation in the circuit level, both modules (LUT S-Box and arithmetic operations) are performed at the same clock cycle, to hide the correlation with processed data, with the total power dissipation is 1.598mW. The power dissipation for AES-128 with LUT S-Box only and the AES-128 with arithmetic operation performed in parallel with LUT S-Box are depicted in Fig. 9. The grey color is indicated as arithmetic operations with LUT S-Box dissipates higher power as to overshadow the LUT S-Box as plotted in black color. As resulted from implementation of the controller, the static power dissipation is remined the same as in the LUT S-Box (~0.95mW) which implies the leakage current is negligible during the performance of arithmetic operations.

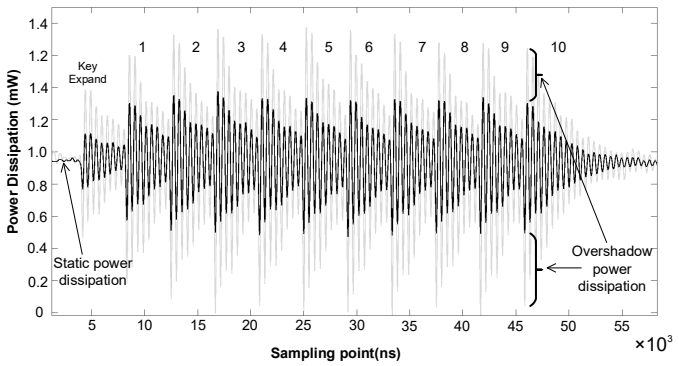


Fig. 9: Power dissipation measurement of 10 rounds iterations AES-128 with LUT S-Box and our proposed arithmetic hiding.

It is worthwhile to note that although in our proposed AHLUT S-Box, the power dissipation overhead is ~2× higher than LUT S-Box, but it is much lower than conventional S-Box implementation, which is 5.6× lower than conventional S-Box. Table III depicts the power dissipation of three S-Box implementations, LUT S-Box, proposed AHLUT S-Box and conventional S-Box.

TABLE. III. POWER DISSIPATION OF THREE DIFFERENT S-BOX TOPOLOGIES

S-Box Topology	Power dissipation (mW)	Normalized Power
LUT S-Box	0.785	0.5×
Conventional S-Box	18.37	11.56×
Proposed AHLUT S-Box	1.589	1×

In this experiment, the CPA attack is performed based on the proposed AHLUT S-Box and compare the result against LUT S-Box. Fig. 8 depicts the number of traces required to reveal the most difficult sub key of AES-128. It shows that the Fig. 10(a) requires 516 power traces to reveal the secret key while with our proposed AHLUT S-Box requires 1,751 power traces to reveal the same secret key as depicted in Fig. 10(b). Based on the CPA attack on single byte secret key, our proposed arithmetic hiding features 3.4 × secured than LUT S-Box implementation with ~2× power dissipation overhead.

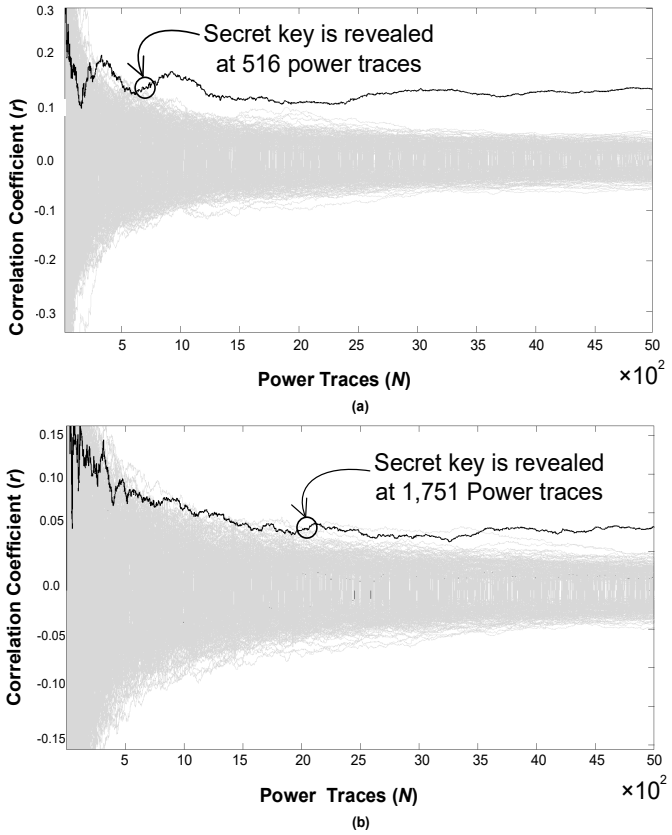


Fig. 10: Evaluation of the CPA as security features of (a) LUT S-Box and (b) our proposed AHLUT S-Box

Fig. 11 depicts the CPA and CEMA attack based on the proposed AHLUT S-Box. It shows that the 16-byte sub-secret key has been successfully revealed at 38×10^3 and 44×10^3 of the power and EM traces respectively. In this context, the security features of the AES-128 has been increased against CPA and CEMA by 73× and 25× respectively when compared with conventional S-Box.

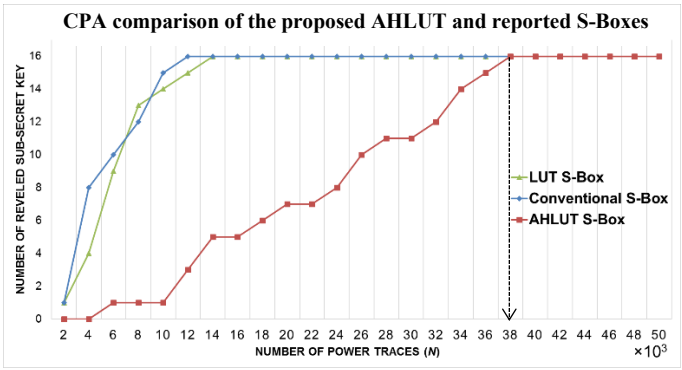


Fig. 11: Comparison of proposed AHLUT S-Box and reported hiding based on CPA attack

When comparing with reported hiding counterparts, SCCE [9] and iSCE [10], which can only protect the AES-128 algorithm against CPA attack, our proposed AHLUT S-Box can protect the AES-128 algorithm against EM based attack (CEMA) as well as in the CPA. Fig. 10 depicts the performance result of the CEMA attack for various hiding techniques. As shown in the Fig. 12, our proposed AHLUT S-Box outperforms the reported hiding techniques, which is requires 44×10^3 EM traces to reveal all the 16-byte secret key. The result is 5.5×, 4.8× and 44× higher than SCCE, conventional S-Box, iSCE and LUT based S-Box respectively. This is due to the physical leakage information (EM emanation) is generated by arithmetic operation, can dominate the EM generated by LUT S-Box and reduce the correlation between the EM signal of the LUT S-Box and the processed data of the AES-128 algorithm.

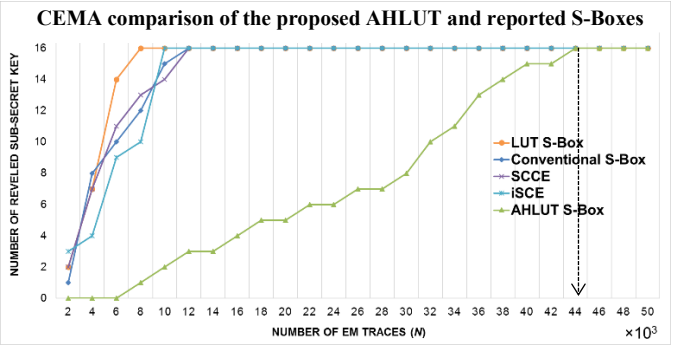


Fig. 12: Comparison of proposed AHLUT S-Box and reported hiding based on CEMA attack

5. Conclusions

We have proposed an AHLUT S-Box in the AES-128 cryptographic algorithm implementation to countermeasure against SCA. There are three key features in our proposed AHLUT S-Box. First, the arithmetic hiding performs four types of arithmetic operations such that their total physical leakage information sufficiently overshadows the correlated physical leakage information of the S-Box operation. Second, the AHLUT S-Box pre-stores all the 256 bytes of possible output values based on the conventional S-Box and selects a corresponding output value with respect to the input accordingly. Third, we propose a methodology to determine a minimum number of the arithmetic operations to sufficiently overshadow the physical leakage information of the S-Box operation. Based on the measurement results of performing AES-128 algorithm on Sakura-X FPGA encryption-board and in term of power dissipation, our proposed AHLUT S-Box dissipates 1.6mW and features 11.56× lower power dissipation than the conventional S-Box. In term of security of the CPA attack, it requires 73× more power traces to reveal the secret key for our proposed AHLUT S-Box than the conventional S-Box. As for the non-invasive CEMA attack, it requires 25× more EM traces for our proposed AHLUT S-Box.

Acknowledgment

This research work was supported by Agency for Science, Technology and Research, Singapore, under SERC 2013 Public Sector Research Funding, Grant No: SERC1321202098. The authors thank A*STAR for the kind support in funding this research.

References

[1] S. Mangard, E. Oswald, and T. Popp, *Power Analysis attacks: Revealing the secrets of smart cards*. 2007.

[2] Y. Hori, T. Katashita, A. Sasaki, and A. Satoh, "SASEBO-GIII: A hardware security evaluation board equipped with a 28-nm FPGA," in *1st IEEE Global Conference on Consumer Electronics 2012, GCCE 2012*, 2012, pp. 657–660.

[3] S. Ghosh and I. Verbauwhede, "BLAKE-512-based 128-bit CCA2 secure timing attack resistant McEliece cryptoprocessor," *IEEE Transactions on Computers*, vol. 63, no. 5, pp. 1124–1133, 2014.

[4] L. Zhuang, F. Zhou, and J. D. Tygar, "Keyboard acoustic emanations revisited," in *Proceedings of the 12th ACM conference on Computer and communications security - CCS '05*, 2005, vol. V, no. November, p. 373.

[5] K. Tiril and I. Verbauwhede, "Charge Recycling Sense Amplifier Based Logic: Securing Low Power Security IC's against DPA," 2004, pp. 179–182.

[6] K. Tiri *et al.*, "AES-based cryptographic and biometric security coprocessor IC in 0.18-µ;m CMOS resistant to side-channel power analysis attacks," *IEEE Symposium on VLSI Circuits, Digest of Technical Papers*, vol. 2005, pp. 216–219, 2005.

[7] M. Bucci, L. Giancane, R. Luzzi, and A. Trifiletti, "A flip-flop for the DPA resistant three-phase dual-rail pre-charge logic family," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 20, no. 11, pp. 2128–2132, 2012.

[8] K.-S. Chong *et al.*, "Counteracting Differential Power Analysis: Hiding Encrypted Data from Circuit Cells," in *2015 IEEE International Conference on Electron Devices and Solid-State Circuits (EDSSC)*, 2015, pp. 297–300.

[9] C. Tokunaga and D. Blaauw, "Securing encryption systems with a switched capacitor current equalizer," *IEEE Journal of Solid-State Circuits*, vol. 45, no. 1, pp. 23–31, 2010.

[10] N. Miura, D. Fujimoto, R. Korenaga, K. Matsuda, and M. Nagata, "An Intermittent-Driven Supply-Current Equalizer for 11x and 4x Power-Overhead Savings in CPA-Resistant 128bit AES Cryptographic Processor," in *IEEE Asian Solid-State Circuits Conference*, 2014, pp. 9–12.

[11] T. Güneysu and H. Handschuh, "SoC It to EM: ElectroMagnetic Side-Channel Attacks on a Complex System-on-Chip," in *Cryptographic hardware and embedded systems – CHES 2015: 17th international workshop Saint-Malo, France, september 13–16, 2015 proceedings*, 2015, vol. 9293, pp. 620–640.

[12] A. A. Pammu, K.-S. Chong, and B.-H. Gwee, "Secured Low Power Overhead Compensator Look-Up-Table (LUT) Substitution Box (S-Box) Architecture," in *IEEE International Conference on Networking, Architecture and Storage (NAS), Aug. 2016*, 2016, pp. 1–7.

[13] A. A. Pammu, K.-S. Chong, K. Z. L. Ne, and B.-H. Gwee, "High Secured Low Power Multiplexer-LUT Based AES S-Box Implementation," in *2016 International Conference on Information Systems Engineering (ICISE)*, 2016, pp. 3–7.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 427-434 (2017)

www.astesj.com

Special issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Service Productivity and Complexity in Medical Rescue Services

Markus Harlacher*, Andreas Petz, Philipp Przybysz, Olivia Chaillié, Susanne Mütze-Niewöhner

Institute of Industrial Engineering and Ergonomics, Work Organization Department, RWTH Aachen University, 52062 Aachen, Germany

ARTICLE INFO

Article history:
Received: 05 April, 2017
Accepted: 07 May, 2017
Online: 24 May, 2017

Keywords:
Service Productivity
Service Complexity
Construct Reliability
Construct Validity

ABSTRACT

The increasing impact of services on economic wealth and income in high-wage countries calls for a detailed examination of service productivity and complexity in current research projects. This paper describes the development of a measurement model of service productivity and complexity within the domain of medical service provision. While the measurement model of service productivity is partly literature-based, a novel construct is developed for service complexity. The reliability and validity of these constructs have been analyzed by a survey with 454 employees of German medical rescue service providers. Reliability of the construct is estimated by analyzing Cronbach's alpha, average variance extracted, variance inflation factors and the significance, weights and loadings of the items. Furthermore, the relationship between the partial productivities (capacity, internal and external) as well as complexity and service performance is analyzed by using the structural equation modeling approach with partial least square method (PLS-SEM). The results indicate that the developed constructs have sufficient reliability, although the examination of validity highlights difficulties in quantifying service productivity and complexity. While a significant and substantial effect of internal and external efficiency on service performance could be proved, complexity influences performance only marginally and positively while capacity efficiency has no effect at all. Thus, further research is needed in order to foster the measurement model and to investigate possible nonlinear relationships between the constructs in more detail.

1. Introduction

1.1. Service Productivity and Complexity

The service sector represents the largest part of the economy in the majority of the developed countries. In Germany, the tertiary sector's proportion of the gross domestic product (GDP) was 69% in 2015. In the US, the proportion was even higher, reaching 78%. This highlights the importance of services in overall economy [1]. Focusing the impact services have on economic growth, employment and international competitiveness, scientific studies are essential for a comprehensive understanding of service productivity and complexity [2].

In previous research [3-6] different models describing service productivity and complexity were developed. Thus, different ways to operationalize service productivity were assessed. Based on these results, service productivity and complexity within the domain of medical services are discussed in this paper.

1.2. Rescue Services and Electric Mobility

Due to technological advances in electric mobility, changes in process of provision of emergency services are inevitable. Ambulance services are not only essential for society, they also occur with a high frequency. During 2012/2013 there were more than 14 million rescue missions in Germany. More than 270,000 rescue missions were caused by traffic accidents [7]. Driven by technological change and growing environmental awareness as

*Corresponding Author: Markus Harlacher, Bergdriesch 27, 52062 Aachen, Germany | Email: m.harlacher@iaw.rwth-aachen.de

well as falling prices the number of electrically-powered vehicles increases. In consequence, the increasing number of electric mobility vehicles in road traffic will affect the provision process of ambulance services. For example, the measures to guarantee the personnel protection in road traffic accidents will be more challenging and more varied caused by the higher number of vehicle variants. Moreover, novel injury patterns will occur in the context of road traffic accidents. These changes in service provision must be examined and the employees need to be sufficiently and appropriately trained for safe rescue missions with electric vehicles. Thus, higher acceptance of electric mobility can be reached.

1.3. Objectives and Structure

This paper aims to empirically measure and evaluate the relationship between service productivity and complexity and the effect on medical rescue service performance. For this purpose, a measurement model of service productivity and complexity is developed. Furthermore, reliability and validity of the measurement model is assessed by specific criteria. Following the findings, the process of the rescue service can be adapted for increasing service productivity and performance.

The paper is structured as follows: First, a literature based definition of service productivity and complexity is given. The hypotheses and the study design are presented in chapter 3. The model evaluation in chapter 4 is divided in the measurement model evaluation focusing reliability and validity and structural model evaluation focusing on testing the hypotheses. The results and the limitations of this study are discussed in chapter 5. Finally, key findings for practitioners and academia are presented.

2. Literature Review

2.1. Defining Service Productivity

Starting point of any productivity measure is the goods-based definition of productivity as the ratio of output to input. It measures the performance of the operational factor combination: e.g. number of vehicles produced in one hour. Usually traditional performance indicators of productivity are made out of the same physical dimensions in numerator and denominator. Due to specific characteristics of services like customer integration and their immediate effect on productivity and complexity the traditional concept of productivity is too narrow. Researchers have proposed different concepts for service specific productivity evaluation and service system design.

Production-oriented approaches go back to Levitt [8] and Jones [9] as well as Corsten [10]. They define partial productivities in different, subsequent stages of service provision focusing on capacity utilization and one sided, technical quality management. A more comprehensive view was taken by Johnston and Jones [11] as well as Grönroos and Ojasalo [4]. Grönroos and Ojasalo define service productivity as a function of three elementary measures of performance [4]:

1. Efficiency of transforming input resources to output result for the customer (internal efficiency)
2. Perception of the quality of the service process and of the outcome (external efficiency)
3. Efficiency of utilizing the capacity of the service process (capacity efficiency)

Customer integration and internal processes were identified as the relevant influencing factors of internal efficiency. Customer satisfaction as well as corporate and customer benefits are key figures of external efficiency. Finally, quantitative and qualitative congruence (accordance of supply and demand) defines the core of capacity efficiency [5, 6].

2.2. Defining Service Complexity

Complex services, like emergency medical services, consist of a variety and diversity of interdependent system elements [5]. Service complexity as the property of these elements assumes a variety of emergent states in a defined period of time [12]. This in turn causes variety, makes complex services hardly predictable and uncertain in their behavior [13].

Blockus [14] defines service complexity as the characteristic of the system "service provider" concerning the variety, diversity, interdependencies and dynamic of the three dimensions of service: potentials, processes and outcomes. The "service potential" dimension describes the resources of a service provider for fulfilling the demand for services. Considering Bruhn and Georgi [15] these particularly involve the employees with their administrative and operative abilities, capacities and their qualifications including applied technologies and physical environment such as subsidiaries, information leaflets and blank forms.

The "service processes" dimension includes the complexity of direct and indirect processes of service provision as well as the autonomous customer process. The direct processes regard the actual service provision taking customer into account. The indirect processes are supporting processes as planning and procuring the resources or marketing the services. Autonomous customer processes are to be executed by the customer without direct involvement of the service provider. Moreover, the service processes dimension considers the complexity of external factors. External factors are characterized by not being predictable and influenceable by the service provider. These include the customer itself, its mobile and immobile properties as well as necessary permissions and information for service fulfillment. [14]

The "service outcomes" dimension refers to service results, which can be sold as market products. It considers the complexity of the applied service portfolio and of the single investigated service itself. [14]

3. Method

3.1. Study Design

The data for testing the hypothesized relationships were gathered by a questionnaire survey. 454 persons employed in the knowledge-intensive medical rescue service sector in Germany participated in the survey. A structural equation modeling method (SEM) was used for statistical analysis of the service productivity model. An empirical validation study was carried out and analyzed using partial least squares (PLS) path modeling as well as SPSS (v.21) for data preparation, data cleaning and descriptive statistics. The measurements and the structural model were analyzed using the SmartPLS 3 software package [16].

3.2. Hypotheses

Productivity describes the relation between process result and its means deployed. Considering the economical definition of productivity, service productivity and service performance are correlated positively. Thus, measures of efficiency should also correlate positively with service performance. The influence of the single efficiencies still remains uncertain.

An increasing internal efficiency means transforming input to output by minimizing input while maintaining the output or maximizing output while keeping inputs constant [11]. Thus service productivity increases and we posit:

H_1: Increasing internal efficiency positively influences service performance.

The better the perceived quality by using a specific amount of input resources, the better is the external efficiency, which results in an improved service performance. A decreasing perceived quality in turn causes a lowered "revenue-generating capability" which has a negative impact on service performance. [17]. Therefore hypothesis 2 is as follows:

H_2: Increasing external efficiency positively influences service performance.

The capacitive efficiency correlates positively with service productivity up to its optimum – demand equal to supply – because the costs for resources decrease for a single service. Assuming that service capacities in rescue services can only be overloaded in extreme situations (as they are highly important for society safety), overloaded resources should not affect the quality of service performance negatively [17]. Thus hypothesis 3 implies:

H_3: Increasing capacitive efficiency influences service performance positively.

A high complexity implies more procedures for handling it [18]. These extra activities disturb the working process and thus reduce service performance. Therefore hypothesis 4 can be derived as follows:

H_4: High complexity due to increased electro mobility influences service performance negatively.

3.3. Questionnaire

In order to operationalize the constructs items were formulated based on existing questionnaires as well as on literature on the dimensions of complexity [4-6]. Items were formulated and evaluated in semi-structured interviews with a group of experts. Furthermore, a pre-test was conducted for optimization purpose.

The structure and content of the questionnaire are depicted in Figure 1. The questionnaire can be divided into four parts. The first two parts collect personal and organization-specific information. The third part deals with already gained experience concerning electric mobility and expected problems caused by increased number of licensed electric vehicles with impact on service productivity and complexity. Finally respondents were asked to evaluate their performance regarding service productivity and complexity in the fourth part.

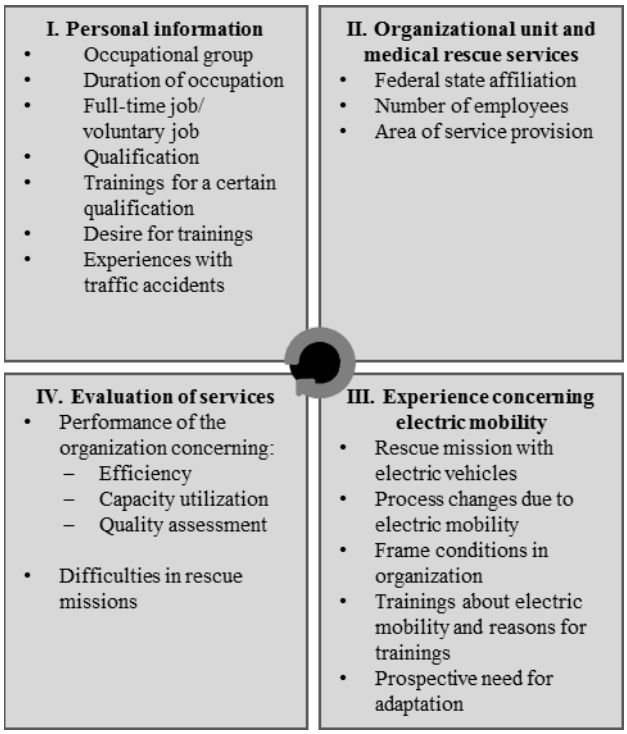


Fig. 1. Structure of the questionnaire.

Measurement Model of Service Productivity

In order to operationalize the constructs of service productivity 12 items were formulated based on the work of Grönroos and Ojasalo and Garrel et al. [4, 6]. In summary, there are five items that quantify internal efficiency. Customer integration is operationalized by an item assessing whether frequent interactions with involved parties in accidents takes place without problems. Quality of internal processes is measured by the following statements:

1. The rescue mission is predominantly smooth.
2. There are delays in the rescue mission caused by insufficient technical equipment.
3. Rescue teams always get the most recent information to fulfill the task.
4. The work structure (e.g. authority and decision making) enables an interference-free operation.

External efficiency is measured by four items. Two items assess customer satisfaction regarding the level of medical and psychological care during the rescue mission. Organizational and customer benefit is mapped by items addressing the level of standardization of the procedures during the rescue process as well as by the aspect of achieving the target time. The operationalization of the quantitative congruence includes bottlenecks due to technical equipment. The qualitative congruence is operationalized by the following items:

1. Problems occur due to unavailability of sufficiently trained employees.
2. The technical equipment is optimized for the tasks of ambulance service.

Measurement Model of Service Complexity

The results of the pre-test did not confirm a differentiation into the three dimensions of potential, process and outcome complexity. Thus, the measurement model of service complexity was revised in the final questionnaire. In order to assess complexity, rescue service processes of accidents with electric vehicles were compared to rescue service processes of accidents with conventional vehicles. Service complexity was measured by the following items:

1. Known algorithm can be used at car accidents with an e-car.
2. It is more difficult to ensure occupational safety on rescue missions with electric vehicles.
3. Handling electrical vehicles in accidents is more ambitious than handling conventional vehicles.

All items were answered on a six-point Likert scale. The response format ranged from ´Does not apply at all` (1) to ´Does fully apply` (6). For assessing the quality of the measurement model, both models of productivity and complexity were analyzed by means of correlation, reliability, and validity.

3.4. Participants

The nationwide survey was conducted online as well as paper-based on training events for rescue personnel. In total, 454 persons have completed the questionnaire. The majority completed the paper-based version of the survey. Less than 150 participants completed the online version, which was promoted solely within the organizations of the project partners (especially Red Cross Germany and the Interdisciplinary Training Center for Medical Education in Aachen).

65.4% of the subjects were directly involved in the medical care of casualties. The majority of them were paramedics and emergency physician. More than 20% of the participants were firemen focusing the technical aspects of rescue. The remaining subjects held an administrative position in a rescue service organization.

About 75% of the participants were employed in rescue service as their main profession. Most subjects had work experience of more than two years, therefore a sufficient level of experience in rescue services can be assumed. 10% of the participants were already previously involved in a traffic accident with electric-powered vehicles. Given the market share of electric cars of only .4% in Germany, the quota of respondents with electric-powered vehicle experience surpasses the expectations on the survey [19].

4. Model Evaluation

For model validation, we considered empirical measures of the relationships between the proposed indicators and constructs as well as between the constructs. We also evaluated the measurement model by means of reliability and validity.

4.1. Measurement Model Evaluation

The questionnaire used in the survey was based on previous research of Garell et al. [6]. The items used to measure the latent variables were defined as reflective and reliability was analyzed accordingly.

The formal accuracy of the model is assessed by Cronbach`s alpha. The results of the reliability analysis of the measurement model of service productivity are visualized in Figure 2. Only external efficiency exhibits a sufficient reliability. Cronbach`s alpha of the total measurement model of service productivity achieves an unsatisfactory value of .407.

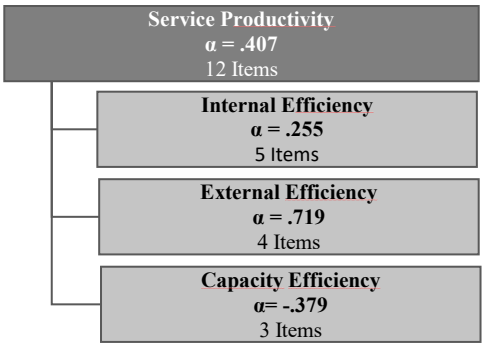


Fig. 2. Cronbach's alpha of the construct service productivity.

Due to the unsatisfactory results of the reliability analysis we examined a formative model as well. Based on the definition of the constructs it seems likely, that while adapting the constructs for the field of medical rescue services the cause-effect-relationship between items and construct changed. Therefore a confirmatory factor analysis is unsuitable. Thus, the structural equation modelling (PLS-SEM) approach is adopted. By using this method we can evaluate the multiple dependencies between productivity, complexity, and actual performance of medical rescue services.

For model validation, we considered empirical measures of the relationships between the proposed indicators and constructs as well as between the constructs. We also evaluated the measurement model by means of reliability.

First, the single reflective measured variable of the self-assessed service performance was analyzed. In order to do so, we looked at composite reliability for evaluating internal consistency. Individual indicator reliability and convergent validity where evaluated by average variance extracted and predictive relevance [20]. Discriminant validity was not investigated due to the fact that the latent variable service performance is the only reflective operationalized construct in the model, so a comparison with other variables is not possible.

Internal consistency reliability is usually measured by Cronbach's alpha. However, due to the PLS-SEM method, the preconditions for applying Cronbach's alpha (equal outer loadings of the indicators) were not satisfied. Thus, composite reliability (ρ_c) was used instead. Composite reliability can vary between zero and one, with a high value indicating high reliability. An acceptable level of .6 to .7 in exploratory research is regarded as satisfactory [20]. The composite reliability of the performance is .635, hence slightly above the required .6. This argues for sufficient internal consistency (measuring the same construct) in the proposed indicators of the self-assessed performance.

Individual indicator reliability is evaluated by the indicators' outer loadings. Their outer loadings should be .7 or higher [20]. In our study, two of three loadings were significantly above .7, thus revealing high individual reliability. For convergent validity, we

considered average variance extracted (AVE). AVE evaluates communality not for a single item but for the whole construct, analogous to individual indicator reliability. AVE indicates the degree to which a construct explains the variance in the indicators. In our study, AVE accounted .44, so the construct explains more than 40% of variance. Table 1 summarizes the results of the assessment.

Table 1. Criteria for reflective measurement model evaluation

Items		$\rho_c \geq .6$	AVE ≥ .5	λ ≥ 0,7
	Self-assessed service performance	**.663**	**.44**	
Perf_1	Please estimate the workload of your organization.			.723
Perf_2	Please estimate the efficiency of your organization.			.098
Perf_3	Please estimate the quality of the performance of your organization as measured by the standard treatment methods (e.g. "Golden Hour", ETC, PHTLS, ITLS).			.887

The majority of the latent variables were operationalized using formative indicators. Compared to reflective indicators, formative indicators define the latent variable as a whole. The indicators describe and specify the content of the construct in a logical manner according to the underlying theory. Formative indicators are the source of the latent construct and therefore are not necessarily intercorrelated. They complement each other within the construct and form a self-contained variable. Thus, reliability measures are inappropriate for evaluating the goodness of the formative measurement model. Instead, content validity is evaluated by collinearity, significance and relevance of the indicators, and discriminant validity at the construct level [20, 21].

Before the relationship and the predictive relevance of the model can be analyzed, the collinearity of the latent variables should be evaluated. In contrast to reflective indicators, where high levels of collinearity between indicators are important, in the case of formative indicators this is counterproductive because they increase standard errors and influence weight estimation negatively. In order to evaluate collinearity, we considered the variance inflation factor (VIF). The square root of the VIF is defined as the degree to which the standard error increases due to collinearity. Removing indicators should be considered in cases where the VIF value is higher than 5 [21]. All of the indicators analyzed showed an acceptable degree of collinearity (VIF<5). Thus, all items were considered for further analysis.

To evaluate the relevance of formative measures, we considered the outer weights of the indicators. In the proposed measurement model, 7 out of 15 items have a significant weight at a significance level of $\alpha = .1$ ($t = 1.65$). Second, we also evaluated the significance of the outer loadings. After analyzing the outer loadings, only 3 out of 15 indicators proved to be non-significant. Two of the items measure complexity and one capacity efficiency. Considering the acceptable degree of collinearity, we retained the items in the model in order to preserve the logical, content-oriented composition of the constructs.

At the construct level, we considered discriminant validity by evaluating the correlations between the latent variables. Scores should be below .9 to ensure that no substantial relationship between different constructs exists [21]. All latent variables were found to be below the given limit. The highest correlation between internal efficiency and external efficiency of provided service (.57) is reasonable and in line with the content. A highly selective specification of the constructs is desirable but not practicable.

Table 2. Criteria for formative measurement model evaluation

Items		VIF<5	weight t>1,65	loading t>1,65
	Complexity			
Comp_1_R	Known algorithm can be used at car accidents with an e-car.	1.015	6.172	5.806
Comp_2	It is more difficult to ensure occupational safety on rescue missions with electric vehicles.	1.404	.223	.124
Comp_3	Handling of electrical vehicles in accidents is more ambitious than handling of conventional vehicles.	1.395	.438	.362
	Capacity Efficiency			
CE_1_R	Problems occur due to unavailability of sufficiently trained employees.	1.227	1.846	1.903
CE_2	The technical equipment is optimized for the tasks of ambulance service.	1.324	1.325	1.604
CE_3_R	There are always bottlenecks with the technical equipment.	1.436	.297	1.798
	Internal Efficiency			
IE_1	The rescue mission is predominantly smooth.	1.242	4.354	10.600
IE_2_R	There are delays in the rescue mission caused by insufficient technical equipment.	1.035	1.648	2.884
IE_3	Rescue teams always get the most recent information to fulfill the task.	1.329	.781	5.106
IE_4	The work structure (e.g. authority and decision making) enables an interference-free operation.	1.430	5.636	13.758
IE_5	The inclusion of other participants (e.g. patients, eyewitnesses, emergency responder) into the rescue process is unproblematically.	1.228	.117	3.677
	External Efficiency			
EE_1	The patient can be treated satisfactorily.	1.690	3.060	11.317
EE_2	We try to rescue crashed persons with standardized procedures.	1.632	4.402	18.146
EE_3	We always manage to keep the planned times (reaching the place of the accident).	1.283	1.235	5.454
EE_4	The patients (eyewitnesses, emergency responder) feel excellently looked after during the rescue.	1.306	.017	3.740

4.2. Structural Model Evaluation

After the measurement model has been evaluated as reliable and valid, the structural model can be assessed accordingly. The structural model is evaluated by the relevance and significance of path coefficients, the coefficient of determination (R^2), effect size (f^2) and predictive relevance (Q^2) [20, 21]. By analyzing the path

coefficient, we examined the hypothesized relationships. The posited relationships were assessed according to the magnitude, significance, and direction of the effect. The results are shown in Figure 3. The path coefficients can take values between -1 and 1, with values close to ±1 representing a strong positive/negative relationship and values close to or equal to 0 indicating no relationship. In addition, the significance is evaluated by computing t-values with the help of the bootstrapping procedure. Based on the t-value, the probability of error is calculated, with t-values of 1.65 (10% probability of error) and above considered to be significant. The results show that three out of four postulated relationships were significant. All significant relationships were found to be positive.

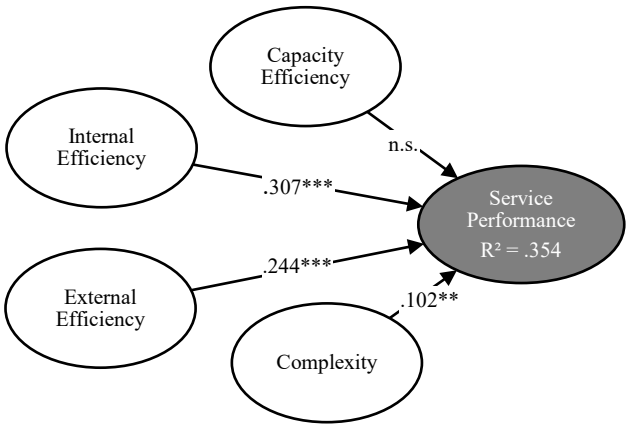


Fig. 3. Path coefficients and probability of error for the structural equation model of medical rescue service performance. *** p<.01; ** p<.05; * p<.1

Thus, hypotheses 1 and 2 can be confirmed. No significant effects were found with regard to hypothesis 3. Contrary to hypothesis 4 the examination of the relation between service complexity and service performance results in a positive interdependence. Bruhn and Blockus [5] explain these improvements of productivity with increased motivation and satisfaction of the employees. Up to a specific level complexity of the task can be motivating and can increase thus the service performance.

The model's predictive accuracy is evaluated by the coefficient of determination (R^2). This coefficient is defined as the squared correlation between a specific endogenous construct's actual and predicted values representing the amount of variance explained by all of the exogenous constructs ranging from 0 to 1, with higher values indicating greater predictive accuracy. There are no generally accepted intervals for good coefficients of determination since they are specific to individual research disciplines [20]. Thus, we consider the intervals from Chin [22] used in explorative research cases. The limit starts at .190, with values above this limit considered to have weak predictive accuracy, values above .330 moderate predictive accuracy and values above .670 very high predictive accuracy.

Self-assessed service performance shows a moderate predictive accuracy and thus is considered to be explained adequately by productivity (internal efficiency, external efficiency and capacitive efficiency) and complexity (R^2 = .354).

The effect size f^2 of exogenous constructs is regarded as another measure for evaluating predictive accuracy. The limits for small, medium and large effects are given as .02, .15 and .35 respectively [20]. The exogenous variables of internal and external efficiency plus capacitive efficiency are found to have a small effect on service performance. Furthermore, complexity shows a non-effect on the service performance (f^2 = .016).

Finally, we evaluated predictive relevance (Q^2). The Stone-Geissers Q^2 value assesses the fitness to rebuild the latent variable out of the associated indicators. Therefore, a value above zero is required. This final assessment criterion was also fulfilled with a value of .277 for the service performance variable.

5. Discussion and Implications

5.1. Hypotheses

In order to get a more sophisticated understanding of the interrelation of service productivity and service complexity, a total of four hypotheses were formulated. The results of the study show the importance of different efficiencies with regard to service performance. Both, internal (hypothesis 1) and external efficiency (hypothesis 2) are crucial to optimize service performance in medical rescue services. Hence, perceived customer satisfaction with service provision, or as in our case patient satisfaction, are just as equally important to service success as are internal processes.

However, capacity efficiency does not have a significant effect on service performance. Thus, hypothesis 3 cannot be confirmed. Since the setting of the described study was in medical rescue service, it was assumed that due to the importance of medical provision there would only be a theoretical resource limit which would be reached only in extreme situations. This assumption, however, does not hold true for the majority of service organizations in Germany. Limited resources can still be a limiting factor in service provision. Therefore, possible effects between capacitive efficiency and overloading of resources cannot be revealed in the present study and have to be addressed in future research in different settings.

Statistical analysis of hypothesis 4 indicated that, contrary to initial expectations, service complexity influences service performance positively. Again, as with hypothesis 3, a possible explanation for this finding might be in the special setting of the study. As a consequence of good education employees of medical rescue providers are well prepared to handle the often highly individual situations, so that increased complexity even with relatively unknown variables as electric vehicles seems to pose a challenge and increase service performance rather than deter it. This finding also has to be reexamined in future research. Moreover it is to be reflected, if complexity is just driven by a rising electro-mobility or if there are other factors that have a more significant impact on medical rescue service complexity.

5.2. Implications for practitioners

The results show a clear dependency between service productivity and performance. Surprisingly capacity efficiency does not influence medical rescue performance. Conversely, increases in complexity seem to have only a marginal though positive contribution on performance.

According to the magnitude and significance of the path coefficients between the latent variables we distinguish a clear order in decreasing importance of the relationships. Internal efficiency is regarded to be of uttermost importance. Therefore practitioners should focus on increasing the efficiency of service provision by reducing disturbing factors and optimizing the process and information flow. Practitioners should install work structures that enable a smooth and interference-free service provision.

Second, external efficiency should be regarded as an important lever for improving medical rescue service performance. Key drivers are not necessarily in time arrival to the location of accident, but the utilization of effective and standardized procedures in order to safely rescue persons directly involved in an accident. The overall goal is to not only medically treat the patient but also taking care of him regarding any other aspects during the entire process of provision.

Last but not least there is – contrary to our expectations – no significant effect of capacity efficiency on performance. Thus, availability of personnel and material resources seems not to be an issue in medical rescue. The results suggest that performance is independent of the capacity utilized. This should be treated with caution, since it is possible that some other interaction effects were neglected or not considered at all in the study. For example it is not known whether the respondents were able to assess the capacity efficiency as the majority was not in a management position. Another possible explanation might be that there were no capacity problems and thus no influence on performance.

Complexity due to increased electric mobility is also of marginal impact on performance of rescue services. It is plausible that existing rescue procedures in Germany are also suitable and can be applied in mission which involves electric driven vehicles. Thus, an increase on complexity does not negatively impact performance. With existing skills and equipment the rescue personnel is able to fulfill a broader and more demanding rescue mission.

5.3. Implication for future research

The results of the analyses point out, that further research is needed. There is still little known about reliable measures for service productivity and complexity, especially in context of medical rescue services. In this paper a first approach for operationalizing service productivity and complexity was made. The results show that further in depth research for every single specific influencing factor on service performance is needed. Furthermore, different skill levels (with and without knowledge about electric driven vehicles as well as standardized procedures) and even different management levels should be considered and evaluated separately. Moreover, further research should also address patient in a dyadic survey.

The effect of complexity and capacity efficiency should be analyzed in more detail. It seems likely that linear models are unable to represent the real relationship between the constructs. Thus, novel approaches should be developed in order to operationalize these constructs as well as to analyze the impact on medical rescue performance.

5.4. Limitations

Despite the complexity of the developed measurement models and the limitation of extent of the questionnaire the reliability of the measurement model are on a sufficient level. Construct validity is also given on a sufficient level. Thus, the measurement of the latent variables has to be improved. Higher level of reliability and validity are desirable. On the one hand further investigations should stronger differentiate between internal and external efficiency of medical rescue services. On the other hand measurement models focusing on technical equipment and employees characteristics and influence on service productivity could be developed and proved.

Medical rescue services are unique caused by its framework conditions e.g. high individuality, federal financing models and deficient suggestibility of the content of the services. Thus, the transferability of the results to other kinds of services should be considered with caution. Moreover, the application to rescue service providers of other nationalities has to be evaluated.

6. Conclusion

Caused by the big impact of services generated in high-wage countries it is of utmost importance to increase service productivity and handle complexity. In order to do this service productivity and complexity have to be measured and evaluated. Based on extensive literature review a measurement model of service productivity was developed. The reliability and validity of this measurement models were demonstrated although there are still possibilities for improvement. A novel measurement model for service complexity was developed and is considered to be reliable and valid. To foster the measurement model and investigate possible nonlinear relationships between the constructs in more detail, further research is needed.

In the structural model evaluation a significant, positive relation between internal efficiency respective external efficiency and service performance is found. The dependency between capacitive efficiency and service performance is not significant. The positive influence of complexity on service performance should be investigated in more detail. Overall, by a focused productivity management the performance of medical rescue services can be increased accordingly.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This research is funded by the German Federal Ministry of Education and Research (BMBF) within project SafetE-car - “Scenarios Supported development of the service system - Secure supply in case of accidents and mishaps with electric vehicles" according to Grant No. 02K12A015 (formerly 01FE13009). Responsibility for the contents of this publication rests with the authors. The authors would like to express their gratitude for this support.

References

[1] Statistisches Bundesamt, “Bruttoinlandprodukt 2015 für Deutschland”, Wiesbaden, p 11, 2016.

[2] S. Bartsch, M. F. Demmelmair, A. Meyer, "Dienstleistungsproduktivität – Stand der Forschung und Zusammenhang zu zentralen vorökonomischen Größen im Dienstleistungsmarketing", Wiesbaden, NJ: Gabler, 2011.

[3] M. Nüttgens, O. Thomas, M. Fellmann, "Dienstleistungsproduktivität: Mit mobilen Assistenzsystemen zum Unternehmenserfolg", Wiesbaden, NJ: Springer, 2014.

[4] C. Grönroos, K. Ojasalo, "Service productivity - Towards a conceptualization of the transformation of inputs into economic results in services", in Journal of Business Research, Vol. 57, no. 4, pp. 414-423, 2004.

[5] M. Bruhn, M.O. Blockus, "Komplexität und Produktivität bei Dienstleistungen", Wiesbaden, NJ: Gabler, pp. 59-87, 2011.

[6] J. Von Garrel, "Dienstleistungen produktiv erbringen: Eine empirische Analyse wissensintensiver Unternehmen in Deutschland", Wiesbaden, NJ: Gabler, 2014.

[7] Statistisches Bundesamt, Einsatzfahrtaufkommen im öffentlichen Rettungsdienst in Deutschland nach Einsatzart in den Jahren 1994 bis 2013", 2014.

[8] T. Levitt, "Production-Line Approach to Service", in: Harvard Business Review, Vol. 50 No. 5, 1972.

[9] P. Jones, "Quality, capacity and productivity in service industries", in: International Journal of Hopitality Management, Vol. 7, No. 2, pp. 104-112, 1988.

[10] H. Corsten, "Produktivitätsmanagement bilateraler personenbezogener Dienstleistungen", in: "Dienstleistungsproduktion: Absatzmarketing, Produktivität, Haftungsrisiken, Serviceintensität, Outsourcing", pp. 43-77, 1994.

[11] R Johnston, P Jones, "Service Productivity – towards understanding the relationship between operational and customer productivity", International Journal of Productivity and Performance Management, Vol. 53, No. 3, pp. 201-213, 2004.

[12] K. Bleicher, "Das Konzept Integriertes Management: Visionen - Missionen – Programme", Frankfurt am Main, NJ: Campus, 2011.

[13] D. Ahlert, H. Evanschitzky, J. Hesse, "Exzellenz in Dienstleistung und Vertrieb: Konzeptionelle Grundlagen und empirische Ergebnisse", Wiesbaden, NJ: Gabler, 2013.

[14] M. O. Blockus, "Komplexität in Dienstleistungsunternehmen - Komplexitätsformen, Kosten- und Nutzenwirkungen, empirische Befunde und Managementimplikationen", Wiesbaden, NJ: Gabler, 2010.

[15] M. Bruhn, D. Georgi, "Services marketing", Harlow, England London New York: Financial Times/Prentice Hall, an Imprint of Pearson Education, 2006.

[16] Ringle, C. M., Wende, S., and Becker, J.-M. 2015. "SmartPLS 3." Boenningstedt: SmartPLS GmbH, http://www.smartpls.com.

[17] K. Ojasalo, "Conceptualizing productivity in services", Helsingfors: Swedish School of Economics, 1999.

[18] C. Bliss, "Management von Komplexität: Ein integrierter, systemtheoretischer Ansatz zur Komplexitätsreduktion". NJ: Springer 2013.

[19] Kraftfahrtbundesamt, "Jahresbilanz des Fahrzeugbestandes am 1. Januar 2017", Retrieved from http://www.kba.de/DE/Statistik/Fahrzeuge/Bestand/b_jahresbilanz.html;jsessionid=6240F6A000F3C10FCD49617694AC2577.live21304?nn=644526 (15.05.2017)

[20] J. F. Hair, G. T. M. Hult, C. M. Ringle, M. Sarstedt, "A primer on partial least squares structural equation modeling (PLS-SEM)". NJ: Sage, 2013.

[21] F. Huber, A. Herrmann, F. Meyer, J. Vogel, K. Vollhardt, "Kausalmodellierung mit Partial Least Squares – eine anwendungsorientierte Einführung", Wiesbaden, NJ: Gabler, 2007.

[22] W. Chin, "The Partial Least Squares Approach to Structural Equation Modeling", in: G.A. Marcoulides, "Modern Methods for Business Research", Mahwah, New Jersey, NJ: Lawrence Erlbaum Associates, 1998.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 435-440 (2017)

www.astesj.com

Special issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Principal Component Analysis Application on Flavonoids Characterization

Che Hafizah Che Noh*, Nor Fadhillah Mohamed Azmin, Azura Amid

Department of Biotechnology Engineering, International Islamic University Malaysia, 53100, Malaysia

A R T I C L E I N F O

Article history:
Received : 05 April, 2017
Accepted : 05 May, 2017
Online: 24 May, 2017

Keywords :
Principal Component Analysis
Fourier Transform Infrared Spectroscopy
Flavonoids

A B S T R A C T

Flavonoid is one of the bioactive compounds that are currently used in pharmaceutical and medicinal industries due to their health benefit. The focus of current research is mainly on the extraction and isolation of bioactive compounds; however non to date has explored on the identification of flavonoids classes under the Fourier Transform Infrared spectroscopy (FTIR). This gap presents an opportunity for the application of statistical analysis which can identify the distinct wavenumbers range of flavone, flavanone and flavonol for their characterization in the FTIR spectrum. Development of algorithm based on principal component analysis (PCA) for the analysis and identification of flavonoids classes based on FTIR spectrum was introduced. Based on the results, five wavenumbers ranges have been identified as the distinct characteristics of flavonol, flavone and flavanone hence used for their identification.

1. Introduction

This paper is an extension of work originally presented at IEEE Student Conference on Research and Development (SCOReD) 2016 [1]. Currently, flavonoids have received much attention especially in medicinal and pharmaceutical industries because of its biological properties such as high antioxidant activity, high radical scavenging, anti-inflammatory, anti-cancer and anti-allergic [2]. Flavonoids are one of the active ingredients used for the formulation of drugs and potential to treat many diseases. Increasing research performed (in vitro) on the importance of flavonoids consumption such as reducing the risk of Parkinson disease [3], may lower risk of ovarian cancer [4] and protecting brain damage [5]. Besides, flavonoids are widely distributed in fruits, plants, vegetables and microorganisms [6].

Flavonoid have C6-C3-C6 general structure backbone which consists of two phenolic structures, ring A and ring B attached to heterocyclic ring C as in Figure 1. There are more than 7000 structures of flavonoids from different subgroups have been reported [7]. All flavonoids share a basic C6-C3-C6 phenyl-benzopyran backbone; however differ in their substituents (type, number and position) and in their insaturation [8]. Individual differences within each group are due to variations in number and arrangement of the hydroxyl groups and their extent of alkylation and/ or glycosylation.

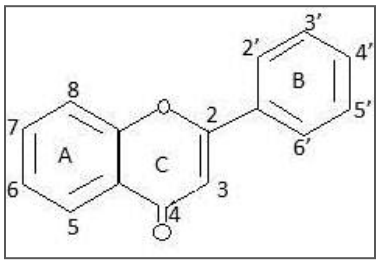


Figure 1: Flavonoids backbone

As the multiple bioactive compounds presence in compound, the characterization and identification of the bioactive compounds remain challenging due to complicated isolation procedures [9]. Several analytical techniques have been developed for the identification of flavonoids such as the application of High Performance Liquid Chromatography (HPLC) [10], HPLC coupled with photodiode array and mass spectrometry detectors [11], nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry (MS) evidences [12], Thin Layer Chromatography (TLC) plate [13] and many more. However, some disadvantages possessed are less sensitive, expensive and the need for chemist to analyze.

*Corresponding Author: Che Hafizah Che Noh, Department of Biotechnology, International Islamic University Malaysia, 53100 Kuala Lumpur, Malaysia. Email: chehafizah8989@gmail.com

Today, the applications of Fourier Transform Infrared spectroscopy (FTIR) has increased in food, pharmaceutical as well as medicinal applications for the identification of active compounds in plants, fruits, vegetables and microorganisms [14,15,16,17] and particularly has become a powerful analytical tool in the study of active compounds presents in samples. In fact, there have been several studies concerning with the characterization, classification and identification of flavonoids using FTIR [18,19,20].

FTIR has become well-accepted method due to its ease of samples preparation, fast, need little samples size, and does not require the use of solvents which is more economical [21].

FTIR spectrums however are difficult to interpret, unless for the expertise since the library spectrum are limited, and the software used to interpret the results somehow expensive [22]. Principal component analysis (PCA) is one of the chemometric methods that can be used to identify the similarities and differences of the chemical information [23]. Infrared spectroscopy, combined with chemometric, has been used for the identification of active compound [24].

Studies have been performed on the identification of flavonoids [11,12], however non to date has explored on the identification of individual flavonoids under FTIR spectrum coupled with chemometric analysis, which is PCA. Despite some studies performed on the identification of compound using few combination of chemometric techniques such as hierarchical cluster analysis [24], PCA, partial least square (PLS) and factor discriminant analysis (FDA) [25], the application of PCA into the characterization model has not been undertaken yet.

Introduction of the characterization model based on computational method provides the valuable tools with the application of PCA into the model algorithm for rapid identification. The proposed method are unique, since adapted both the structural information provided by PCA based on spectrum data of flavonoids structure with the interaction information from the compound, unlike most current methods [26,27,28]. The adapted characterization model is based on nested if and if else statement builds as an algorithm for the characterization.

Therefore, the objectives of this study are to develop an algorithm utilizing the PCA as the base for the interpretation and analysis of FTIR spectrums and to identify the distinct wavenumbers range of flavone, flavanone and flavonol as for their identification under the FTIR spectroscopy.

2. Literature review

2.1. *Fourier Transform Infrared Spectroscopy and Principal Component Analysis Applications*

FTIR offers high speed of analysis and needs little or no sample preparation [29]. Moreover, it also provides information on multiple parameters simultaneously [30]. The chemical bonds in material will vibrate once the infrared radiation that interacts with matter is absorbed hence producing the infrared spectrum. Regardless of the structure of the rest of the molecule, the functional groups tend to absorb infrared radiation in the same wavenumber range [22].

Band assignments for flavonoids identification have been studied in some details [31]. Based on studies, different peak shown due to wavenumber shift of different molecules with the same functional group under their wavenumber range; indicate the uniqueness of the molecules under FTIR spectrum hence used for the fingerprint of the molecules. The correlation shows between the infrared band positions and chemical structure becomes the fundamental information for the analyst to identify the unknown molecules or even the structural identification.

According to [22], the wavenumber range between 3500-3200 cm^{-1} is corresponds to the OH functional group and studies indicated that the overall arrangement of molecular structure will cause the wavenumber shift of the functional groups in their range [31]. Studies performed on phenolic compound for their antioxidant nature using FTIR [32,33]. From the studies, the presence of OH group under FTIR region will cause wavenumber shift depend upon to its oxidant nature.

FTIR spectra on flavonoids structure has been studied extensively [34]. The study highlighted the spectrum wavenumber range which are common for the flavonoid structure is the C=O bond (1630 and 1665cm^{-1}), C-O bond (1000 and 1300 cm^{-1}), and in-plane deformation vibrations of C-H (600 and 980cm^{-1}). Other studies has identified the OH group at the wavenumber at 3300cm^{-1}, 2970cm^{-1} and 2856cm^{-1}; meanwhile C=C at 1644cm^{-1}.

Previous studies shown that appropriate application of statistical analysis or chemometric has provides accurate and rapid characterization of plant extract based on analytical spectroscopy that contain multidimensional information [35]. The correlation factors obtained from the PCA analysis within a set of observed variables provides the pattern of the sample [36]. Henceforth, with the application of PCA, the spectrum analysis can be performed for rapid and easier identification.

The application of PCA from the FTIR spectrums has successfully discriminated the Romanian sea buckthorn berries and leaves groups based on its phenolic contents [37]. Correlation between the content of flavonoid in certain frozen fruits pulps with free radical scavenger activity has been identified using the PCA [38]. From the study, flavonoid has shown to exhibit high scavenging activity against hydroxyl radical.

3. Methodology

The main purpose of this study is to develop an algorithm for flavonoids identification utilizing the PCA based on FTIR spectrums. Therefore, following were the steps performed to achieve the objectives.

3.1. *FTIR Measurements*

Samples of three flavonoids (flavone, flavanone and flavonol) in powder form were purchased from Bumi Pharma Sdn Bhd (Malaysia) that was imported from Sigma Aldrich (UK).

The flavonoids were sent to Centre for Research and Instrumentation, Universiti Kebangsaan Malaysia for FTIR analysis.

(i) Spectrum 400 FT-IR, Perkin Elmer was used for the sample analysis. The wavenumber range set for the FTIR is 4000-650 cm^{-1} which is under mid infrared region. The FTIR was performed using the ATR (Attenuated Total Reflectance) where the solid samples (in powder form) were directly applied to the plate for measurements.

(ii) Using the built in algorithm provided from software package of FTIR, the spectrum was preprocessed using baseline correction to produce horizontal baseline shift and smoothing to reduce the noise.

(iii) The preprocessed data extracted in Excel form which consists of absorbance versus wavenumber (cm^{-1}) was recorded for statistical data analysis.

3.2. *Statistical Data Analysis*

(i) Data matrix, Xn where (n=1…x) which comprises of transposed data matrix of all flavonoids spectrum under selected wavenumber region (n), was analyzed with PCA on a MATLAB platform.

(ii) Commonly data preprocessing procedures such as standardization and normalization are required prior to performing PCA on a data matrix. Standardization is used to center and scale the data meanwhile normalization is generally used for data which have different units and scale. In this study, only standardization was performed since the data have same units. Figure 2 shows the configuration of the data matrix X, mxp (types of flavonoids x spectral range).

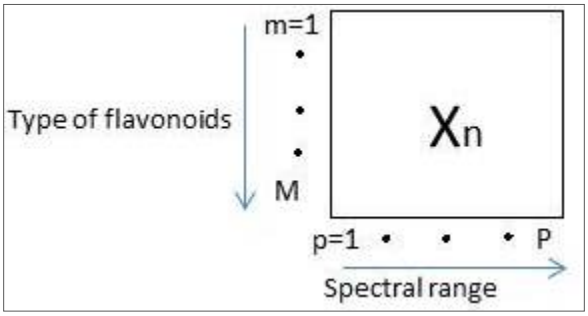


Figure 2: Two dimensional array, X_n of spectrum where n=1…N that represent the number of region, m=1…M that represent type of flavonoids and p=1…P that represent the spectral range. The weighted used for the data is the absorbance value of spectral data corresponds to the wavenumber range.

(iii) The similarities and differences based on chemical structure of the flavonoids provide a basis for the flavonoids characterization based on functional group identification offered by FTIR. The molecules that have same functional group will have the peak at the same wavenumber range regardless the rest of the molecules; however differ in wavenumber shift in the wavenumber range due to the whole structure arrangement. Henceforth, the wavenumber shift and the presence of the peak at certain wavenumber region provide the basis for flavonoids structure classification.

(iv) Eight significant regions were assigned in Table for the identification of flavonoids based on their chemical structure. The similarities and differences identified from the PCA were further utilized for development of algorithm for flavonoids identification using nested if else statement as in Figure 6.

Table: Five regions assigned for classification of flavonoids under PCA

Region	Wavenumbers (cm^{-1})	Structural identification	References
1	4000-3250	Presence of hydroxyl group	[22]
2	3140-3000	Presence of di-substitution	[22]
3	1670-1620	Unsaturation bonds	[22]
4	1650-1600	Conjugation of bonds	[22]
5	1600-1500	Aromaticity	[22]

4. Results

Figure 3 shows the FTIR spectrum for flavone, flavanone and flavonol. From the FTIR spectrum shown in Figure 3, each of flavonoids shows different spectrum due to different structure arrangement. Different peak positions, intensities, widths and shape shown in Figure 3 indicate the structure difference of flavonoids subgroups hence used for the band assignments as in Table.

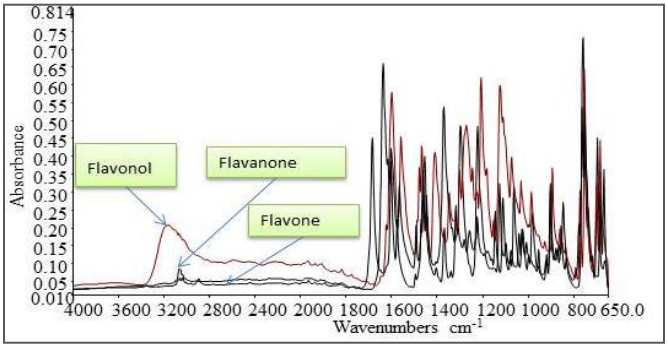


Figure 3: FTIR spectrum in mid IR range for flavonol, flavone and flavanone

As seen in Figure 3, the FTIR spectrum of molecules differs at certain wavenumber range even though the molecules have the same functional groups. The presence of hydroxyl group, compared to the other flavonoids for example, will cause the wavenumber shift at particular wavenumber range. Besides that, the overall structure arrangement such as the presence of different substitution will cause wavenumber shift hence producing different spectrum. That is why, the FTIR is powerful method for identification since no different molecules will have same the spectrum; it is the fingerprint data for the molecule.

Five significant regions have been assigned in Table for their identification based on its structural difference as in Figure 5. Figure 4 shows the results of PCA analysis of flavonoids spectrum based on the assigned regions.

In PCA analysis, the correlation factors that contribute by the PCA analysis provide the pattern of the samples. The sample that is close to each other indicates their similarities meanwhile far to each other indicate their differences. This pattern of analysis provides clear discrimination between samples as a tool of identification.

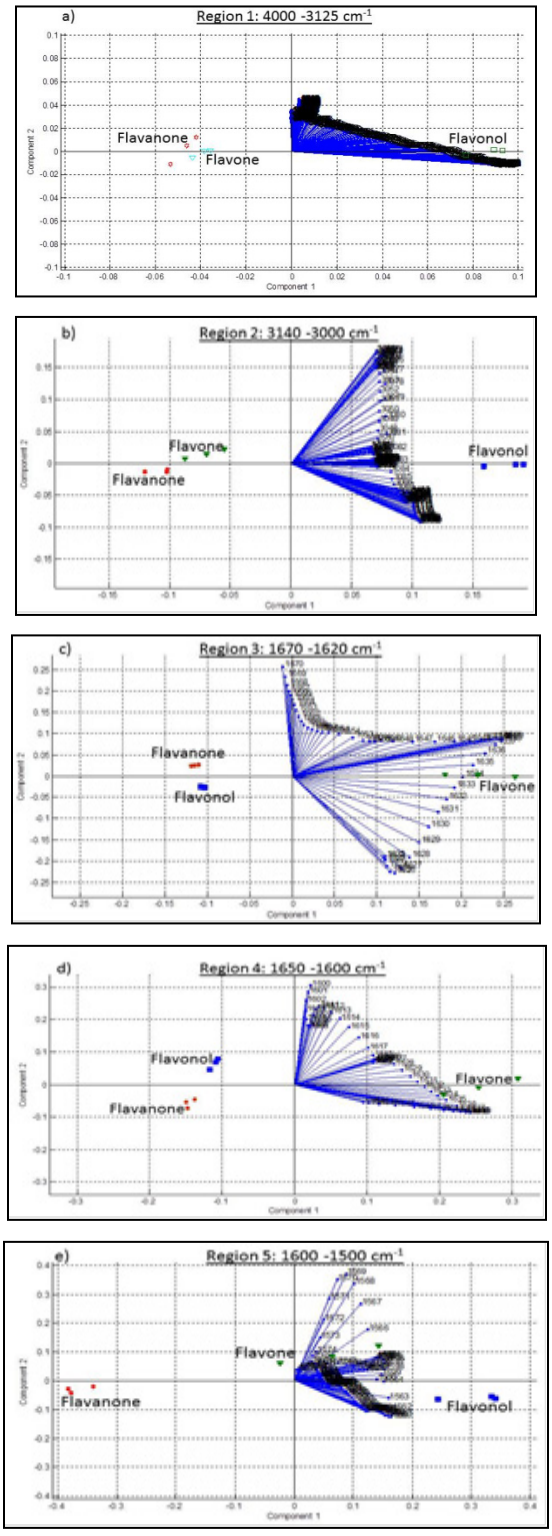


Figure 4. Results of PCA for flavonoids in the region 1 (4a), region 2 (4b), region 3 (4c), region 4 (4d) and region 5 (4e)

5. Discussion

. Flavonoids aglycone can be divided into few subgroups which are flavanone, flavonol, flavone, flavanol, anthocyanidin and isoflavonoid [39]. For flavanone, flavonol and flavone, their variation within the groups is depending upon the presence of OH group at C-3, a saturated single bond between C-2 and C-3, and their conjugation and saturation behavior as in Figure 5 [40].

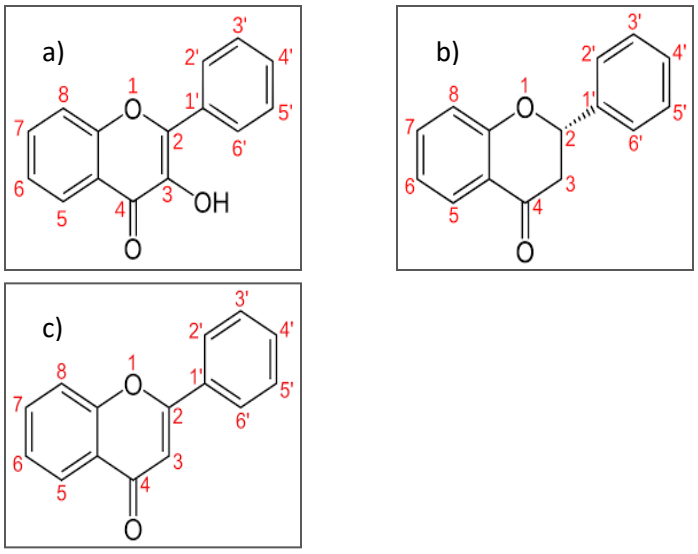


Figure 5. Structure of flavonoids subgroups, a) flavonol b) flavanone c) flavone

The differences exhibit by the flavonoids subgroups structure has lead to different spectrum shown in Figure 3 and results of PCA shown in Figure 5.

Based on Figure 4, the identification of flavone, flavanone and flavonol are focusing on the heterocyclic ring C that represents the major structural difference between the flavonoids. Based on Table, 4000-3125 cm^{-1} has been identified for the presence of hydroxyl group in flavonoids, 3140-3000 cm^{-1} for the presence of di-substitution of heterocyclic ring C, 1670-1620 cm^{-1} and 1650-1600 cm^{-1} for the saturation and conjugation of bonds respectively. Meanwhile, 1600-1500 cm^{-1} indicates the general aromaticity behavior of the flavonoids.

For flavonol, the presence of hydroxyl group at carbon 3 and di-substitution of heterocyclic ring C at carbon 2 and 3 provides the distinct characteristics of structure arrangement as compared to flavanone and flavone. Based on Figure 4 (a) and (b), for region 1 and 2, clear discrimination shown by PCA analysis on flavonol at wavenumbers 4000-3125 cm^{-1} and 3140-3000 cm^{-1}due to the presence of hydroxyl group and di-substitution at carbon 2 and 3 respectively as compared to flavanone and flavone.

Once the flavonol is identified, the flavanone and flavone is identified and distinguish based upon their saturation and conjugation bond as in Figure 5. Based on Figure 5, flavanone contain saturated carbon at 2 and 3 meanwhile, the flavone contain unsaturated carbon at 2 and 3. Besides, compared to flavanone, the conjugation of bond particularly presence in flavone. The arrangement of saturation and conjugation characteristics of flavonol that similar to flavanone provides the discrimination between the flavone as compared to flavonol and flavanone. Based on Figure 4 (c) and (d) in region 3 and 4, clear discrimination is shown by the PCA analysis on flavone at wavenumbers 1670-1620 cm^{-1} and 1650 -1600 cm^{-1} due to unsaturation and conjugation bonds respectively, as compared to flavanone.

The aromaticity characteristics of the flavone, flavonol and flavanone are differs due to their structure arrangement as in Figure 5 that can be an indication of their individual differences. Based on Figure 4(e) in region 5, clear discrimination is shown by PCA analysis on flavanone, flavone and flavonol due to their aromaticity at wavenumbers 1600-1500 cm^{-1} that provides their individual difference.

The significant region based on the structural information of the compounds provides the algorithm for the characterization model as shown in Figure 6. The application of PCA is integrated into the model to provide rapid identification of flavone, flavanone and flavonol.

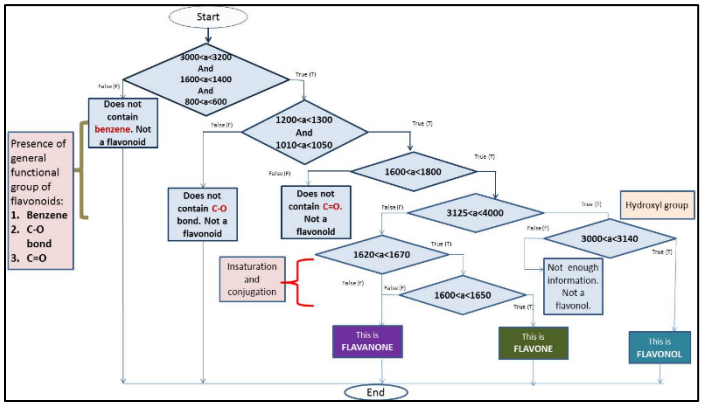


Figure 6. Flow diagram of characterization model algorithm where a is refer to absorbance value

The main contribution in this study is, despite other method which utilized much equipment and statistical method for identification and confirmation, with further exploitation on PCA methodology and the informative spectrum data provided by FTIR which not yet fully utilize, will provides the simple, easier, accurate, less expensive and rapid identification of flavonoids subgroups. The selected region based on PCA will provides significant identification of flavonoids based on its structure arrangement. Henceforth, the isolation procedure for example, will be less time consuming and identification of flavonoids presence in fruits and plants can be optimized.

6. Conclusion and recommendation

Application of PCA into the characterization model for rapid identification of flavonoids based on the FTIR spectrums was presented. The results shown that clear discriminations of flavonoids shown at wavenumbers 4000-3125 cm^{-1} and 3140-3000 cm^{-1} for flavonol, 1670-1620 cm^{-1} and 1650-1600 cm^{-1} for flavone and 1600-1500 cm^{-1} for individual flavonoids.

The research is still in progress where some individual flavonoids subgroups derivatives such as quercetin, myricetin and luteolin to name a few, were tested under the selected regions for flavonoids characterization and identification.

Besides, further exploitation on the advantages offered by PCA methodology will be performed; for example the studies on contribution plot and confidence limit provided by PCA to enhance and validate the results.

It is recommended that further studies to be performed on the identification of other flavonoid subgroups using proposed method to enhance the identification of flavonoids that is widely distributed in plants and fruits.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

Authors acknowledge to International Islamic University "IIUM" and Ministry of Higher Education "MOHE" for the research grant "RAGS14-050-0113".

References

[1] C. H. C. Noh, N. F. M. Azmin, A. Amid, A. L. Asnawi, S. N. Ibrahim, "Development of Algorithm for Flavonoids Classes Characterization" in 14th International Conference Student on Research and Development, Kuala Lumpur Malaysia, 2016.

[2] S. K. Kumar, M. Suresh, S. A. Kumar, P. Kalaiselvi, "Bioactive compounds, radical scavenging, antioxidant properties and FTIR spectroscopy study of Morinda citifolia Fruit Extracts" International Joutnal of Current Microbiology and Applied Science, 28-42, 2014.

[3] X. Gao, A. Cassidy, E. B. Rimm, A. Ascherio, "Habitual intake of dietary flavonoids and risk of Parkinson disease" AAN Enterprise, Inc., 1138-1145, 2012.

[4] A. Cassidy, T. Huang, M. S. Rice, E. B. Rimm, S.S. Tworoger, "Intake of dietary flavonoids and risk of epithelial ovarian cancer" American Society for Nutrition, 1-8, 2014.

[5] S. L. Xu, C. W. C. Bi, R. C. Y. Choi, K. Y. Zhu, A. Miernisha, T. T. X. Dong, K. W. K. Tsim, "Flavonoids induce the synthesis and secretion of neurotrophic factors in cultured rat astrocytes: A signaling response mediated by estrogen receptor" Hindawi Publishing Corporation, 1-10, 2013.

[6] S. Dharmaraj, "Marine Streptomyces as a novel source of bioactive substances" World J Microbiol Biotechnol, 2123-2139, 2010.

[7] O. M. Andersen, K. R. Markham, Flavonoids: Chemistry, Biochemistry and Applications, Boca Raton: CRC Press, 2006.

[8] P.F. Pinheiro, G. C. Justino, "Structural analysis of flavonoids and related compound: A review of spectroscopic applications" in Phytochemicals-A Global Perspective of Their Role in Nutritrion and Health, D. V. Rao, Eds. Europe: InTech, 33-56, 2012.

[9] S. Sasidharan, Y. Chen, D. Saravanan, K.M. Sundram, L. Y. Latha, "Extraction, isolation and characterization of bioactive compounds from plants' extracts" Afr J Tradit Complement Altern Med., 8(1), 1-101, 2011.

[10] S. Durum, R. N. Allam, M. A., A. S. E. Shazely, "Identification of phenolic acids and flavonoids compounds from some varieties Triticum Aestivum and Triticum" 6(3), 521-530, 2016.

[11] Z. Daqiu, T. Wenhui, H. Zhaojun, T. Jun, "Identification of flavonoids and expression of flavonoid biosynthetic genes in two coloured tree peony flowers" Biochemical and Biophysical Research Communications,450-456, 2015.

[12] W. Lingrong, W. Dan, J. Yueming, P. N. K., L. Sen, J. Guoxiang, H. Jirui, Z. Mouming, L. Wei, Y. Bao, "Identification of flavonoids in litchi (Litchi chinensis Sonn.) leaf and evaluation of anticancer activities" Journal of Functional Food, 555-563, 2014.

[13] K. Samariya R. Sarin, "Identification of flavonoids from Cyperus Rotundus Linn. in vivo and in vitro" Journal Of Drug Delivery & Therapeutics, 3(2), 109-113, 2013.

[14] Y. S Park, B. G. Heo, K. S. Ham, S. G. Kang, Y. K. Park, A. Nemirovski, "Analytical determination of bioactive compounds as an indication of fruit quality" Journal of AOAC International, 1725-1731, 2012.

[15] T. S. Preetha, S. N. Mohan, N. M. P., S. R. Deepthi, "Chemical fingerprinting of Spilanthes Scmella L.(Murr.) - An acutely theatened medicinal plant of pharmaceutical importance by HPLC, FTIR and UV-VIS spectoscopic tools" World Journal of Pharmacy and Phamaceutical Sciences, 1275-1287, 2014.

[16] R. Sahu, J. Saxena, "Ultraiolet-Visible and Fourier Transform Infrared Spectroscopic studies on non-conventional species of Curcuma" Indian Journal of Advances in Chemical Science, 300-302, 2014.

[17] S. Dharmaraj, "Marine Streptomyces as a novel source of bioactive substances" World J Microbiol Biotechnol. 2123-2139, 2010.

[18] D. Mrinmay, K. Ashok, D. Jyotirmoy, S. R. Durga, "Isolation and characterization of a flavonoid from ethanolic extract Of Portulaca Quadrifida Linn" American Chemical Society, 5 (1) , 13-20, 2017.

[19] K. S. Amit, S. Raghvendra, S. Manimegalai, "Identification of flavonoids in methanolic extract of Caryota urens (Fish Tail Palm): A phytochemical screening involving structure analysis by FTIR spectroscopy" Research Journal of Phytochemistry, 9 (3), 127-136, 2015.

[20] S. Ranjana, D. M. Vijay, "FTIR studies and spectrophotometric analysis of natural antioxidants, polyphenols and flavonoids in Abutilon indicum (Linn) Sweet leaf extract" Journal of Chemical and Pharmaceutical Research, 7(6),205-211, 2015.

[21] M. Martinez, T. Velazquez, G. Revilla, N. Abarca, and E. Perez, "Application of MIR-FTIR spectroscopy and chemometrics to the rapid prediction of fish fillet quality" CyTA Journal of Food, 369-377, 2014.

[22] B. C. Smith, Infrared Spectral Interpretation: A Systematic Approach, United States: CRC Press, 1999.

[23] A. A. F. Zielinski, C. W. I. Haminiuk, A. Alberti, A. Nogueira, I. M. Demiate, and D. Granato, "A comparative study of the phenolic compounds and the in vitro antioxidant activity of different Brazilian teas using multivariate statistical techniques" Food Research International, 246-254, 2014.

[24] G. Seher, S. Mete, G. Erik, K. Irfan, S. Feride, "Differentiation of Anatolian honey samples from different botanical origins by ATR-FTIR spectroscopy using multivariate analysis" Food Chemistry, 234–240, 2015.

[25] I. Gouvinhas, J. M. M. M. D. Almeido, T. Carvalho, and A. I. R. N. A. Barros, "Discrimination and characterisation of extra virgin olive oils from three cultivars in different maturation stages using Fourier transform infrared spectroscopy in tandem with chemometrics" Food Chemistry, 174-226, 2015.

[26] A. Hirri, M. Gammouh, A. Gorfti, F. Kzaiber, M. Bassbasi, S. Souhassou, and A. Balouki, "The use of Fourier transform mid infrared (FT-MIR) spectroscopy for detection and estimation of extra virgin olive oil adulteration with old olive oil" Sky Journal of Food Science, 60–66, 2015.

[27] M. Martinez, T. Velazquez, G. Revilla, N. Abarca, and E. Perez, "Application of MIR-FTIR spectroscopy and chemometrics to the rapid prediction of fish fillet quality" CyTA Journal of Food, 369-377, 2014.

[28] M. Traore, K. Joeri, and M. C. Antonio, "Application of FTIR spectroscopy to the characterization of archeological wood" Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 63-70, 2016.

[29] R. M. Pop, A. D. Buzoianu, I. V. Rati, and C. Socaciu, "Untargeted metabolomics for sea buckthorn (Hippophae rhamnoides ssp. carpatica) berries and leaves : Fourier Transform Infrared spectroscopy as a rapid approach for evaluation and discrimination" Not Bo Horti Agrobo, 545-550, 2014.

[30] S. A. Kustrin, D. W. Morton, and A. P. Md. Yusoff, "The use of Fourier Transform Infrared (FTIR) spectroscopy and Artificial Neural Networks (ANNs) to assess wine quality" Modern Chemistry & Applications, 2329-6798, 2013.

[31] M. Rochana, R. Rishob, G. Subham, M. Sougata, "Synthesis characterization and study of antioxidant activity of luteolin – magnesium complex " 21-26, 2016.

[32] J. Joby, S. Sudheesh, K. T. M. Sumesh, J. Sony J, J. V. E., "A comparative evaluation of anticancer activities of flavonoids isolated from mimosa pudica, aloe vera and phyllanthus niruri against human breast carcinoma cell line (mcf-7) using mtt assay" International Journal Of Pharmacy And Pharmaceutical Sciences, 6 (2), 2014.

[33] N. Rajalaxm, P. Ashok, S. Luna, S. Santilata, "Phytochemical analysis of antioxidants from abutilon indicum l. and paederiafoetida l. supported by TLC, FTIR and NMR studies" International Journal of Science and Research (IJSR), 1312-1318, 2015.

[34] S. M. Manki, A. A. Omotayo, G. L. B. Priscilla, T. M. Wilfred, J. Nazeem, A. S. Everlyne, I. I. Emmanuel, "Electroanalytical determination of catechin flavonoid in ethyl acetate extracts of medicinal plants" Int. J. Electrochem. Sci., 4, 1497 – 1510, 2009.

[35] G. Royston, V. Y. Emma, K. H. James, M. S. Ian, "Chemometric discrimination of unfractionated plant extracts analyzed by electrospray mass spectrometry" Phytochemistry, 859–863, 2003.

[36] S. Kallithraka, I. S. Arvanitoyannis, P. Kefalas, and A. El-zajouli, "Instrumental and sensory analysis of Greek wines ; implementation of principal component analysis (PCA) for classification according to geographical origin" Food Chemistry, 501-514, 2001.

[37] Pop, R. M., Buzoianu, A. D., Rati, I. V., Socaciu, C., "Untargeted metabolomics for sea buckthorn (Hippophae rhamnoides ssp. carpatica) berries and leaves : fourier transform infrared spectroscopy as a rapid approach for evaluation and discrimination" Not Bo Horti Agrobo, 545-550, 2014.

[38] L. C. Vissotto, E. Rodrigues, R. C. Chisté, M. D. T. Benassi, and A. Z. Mercadante, "Correlation , by multivariate statistical analysis , between the scavenging capacity against reactive oxygen species and the bioactive compounds from frozen fruit pulps" Food Science and Technology, 57–65, 2013.

[39] S.V. Knezevic, B. Blazekovic, M. B. Stefan, M. Babac, "Plant polyphenols as antioxidants influencing the human health" in Phytochemicals as Nutraceuticals- Global Approach to Their Role in Nutritrion and Health, D. V. Rao, Eds. Europe: InTech, 155-180, 2012.

[40] J.K. Praisan, C. C. Wang, S. Barnes, "Mass spectrometric methods for the determination of flavonoids in biological samples" Free Radical Biology & Medicine., 37, 1324-1350, 2004.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 441-448 (2017)

www.astesj.com

Special issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

A Reconfigurable Metal-Plasma Yagi-Uda Antenna for Microwave Applications

Giulia Mansutti[*,1], Davide Melazzi[2], Antonio-Daniele Capobianco[1]

[1]*Department of Information Engineering, University of Padova, Via G. Gradenigo 6/B, 35131, Padova, Italy*

[2]*CISAS G. Colombo, University of Padova, Via Venezia 15, 35131, Padova, Italy*

ARTICLE INFO

Article history:
Received: 05 April, 2017
Accepted: 03 May, 2017
Online: 24 May, 2017

Keywords:
Plasma Antennas
Yagi-Uda Antenna
Microwave Antennas
Reconfigurable Antennas

ABSTRACT

This paper is an extension of the work originally presented at the European Microwave Conference (EuMC) about a reconfigurable hybrid metal-plasma Yagi-Uda antenna operating at 1.55 GHz: this antenna consists of metallic reflector and active element and two plasma directors. The conference work showed through full-wave numerical simulations (CST Microwave Studio) how it is possible to achieve reconfigurability with respect to the gain by turning on/off the plasma discharges. However the model that was used to represent the plasma discharges was quite ideal, so one comment that was provided questioned the actual possibility of achieving reconfigurability in a real system.
Consequently we performed extensive measurements of different plasma discharges and thanks to the collected data, we noticed some important differences between the full-wave numerical model of the plasma that we used in the conference paper and the actual plasma discharges that were generated in the experimental setup: the dielectric vessel and the metallic electrodes used respectively to confine and generate the plasma have an influence on the radiation pattern of the antenna and so they must be included in the design procedure; the cylindrical plasma discharge is much easier to realize when the cylinder diameter is at least 3mm; and finally the collision frequency of the plasma in realistic cases is pretty higher than the one adopted in our previous work.
Therefore this work presents a feasibility study of a more detailed and realistic model of our antenna with respect to the plasma discharges. We will show that reconfigurability can still be achieved through a proper design of the overall antenna, thus paving the way to an actual realization of the proposed reconfigurable Yagi-Uda.

1. Introduction

Traditional antennas exploit metallic components to radiate and/or receive electromagnetic waves. In recent years there has been an increasing interest in another type of antennas, namely Gaseous Plasma Antennas (GPAs), whose radiating properties rely on partially or fully ionized gas rather than on metallic conductors [1-8]. The gas is confined into dielectric vessels such as glass cylinders (but the enclosure can be of any shape, e.g., monopoles and loops, and of any dielectric material) and the plasma can be turned on and off very rapidly (on time scales the order of milliseconds or microseconds) by different excitation means [3]. In the off state, the plasma reverts to a neutral gas that is confined in a dielectric holder, thus being transparent to any electromagnetic (EM) waves; when the plasma is turned on, it behaves as an electrical conductor that is able to transmit and receive EM signals.

1.1. Gaseous Plasma Antennas: Advantages

Plasma antennas have a number of potential advantages over conventional metal antennas since their properties can be electrically (rather than mechanically) controlled. This makes plasma antennas appealing for a number of applications, some of which are briefly discussed in the followings [4].

First of all, when a plasma discharge is unenergized, the gas reverts to its neutral state and so it stops behaving as a conductor. This feature makes plasma antennas difficult to be detected by radars, once the plasma vessel and excitation system are properly designed and if the antenna is energized only when the

[*]Corresponding Author: Giulia Mansutti, Department of Information Engineering, University of Padova, Via G. Gradenigo 6/B, 35131, Padova, Italy | Email: giulia.mansutti@gmail.com

communication takes place. This property leads to a potential interest in plasma antennas for military applications.

Secondly, plasma antennas are transparent to EM waves whose frequency is above the plasma frequency; this feature makes plasma antennas interesting when dealing with appli- cations that need many antennas (or antenna arrays) used for different communication systems that must be stacked together due to space constraints (e.g., shipborne maritime communications) and that operate at different frequencies. In this case plasma elements are an interesting alternative to conventional metallic ones since they can provide significantly lower mutual coupling values among the different arrays; as a matter of fact, unenergized plasma elements do not interfere with active elements, and the coupling among plasma antennas working at different frequencies is significantly lower than the one provided by metallic antennas.

Additionally, plasma elements can be energized and de-energized on time scales of microseconds; as a result, the signal degradation caused by antenna ringing can be greatly reduced by using a plasma antenna transmitting at high rates and that is switched off at the end of each bit.

Finally, the radiation properties of a plasma antenna can be changed by electrically controlling the plasma parameters (e.g., density), which can be achieved by tuning the power delivered to the plasma. Therefore plasma antennas can be rapidly reconfigured with respect to their radiation pattern, input impedance, and working frequency.

1.2. Gaseous Plasma Antennas: Issues

Nevertheless, the usage of plasma discharges as antennas in communication systems can be hindered by several issues such as the noise introduced by the plasma, the complexity of the plasma antenna, and the availability of suitable technology solutions to generate the plasma.

The noise introduced by the plasma is strongly related to the way the plasma is generated, and this fact has been holding back the use of plasma antennas for many years. Different plasma generation techniques are available: laser-initiated atmospheric discharge [5], DC/AC discharge [6], and radio-frequency (RF) surface-wave plasmas [7]. Both DC and low-frequency AC methods introduced a rather high noise level when generating the plasma. RF surface wave generation and the pulsing power technique made it possible to obtain not only lower noise levels, but also higher plasma density and reduced power consumption [3].

Another obstacle to the exploitation of plasma antennas can be identified in the inherent complexity of plasma antennas. Specifically, a plasma antenna comprises not just the plasma, but also the equipment needed to generate it (e.g., the metallic electrodes in the case of a RF surface-wave-generated plasma). Furthermore, the plasma density in an actual plasma discharge is nonuniform, with a nonuniformity that depends on the plasma generation technology; before considering the actual realization of a plasma antenna, it is therefore necessary to develop a representative prototype of the plasma discharge the antenna will be comprised of, and to measure the plasma density therein. As a result, the presence of electrodes, and the nonuniform plasma distribution within the discharge are expected to affect the radiation properties of the actual plasma antenna, and are to be taken into account in the design, and realization of an actual plasma antenna.

Other critical aspects towards the realization of an actual plasma antenna, are the availability of the proper technology to ignite, and sustain the plasma, and the capability of developing customized dielectric vessels whose shape and dimensions must be carefully chosen according to the antenna type (e.g., monopole, dipole) and operating frequency.

1.3. Original Contribution and Paper Organization

In the proof of concept of a hybrid metal-plasma Yagi-Uda antenna presented in [1], two plasma discharges have been used as antenna directors to achieve reconfigurability with respect to the gain; specifically, the gain has been controlled by turning either on or off the plasma directors. In the full-wave numerical preliminary study that was performed in CST Microwave Studio [8], the plasma directors were ideally modeled as two cylinders made of a uniform dispersive Drude material, i.e., neglecting the presence of the dielectric tubes into which the plasma is confined and of the plasma generation equipment, i.e., the metallic electrodes attached at the two edges of the dielectric cylinder. Moreover we have modeled the off state of the plasma directors simply excluding the plasma elements from simulations. Therefore the comment that was made in the review process was related to the actual possibility of our proposed antenna to achieve reconfigurability in a more realistic scenario.

The aim of this work is to present a feasibility study of a realistic version of our hybrid metal-plasma Yagi-Uda antenna with some major differences with respect to the ideal model provided in [1]: (i) the plasma elements comprise the dielectric tubes and the electrodes; (ii) the collision frequency and plasma dimensions have been set now to more realistic values obtained through experimental characterizations of different plasma discharges; (iii) the off-state of the plasma directors is modeled including the dielectric tubes filled with neutral gas (i.e., unenergized plasma) in the simulations; and finally (iv) the overall design of the Yagi-Uda antenna was revised according to the above-mentioned differences.

The rest of the paper is organized as follows. In Section II the model of our new version of the hybrid metal-plasma Yagi-Uda antenna is described in details; Section III is dedicated to the results and related discussion with a particular focus on the comparison with our previous work [1]; finally in Section IV conclusions are reported.

2. Antenna Design

The aim of this work is to investigate the capability of achieving reconfigurability for a realistic metal-plasma Yagi- Uda antenna whose reflector and active element are made of metal while the directors are plasma discharges (as the one depicted in Figure 1). Therefore the substantial difference with respect to the work that was made in [1] involves the modelization of the plasma directors: more realistic values of the discharge diameter and collision frequency have been chosen according to typical values that were obtained from measurements of different implemented plasma discharges prototypes; moreover the dielectric vessels and electrodes used to generate the plasma are present.

The starting point for our analysis was then the modelization of the ideal plasma discharge, i.e. the plasma material that is confined into the glass vessel in the real plasma director. A weakly ionized plasma has been assumed in this work and it has been

represented by means of a cold fluid model [9] by making the following assumptions.

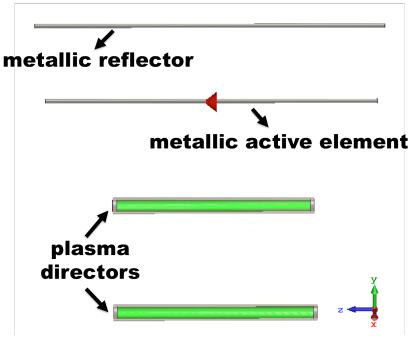


Figure 1. Design of the realistic Yagi-Uda antenna in CST Microwave Studio [8].

The first one is that the thermal motion of the ions inside the gas is negligible. Secondly, we assumed one single collision frequency that accounts for all the dissipation processes, i.e., the charge-to-charge and charge-to-neutral collisions. Finally, we assumed that the electron density and spatial distribution are temporally constant, that is equivalent to assume that the electric field intensity of the incident EM wave is small enough.

The weakly ionized gaseous plasma is modeled as a dispersive material that is suitably represented (both theoretically and in CST) as a Drude material. This provides the expression of the relative permittivity of the plasma ε_r as a function of the plasma parameters and of the operating frequency of the antenna:

$$\epsilon_r = \left(1 - \frac{\omega_p^2}{\omega^2 + \nu^2}\right) - j\frac{\nu}{\omega}\left(\frac{\omega_p^2}{\omega^2 + \nu^2}\right) \qquad (1)$$

where $\omega = 2\pi f$ is the angular frequency of the incident EM wave (i.e., the signal), ν is the momentum-transfer collision frequency of electrons with heavy particles (ions/atoms), $\omega_p=[n_e e^2/m_e\varepsilon_0]^{1/2}$ is the electron plasma frequency, m_e and e are the electron mass, and charge respectively, n_e is the local plasma electron density that is assumed to be uniform within the discharge and ε_0 is the free space permittivity.

Some of these parameters are unchanged with respect to [1]: the operating frequency of the antenna is 1.55 GHz, and the plasma frequency is $\omega_p = 178.4 \cdot 10^9$ rad/s since it has been shown by experimental measurements that this is an easily achievable value in real scenarios.

Regarding the collision frequency that accounts for the dissipation processes inside the plasma, this has been chosen significantly different from [1]: it is $\nu = 2.5067 \cdot 10^9$ s^{-1} instead of $\nu = 0.10328 \cdot 10^9$ s^{-1}, i.e., much higher, since in actual realization of plasma discharges this is a more realistic value.

Therefore the values of ε_r are significantly different with respect to [1] and this can be clearly evinced from Figure 2 in which the real and imaginary parts of the relative permittivity are reported for the two scenarios. At the frequency of interest 1.55 GHz the real and imaginary parts of ε_r are $\mathrm{Re}\{\varepsilon_r\} = -314$ and $\mathrm{Im}\{\varepsilon_r\} = -81$ in this work, while they were $\mathrm{Re}\{\varepsilon_r\} = -334$, $\mathrm{Im}\{\varepsilon_r\} = -4$ in [1]. There is an important difference between the values of $\mathrm{Im}\{\varepsilon_r\}$ in the two cases: the great increase in the imaginary part of the permittivity (that translates in an increase in the losses of the antenna) is caused by the substantial change in the collision frequency that is much higher now and that indeed accounts for the dissipation processes inside the plasma.

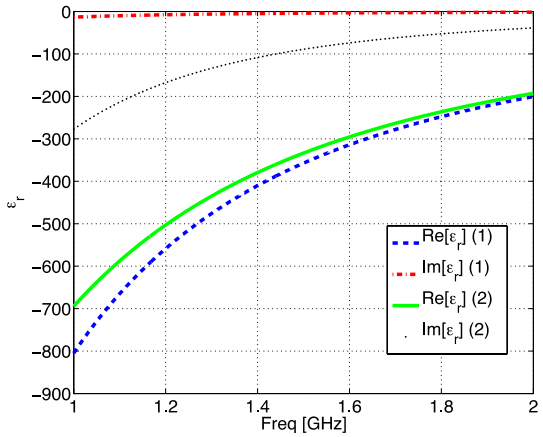


Figure 2. Comparison between the relative dielectric permittivity of the dispersive plasma in this work and in [1] (respectively (2) and (1) in the legend): imaginary part (dotted black line for (2), dash-dotted red for (1)) and real part (solid green for (2), dashed blue for (1)).

Now as far as the plasma directors dimensions are concerned, we chose to change the diameter of the discharge from 2.5 mm to 3 mm since in real cases it is more feasible to confine plasma in vessels with a larger diameter, and since 3 mm has proven to be a feasible dimension as it was proven when by actually realizing such a plasma discharge.

The length of the discharge was set according to a simplified model described in the followings. The cylindrical plasma column behaves as an open waveguide that can sustain surface- wave propagation and that therefore radiates significantly due to the breaking of the translational symmetry: this happens during the antenna operation mode, i.e., when the angular frequency $\omega = 2\pi f$ is much lower than the plasma frequency ω_p. This behaviour of the cylindrical plasma bar has been verified through a simplified analytical model according to which, for an infinitely long column of uniform non-magnetized plasma endowed with relative dielectric permittivity ε_r, the dispersion relation for the surface wave that propagates along the column is given by [7]:

$$\frac{k_{\perp p}}{\epsilon_r}\frac{I_0(k_{\perp p}r_p)}{I_1(k_{\perp p}r_p)} + k_{\perp 0}\frac{K_0(k_{\perp 0}r_p)}{K_1(k_{\perp 0}r_p)} = 0 \qquad (2)$$

where r_p is the plasma radius, $k_{\perp p} = (\beta^2 - \omega^2\varepsilon_r/c^2)^{1/2}$, $k_{\perp 0} = (\beta^2 - \omega^2/c^2)^{1/2}$ are the transverse propagation constants in the plasma and in free space, and β denotes the propagation constant along the plasma column; and finally, I_i , and K_i , $i = 0, 1$, are modified Bessel functions of the first and second kind, respectively.

The solution of (2) yields the so-called dispersion relation $\omega = \omega(\beta)$ of the surface wave. Since the plasma column has finite length, it is expected that suitable combinations of progressive and regressive surface waves along the plasma-air interface can satisfy the boundary conditions at the ends of the column, and thus give rise to standing surface waves or modes of the open cavity resonator.

Exploiting the dispersion relation it is possible to calculate the resonance frequency of a plasma column with given dimensions and therefore it is possible to calculate the length of the plasma column in order to have it resonating at a desired frequency.

According to traditional RF antenna design rules, the column should resonate at a given frequency f when its length is approximately half the wavelength in free space λ_0. But since plasma is a dispersive medium, the plasma column in the UHF band shows an effective wavelength that is shorter than the one in free space, that means that the resonance frequency for a column of length L is downshifted. Therefore, the plasma column is significantly shorter than its metallic counterpart: its length has been ultimately set to 52mm (instead of $\lambda_0/2$ = 96.8mm), a value obtained starting from $\lambda_{eff}/2$ and adjusted in order to optimize the performances of the overall Yagi-Uda antenna.

As far as the Yagi-Uda antenna model is concerned, its dimensioning follows the general rules of RF antenna design [10]: the reflector length is $L_{rfl} = 0.5\lambda_0 = 90.91$ mm and the active dipole is $L_{dpl} = 0.47\lambda_0 = 85.45$ mm. We considered a classic metal Yagi-Uda antenna in order to compare the performance of our solution with those of the standard design: in this case the metal directors length is approximately $L_{dir} = 0.4\lambda_0 = 73.82$ mm. The distance between the reflector and the dipole is $s_1 = 0.25\lambda_0 = 45.45$ mm and the spacing between the dipole and the directors and the two directors is $s_2 = 0.34\lambda_0 = 61.82$ mm. All the elements have the same radius, i.e., $r = 0.003\lambda_0 = 0.54$ mm. In CST Microwave Studio the dipole is fed through a 50 Ω discrete face-port that is placed in a gap g = 1.4 mm.

As far as the simulative model of the plasma directors is concerned, we decided to include the dielectric vessels and the metallic electrodes used to generate the plasma in the CST simulations. The gaseous plasma is confined into a glass cylinder (pyrex material from CST library) of internal radius $r_{int} = r_{plasma} =$ 3 mm and thickness of 0.6 mm: these values have been chosen according to real glass tubes that have been realized to confine the plasma in an experimental setup. Finally the glass cylinder is closed at the two extremities by two metallic electrodes of 1mm thickness that resemble those used to generate the plasma by RF surface wave technique. A detail of the director design in CST Microwave Studio is reported in Figure 3.

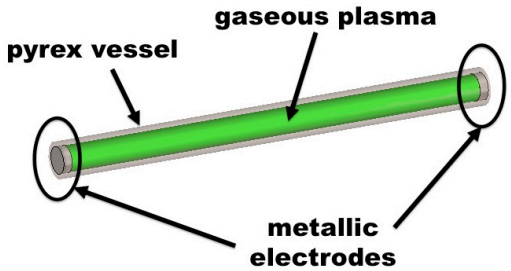


Figure 3: Design of the realistic plasma director in CST Microwave Studio [8].

In order to optimize the properties of the Yagi-Uda antenna when the two discharges are used as directors, we increased of 4 mm the distance between the two directors (this value was obtained performing optimization through a parameter sweep in CST). The last thing to point out about the model of our antenna concerns the off state of the plasma directors: in [1] when one or both the plasma bars were un-energized, they have been simply excluded from the simulations; in this work instead, we included the glass cylinders and metal electrodes in the simulations, replacing the plasma cylinder with vacuum when the plasma is off.

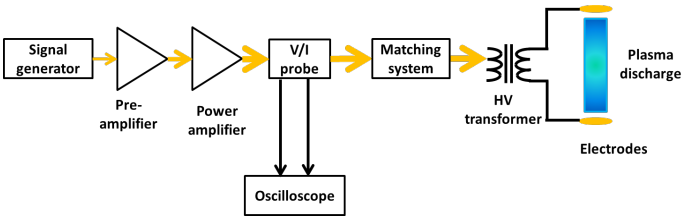


Figure 4. The scheme of the RF plasma generation system.

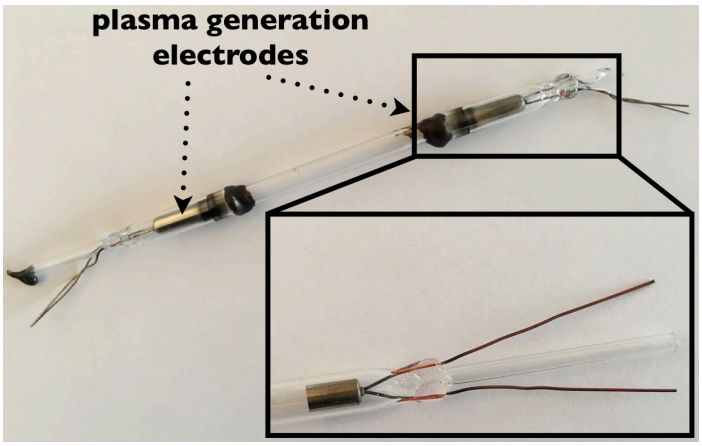


Figure 5. The experimental plasma discharge that has been used as director.

3. The Experimental Setup of the Plasma Director

To realize the plasma directors, we have relied on a ca- pacitive discharge [6] that has been driven by 50 Ω Radio Frequency (RF) power source; the RF power has been fed in the 1.8 ÷ 15 MHz range of frequencies through a matching network for efficient power transfer to the plasma, as illustrated in Figure 4. The signals are carried by means of RG-58 coaxial cables, while RF power is conveyed via RG-214 coaxial cables. The plasma has been generated by a couple of high voltage electrodes (see Figure 5) that have been placed internally at the far ends of the cylindrical pyrex vessel while being sealed with the prescribed argon gas atmosphere of 2 mbar. The pyrex vessel has an inner diameter of 3 mm, and an outer diameter of 4 mm; the two electrodes have been placed at a distance such that it is possible to realize a plasma column like the one identified in the antenna design phase.

To estimate the actual plasma density distribution within the plasma discharge, we have relied on a microwave interferome- ter [11], which is capable of plasma density measurements regardless of the gas type; specifically, the plasma density value is related to the phase shift produced on a microwave signal travelling through the plasma. The instrument works in the $10^{16} \div 7 \cdot 10^{19}$ m^{-3} range of plasma density values, is movable along the plasma source, and measures density in a plasma slab that is extended across the diameter of the source (see Figure 6). The plasma frequency value that was used in this work (see Section II) agrees with the measurements that were performed.

4. Results and Discussion

In this section the main results concerning the realistic metal plasma antenna are reported and compared with those obtained in

[1] for its ideal counterpart. The results were obtained through full-wave numerical simulations by means of CST Microwave Studio Suite [8].

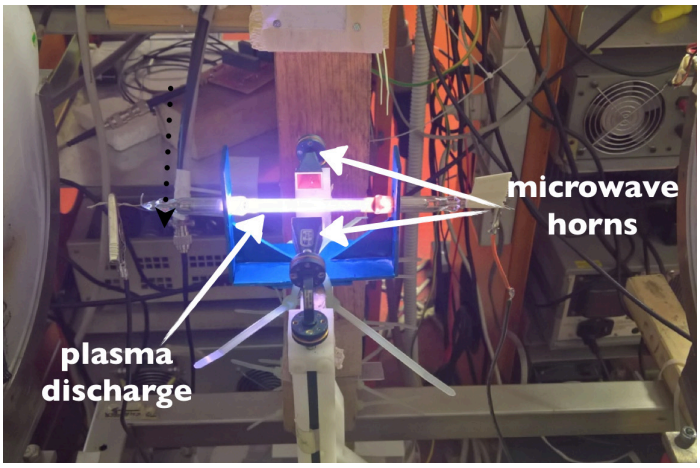


Figure 6. The experimental set-up with the plasma discharge between the two horns of the microwave interferometer for plasma density measurement.

The analysis was carried out in the frequency range 1 ÷ 2 GHz and concern mainly two metrics: (a) the reflection coefficient and the input impedance, (b) the radiation properties of the antenna, i.e., farfield gain and its reconfigurability together with radiation efficiency.

4.1. Reflection Coefficient and Input Impedance

As a first step, Figure 7 reports the absolute value of the reflection coefficient $|S_{11}|$ for the Yagi-Uda antenna with zero and two metal directors, for the hybrid Yagi-Uda with two ideal plasma directors [1] and for the antenna with the two realistic directors (as described in Section II). These three configurations were chosen as representative since they account for the starting condition, i.e., the antenna with no directors that is designed to resonate at f = 1.55 GHz, and the condition of maximum possible detuning for the antenna, i.e., when two directors are present; the other cases involving one director are omitted for ease of read of the graphs but are similar to and in perfect agreement with the results shown in the figure. As it can be seen, the absolute value of the reflection coefficient at the operating frequency is well below −15 dB in all the three cases, thus confirming that the realistic plasma discharges that were conceived in this work do not detune the antenna. In Figures 8-9 the real and imaginary parts of the input impedance Z are reported again for the classic metal antenna, for the ideal hybrid Yagi-Uda and for the realistic hybrid antenna both with two directors. It can be seen how at the operation frequency f = 1.55 GHz, Im{Z} is almost zero for the three antennas, while Re{Z} changes: it is 45 Ω for the classic antenna and for the hybrid ideal one, while it is 43 Ω for the realistic one. It can be noticed however how also in the case in which a realistic discharge is considered the idea of using a metal dipole as the active element of the antenna proves to be a good solution in order to avoid the matching issues caused by the low input impedance of plasma antennas [12].

4.2. Radiation Properties and Reconfigurability

As far as the reconfigurability of the antenna is concerned, Figures 10-11 depict the gain of the hybrid realistic Yagi-Uda antenna with zero, one and two directors in linear scale (for ease of read) in the E-plane (xz-plane) and H-plane (xy-plane) respectively.

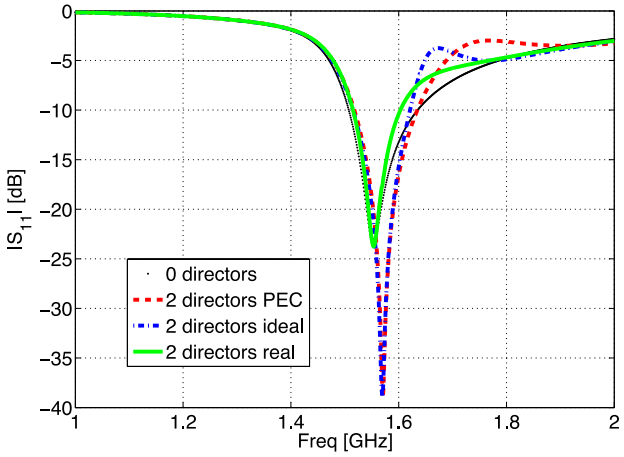


Figure 7. Absolute value of the reflection coefficient $|S_{11}|$ for the Yagi-Uda antenna with: no directors (dotted black line), two PEC directors (dashed red line), two ideal plasma directors (dash-dotted blue line) and two realistic plasma directors (solid green line).

As stated in Section II, when one or two directors are switched off, the vessel and electrodes have been included in the simulations and nonetheless we can see how reconfigurability with respect to the gain can still be achieved: the higher the number of active plasma directors, the higher the gain.

Then we compared the gain of the realistic hybrid antenna both with its ideal counterpart [1] and with the classic metal antenna. Figures 12-13 report the gain for the three antennas with one director switched on, on the E-plane and H-plane respectively: it can be seen how the gain of the antenna with the realistic plasma director is lower than those of the antenna with the ideal plasma director and than the one with the metal director (5.13 vs 6.02 and 6.25). Nonetheless the plot is in linear scale and therefore there is around 1 dB of difference between the maximum achievable gain in the three cases. The fact that the gain with the realistic plasma director is lower than in the case of the ideal one is reasonable considering that the collision frequency that accounts for the dissipation processes is considerably higher in the realistic case.

Figures 14-15 report the same plot but this time with both the plasma directors turned on. Analogous considerations can be made as in the previous case: the gain for the realistic antenna is 5.6 that is lower than that of the ideal one (6.63) and of course lower than the one of the classic Yagi-Uda (7.1); but again, the values are in linear scale so the difference between the gain of the ideal and realistic hybrid antenna is less than 1 dB.

The last case concerns the antenna when both the directors are turned off: in the ideal case, the directors were excluded from simulations, while in the realistic case, the dielectric vessels and electrodes are included. The results are not reported graphically since the two radiation patterns are almost identical. This proves, as it was expected, that the main factor that influences the radiation pattern of the antenna is the plasma collision frequency, i.e., the parameter that accounts for the dissipative processes and that in this work is significantly higher than in [1].

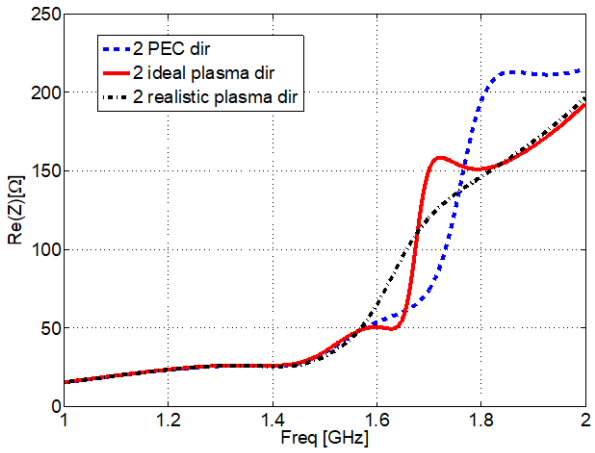


Figure 8. Real part of the input impedance of the Yagi-Uda antenna with: two PEC directors (dashed blue line), two ideal plasma directors (solid red line) and two realistic plasma directors (dash-dotted black line).

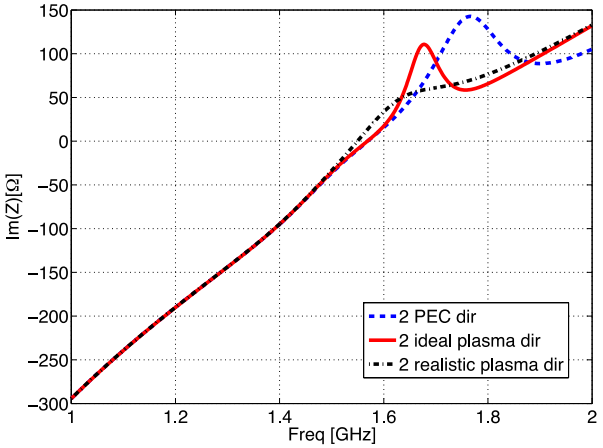


Figure 9. Imaginary part of the input impedance of the Yagi-Uda antenna with: two PEC directors (dashed blue line), two ideal plasma directors (solid red line), and two realistic plasma directors (dash-dotted black line).

As far as the efficiency is concerned, as expected this is lower in the realistic case in which losses are higher (this translates in a greater value of the collision frequency as commented previously): the efficiency is approximately 100 % for the metal Yagi-Uda, 98 % for the ideal hybrid Yagi-Uda and 80 % for the realistic one.

5. Conclusions

In this work the study and design of a reconfigurable metal-plasma Yagi-Uda antenna for microwave applications has been presented: plasma discharges have been suitably designed to be used as directors of the antenna in order to achieve reconfigurability with respect to the gain. These plasma directors were designed starting from [1] and they were modified in order to resemble more realistic plasma discharges: the dielectric vessel and metallic electrodes were included in the model and moreover the dimensions and plasma parameters of the discharges were set according to experimental measurements of real plasma antennas.

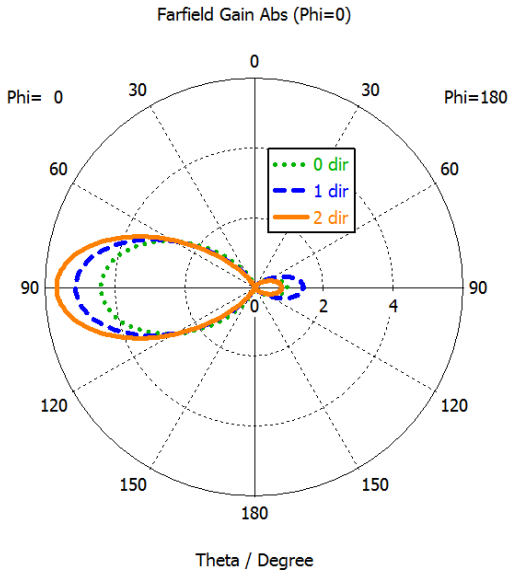


Figure 10. Farfield gain of the realistic hybrid antenna on the E-plane, i.e., xz-plane (see Figure 1) with zero (dotted line), one (dashed line) and two directors (solid line).

It has been shown that the performance of the antenna is very good as compared to [1]. Moreover it is clear from the results that reconfigurability with respect to the gain can still be achieved even with realistic plasma discharges used as directors. Of course, the radiation pattern is affected by the considerably higher value of the collision frequency, i.e., by the fact that dissipative processes are stronger in this case.

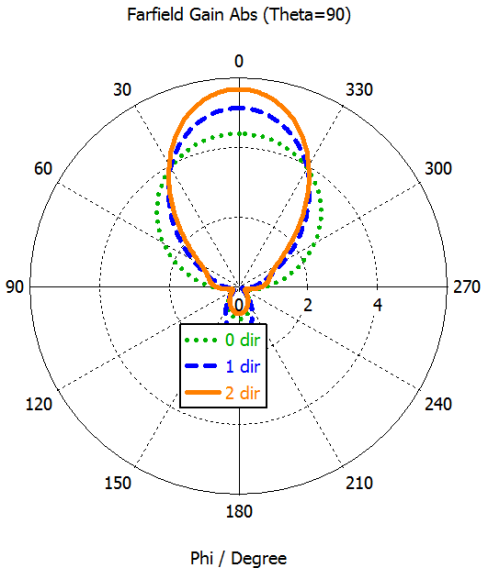


Figure 11. Farfield gain of the realistic hybrid antenna on the H-plane, i.e., xy-plane (see Figure 1) with zero (dotted line), one (dashed line) and two directors (solid line).

As a future step, we aim at considering plasma discharges with a non-uniform density, i.e., a density that changes axially along the cylindrical vessel and that resembles even better the behaviour of a realistic plasma discharge.

Therefore this study paves the way to the actual realization of the conceived hybrid metal-plasma Yagi-Uda antenna. However,

it must be pointed out that, in order to characterize and measure our proposed antenna not only in terms of its reflection coefficient but also in terms of its radiation pattern, there are still some issues to face, mainly related to the rather complicated setup that is needed to ignite and sustain the plasma inside the two directors.

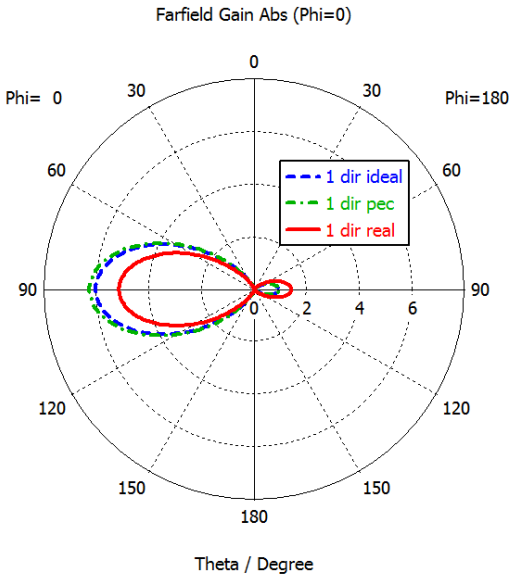


Figure 12. Farfield gain of the Yagi-Uda antenna on the E-plane, i.e., xz-plane (see Figure 1), with: one PEC director (dash-dotted green line), one ideal plasma director (dashed blue line) and one realistic plasma director (solid red line).

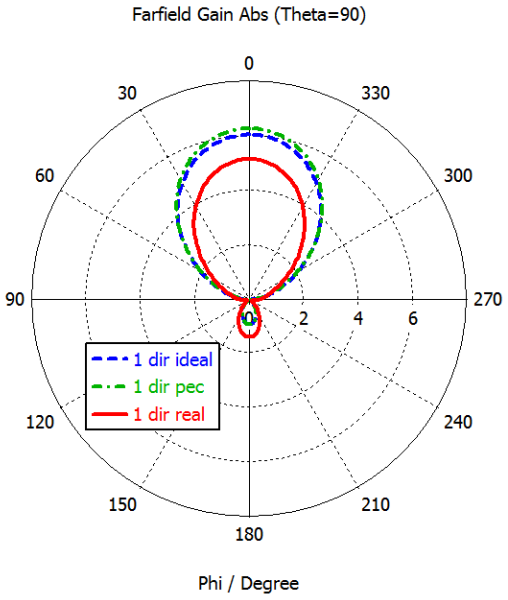


Figure 13. Farfield gain of the Yagi-Uda antenna on the H-plane, i.e., xz-plane (see Figure 1), with: one PEC director (dash-dotted green line), one ideal plasma director (dashed blue line) and one realistic plasma director (solid red line).

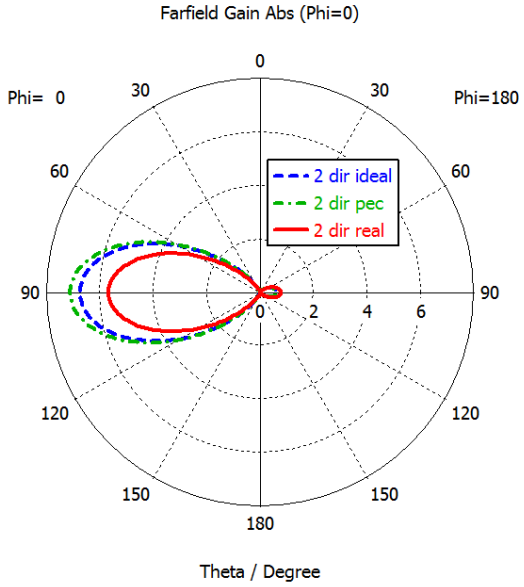


Figure 14. Farfield gain of the Yagi-Uda antenna on the E-plane, i.e., xz-plane (see Figure 1), with: two PEC directors (dash-dotted green line), two ideal plasma directors (dashed blue line) and two realistic plasma directors (solid red line).

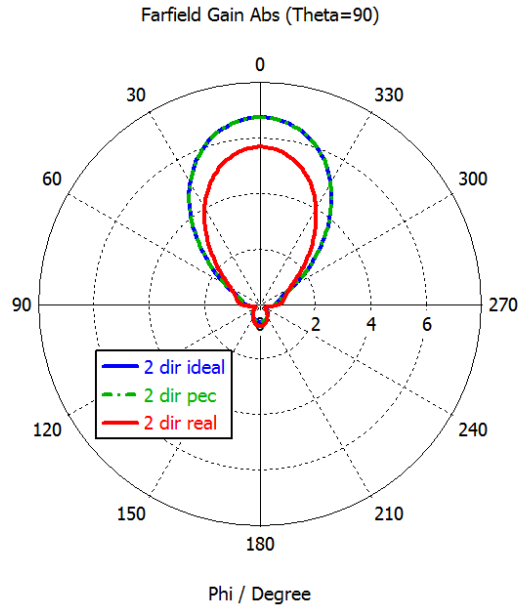


Figure 15: Farfield gain of the Yagi-Uda antenna on the H- plane, i.e., xz-plane (see Fig.(1)), with: two PEC directors (dash-dotted line), two ideal plasma directors (dashed blue line) and two realistic plasma directors (solid red line).

Acknowledgment

The authors acknowledge financial support from the University of Padova, project CPDA150022.

References

[1] G. Mansutti, D. Melazzi, and A.-D. Capobianco, “Hybrid metal-plasmaYagi-Uda antenna for microwave applications,” 46th European Microwave Conference (EuMC), London, UK, Oct. 2016, pp. 727-730.

[2] G. G. Borg, J. H. Harris, N. M. Martin, D. Thorncraft, R. Milliken et al. “Plasma as antennas: Theory, experiment and applications,” Phys. Plasmas, vol. 7, no. 5, pp. 2198–2202, May 2000.

[3] T. Anderson, ”Plasma Antennas,” Artech House, 2011.

[4] J. P. Rayner, A. P. Whichello, and A. D. Cheetham, “Physical characteristics of plasma antennas,” IEEE Trans. Plasma Sci., vol. 32, no. 1, Feb. 2014.

[5] T. Dwyer, J. Greg, D. Murphy, J. Perin, R. Pechacek, and M. Raileigh, “On the feasibility study of using an atmospheric discharge plasma as an RF antenna,” IEEE Trans. Antennas Propag., vol. 32, pp. 141–146, 1984.

[6] M. Lieberman and A. Lichtenberg, “Principles of Plasma Discharges and Materials Processing”, ser. Second Edition. John Wiley & Sons, Inc., 2005.

[7] M. Moisan, A. Shivarova, and A. Trivelpiece, “Experimental investigations of the propagation of surface waves along a plasma column,” Physics of Plasmas, vol. 24, p. 1331, 1982.

[8] CST Microwave Studio 2016, https://www.cst.com/products/cstmws.

[9] D. Swanson, “Plasma Waves,” Institute of Physics, Second ed., 2003.

[10] C. A. Balanis, Antenna theory: analysis and design, 3rd edition, Wiley,

[11] O.Tudisco et al., ”A microwave interferometer for small and tenuous plasma density measurements,” Rev. Sci. Instrum. 84, 2013.

[12] D. Melazzi, P. De Carlo, M. Manente, V. Lancellotti, and D. Pavarin, “Numerical results on the performance of gaseous plasma antennas,” International Conference on Electromagnetics in Advanced Applications (ICEAA), September 2015, pp. 569 - 572.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 449-459 (2017)
www.astesj.com
Special issue on Recent Advances in Engineering Systems

ASTESJ
ISSN: 2415-6698

Verifying the Detection Results of Impersonation Attacks in Service Clouds

Sarra Alqahtani*, Rose Gamble

Tandy School of Computer Science, University of Tulsa, Tulsa, OK, USA

ARTICLE INFO

Article history:
Received : 05 April, 2017
Accepted : 04 May, 2017
Online: 24 May, 2017

Keywords :
Impersonation
Cloud
Cybersecurity

ABSTRACT

A web service impersonation is a class of attacks in which an attacker poses as or assumes the identity of a legitimate service to maliciously utilize that service's privileges. Providing security for interacting cloud services requires more than user authentication with passwords or digital certificates and confidentiality in data transmission. In this paper, we focus on the service cloud model, which facilitates the composition and communication among web services owned by different cloud vendors. We develop a distributed mechanism to detect and mitigate impersonation attacks against web services in the cloud. The detection approach monitors the behavior of each service and identifies anomalies as a potential impersonation attack if it deviates significantly from the expected behavior. To verify the impersonation attack, we deploy a cloud-based verification technique, misleading suspicious services with useless responses. The experimental results show that modeling request behavior reliably detects a significant number of impersonation attempts, with a performance degradation that is a reasonable trade-off.

1. Introduction

A web service impersonation is a class of attacks in which an attacker poses as or assumes the identity of a legitimate service to maliciously utilize that service's privileges. This attack can cause a serious threat to the security of service clouds, i.e. those clouds that allow the provisioning of multiple vendor services to dynamically compose an application to answer a client request. By impersonating a legitimate service, the intruder can maliciously access the victimized a service's resources. For the service cloud, this attack class is a version of identity theft. Such attacks completely undermine traditional security mechanisms due to the trust imparted to service credentials once they have been authenticated. Many attempts have been made at detecting this kind of attack (also called masquerader or identity spoofing attacks) against legitimate users in clouds. In this dissertation, we have developed a distributed mechanism to detect and mitigate impersonation attacks against web services in the cloud [1].

When using the cloud, an end user's credentials to authenticate request messages between web services lacks the verification of the origin of the request (i.e. the requester service). Thus, this vulnerability can be used by the attacker after stealing the credentials to create fake services. Since the credentials represent a long-term authentication tool, the attacker can attack a web service for unlimited time and the owner of the credentials would not discover the problem until the damage is done.

In our proposed architecture (Figure 1) [1], we assume a secure session with a security token issued by a Security Token Service (STS) to mitigate the general vulnerability of spoofing the user credentials. The STS is trusted by both the client and the web service to provide interoperable security tokens. The client sends an authentication request, with accompanying credentials, to the STS. The STS verifies the credentials presented by the client, and then in response, it issues a security context token (SCT) that provides proof that the client has authenticated with the STS. The SCT is built on a SAML standard format for exchanging authentication and authorization data between different parties [1, 2]. The client presents the security token to the service. The service then verifies the token with the STS, which proves that the client has been successfully authenticated. For proper use of the token, it is expected that all composed, trusted services communicate with the STS to perform token delegation and validation at each service request and response, but only client services issue or cancel a token. The SCT holds information to specify its scope, creation, and expiration time to develop the basis for encrypting and signing subsequent message exchanges, which results in efficient and secure communications between services in an application. The cloud services and STS use X.509 certificates to sign and encrypt their messages. The scope of the issued SCT is limited to the designated STS regardless of whether the sender service specified

*Corresponding Author: Sarra Alqahtani, Tandy School of Computer Science, University of Tulsa, Tulsa, OK, USA | Email: sarra-alqahtani@utulsa.edu

the scope in the initial request. This prevents the sender from using the SCT to directly access a recipient without communicating first with STS. Therefore, a token's lifetime is limited to its session active time.

The proposed architecture introduces the concept of scoped distributed Security Management Databases (SMDBs) to store audit logs, meta-information about the infrastructure of each scope, history of security incidents, incident reports, related security policies, and SLAs [2]. Scoped detectors use SMDB content to identify security threats exploiting the scoped content. We have shown in [3] how CAPEC attack patterns fit into the evidence collection and evaluation strategy using the proposed architecture.

For the proposed system, we implement our solution using Apache Web Service Security for Java (WSS4J) to secure the deployed cloud services and their messages within a provisioned application given the primary security standards. WSS4J supports the encryption of messages using X.509 certificates of services. Furthermore, it prevents replay attacks that can happen when the attacker sniffs a message and resends it to the same recipient. Apache WSS4J makes all security tools available for developers to secure their services against impersonation attacks by using strict authentication standards. When a service passes the identity authentication it is fully trusted in the application. However, if the service identity has been compromised for any reason, WSS4J cannot detect the impersonation attack or even log its traces in the log file. Consequently, the impersonated service will be authenticated and the attacker can illegally gain access to potentially private information.

The impersonation attack is not only limited to the SOAP messaging protocol in service clouds. Any type of communications between services that requires request authentication is vulnerable to impersonation attacks. Authentication in REST services does not have a standard policy or rules. However, the most common used approaches for REST request authentication are HTTP AUTH and tokens. In HTTP AUTH, the client service needs to compute the Base64 encoding of its credentials and include them in each future HTTP request to the server using the "Authorization" HTTP header. This header can be encrypted using the server public key and the client private key. This approach is very similar to the approach used in the SOAP protocol by using the SOAP header to carry authentication information to the destination. Both HTTP and SOAP headers can be faked by the attackers who illegally obtain a copy of the legitimate client service's identity.

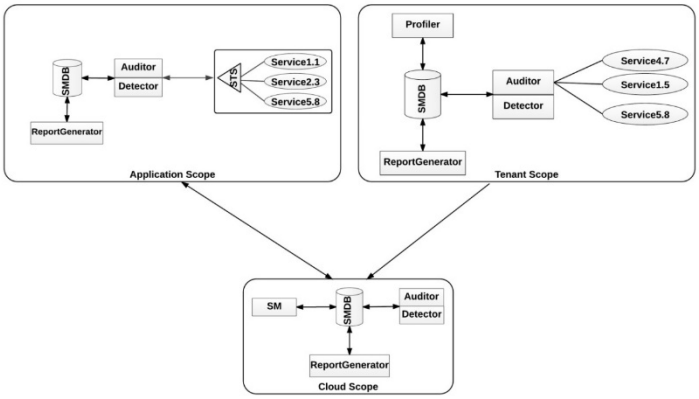
The second authentication approach used in RESTFUL clouds is to create a dedicated login service similar to a STS that accepts credentials and returns a token. This token should then be included, as a URL argument, to each following request. The attacker can also get a valid token by using the stolen credentials. Consequently, regardless of the type of communication messaging used between services, it is possible for an attacker who has a stolen identity to successfully pass the authentication step and get trusted in the target server.

The Amazon REST APIs use a custom HTTP scheme based on a keyed-HMAC (Hash Message Authentication Code) to authenticate REST requests. To authenticate a request, the client needs first to concatenate selected elements of the request to form a string. Then, it must use its AWS secret access key to calculate the HMAC of that string. Informally, this process is called "signing the request," and the output of the HMAC algorithm is called the signature, because it simulates the security properties of a real signature. Finally, the client, either a human user or another service, has to add this signature as a parameter of the request by using the syntax described by the server. The risk stems from the leaked AWS secret access key, which happened recently for many customers of Amazon clouds. Hence, the REST messaging architecture is also vulnerable to impersonation attacks if the authentication identity has been compromised. In this paper, though we will focus on detecting the attack by using SOAP protocol, the solution design can be extended to work with REST services.

In this paper, we hypothesize that the legitimate service can obtain sufficient information from the cloud regarding the successor service classes that it often communicates with when provisioned for certain application. An impersonated service, on the other hand, would likely have a more chaotic and random behavior by communicating more extensively with known successor services and more broadly to services not often communicated with in a manner that is different than the victim service being impersonated [4]. Thus, our detection approach focuses on monitoring a service's behavior in real time to determine whether current service actions are consistent with the service's profiled behavior.

To reduce the false positive rate, which is the main drawback of the anomaly detection systems, we combine the detection of anomalous behaviors with a cloud-based verification technique. The verification technique prevents the risk of revealing private data to impersonators and, at the same time, verifies the identity of the malicious service. When the verification algorithm positively verifies the detection results, it returns a fake response to delude the attacker.

The major contribution of this paper is our distributed framework for auditing and detecting impersonation attacks in service clouds. We design a distributed impersonation detection approach within the framework that is based on profiling each service's request behavior. This behavior reveals that monitoring the general features of the SOAP request messages between web services can achieve a reasonable detection rate with minimal false positive alarms. In order to increase the accuracy of the detection system, we design an aging method to determine when the profiles need to be updated and how to adjust them so they are more heavily influenced by the most recently observed behavior. Moreover, we manifest the false positives and decrease the impersonator live time inside the cloud by re-authenticating the suspicious service using a novel verification technique. Since the credentials of the suspicious service may be spoofed by the impersonator, our verification technique uses a predefined fake response to validate the identity of the suspicious service that exhibits a significant behavior deviation. If the suspicious service accepts the fake response by the suspicious service, then the detection system has correctly detected the impersonator. Otherwise, the suspicious service is considered to be a legitimate service that has been incorrectly classified as an impersonator because it changed its behavior without updating its profile.



Our experimental results show that modeling service behavior reliably detects a high percentage of impersonated services with no false positives. The limited set of statistical values used for service behavior modeling results in an acceptable performance. However, in the case of mimicry attacks where the impersonator fully mimics the victim service's behavior, the detection algorithm cannot detect the malicious requests. Resolving such attacks and improving the algorithm performance are future work.

The paper is organized as follows. The threat model is described in Section 2. The literature review of impersonation detection techniques is presented in Section 3. Section 4 demonstrate the detection approach used in this research. Section 5 discusses the detection results. Section 5 concludes the article with some suggestions.

2. Threat Model

2.1. Vulnerabilities

The standard approaches of identity verification including authentication and encryption have been shown to fail to detect impersonated services for a variety of reasons, including insider attacks, misconfigured services, faulty implementations, buggy code, and the creative construction of effective and sophisticated attacks not envisioned by the implementers of security procedures. Attack methods such as phishing, fraud, and exploitation of system vulnerabilities still achieve results in service clouds. Credentials and passwords are often reused, which amplifies the impact of such attacks.

The Heartbleed bug leaked private information such as encryption keys and credentials from the service's owner server. Leaked secret keys allow an attacker to decrypt any past and future traffic to the protected services and to impersonate the services as well. Therefore, any protection given by the encryption and the signatures in the X.509 certificates can still be bypassed. Mitigating bugs like Heartbleed is out of the service owner and cloud provider control, which means we need to design another line of defense to mitigate the risk of "bleeding" the encryption keys (X.509) and service credentials. In this article, we introduce a preventative approach that distinguishes between requests coming from legitimate services and those coming from fake services with impersonated or spoofed credentials to block them and alert the victim's owner to revoke the stolen keys.

Different incidents happened in Amazon cloud prove that the classical authentication and authorization techniques are not safe enough to fully trust all services holding valid identities. In April 2010, Amazon experienced a Cross-Site Scripting (XSS) bug that allowed attackers to hijack credentials from the site. In 2009, numerous Amazon systems were hijacked to run Zeus botnet nodes. Also, in April 2014, some of AWS clients claimed about very high bills due to their credentials were compromised.

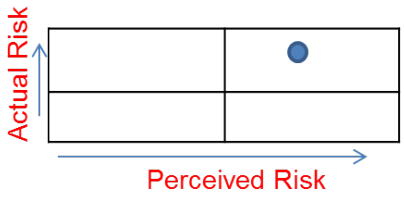
2.2. Target Assets

The expected target of the impersonation attack is the legitimate service's identity.

2.3. Impact

According to the CSA in [5], the risk analysis technique CIANA classifies the impact of service hijacking or impersonating as a combination of authenticity, integrity, confidentiality, non-repudiation, and availability threat. On the other hand, STRIDE risk analysis puts the service impersonation under tampering with Data, repudiation, information disclosure, elevation of privilege, and spoofing identity. Therefore, service impersonation attack causes a very high risk on the cloud from all aspects of security.

According to a survey conducted by CSA the extended CSA Top Threats Working Group [5] , the impersonation threat is still relevant to clouds by 87%. Also, the data leakage was ranked as the third top threat in 2013.The risk matrix for service impersonation is depicted in Figure 2.



2.4. Scenarios

The attacker who exploits a bug similar to Heartbleed and has a copy of the encryption public and private keys and the credentials of the victim service (we use identity to describe these information) can effectively perform a service impersonation attack against the victim service. Then, it can illegitimately use the stolen identity to gain access to other services in the service cloud by following the listed steps:

1. The attacker builds a fake web service by using the stolen identity to impersonate the victim service.
2. During an active session of an application that has the victim service provisioned within it, the attacker uses sniffer software to steal a valid token sent from the STS to the victim service. The intercepted token is encrypted by the public key of the victim service. Since the attacker has the encryption keys of the victim service, he can decrypt the token.
3. The attacker uses the stolen token along with the victim identity to ask the STS to delegate the token to the target service of the attack. The target service must be provisioned within the same application with the victim service.
4. The STS authenticates the impersonated service and replies with a delegated token.
5. The attacker finds the interface of the target web service and its entry points (i.e., methods) by discovering its public WSDL file.
6. The attacker uses the delegated token and the stolen identity to create and encrypt a request to the target service.

7. The target web service validates the token of the impersonated service with the STS.
8. The STS authenticates the token and sends the confirmation to the target service.
9. The target service decrypts and checks the structure of the message.
10. The target service responds to the impersonated service with a valid encrypted response SOAP.
11. The attacker decrypts the received response using the victim private key.

With this scenario, private information can be revealed to illegitimate services by trusting the presented identity without doing a more thorough check. The success of this attack can trigger several attacks, such as DDoS attacks as we have explained in [6], to abuse the availability of the target service and the integrity of the victim service.

3. Related Work

There are two general approaches for intrusion detection: misuse detection and anomaly detection [7]. The misuse detection uses the knowledge accumulated about attacks and checks for signatures of these attacks while the anomaly detection builds a reference model of the usual behavior of the system being monitored and checks for deviations from the observed usage. The false positive rates of misuse detection are lower than that of anomaly detection [8] but anomaly detection has the advantage of detecting previously unknown attacks. In impersonation attack detection, the intrusion detection system cannot get the signatures of attackers until the attackers have been detected by the system. Therefore, the attack model cannot be constructed in advance; hence we prefer to use anomaly detection approach in this context.

In the literature, anomaly detection was implemented in a variety of approaches. These approaches are usually categorized into two groups, i.e. statistical approaches and machine learning approaches [8]. In statistical approaches, anomaly detection systems usually watch behaviors of observed objects to comprise statistical distributions as a set of trained profiles during the training phase. These systems then apply the set of trained profiles by comparing them against a new set of activities of observed objects during runtime. An anomaly is detected if there is a significant deviation resulted from the observation. In general, any incident whose frequency goes beyond a predefined standard deviation from statistical normal ranges raises an intrusion alarm [7].

Machine learning based approaches tend to reduce the supervision costs during the training phase of statistical approaches by enabling detection systems to learn and improve their performance on their own. Neural networks and Hidden Markov Model have been proved to be useful techniques at the network traffic level as shown in in [9] and [10]. However, using ML algorithms in a highly dynamic environment like service clouds have several drawbacks such as increasing performance overhead, storage requirements, and computational expense [11].

In the context of this paper, we focus on statistical approaches for detecting impersonation attacks. In [8,12-15], statistical algorithms are applied to detect anomalous patterns in the system. They use the normal behavioral profiles of the monitored object, which is either a user or system process. Unlike these systems, this research applies anomaly detection at the web service level. Rather than profiling the normal behavior of users or systems, our approach profiles the normal request behavior of web services. Hence, the detection system deals with the web service as an independent object that can be impersonated irrespective of its owner's behavior.

Researchers in [16] provide an intrusion detection framework for cloud systems targeting masquerader attacks. Each user has a profile to model its behavior and make it available across the cloud. To measure the deviation between an activity initiated by a normal user and activity initiated by the masquerader, they use the Data-Driven Semi-Global Smith Waterman alignment algorithm. The objective is to compute the best alignment score, by aligning the active user's session sequence (e.g., mouse movements, system calls, opened windows titles, etc.) to the previous stored sequences for this user. It achieves good results in detecting masqueraders in clouds with detection accuracy of 88.4 % and a low false positive rate of 1.7 %. However, the focus on monitoring human users is insufficient for detecting service impersonations in service clouds. Service clouds are mainly composed of collaborative services from different administrative domains with different security policies. Researchers in [16] do not address attacks that might be triggered by impersonated services with legitimate identities.

The detection approach presented in [17] detects insider masqueraders in file systems. This approach tracks and measures changes in user behavior, alerting on any significant changes. They use one-class support vector machines to develop user behavior models and a set of data features related to the file systems including: the process name, the process path, the parent of the process, and the type of process action, the process command arguments. Their experiments show that modeling search behaviors of genuine users reliably detects all masqueraders with a very low false positive rate of 1.1%. Despite the fact that they do not test this approach on clouds, it may be applicable for only storage clouds. Hence, their detection system is not applicable for the service cloud model we work with. However, our verification technique is inspired by their proposed decoy technique, which launches disinformation attacks against malicious insiders, preventing them from distinguishing the real sensitive data from fake worthless data. They place traps within the file system which are documents downloaded from the Internet including several types of useless documents such as tax return forms, medical records, credit card statements, e-bay receipts, etc. The decoy files are downloaded by the legitimate user who owned the system and placed in highly-conspicuous locations that are not likely to cause any interference with the normal user activities on the system. A masquerader, who is not familiar with the file system and its contents, is likely to access these decoy files, if he or she is in search for sensitive information. Their decoy technology is more suitable for file systems or storage clouds more than service clouds. Hence, we design a service-cloud-based decoy technique as our verification technique to mislead the impersonators with fake information.

An anomaly based network intrusion detection system for Remote to Local (R2L) attacks is proposed in [15]. R2L attacks usually exploit a vulnerability in a service at the target machine to elevate the attacker's privileges. This attack is similar to the impersonation attacks occurring in clouds. The proposed detection system consists of two logical modules, the Packet Processing Unit (PPU) and the Statistical Processing Unit (SPU). The PPU has to extract service requests from the stream of packets on the wire to pass them to the SPU. The task of the SPU is to read the service

requests and extract suitable statistical input data. Requests are divided into several groups where each group contains request types with similar statistical properties such as Get and Post protocols. The requests of each group are then analyzed independently. The properties of a request used to determine its anomaly score include: type of request, length of request, and payload distribution. The anomaly score of a service request is the weighted sum of the three scores computed for each of the previously mentioned properties. Finally, the anomaly score is compared to a threshold that can be manually set by the security administrator. The presented detection system has detected all anomaly requests with a very slight false positive rate. Since the anomaly score technique proposed for network packets, we adjust it to detect impersonators in service clouds by computing the anomaly score of SOAP request messages exchanged between services.

In [14], the detection system concentrates on profiling the behavior of system processes by studying the set of system calls made by the program. The designed algorithm compares the sequence of system calls captured during online usage to the normal sequences profiles during the training phase. The detection accuracy was 72.2% with a false positive rate of 2.1%.

4. Impersonation Attack Detection Approach

When dealing with the impersonation attack detection, it is important to mention that we assume the impersonator has already obtained the required security parameters including the victim service's credentials, the security token, and the encryption keys to get involved in the cloud application and communicate with its services (explained above). We also assumed that the attack targets are only web services. Hence, SMDBs, detectors and other entities are assumed to be sufficiently secure.

When presenting the stolen identity, the impersonated service appears as a legitimate service with the same access rights as the victim service. Ideally, monitoring a service's behavior after being granted access is required in order to detect such attacks. We find that certain types of service activities can reveal the service intent. For instance, SOAP requests used for web service communication offer an interesting behavior to monitor. A service request contains the sender supplied data which is sent over the network to another provisioned service in the same application session to perform a single task on behalf of that sender, which is either a human user or another service. This data shows the intent of the service trying to gain secure information from different services in a very short period of time before the stolen security token expires and before the victim finds out about its stolen identity.

The developed approach composes of three steps. The output of each step represents the input to the next step (Figure 2):

A. Auditing services' messages by the auditing system during the training phase.

B. Profiling the request behavior of each service in a profile document before running the detection algorithms during runtime using log files.

C. Detecting any significant change between the profiled and live behaviors then verifying the identity of the malicious service by using the verification technique.

To assess the feasibility of our approach within the service cloud model, we craft an easy to understand case study with a deployed service cloud hosting 25 distinct web services with 15 tenants owning and offering these services. We implement a session manager randomly chooses among appropriate services and composes them to provision an application that satisfies a user's request. A service cloud receives the traveler requirements with the preferred maximum price and the cloud returns the most suitable travel package with allowed payment. Table 1 shows the services and tenants used in the case study to evaluate the proposed impersonation attack detection approach. The cloud employs different meta-services to manage the travel reservations. These services are a Travel Agency service, Transportation services, and Living services.

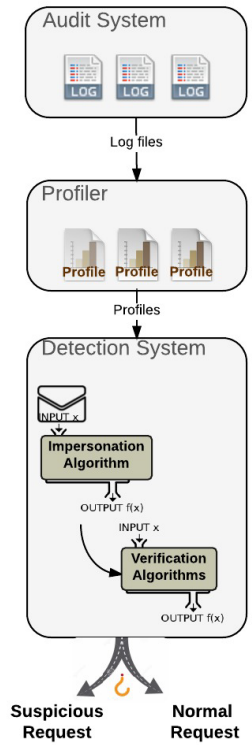


Figure 3: Impersonation detection approach

Table 1: The available services in the case study

Tenant Class	Web Services
Banks	CreditChecking, Payment, CreditCard, paymentService, CreditVerification,CreditPayment, CheckBalance
Airlines	FlightFinder, FlightService, CheapFlight, TicketFinder, FlightFinder
Car Rentals	RentalAgent, RentalService, CarRental, RentalService, RentCar
Hotels	RoomBooking, RoomLocator, Accom, AvaRoom, BookingService, FindRoom
Weather	WeatherRetriever, Forecast

The Session Manager (SM in Figure 1) dynamically provisions services from tenants into a service chain to simulate a travel management system as in Figure 4. Given a client request, the provisioned services buy a flight ticket to a specific destination at a specific time, reserve a hotel room, rent a car, retrieve the forecast for the desired travel dates, and pay the total price from the client's bank account. The resulting service chain consists of at least 9 services from at least 5 tenants.

4.1. Auditing System

In our service cloud architecture [3], the impersonation detection algorithm is distributed across a tenant scope, session or application scope, and cloud scope as depicted in Fig. 1. This distribution allows for load balancing among services logging events, tenants generating profiles of all of its hosted services, applications executing service chains to satisfy clients' requests and detect impersonation attempts, and the cloud verifying the extent of the attack.

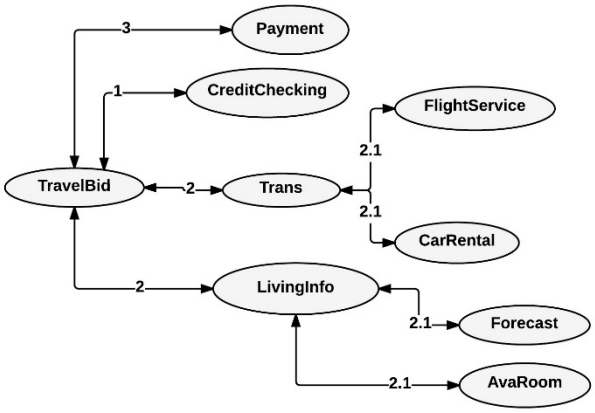


Figure 4: A scenario of Travel Package Bidder

Event logging at each service is done by its tenant's Auditor. The Auditor is a web service interceptor attached to each service to monitor its request and received messages. By intercepting SOAP requests, the Auditor is able to collect important information about the recipient to sufficiently check the sender's request behavior. Such information includes recipient services and their business classes, requested methods in these recipients, request size, and other critical details. The advantage of this approach is that the logging process created and managed independent of the service process code to protect the auditing process from intruders. Since the tenant auditor would not be able to log requests issued by the impersonated service, checking the victim service's log file for the suspicious events helps to detect the attack at early stages.

At the application scope, the STS log file retains messages communicated by the client and services as it manages the creation, delegation, validation, and cancellation of the security token. By logging these details, we can corroborate the STS log file with the victim service's log file (the service whose credentials have been stolen) to reveal any missing events [1] as significant evidence of an impersonation attack.

The cloud scope gathers centralized forensic investigation information and examines it for impersonation attacks. During the detection process, the cloud scope correlates alerts received from any affected application scope and compares the malicious request of the impersonated service with the victim service's log file. The final decision about the malicious request is logged by the cloud Auditor at the cloud scope.

4.2. Profiler

The detection system detects an impersonated service by comparing its observed request to its expected behavior using certain metrics that are defined in its profile. The expected behavior can be deduced during the training period. Since the manual creation of expected behavior is cumbersome, we use the training period for each service to collect information about its behavior. During training, the auditing system logs enough information about request events for each service. Extracted data about requests from a specific service's log file are not classified in advance, meaning that the frequent requests represent the general behavior of the service. Hence, during runtime, received requests are evaluated against a summary of the training data (i.e. profile) to measure the deviation from the normal behavior.

Profiling the services' behavior is done in two phases: profile generation and profile aging. Each phase is discussed in detail in next subsections.

4.2.1. Profile Generation

Instead of building the detection system by evaluating the received request during runtime against all requests in the sender's log file, we summarize these records in profile documents. Thus, the detection time dramatically decreases which leads to an improvement in the performance. The Profiler uses the logged data at the tenant scope to generate the service's request behavior profile. Then, SM makes the services' profiles available to the involved application detector.

The profiler depends on a statistical model to generate the request behavior profile from the service log file. This model consists of different attributes extracted from the SOAP request without processing its payload. Hence, the model reduces the cost of decrypting payloads of malicious requests that would be declined at the end. The used attributes are recipient's class, SOAP size, method, and historical incidents. We found these attributes sufficient to reveal any deviation from the normal request behavior.

1) Recipient's Class:

We hypothesize that each service has a logical "preferred" set of business classes where it sends requests based on its business purpose. For example, flight reservation services always communicate with different services in the banking and airline classes to buy a ticket for the client. In the statistical model, we compute how frequent the recipient's class has been accessed by the sender during the training phase. The probability of relevant frequency is used to compute the anomaly score (AS) of recipient's class as follows:

$$AS_{class} = - \log_2 (P[c]) \quad (1)$$

Equation (1) assigns higher anomaly scores to requests sent to recipients with less frequent classes. The probability P[c] represents the relative frequency that a request with class c has occurred during the training period of the sender. In order to prevent too high anomaly scores (or even infinite values) for the class that has been requested by the sender very infrequently (noise) or not at all during the training period, the probability of each request class needs to be set to a minimum value depends on the service characteristics and it might be adapted from time to time. Since this minimum value yields to a maximal anomaly score for classes that never received requests from the sender, it is necessary to maintain the same maximum anomaly score across all attributes of SOAP requests to keep the total anomaly score consistent.

2) SOAP Size:

The size of a SOAP request can be a good indicator of the reliability of its source. The structure of a SOAP request consists of optional elements that the service can include, such as SOAP Header and SOAP Fault. Usually, the size of the SOAP message does not vary much between requests that are generated by services in a certain class to a specific recipient. The situation looks different when requests carry input wanting to obtain information from the target recipient. For instance, the attacker may send a request with nested queries to get more information from the recipient's database. To assess this attribute properly, we correlate SOAP size with either the recipient service or class based on some properties of the observed request.

The sender's profile keeps track of all recipient services and their classes that have ever received requests from that sender along with the average size of these requests. If the current recipient of the observed request has been accessed by the sender over an acceptable rate of requests then the available information about the requests sent by the sender to this recipient would be sufficient to evaluate the size of the received request. Otherwise, the size of requests received by the current recipient's class (received by any service in this class) would be used to compute the size attribute.

We calculate the anomaly score for SOAP size as in [15], by using the mean (μ) and the standard deviation (σ) of the sizes of the requests that have been sent by the sender during its training period to either a specific recipient or class. These two values are stored in the sender's profile for each recipient and each class that has been previously accessed by the sender. The following formula is used to check the anomalousness of a request with size Z during run time. The anomaly score grows exponentially as the request size increases. In order to tolerate a reasonable amount of deviation of the size attributes, we set x variable as the base and multiply σ with a constant factor of y. The choice of x and y is the recipient's responsibility based on its features and needs. The x and y variables are stored in the application SMDB and available for the detection algorithm.

$$AS_{size} = x \wedge ((Z-\mu) / (y*\sigma)) \quad (2)$$

This equation assigns anomaly scores greater than x only to requests that have SOAP size longer than the average. This is consistent with our assumption that malicious request increases the total size of the soap message. We limit the maximum value of AS_{size} to the same value in AS_{class} to avoid having infinite anomaly scores.

3) Method

The WSDL file for a web service shows its business logic and policies. The business logic is represented by the operations or methods that are available for the client to call along with information about their required parameters, data types, maximum values etc. Since this information is publicly available, there is no reason for the attacker to restrict himself from calling any method while he has valid credentials. However, for the victim service, there should be a noticeable behavior towards the recipients' methods.

Assume the recipient is a credit card service with two methods; one to check the sufficiency of an input card number to cover a certain payment and another method to return the available balance of the received card number. If the recipient receives a call from the impersonated service to the balance returning method given information that the victim service is used to frequently call the sufficiency checking method then the detection system can consider this call as a serious flag of an impersonation attack.

To compute the anomaly score of the invoked method in the observed request, we correlate this method with the probability of relevant frequency of its service which is the current recipient of the request. However, the dynamic nature of the provisioning process in service clouds can lead to provisioning some services offering the same business service less frequent than others. For recipients that have been provisioned and accessed by the sender more than a predefined threshold, the anomaly score of their methods would be computed using their probabilities in equations (3) and (4) below.

The anomaly score (4) for the called method is calculated for each method in the recipient WSDL that was called by the sender using the conditional probability in (3) if P(rec) is greater than a preset threshold. The anomaly score is stored in the sender's profile. We also limit the maximum value of AS_{meth} to the same value in AS_{class} and AS_{size} to avoid having infinite anomaly scores.

$$P(meth|rec) = (P(meth \text{ and } rec))/(P(rec)) \quad (3)$$

$$AS_{meth} = - \log_2 (P[meth|rec]) \quad (4)$$

Other recipients with less provisioning probability do not have sufficient information about their methods in the sender's profile that can be used by the detection system to determine the anomaly scores of their methods. Hence, using equations (3) and (4) to compute the anomaly score of the invoked method for infrequent recipients would introduce a significant bias into the final anomaly decision. Instead, owners of infrequent recipients need to set anomaly scores to their methods representing their acceptable risk degree of revealing information by each method if it has been accessed by an impersonated sender.

4) Historical Impersonation Incidents

Since the cloud hosts many services owned and operated by different tenants, the cloud provider must not assume that all tenants are careful about the security configuration of their services. Thus, the detection system needs to consider the historical incidents of impersonations occurred for each service when examining its requests. The service profile provides the probability that the service historically reveals to be a victim of an impersonation attack.

To calculate the impersonation probability for each service, information about the historical incidents that occurred for the service is accumulated in the tenant SMDB. Then, Profiler exploits this information with the information extracted from the service log file about the total number of requests to compute the probability of the service being a victim of previous impersonation attacks:

$$P(s) = (\text{total impersonated requests}) / (\text{total requests}) \quad (5)$$

We use P(s) to compute the anomaly score for the service history in (6). We weight the probability to have a maximum value similar to other attributes. The higher the anomaly score of the service history the more likely that the service is currently involved in an impersonation attack.

$$5) \quad AS_{history} = P[s] * max \quad (6)$$

The behavior profile for each service has a summary of its request messages logged in its log file using equations (1)-(6). These values would be preserved in the tenant SMDB upon the service registration and provisioned to the relevant applications' SMDBs by the SM.

4.2.2. Profile Aging

When the detection system produces more false alarms than that logged for a previous period for a certain service, that means either the request behavior of this service has dramatically changed or the training data used to generate the profile became old and misrepresents the current behavior. Hence, we propose a strategy indicating when and how to update behavior profiles for services to increase the accuracy of the detection system.

We combine two conditions to determine the profile age and trigger an updating. The first one is a chronological condition for the profile to be updated periodically based on the service owner's policy. The updating process is automatically done at the tenant scope by Profiler using the service's log file. Secondly, when the detection system causes too many false alarms for a certain service before reaching the next updating cycle, the cloud scope (Figure 1) triggers the update at the tenant scope to use the most recent requests in the log file to model the service's profile.

During the updating period for a specific service, the service becomes unavailable for provisioning in cloud applications. Profiler checks the availability of enough requests in the service log file to replace the currently used training data since the last time this data has been collected. If there is enough data, the most recently logged requests are used to generate the profile. Otherwise all recent requests substitute the oldest requests in the training dataset. This method of aging has the effect of creating a moving time window for the profile data, so that the behavior is influenced strongly by the most recently observed behavior. Thus, the detection system adaptively learns services' behavior patterns; as services alter their behavior, their corresponding profiles change.

4.3. Detection System

Although any anomaly detection algorithm can be used to model normal behavior, the algorithm to be used for a significant amount of data in service clouds during real-time must be very efficient. Instead of ML algorithms, we develop an anomaly detection algorithm, which in our experiments leads to a high rate of accuracy and an acceptable performance overhead. In next sections, we explain in details our algorithms to detect suspicious requests and verify their owners' identities.

4.3.1. Detection Process

Inspired by the anomaly detection scoring system in [15], our algorithm does not detect data that represent an attack in the training period because it does not label such data as an anomaly; it assumes all data is normal. It is based on the premise that low probability values with a low likelihood of certain attributes are regarded as noise, and hence. Given that attack data is a minority of the training data then they would be recognized as attacks by the detection algorithm implicitly as low probability events and treated as anomalies at run time. "If attacks were prevalent and high probability events, then they are normal events, by definition" [18].

Upon each request, the application detector retrieves the sender's profile from the application SMDB to examine the impersonation attack. The deviation rate between the received request and the normal behavior is calculated by using the statistical model previously explained.

The total request anomaly score is a value that specifies the extent of the deviation of the received request from the expected values of anomaly scores specified by the sender's profile. It is a compound value derived from the attributes that have been computed in equations (1), (2), (4), and (6) and stored in the profile as follows:

$$AS = w_1\, AS_{class} + w_2\, AS_{size} + w_3\, AS_{meth} + w_4\, AS_{history} \quad (7)$$

The anomaly score for each attribute can be weighted to reflect its importance for the detection system using weight variables w1, w2, w3, and w4. This total score is compared to a threshold that can be chosen by the security administrator. A lower threshold means it is more likely that requests from impersonated services are detected with the disadvantage of an increasing number of false alarms. On the other hand, a higher threshold would allow for malicious requests to pass without detection.

ALGORITHM 1 performs the detection process. Line 24 indicates that the detector of the affected application sends the adequate information about the detected request to the cloud scope to take the proper action.

New services may not have a performance record in the cloud. If they are impersonated by attackers during the training phase; the detection system assumes their abnormal behaviors are normal behaviors. Therefore, once a new service arrives, the cloud must assume a certain amount of risk for some acceptable time if it provisions a new service into an application, given that the use of the new service can cause unrecoverable damage or too many false positives. Thus, the cloud needs to set constraints on the use of new services to determine their trustworthiness. In the proposed architecture (Figure 1), the SMDB of each tenant has an archive of the historical security incidents that have occurred on each service owned by this tenant. This history can be used by the cloud provider to evaluate the trustworthiness of the tenant in order to accept its new services in the provisioning process.

Each application has in its SMDB a white list of services that are within their training phase. The application detector skips checking this new service against ALGORITHM 1. Otherwise, the detector would generate an alert for each request instantiated by every new service which increases the false positive rate of the detection system.

```
ALGORITHM 1: IMPERSONATION_DETECTION
Input:
1.  Sender's profile (prof).
2.  Received request (req).
Output:
3.  Anomaly score (AS) for the received
request.
Begin:
4.  initialize soapSize, recipient, method,
class,μ,σ,AShist,ASmeth,ASclass,ASsize,
anomalyThreshold
5.  method =  req.getSoapAction()
```

```
6.  recipient = req.getRecipient()
7.  soapSize = req.getSoapSize()
8.  class = SMDB(recipient)
9.  if ( find(class in prof) {
10. ASclass= prof.get(ClassAS)          (1)
11. Ashist=prof.get(HistoryAS)          (6)
12. μ = get (Mean, prof)
13. σ = get (standardDeviation, prof)
14. ASsize=x^((soapSize-μ))/(y*σ)       (2)
15. if (find(recipient in prof)
16.     ASmeth=prof.get(MethodAS)       (4)
17. else
18.     ASmeth=max
19.     AS=w1*ASclass+w2*ASsize+w3*ASmeth+
        w4*AShist(7)
20.}
21.else {
22. AS= 0.4* 15 + 0.2* zero + 0.2* zero +
    0.2* zero }
23.if ( AS> anomalyThreshold)
24.  send alert to Cloud scope
25.else
26.  pass req
End
```

4.3.2. Verification Process

The impersonation detection algorithm gives an alert about a malicious request deviating from the normal behavior of the legitimate sender. To verify the identity of the suspicious sender, we use a verification technique. This technique confounds and confuses an impersonator into believing they have useful information when its identity is not verified. We integrate this technique with the service behavior profiling to secure the service in the cloud from being impersonated.

Upon registration, the verification technique requires each service to present a faked version of its response to the cloud SMDB. Then, the SMDB provisions this response to the applications that this service involved in. In addition, the cloud SMDB provisions a random code to each application for verification purposes. Hence, whenever a malicious request to the recipient service is detected by the application detector using ALGORITHM 1, the fake response is returned by the application detector in such a way as to appear completely legitimate and normal. The detector is the only party in the application has the privilege to access the code and fake responses reside in the application SMDB. Therefore, whenever a service has been impersonated, the attacker has no way to neither realize that the received response is faked nor use the application code to verify his identity.

When the sender is correctly identified as an impersonated service, it accepts the faked response since it does not have the ability to communicate with the application detector to distinguish between real and fake responses (Figure 5.A). As a result of this acceptance, the application detector terminates the application after a predefined period of time and delivers the information about the detected request as a security alert to the cloud scope for more investigation. The cloud scope then asks the tenant that owned the detected sender for its recent log file. If the detected request is not recorded in the original sender log file then the decision made by the application detector is correct otherwise it is a sign that the profile of the falsely detected needs to be updated.

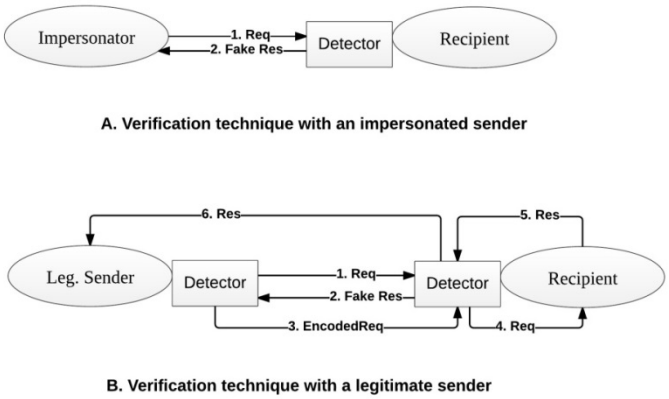


Figure 5: The verification process after detecting an impersonator

For the legitimate sender, the application detector would receive the fake response and readily recognize it and send the same request again along with the application random code to the same recipient (Figure 5.B). This encoded request is the application detector way to verify the sender's identity and issue a false positive alarm. The verification technique here serves two purposes: (1) validating whether the sender is authorized when abnormal request is detected, and (2) confusing the impersonator with useless information.

To incorporate the verification technique with the defined detection system, we need to change ALGORITHM 1 at line 24 to call ALGORITHM 2 instead of the cloud scope as follows:

VERIFICATION (req)

ALGORITHM 2 details how the detection system sends useless information as a response to malicious services. In line 5, the detector checks if the received request is a repeated request from a legitimate sender having the predefined code to verify its identity. In Line 7 the request is passed to the recipient after decoding it. Line 9 shows that the fake response for each service is extracted from the application SMDB. The fake response is passed in Line 10 to the sender.

```
ALGORITHM 2: VERIFICATION

Input:

1.   The malicious request (req).
Output:

2.   Detector's action against the malicious
  request.
Begin:

3.   sender = req.getSender()
4.   recipient = req.getRecipient()
5.   if (req.hasEncodedData) {
6.       Req.removeEncoded()
7.       pass req }
8.   else {
9.       fake_info= SMDB.getFake(recipient)
10.  pass fake_info  }
End
```

ALGORITHM 3 shows how the detector reacts when receiving a fake response by one of the application services. Line 4 declares that the detector uses a single code provisioned by the cloud scope to resend the legitimate requests. The detector deploys the verification technique either by sending or responding to the fake

response without getting the recipient service involved, which reduces the performance overhead and keeps the service black-boxed in the cloud without changing its internal code.

```
ALGORITHM 3: RECEIVING_FakeRes
Input:
1. fake response (res).
Output:
2. req: a new request holding the old request
and the code.
Begin:
3. if (res is fake){
4.   req.update(code)
5.   pass req }
6. else
7.   pass res
End
```

5. Results and Discussion

The developed framework is deployed in Azure cloud. JBoss server is used to deploy the web applications while Apache CXF Framework is used to develop web services and the STS. The application Detector and Auditor are implemented as CXF inbound and outbound interceptors attached to the STS in order to manipulate each SOAP request and response after being authenticated.

To generate behavior profiles for services in the Travel Agency case study, we deployed the framework 300 times to respond to client requests, randomly creating 300 service compositions (applications) from the available services. We used the resulting log files to generate profiles from while recording the generation time. The average time for each service to generate its behavior profile was about 3.2 seconds.

We have conducted three different tests on the designed case study: no attack check, anomaly check, and attack check. We have run the framework without attacks for legitimate services and anomalous services that are changed their behaviors without updating their profiles to measure the false positive rate. In addition, we designed several attacks to measure the detection accuracy rate of the detection algorithm. The designed tests have been deployed on randomly generated applications with a distinctive percentage of 94%. The anomaly threshold was ranging between 3 and 5. We used 0.4, 0.2, 0.2, and 0.2 to weight ASclass, ASsize, ASmeth, and AShistory respectively. The recipient's class anomaly score has slightly more weight since it primarily represents the business logic of the sender. The threshold is computed during the training phase and set to a value that would cause 10 false alarms per 100 applications when the system would receive the training data itself as input. A lower threshold would detect more attacks with the disadvantage of an increasing false positive rate. Hence, the threshold should be set to the lowest value possible provided the acceptable false positive rate. This decision depends on the type of traffic that is seen on the service clouds and SLAs between cloud providers and consumers deciding how many false alarms are considered acceptable [15].

For the test with no attacks, the framework responded with random applications to 100 client requests without involving any attacks. The services here behaved normally as in their profiles. The false positive rate was 1.7%. We have decreased the default threshold of 3 to 2.7 and the false positive rate slightly increases to 1.715%.

During the test with behavior anomalies, the same requests were tested on other random applications with some modifications on the legitimate services' behaviors. These legitimate services were enforced to dramatically change their behavior during runtime before the profile's update time. The changes included accessing new methods, sending large message sizes, and accessing rarely accessed classes. The false positive rate during this test was 28.75%. The false alarms have been exclusively caused by requests that called methods that had not been accessed during the training period. However, this result was generated without using the verification technique. With the verification technique, our detection system eliminated all false positives with a considerable performance overhead as depicted in Figure 6. The detection system needs to support variable security levels with different overhead for different applications. For example, if a banking application is running in the service cloud then there should be possibility to provide highly secure detection for this application by running the verification technique. In case of online games and voice over IP (VoIP) applications, performance and quality of service (QoS) could be of higher priority which requires disabling the verification technique.

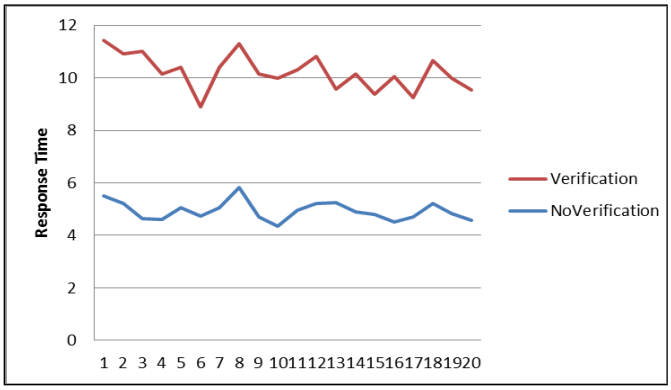


Figure6 : Verification technique overhead

After exploring the false positive rate, we attacked the case study in four distinct tests to measure accuracy. Each test had 100 client requests to initiate 100 different applications each has an impersonated service sending a malicious request to another service in the same application. The attacks mimicked the victim's behavior at different degrees. The victim services were meta-services since they have request behavior profiles while the tenants' services did not have profiles in this test. Any attack targeting a tenant's service would be detected immediately since the expected normal behavior for such services does not include sending requests to any service in the cloud. Table 2 lists the average anomaly scores for statistical properties of malicious requests and the detection accuracy of the impersonation attacks for all four cases. All tests were conducted twice; one with victim services having no historical impersonation attacks and one with a moderate history of attacks.

Table 2: The detection results for the attack tests

Test	Avg (ASclass)	Avg (ASsize)	Avg (ASmeth)	Avg (AShistory)	Accuracy
T1	6.6	1.1	1.9	0(1.5)	92%
T2	1.2	11.3	0.15	0(1.5)	67%(69%)
T3	1.2	11.3	6.2	0(1.5)	97%
T4	1.7	1.8	0.14	0(1.5)	0%

In the test T1, different impersonated services used the legitimate credentials of the victim services to send requests to services from classes that never been or rarely accessed by the victims before. The detection accuracy in this test was high (92%) because the Recipient's Class attribute has a highest weight in the anomaly detection equation among other attribute in the statistical model. However, any attribute can be differently weighted to reflect its security importance in the cloud.

T2 focused on revealing more information from the attacked services by embedding multiple queries in one request. 67% of the attacks have been correctly detected. On the other hand, the detection accuracy increased when the victim services had previously faced impersonation attacks. Furthermore, the detection accuracy reached 97% when the attacks targeted both the message size and the method at the same time as shown in T3 in Table 2. The last group of attacks fully mimicked the victim services which cannot be detected by the proposed detection system using the victim behavior profile. Thus, we need to extend the algorithm to cover such mimicry attacks in future work.

From the computational perspective, the presented algorithm accelerates the detection and profile update operations by implementing these operations in distinct scopes. The profile generation process is done before runtime at the tenant scope by the Profiler to reduce the computational load during the detection live time. The application scope holds the detection process during runtime by using the generated profiles. This results in shorter impersonator live time inside the cloud because it is limited to the active session lifetime at the application scope. As a counterpart, this approach needs a fast periodic update of the service's behavior profile and this introduces some overhead in the cloud network. Furthermore, the verification technique adds some overhead. We believe that updating the behavior profiles very frequently would decrease the need to trigger the verification technique which would eliminate much of the overhead.

Conclusions

The presented detection process depends on two parameters: the anomaly score computed by the detection algorithm using the behavior profiles of the services sending requests, and the fake response used by the verification technique to validate the malicious sender's identity.

In future work, some questions regarding the detection approach needs to be further investigated to improve the detection rate including: How much data is required in order to properly train each service? Should behavior profiles be based on individual service, or should services from the same class be grouped together?

Conflict of Interest

This manuscript is an extension of a conference paper that has been published in December 2016. All authors listed have contributed sufficiently to the project to be included as authors, and all those who are qualified to be authors are listed in the author byline. To the best of our knowledge, no conflict of interest, financial or other, exists.

References

[1] R. Gamble, S. AlQahtani, "Mitigating service impersonation attacks in clouds," Future Technologies Conference (FTC), San Francisco, CA, 2016.

[2] S. Alqahtani, R. Gamble, I. Ray, "Auditing requirements for implementing the Chinese Wall model in the service cloud," in World Congress on Services, San Jose, 2013.

[3] S. Alqahtani, R. Gamble, "Embedding a distributed auditing mechanism in the service cloud," in IEEE World Congress on Services,2014.

[4] M, Salem, S.J. Stolfo, "Modeling user search behavior for masquerade detection," in the the 14th International Conference on Recent Advances in Intrusion Detection, 2011.

[5] (CSA), Cloud Security Alliance, "The notorious nine: cloud computing top threats in 2013," Available from: http://www.cloudsecurityalliance.org/topthreats.

[6] S. Alqahtani, R. Gamble "DDoS attacks in service clouds", in 48th HICSS Conference, Hawaii, 2015.

[7] X. Zhao, G. Hu, Z. Wu Z, "Masquerade detection using support vector machines in the smart grid," in Computational Sciences and Optimization (CSO), 2014.

[8] W. Sha, Y. Zhu, M. Chen, T. Huang, "Statistical learning for anomaly detection in cloud server systems: A Multi-Order Markov Chain framework," in Cloud Computing, IEEE Transactions, 2015.

[9] S. Vanakumar, A. Kumar, S. Anandaraj, S. Gowtham, "Algorithms based on artificial neural networks for intrusion detection in heavy traffic computer networks," in the International Conference on Advancements in Information Technology With workshop of ICBMG 2011, Singapore.

[10] C.V. Raman, A. Negi, "A Hybrid method to intrusion detection systems using HMM," Springer Berlin Heidelberg, pp 389-396, 2005.

[11] S. Omar, A. Ngadi, H.H Jebur, "Machine learning techniques for anomaly detection: an overview," International Journal of Computer Applications (0975 – 8887) 79 (2), 2013.

[12] H.A, Kholidy, F. Baiardi, "CIDS: a framework for intrusion detection in cloud systems," in: Information Technology: New Generations (ITNG), 2012.

[13] D. Gao D, M.K. Reiter MK, D. Song, "Behavioral distance for intrusion detection," in the 8th international conference on Recent Advances in Intrusion Detection, Seattle, 2006.

[14] A.K. Ghosh, A. Schwartzbard, M.Schatz, "Learning program behavior profiles for intrusion detection," in the 1st conference on Intrusion Detection and Network Monitoring, Santa Clara, 1999.

[15] C. Krügel C, T. Toth, E. Kirda, "Service Specific Anomaly Detection for Network Intrusion Detection," in ACM symposium on Applied computing, 2002.

[16] H.A. Kholidy, F. Baiardi F, S. Hariri, "DDSGA: A Data-Driven Semi-Global alignment approach for detecting masquerade attacks," IEEE Transactions on Dependable and Secure Computing, **12**(2):164-178, 2015.

[17] S.J. Stolfo, M.B Salem, A.D Keromytis, "Fog Computing: Mitigating Insider Data Theft Attacks in the Cloud," in Security and Privacy Workshops (SPW), 2012.

[18] S.J. Stolfo, F. Apap,E. Eskin, K. Heller, S. Hershkop, A. Honig A, K. Svore, "A Comparative Evaluation of Two Algorithms for Windows Registry Anomaly Detection," Journal of Computer Security,659-693, 2005.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 460-468 (2017)

www.astesj.com

Special issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Image Segmentation Using Fuzzy Inference System on YCbCr Color Model

Alvaro Anzueto-Rios[*1,2], Jose Antonio Moreno-Cadenas[1], Felipe Gómez-Castañeda[1], Sergio Garduza-Gonzalez[1,2]

[1]*Solid-State Electronics Section (VLSI-Lab), Department of Electrical Engineering, CINVESTAV-IPN, 07360, Mexico*

[2]*Biomechanics Laboratory, Bionics Department, UPIITA-IPN, 07340, Mexico*

ARTICLE INFO

Article history:
Received : 05 April, 2017
Accepted : 02 May, 2017
Online: 24 May, 2017

Keywords :
Fuzzy Logic and Fuzzy Set
Fuzzy rules
Color Model
Segmentation
Chrominance and Luminance

ABSTRACT

This paper This paper presents a reliable method for image segmentation using a fuzzy inference system. The Fuzzy Membership function is applied on the YCbCr color space. Triangular membership functions are used in the input of the fuzzy system, Mamdani type fuzzy inference system is applied and for the output universe, singleton-type functions are used; to get the accurate output value, the Weighted Average Method (WAM) is applied. The YCbCr color space is used as feature space. One of the reasons being that it is standardized for the transmission and reception of digital video, (ITU-R Recommendation BT.601-5), and implemented by most of the sensors used in the acquisition of video. The fuzzy membership functions characterize the different membership levels between hue and Chroma from the YCbCr color model. The fuzzy inference system classifies data and generates regions of pixels with a homogeneous color level in the output images. The proposed method was also compared with another system segmentation using Euclidean distance applied to the RGB color space. The best results were obtained in the YCbCr color space. In such model, the changes of hue in presence of illumination variations are considered so that it has a better performance in the segmentation task. In addition, the processing time was lower in this color space.

1. Introduction

Color perception begins with a chromatic light source -which is electromagnetic radiation with wavelengths between 400 and 700 nm approximately. Part of the radiation, which is reflected on the surface of objects, is a scene and the resulting reflected light reaching the human eye, generates the feeling of color perception [1-4].

Brightness is the luminance or relative darkness of the color, hue is the color reflected or transmitted through an object and, saturation (also called chroma) represents the contribution of white color existing in proportion to the hue, or more precisely, the color purity [2]. This is the basis of the YCbCr color space.

The color segmentation in the computer vision field is important for pattern recognition and image processing tasks purposes. Segmentation is the process of partitioning an image into regions of interest (sets of pixels), with homogeneous features; namely, they can share one or more characteristics, such as shape, color, texture, lighting, etc.

The color of the shapes within an image can be used to simplify the identification and extraction of the regions of interest. They can be recognized as separate objects.

Image segmentation methods have recently been proposed, many of them based on Neural Systems and Fuzzy Logic. A segmentation system based on neural networks has been proposed by Ganesan et al. [5], they performed an unsupervised competitive neural network based in Kohonen's self-organizing maps (SOM). They used color satellite images. The images are segmented in RGB and HSV color space. The main problem of his work is the high computational cost because to obtain the value of a pixel, with the structure presented, 48 connections are necessary, increasing with this the number of operations. In addition, the weights of each node should be initialized before training data and

*Corresponding Author: Alvaro Anzueto-Rios, Ave. IPN # 2580. Col. Barrio La Laguna Ticoman, Del. Gustavo A. Madero, 07340, Mexico City.
Email: aanzuetor@ipn.mx

the performance of the training depends on the initial data. If the initial data is not adequate, the process will have to be repeated.

The YCbCr color model has been used for skin detection and, in combination with the fuzzy model, has presented great performance for image segmentation [6, 7]. The Cb and Cr components determine the skin tonality. Tonalities are represented by histograms. Fuzzy inference models to determine the pixels representing the possible regions of skin within an image can be applied to them.

When looking for adaptability in fuzzy systems to model information in an unsupervised way, different proposals have been presented for the image segmentation using the classical algorithm of Fuzzy C-Means (FCM) [8-10]. G. Liu et al. [8] introduced a dissimilarity function by combining region-based and pixel-based distance function, in order to enhance the relationship between pixels which have similar local characteristics. This idea is retaken in this work, since, for the distribution of fuzzy sets, local information is considered.

Liu and Wang [11] developed a method of fuzzy color recognition and segmentation of robot vision scene based on the combination of fuzzy color quantization and edge extraction of HSV color space. This work was supported in part by the same idea; the triangular membership functions in the input of the system and the Mamdani-type inference are used. In contrast with the work mentioned above, we used YCbCr color space, fuzzy singletons are used as output function and the output value is obtained by the Weighted Average Method (WAM) [12, 13].

One of the reasons for using the YCbCr color space is that it is standardized for the transmission and reception of digital video, (ITU-R Recommendation BT.601-5 [14]), therefore, most of the sensors used in the acquisition of video implement it. We used a device that has as main output signals, the YCbCr color space, eliminating the need to perform the transformation to another color space and improving the response time.

The fuzzy singleton function and WAM are used to simplify the mathematical calculation of the output value for the fuzzy inference system.

Distribution of input membership functions for classification of hue is made on the Cb and Cr components of YCbCr color space (see Figure 1).

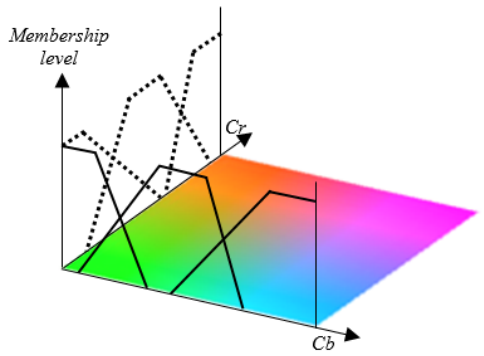


Figure 1: Distribution of membership functions in the space of chroma Cb and Cr

This paper is organized as follows. In Section 2, we give a brief introduction to the YCbCr color space. The basis of fuzzy logic is discussed in section 2. Section 4 presents the fuzzy rules used by the system. The application of segmentation system is presented in Section 5. The discussion of the results is made in Section 6. Section 7 presents the conclusions.

2. YCbCr Color Model

It is a nonlinear encoding of the RGB color model. The YCbCr model components determine the luminance and chrominance. "Y" is the contribution of light in the pixel; "Cb and Cr" are the blue and red chrominance components, respectively. The combination of Chroma values are different colors (color coordinate system), but the luminance value "Y" determines the brilliance of color, specifically, for low values of "Y" the colors will be dark; otherwise, the colors will look very bright with a hue close to white. The three cases for the variation of illumination "Y" are shown in Figure 2, as it can be seen, the color coordinate system changes its hue.

The YCbCr space is commonly used in vision-based hardware systems and digital video transmission. It is defined for standard-definition television and digital transmission purposes by the ITU-R BT.601. One of the main features of this model is that the brightness component "Y" is independent from the chroma components; because of this, this color space can be used to solve illumination variations in non-controlled environments.

The below mathematical expression is given following the recommendations of the ITU-R BT.601 standard (chapter 3.5.2 [14]) for a Cr and Cb re-normalized system.

Equation (1), describes the RGB color space conversion to the YCbCr space. Where matrix A represents the re-normalized coefficients of the horizontal vectors Y, Cb and Cr respectively. C is the matrix that re-measures the data, transforming the 220 and 225 levels, to 256 (8-bit digital representation). Before applying (1), the lowest value for Y, Cb and Cr is 16 and the highest value for Y is 235 and Cb, Cr is 240. After applying (1), the lowest value is 0 and the highest value is 255, for the three components. This mathematical procedure is similar to the function RGB2YCBCR included in the MATLAB's digital image processing toolbox. This value transformation, 256 values, will be used in this paper.

$$\begin{bmatrix} Y \\ Cb \\ Cr \end{bmatrix} = \begin{bmatrix} 16 \\ 128 \\ 128 \end{bmatrix} + \left([A^t \times C]^t \times \begin{bmatrix} R \\ G \\ B \end{bmatrix} \right) \quad (1)$$

Where:

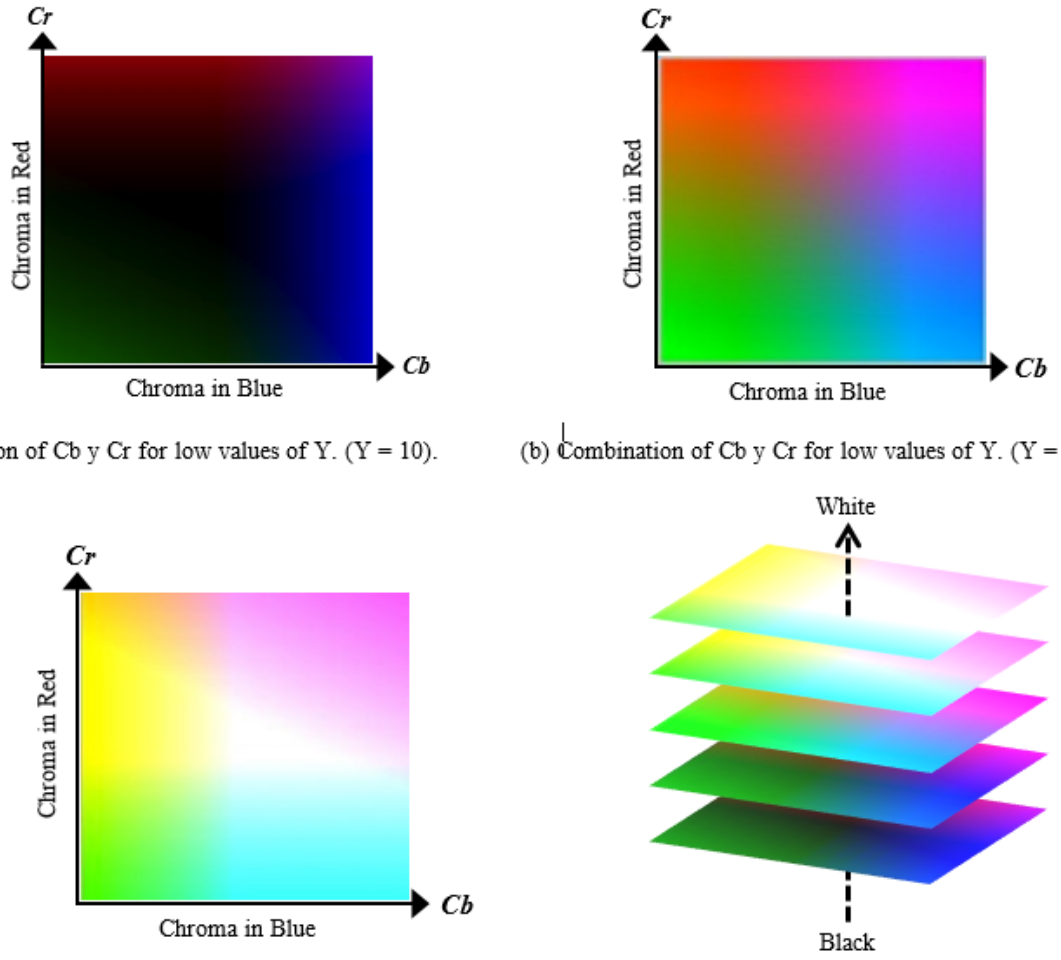
$$A = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ -0.169 & -0.331 & 0.500 \\ 0.500 & -0.419 & -0.081 \end{bmatrix} \quad (2)$$

$$C = \begin{bmatrix} {}^{220}/_{256} & 0 & 0 \\ 0 & {}^{225}/_{256} & 0 \\ 0 & 0 & {}^{225}/_{256} \end{bmatrix} \quad (3)$$

3. Fuzzy Logic

In the YCbCr color space, the transition between a color and another one does not have a defined borderline; namely, it changes gradually from one hue to another. Due to the fact that fuzzy systems work with uncertain data, this system is very appropriate.

A fuzzy set is associated with a linguistic value that is defined by a word, an adjective or a linguistic label. In the fuzzy set, the membership function can take values in the interval (0, 1), making



(a) Combination of Cb y Cr for low values of Y. (Y = 10).

(b) Combination of Cb y Cr for low values of Y. (Y = 128)

(c) Combination of Cb y Cr for low values of Y. (Y = 240).

(d) Colors through "Y".

Figure 2: Color coordinate system obtained by combination of Chroma. Value of low illumination (a), value of medium illumination (b), value of high illumination (c), colors through "Y"

a gradual transition between zero and one, and not discreetly one such as that in classical sets.

The fuzzy set in a universe of discourse can be defined as shown in (4).

$$color = \{(x, \mu_{color}(x) | x \in U)\} \quad (4)$$

Where $\mu_{color}(x)$ is the membership function of the variable x, U is the universe of discourse. The closer the color of the set membership to value 1, the greater the variable x belonging to the set color will be.

$$\mu(x) \begin{cases} 0 & if\ x \leq a \\ x-a/b-c & if\ a \leq x \\ 1 & if\ b \leq x < c \\ d-x/d-c & if\ c \leq x < d \\ 0 & if\ x > d \end{cases} \quad (5)$$

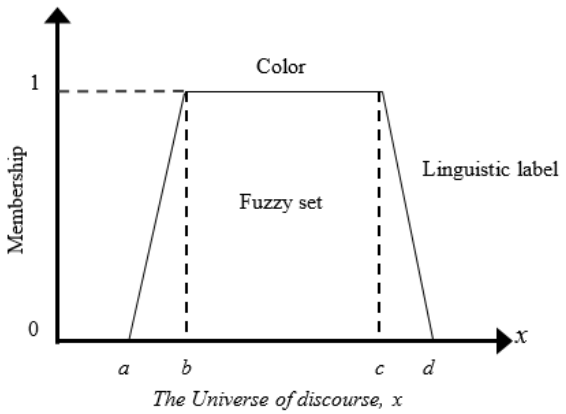


Figure 3: Definition of Trapezoidal Fuzzy Set

Equation (5) defined a trapezoidal fuzzy set type (see Figure 3) which describes the different membership levels [12]. The parameters (a, b, c, d) determined the slopes in the boundary lines of the fuzzy sets. As a fuzzy set, it can be inferred that the maximum level of membership is equal to one, the slope of a line determines the different membership levels and the non-membership is equal to zero. For this system, the overlap in the

fuzzy sets describes the inaccurate transition between the hues, as detailed in Figure 4.

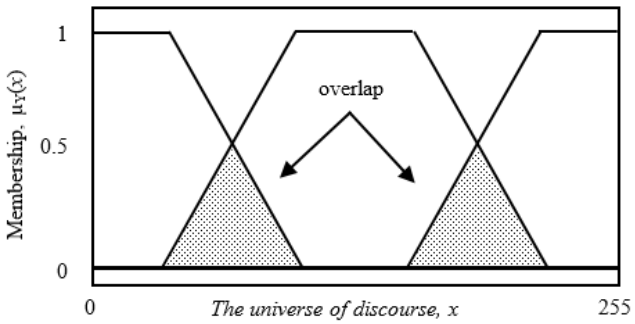


Figure 4: Overlap in fuzzy sets, inaccurate transition of the hue

4. Fuzzy Rules

A fuzzy system inference uses the expert's knowledge to generate a knowledge base, which will give the system the ability to make decisions. This knowledge base is modeled by creating a rule set, which generates the system output; in this case, it is the membership of pixels within an image above a set or another.

As mentioned, the component "Y" represents the contribution of light existing in a color. In Figure 5, a graphical diagram of the inference of nine fuzzy rules when the colors are considered dark can be seen. That is, when luminance "Y" has a low value. (The first nine rules in Table 1).

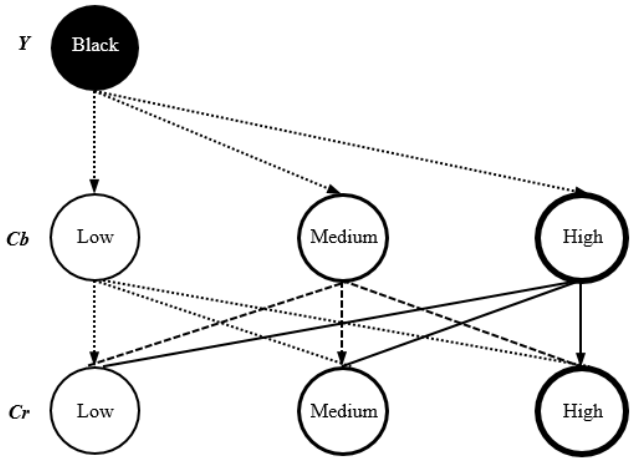


Figure 5: Graphic description of the fuzzy rules

In Table 1, we can see twenty-seven diffuse rules, which are all combinations when we consider three fuzzy sets for each component of the color space YCbCr.

To obtain a crisp output or output image (segmented regions), we need a defuzzification process. The inputs of the defuzzification process are the values of fuzzy membership functions Y, Cb, and Cr, and the output of the defuzzification process is a single number (or the value of one pixel). Many defuzzification techniques have been proposed in literature [12, 13]. The most commonly used method is the WAM (also named centroid). In WAM, the defuzzifier determines the weighted mean (or the center of gravity) of the fuzzy membership functions implicated in each of the knowledge base rules (Table 1). It is given by the algebraic expression in (6).

$$output = \frac{\sum_{i=0}^{n} \mu(A)_i . A_i}{\sum_{i=0}^{n} \mu(A)_i} \quad (6)$$

Where $\sum$ denotes an algebraic sum and A represents the different colors expected at the output, see Table 1.

Table 1: Combination that generate the knowledge base of fuzzy rules.

No. Rule	Description			Output
	Y	Cb	Cr	
1	***Black***	***Low***	***Low***	***Black***
2	***Black***	***Low***	***Medium***	***Black***
3	***Black***	***Low***	***High***	***Black***
4	***Black***	***Medium***	***Low***	***Black***
5	***Black***	***Medium***	***Medium***	***Black***
6	***Black***	***Medium***	***High***	***Black***
7	***Black***	***High***	***Low***	***Black***
8	***Black***	***High***	***Medium***	***Black***
9	***Black***	***High***	***High***	***Black***
10	Color	Low	Low	Green
11	Color	Low	Medium	Yellow
12	Color	Low	High	Red
13	Color	Medium	Low	Cyan
14	Color	Medium	Medium	Gray
15	Color	Medium	High	Crimson
16	Color	High	Low	Blue
17	Color	High	Medium	Violet
18	Color	High	High	Magenta
19	White	Low	Low	White
20	White	Low	Medium	White
21	White	Low	High	White
22	White	Medium	Low	White
23	White	Medium	Medium	White
24	White	Medium	High	White
25	White	High	Low	White
26	White	High	Medium	White
27	White	High	High	White

5. Segmentation of Hue Using Fuzzy Logic

The YCbCr color space has a color chart consisting of the combination of the Chroma belonging to the red and blue (Cb and Cr) and different levels of luminance (Y), see Figure 2.

For the universe set "luminance", we consider three linguistic labels, "Black" which represents the low contribution of light, "Color" where it is considered the optimal lighting to perceive the tone of a color, and "White" representing a high contribution of light, see Figure 6.

In order to reach the perception of color, the presence of luminance without reaching saturation will be necessary.

Different combinations of the components "Cb and Cr" typify the color. By taking this into account, the fuzzy sets are generated in their combination representing the different colors that the system can classify.

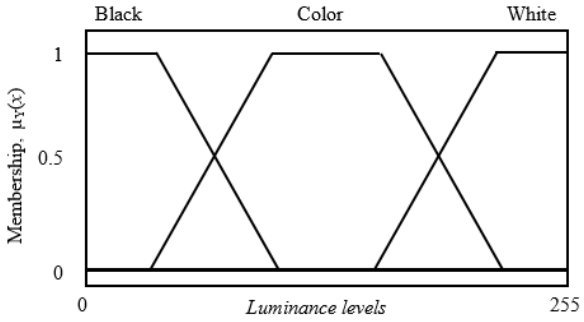
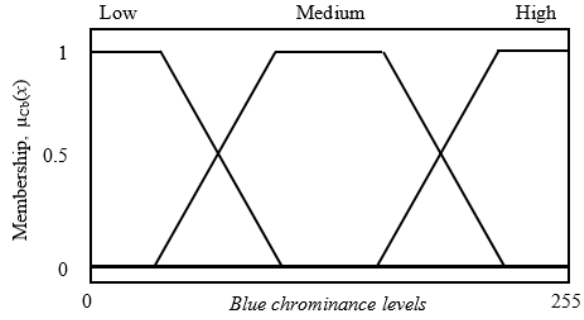
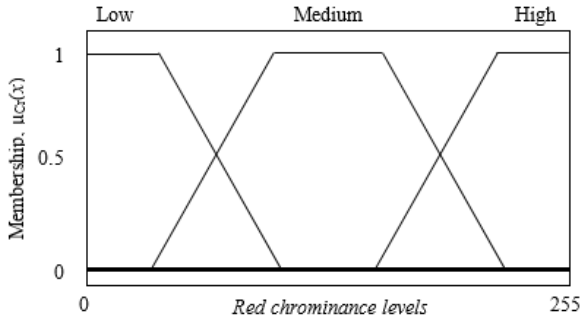


Figure 6: Fuzzy sets for the universe set. "luminance"; Y

Three fuzzy sets are defined, each with their linguistic label respectively for each Chroma component. "Low" when the Chroma saturation in blue or red is low, "medium" when the Chroma saturation is medium, and "high" when the Chroma saturation is high, as detailed in Figure 7.



(a) Fuzzy sets for blue chrominance, Cb



(b) Fuzzy sets for red chrominance, Cr

Figure 7: Fuzzy sets of chrominance. (a) Blue chrominance *Cb*, (b) Red chrominance *Cr*

The values values (a, b, c, d), which are parameters of (5), represent the form of all the fuzzy sets in a universe set (see Figure 8). By varying these parameters, different slopes in the boundary lines of the fuzzy sets are generated, modifying the zone that represents uncertain data.

Also, it is possible to find a particular hue value by modifying these parameters. For example, if we are looking for red color and its different hue values (the red color region is in the upper left corner in the color palette as shown in Figure 8), it is necessary to expand the region covered by the "Low" set in "Cb" and the "High" set in "Cr", which means that the value of the parameter "a" in "Cb" should be increased and the value of the parameter "d" in "Cr" should be decreased. The region of interest is then amplified as shown in Figure 9.

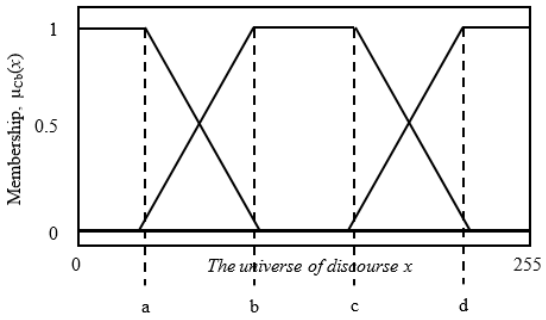


Figure 8: Parameters for controlling the distribution of the fuzzy sets (a, b, c, d)

The numerical values of (a, b, c, d) parameters, which are used in the fuzzy sets for red segmentation, are shown in Table 2.

Table 2: Value of the parameters (a, b, c, d) for the fuzzy sets of Y, Cb, Cr

	Y	Cb	Cr
a	30	95	85
b	60	105	95
c	200	195	145
d	235	205	155

One of the images tested with the proposed fuzzy inference system is shown in Figure 10. It contains an object with a red hue. In order to explain the different steps that the fuzzy logic segmentation system follows, a pixel will be taken from the processed image.

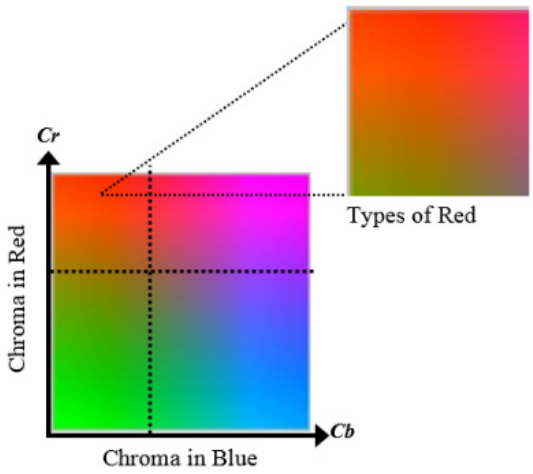


Figure 9: Color model YCbCr. Search of red in the color pallet of the chrominance (Cb and Cr)

This selected pixel corresponds to a reddish object, and it has the following values for Y=143, Cb=94 and Cr=188. The membership values obtained by processing the pixel described above are shown in Table 3.

The membership value shown in Table 3 are introduced in (4) to obtain the numerical values of the system output. Variables named "S" represent the output fuzzy sets (fuzzy Sigleton's), each of them are located at the expected output value; for example, S10 represents the red color, suggesting that the pixel in the output

image will have R = 255, G = 0, B = 0 as a numerical value *as long as* the numerical value obtained in (3) is similar to S10. Table 4 lists the values of all the fuzzy singleton's.

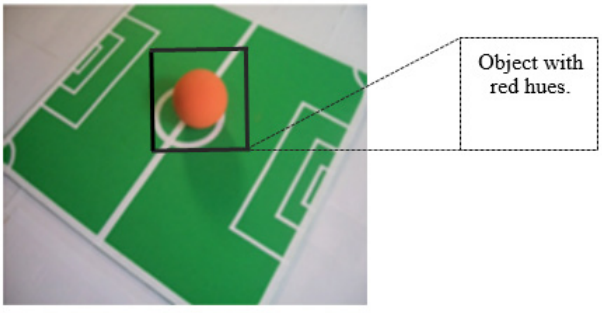
Table 3: Membership Value for Each Fuzz Set

Fuzzy Sets μ	Membership Value
μ_{BlackY}	0
μ_{ColorY}	1
μ_{WhiteY}	0
μ_{LowCb}	1
$\mu_{MediumCb}$	0
μ_{HighCb}	0
μ_{LowCr}	0
$\mu_{MediumCr}$	1
μ_{HighCr}	0

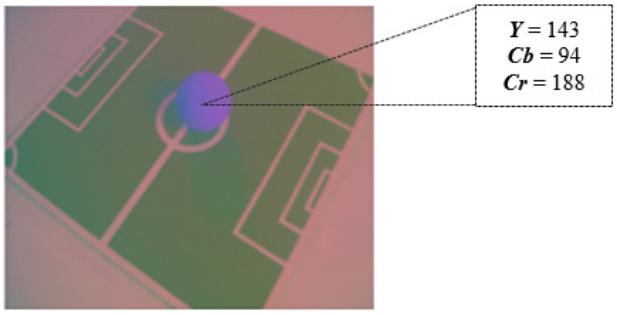
Table 4: Fuzzy Singletons Value (Output Functions)

	Output Numerical Value
S1	10
S2	33
S3	56
S4	79
S5	102
S6	125
S7	148
S8	171
S9	194
S10	217
S11	240

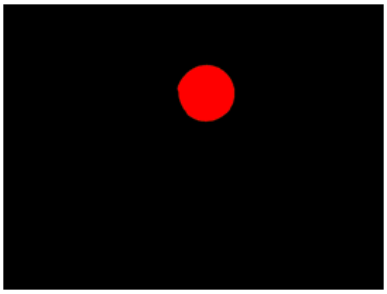
To form the output image, each pixel is is processed in the same way. The used method for image processing is shown in Figure 10. At first, the original image is in RGB color space (Figure 10 (a)). Then, the image is transformed into YCbCr color space (Figure 10 (b)) and finally, the fuzzy inference system is applied to the result image (Figure 10 (c)). It is worth noting here that the output singleton sets are represented again in RGB color space to be displayed on screen.



(a) RGB color space image showing the segmentation area



(b) YCbCr color space image. Choosing a point to be processed



(c)Output image obtained by the segmentation system

Figure 10: Graphic sequence of the fuzzy segmentation system (a), (b), (c)

6. Experimental Evaluation

To evaluate the effectiveness of the segmentation system by fuzzy logic, the segmentation of the green hue in the original image of Figure 11(a) was proposed. This photo was taken in the laboratory, in a closed room, under semi-controlled lighting conditions and the changes in illumination inside the photo can be noticed. The image contains a green area that simulates the conditions of a (small) soccer field, a red ball and white lines that define the green area.

This particular photo was chosen because our segmentation system will be a part of the vision system of a robotic prototype that serves as a football player.

The segmentation system was designed considering only the red, green and blue contributions in the RGB color space model, in order to define the color of an object and matching the results from the opposite system. The classification in the RGB segmentation was made by obtaining the Euclidian distance between each one of the image's pixels and the mean of the pixels (red, green, blue) that defines the green color of the image.

The code was developed based on Matlab programming language in its version R2010a (see Appendix A), on a computer with an Intel Core2 Duo T6400 2 GHz. Images were obtained with a Kodak Easy Share CD14 with a resolution of 1280x960, JPG format.

The execution time of the segmentation system when using RGB color space was of 1 min.15 sec. For YCbCr color space it was 55 sec.

The difference between the RGB and YCbCr execution times is mainly due to the fact that, in the RGB color space, each component contributes to obtain a color hue so that the three components are an input of the fuzzy system; in the YCbCr, the component "Y" determined, as a first step, if the pixel has low lighting (black), high lighting (white) or if the lighting is good

enough to perceive the hue of a color. The pixel Chroma components to determine the membership level are discussed only in the latter case.

Figure 11(c) shows the result of the segmentation of the RGB color space.

For segmentation in YCbCr, we transform the original image into this color space, due to fact that the image is in RGB color space, Figure 11(a), the rgb2ycbcr function was used for this purpose, which is part of the "image processing" toolbox of Matlab, (see Figure 11(b)). In Figure 11(d) the result of the segmentation can be observed.

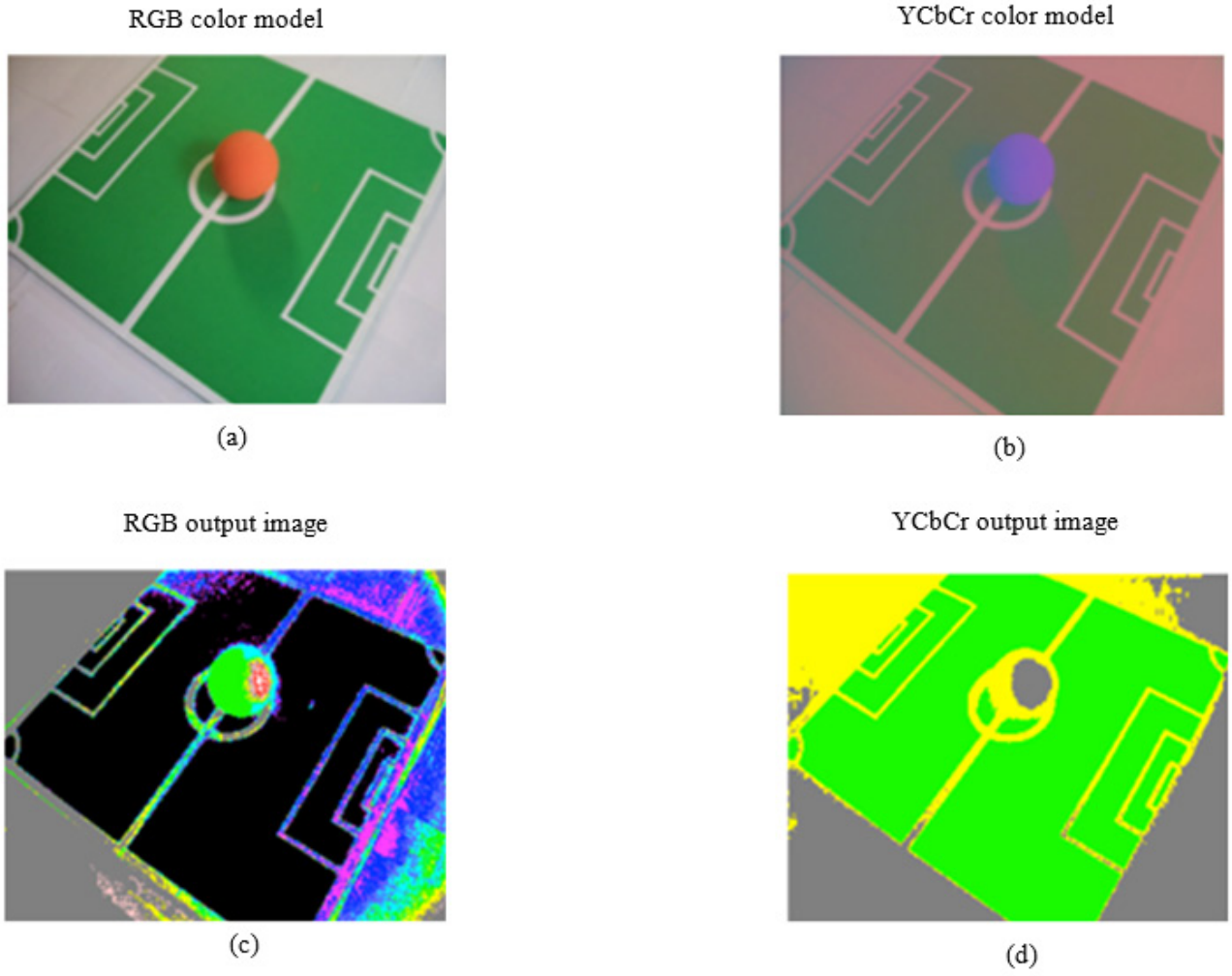


Figure 11: Segmentation process in the RGB and YCbCr color spaces, (a) Original image in RGB color space, (b) Original image transformation into YCbCr color space, (c) Segmentation of green color in RGB, (d) Segmentation of green color in YCbCr

The change in hue due to the variation of lighting was considered in the YCbCr color space, so that, when the images in Figure 11(c) and Figure 11(d) were compared, as expected, we observed a better performance of the segmentation system in YCbCr.

The variation of lighting is not considered for the RGB, so that, the pixels in the top center of the image Figure 11(c), were classified as non-green hue.

In Figure 11(a), three main regions can be seen, one of them is the green region (region of interest), and two more regions that are covered with yellow and gray pixels. The last two regions are the result of the membership levels to the fuzzy sets of the remaining pixels.

The values for the parameters (a, b, c, d) of the fuzzy sets which are used for the segmentation of green color in YCbCr are shown in Table 5.

Table 5: Value of the Parameters (a, b, c, d) for the Fuzzy Sets of Y, Cb, Cr

	Y	Cb	Cr
a	30	95	85
b	60	105	95
c	200	195	215
d	235	205	225

Another important point to mention is that the equations to conform the fuzzy sets are relatively simple and only require 4 parameters to generate the fuzzy sets that cover a universe set. We have taken advantage of this simplicity and it was implemented in a segmentation system in an FPGA.

A test showing how to generate an output image with three hues by adding the segmentation of green and red color from an input image is shown in Figure 12. The image obtained in Figure 12(d), may be used for classification and recognition of objects contained therein.

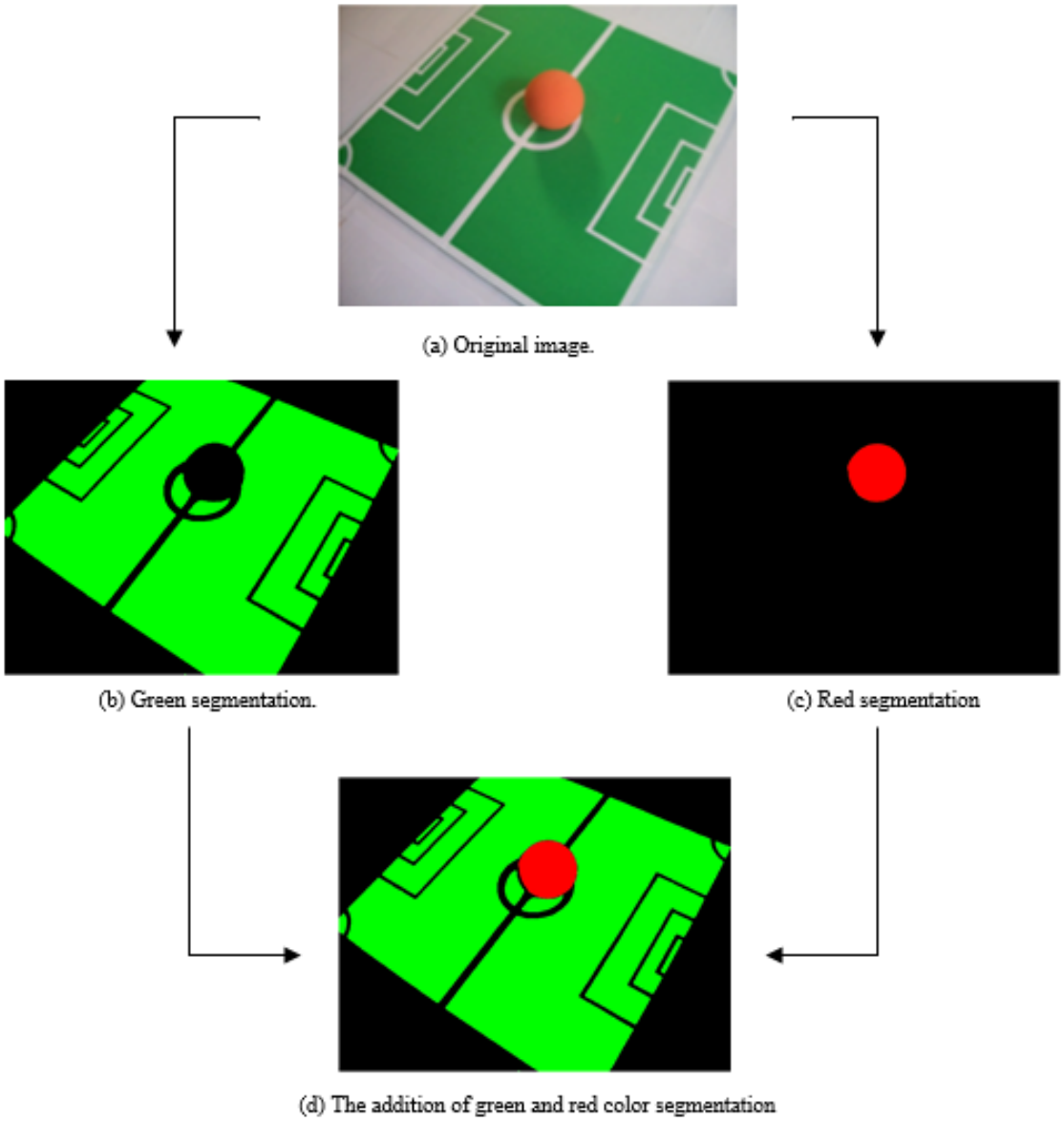


Figure 12: Segmentation process, (a) Original image, (b) Green segmentation, (c) Red segmentation, (d) Addition of b and c

7. Conclusion and Future Works

This paper proposes an application of *segmentation of regions in images* using fuzzy logic. It exploits the ability of fuzzy sets for modeling accurate data to interpret the transition from one hue to another by using the blue and red chromas in the YCbCr color space.

Trapezoidal sets are proposed for mapping of the chroma to membership levels and is considered the ITU-R Recommendation BT.601-5 in order that the system can be applied to most sensors for the acquisition of digital video.

The combination of the YCbCr color space with the fuzzy inference system demonstrated high adaptability to model different color hues, so different color hues can be segmented into an image.

By the simplicity of the equations, both to generate the fuzzy sets and the calculation of the value of the output of the system, it is possible to have a system of low computational cost and high speed to obtain a response.

MATLAB programing language was used in its version R2010A to develop the system and was applied in two color

spaces, RGB and YCbCr, execution times were 1 min. 15 secs and 55 secs, respectively.

Due to the simplicity of the formulas that are used to develop the fuzzy inference system, the implementation in hardware is feasible. In our laboratory, we have implemented the proposed system, using the Spartan 3E development board which contains an FPGA xc3s500e from XILINX Company. The proposed segmentation system has been applied in the object tracking and orientation of robots. Different colors and shapes have been presented to a robot and the execution times obtained allow working with images with continuous sequence.

Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgment

Alvaro Anzueto-Rios thanks CONACYT for the scholarship to finish his doctoral studies. He also thanks the Institute Polytechnic National (IPN Mexico) and UPIITA for the support to complete this paper.

References

[1] R. C. Gonzalez, R. E. Woods, Digital Image Processing, Pearson-Prentice Hall, 2008.

[2] J. C. Russ, The Image Processing Handbook, CRC Press Taylor and Francis Group, 2011.

[3] W. Burger, M. J. Burge, Principles of Digital Image Processing, Springer-Verlag, 2013.

[4] O. Marquez, Practical Image and Video Processing using MATLAB, Wiley-IEEE Press, 2011.

[5] Ganesan, K. B. Shaik, B. S. Sathish, V. Kalist, "Neural network based SOM for multispectral image segmentation in RGB and HSV color space", International Conference on Circuit, Power and Computing Technologies (ICCPCT 2015), Nagercoil India 2015.

[6] F. Z. Chelali, N. Cherabit, A. Dejeradi, "Face Recognition System Using Skin Detection in RGB and YCbCr color space", 2nd World Symposium on Web Application and Networking (WSWAN 2015), Sousse Tunisia, 2015.

[7] M. Banerjee, S. Muhuri, "A Robust Face Detection Algorithm Based on Fuzzy Color Histogram Model" International Journal of Graphics & Image Processing (IJGIP), **4**(4), 296-303, 2014.

[8] G. Liu, Y. Zhang, A. Wang, "Incorporating adaptive local information into fuzzy clustering for image segmentation" IEEE Transaction on Image processing, **24**(11), 3990-4000, 2015.

[9] T. Chaira, "A novel intuitionistic fuzz c means clustering algorithm and its application to medical images" Applied Soft Computing, **11**, 1711-1717, 2011.

[10] E. Rajaby, S.M. Ahadi, H. Aghaeinia, "Robust color image segmentation using fuzzy C-means with weighted hue and intensity", Digital Signal Processing, **51**, 170-183, 2016.

[11] C. Liu, L. Wang, "Fuzzy color recognition and segmentation of robot vision scene", 8th International Congress on Image and Signal Processing (CISP), Shenyang China, 2015.

[12] J. T. Ross, Fuzzy Logic with Engineering Applications, John Wiley & Sons, 2010.

[13] S. N. Sivanandam, S. Sumathi, S. N. Deepa, Introduction to Fuzzy Logic using Matlab, Springer-Verlag, 2007.

[14] ITU-R Recommendation BT.601-5: Studio encoding parameters of digital television for standard 4:3 and wide-screen 16:9 aspect ratios.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 469-478 (2017)

www.astesj.com

Special issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Segmented and Detailed Visualization of Anatomical Structures based on Augmented Reality for Health Education and Knowledge Discovery

Isabel Cristina Siqueira da Silva*, Gerson Klein, Denise Munchen Brandão

Informatics Faculty, Exacts and Technology School, UniRitter Laureate International Universities, 91849-440, Brazil

ARTICLE INFO

Article history:
Received: 05 April, 2017
Accepted: 02 May, 2017
Online: 24 May, 2017

Keywords:
Augmented reality
Interaction design
Health education
Anatomy visualization
Knowledge Discovery

ABSTRACT

The evolution of technology has changed the face of education, especially when combined with appropriate pedagogical bases. This combination has created innovation opportunities in order to add quality to teaching through new perspectives for traditional methods applied in the classroom. In the Health field, particularly, augmented reality and interaction design techniques can assist the teacher in the exposition of theoretical concepts and/or concepts that need of training at specific medical procedures. Besides, visualization and interaction with Health data, from different sources and in different formats, helps to identify hidden patterns or anomalies, increases the flexibility in the search for certain values, allows the comparison of different units to obtain relative difference in quantities, provides human interaction in real time, etc. At this point, it is noted that the use of interactive visualization techniques such as augmented reality and virtual can collaborate with the process of knowledge discovery in medical and biomedical databases. This work discuss aspects related to the use of augmented reality and interaction design as a tool for teaching anatomy and knowledge discovery, with the proposition of an case study based on mobile application that can display targeted anatomical parts in high resolution and with detail of its parts.

1. Introduction

This paper is an extension of work originally presented in International Conference of the Chilean Computer Science Society (SCCC) [1].

The computer is a facilitator of teaching for different sciences, since the static and two-dimensional image printed in the books can be added movement and a new dimension, allowing a better representation of scientific concepts and favoring meaningful learning. In this perspective, interactive digital graphics applications can aid in the teaching-learning process of concepts, contents and chemical skills as well as the knowledge discovery process, due to the fact of stimulating self-learning [2] [3]. In the Health area for example, we have a great concentration of studies looking for viable computational solutions for the most diverse types of information management problems. The analysis and cross-checking of information are able to provide evidence to understand the development of diseases, assist health students and professionals to make decisions to contain the epidemics and improve the care for patients, avoiding extra costs with hospitalizations.

The individuals that born after 1960 having grown up surrounded by all sorts of new technologies, and their minds have undergone cognitive modifications brought about by new technologies and digital media, which have propelled a new range of needs and preferences on the part of younger generations, particularly in the area of learning. Prensky [4] says that today's students and professionals represent the first generation to grow with the technology of interactive graphics applications. They spent their whole lives surrounded by emergent technologies such as augmented reality, ubiquitous learning (u-learning), mobile learning (m-learning), serious games and learning analytics for improving the satisfaction and experiences of the users in enriched multimodal learning environments [5]. As a result of this ubiquitous digital environment and the large amount of interaction with it, students today think and process information fundamentally different from their predecessors.

*Corresponding Author: Isabel Cristina Siqueira da Silva, UniRitter Laureate International Universities, 91849-440, Brazil, +55 51 32303333
Email: isabel.siqueira@gmail.com

In the same way that the brains are programmed to acquire the ability to write and read, the constant use of digital technologies reprograms the brain to think from the logic of these devices. This means that a large part of the problem related to the learning of digital natives resides in the fact that it is very laborious to retrain the brain to construct thought from the linearity of reading and writing, and that this thought was already modified in function of Television and the advent of computer graphics [6]. According to Mazzarotto and Battaiola [7], three aspects are identified as possible implications of the use of such applications in the learning process: increasing learner motivation and interest in learning activity, increasing retention of subjects and improving skills Reasoning and high complexity thinking. The increase in motivation and interest would be the result mainly of the elements of curiosity, fantasy and control offered by the interactive graphic application. In addition to functionality and usability, users seek pleasure in the activity, and this search is key to choosing to purchase and use a product.

Visualization and interaction techniques can aid the analysis of data, since they amplify the process of insight on related data, transforming them and/or amplifying them in the form of images [8] [9]. When an image is analyzed, a cognitive process begins, where perceptual mechanisms are activated that aim to identify patterns and segment elements [10]. The correct mapping of data to images is crucial, since one can discard relevant information or exceed the amount of irrelevant information [11] [12]. Thus, it is noted that the graphical representation should limit the amount of information that the user receives, while keeping him/her aware of the total information space and reducing the cognitive effort [13].

In this context, Augmented Reality (AR) allow the viewing of virtual objects combined with scenes from real environments through computers equipped with a video camera, or even from portable devices, such as tablets and cell phones [14]. There are three basic ideas that make up AR: immersion, interaction and involvement. Immersion is linked with the feeling of being inside environments. Interaction is the computer's ability to detect user input and instantly modify the virtual world and actions on it. Involvement is connected with the degree of motivation for engaging a person with certain activity.

The health sciences have benefited from the power of these tools [15]. By employing virtual technology to teach anatomy, the challenges are different from traditional teaching with real anatomical parts and/or two-dimensional atlas (2D), since virtual anatomy requires three-dimensional (3D) visualizations with high resolution and interaction with them in real time [16] [17]. Students can use AR to build new knowledge based on interactions with virtual objects that bring underlying data to life [18]. Virtual objects can be manipulated in a direct way, which allows the user to interact with them in a free and natural way, as in the real world. This system not only helps students learn anatomical details, but also provides the enhancement of 3D structures that cannot be replicated by a 2D atlas. It is speculated that working directly with AR can help students learn the complex structure of anatomy better and faster than using traditional methods (cadaver dissection and the use of paper atlas).

Thus, AR can amplify the associations between anatomy-physiology and movement through its use in mobile devices, allowing greater flexibility of interaction with virtual anatomical models. However, there are not many software solutions involving AR and visualization of anatomy, human and animal, aimed at mobile devices and, among those that exist (see section 2), it is not possible to visualize and interact with detailed and segmented anatomical structures. In addition to these questions, it was identified that the response time of such tools, coupled with the quality of 3D models, does not favor their use in the classroom as an aid to teaching activities by teachers.

Based on these questions, this work aims to propose the use of AR to visualize segmented anatomical structures, in high resolution, detailing their parts textually next to the image and using cut planes, in mobile devices. Interaction questions with segmented anatomical pieces are also studied, in order to facilitate their study and learning with students in the areas of health sciences. Such proposals were validated in two case studies involving the anatomy of human bone structures and animal organs. The first case study has already been validated alongside potential users through usability testing while cognitive specialists evaluated the second case study.

This article is organized as follows. Section 2 presents works related to the topic discussed in this article. Section 3 describes the proposed tool while section 4 presents details of the validation experiment with users and results. Finally, section 5 presents the final considerations.

2. Related Work

Different studies and applications are being developed with AR focused on teaching concepts related to Health.

Ten years ago, Gomes et al. [19] were already investigating the application of AR for educational purposes in the area of Medicine. Using free tools (Linux and ArToolkit[1]), they presented examples of overlapping of organs on markers in real scenes, allowing the visualization and manipulation of the respective organ with differentiated aspects. To do so, they used sound allied to the image, making the learning more dynamic and realistic, although they did not explore mobile technology, mainly due to the technological restrictions of the time.

Safi et al. [20] indicate that, in relation to the monitoring of specific diseases, such as skin cancer, AR can help to monitor and identify problems early in order to initiate treatment or prevent disease progression. High-risk individuals may present with disturbing and abnormal skin lesions, and, in this context, AR can help the professional to quickly compare the results and the change of these lesions over a period of time based on the data (texts and images) of the patient. Minimally invasive surgeries may also be possible with AR, allowing surgeons to access information more quickly without having to shift their attention to the patient being treated. Vital statistics of this can be kept in their field of view while operating, without taking risks of distraction.

Coles et al. [21] present a virtual environment for training femoral palpation and needle insertion, the opening steps of many interventional radiology procedures. A novel augmented reality simulation has been developed that allows the trainees to feel a virtual patient using their own hands. The authors mimic the palpation haptics effect by two off-the-shelf force feedback devices have been linked together to provide a hybrid device that gives five degrees of force feedback. This is combined with a

[1] https://www.hitl.washington.edu/artoolkit

custom built hydraulic interface to provide a pulse like tactile effect. The simulation has undergone a face and content validation study in which positive results were obtained. During validation tests, experts in the field rated the device as closely reproducing the fine tactile cues of a real life palpation.

Reitzib [22] also agree with these observations, claiming that working directly with AR can help future professionals learn the complex structure of anatomy better and faster than using traditional methods (cadaver dissection and the use of paper atlas). Patient 3D images, constructed from 2D images of computed tomography scan and/or magnetic resonance imaging can be visualized overlapping their bodies with accurate visual information and real-time data retrieval.

Bacca et al. [23] contribute to existing knowledge in AR in educational settings by providing the current state of research in this topic. This research has identified relevant aspects that need further research in order to identify the benefits of this technology to improve the learning processes. In this context, the authors highlight some aspects as learning gains, motivation, interaction, collaboration and student engagement and positive attitudes. However, very few systems have considered the special needs of students in AR and there is a potential field for further research.

Jenny [24] employed Augment[2] software to defend her thesis involving the benefits of AR employed in the fields of Dentistry and Medicine. In dentistry, the AR was used in courses and in practical classes with the purpose of demonstrating dental models for the students and enabling them to compare their sculpture with reference models. It was observed that AR facilitated communication between teachers and students, clarifying the images and texts found in the textbooks.

In relation to the use of AR in Veterinary Medicine, there are still few published papers about it.

Parkes et al. [25] present a mixed reality simulator was developed to complement clinical training. Two PHANToM premium haptic devices were positioned either side of a modified toy cat. Virtual models of the chest and some abdominal contents were superimposed on the physical model. Seven veterinarians set the haptic properties of the virtual models; values were adjusted while the simulation was being palpated until the representation was satisfactory. Feedback from the veterinarians was encouraging suggesting that the simulator has a potential role in student training.

Lee et al. [26] developed an intravenous injection simulator, based on AR, to train veterinary students in the performance of canine venous puncture. The 3D models used were developed from images derived from computed tomography of beagle dogs, converted to stereo lithography format. Another application of AR in Veterinary Medicine is reported in [27] where the University of Liverpool School of Veterinary Science is using such technology to allow students to analyze the internal anatomy of animals from smartphones. The team developed a 3D image of an equine heart for the study in question.

In terms of applications, we have the application based on AR proposed by Chien et al. [15], which allows visualization of anatomical details in 3D, allowing users to disassemble, assemble and manipulate such details for educational purposes. In the article, the technologies employed are not informed and the application was not made available for mobile devices. On the other hand, the BARETA (Bangor Augmented Reality Education Tool for Anatomy) [28] system, in turn, combines AR with 3D models from rapid prototyping technology, aimed at teaching anatomy. Likewise, the project LearnAR[3] proposes an online tool, based on AR, aimed at teaching in health disciplines. Anatomy4D[4] is an application that was developed for mobile devices using AR concepts

The works discussed in this section do not allow the visualization of detailed and segmented anatomical as also parts individualized and intuitive interaction with them, a fact that motivates the researches of this work.

As conclusion of our study of related works, we observed that, although they address the use of AR in educational settings, they mostly focus on the specific applications for a human or animal body part. The visualization of detailed and segmented anatomical as also parts individualized and intuitive interaction with them is not well explored.

3. Segmentation and Visualization of Human and Animal Anatomical Structures with Augmented Reality

Understanding the complexity of human and animal anatomy something that requires a lot of time and study. It is noted that the teaching methods used are mostly based on 2D atlas studies in addition to the traditional use of cadavers or mannequin models, through which students learn the forms and the location of structures of the human/animal body, correlating them with their functions [16]. However, the use of corpses has been aggravated in terms of availability and storage due to the restricted time of access to the material [17]. Besides, the use of 2D atlases compromise the understanding of the real 3D structure of anatomical pieces while multiple needle and wire punctures can cause rapid degradation of mannequin models [21].

In the previous section, there is a series of papers discussing the use of RA as a possible solution for such questions, but there is not found a variety of solutions that allow the combination of AR to the segmented visualization of 3D anatomical parts, with intuitive interaction and support for use on mobile devices. Based on these observations, we propose a tool for collaborate with the

[2] http://www.augment.com/ technology/

[3] http://Learn.org

[4] http://site.daqri.com/project/anatomy-4

teaching area in Health and knowledge discovery that implements some functionality not observed in the related works studied.

3.1 Requirements Gathering

For the requirements gathering regarding the present research, were interviewed teachers who work in undergraduation disciplines of the area of Health involving the teaching of concepts of human and animal anatomy together with the accomplishment of the contextual analysis of the same acting in the classroom with their students. The institution where the case study was carried out has different resources used in the classroom, such as tablets and computers, which have software and virtual atlas applications of anatomy, embryology and histology, imaging resources - such as RX, tomography, Resonance, human models in resin that replace the use of corpses to study human anatomy, anatomy and histology books, among others. Active methodology techniques such as body painting and body projection are also used for visualization and learning dynamics.

As a pedagogical proposal for the teaching of anatomy in undergraduation Health courses, teachers point out that the use of AR may help to identify anatomical structures, amplifying the associations between anatomy, physiology and movement through their use in mobile devices, in order to allowing greater flexibility interaction with the anatomical models. Such benefits are amplified with the possibility of segmented and detailed visualization of anatomical parts, in high resolution, and with intuitive interaction in real time. Thus, the proposed visualization has such characteristics.

3.2 Segmented and Detailed Visualization

From the requirements gathering, the following features were defined for the proposed visualization:

- Visualization and interaction with the use of AR aimed at assisting the teaching of human and animal anatomy, based on real case studies, facilitating their adaptation and use in the classroom according to the demands and disciplines that involve the study of the anatomy;
- Support for use in mobile devices, in order to provide users with flexibility of use (students and teachers);
- Use of 3D models with adequate degree of definition and realism;
- Trace based on markers;
- Intuitive and easy-to-understand interface;
- Possibility of segmented and detailed visualization of a human and/or animal anatomical component.

Figure 1. 3D model of the humerus bone designed over real scene.

For reading 2D markings in real scenes through devices that have camera implementation, the framework Vuforia[5] was used. Vuforia accesses the database remotely, not overloading the execution of the application. The Vuforia was integrated with the game engine Unity[6] in order to generate the AR functionalities, connecting virtual images with real scenes and allowing interaction with them in real time. The 3D models were obtained from public repositories and/or builded in Blender[7] software from 2D pictures.

Four main views were made available:

- Simple anterior view;
- Simple posterior view;
- Detailed anterior view; and
- Detailed posterior view.

In the "previous view", the front part of the projected structure is shown (Figure 1) while in the "anterior view", the back part is seen. The difference of both to the detailed ones is that, in the latter, we have the description of the main bone accidents that compose a structure as showed in Figures 2 and 3. Clicking on any of the viewing options in the menu will trigger the camera of the mobile device that will capture the actual scene. As in all applications involving AR, after finding the marker in the scene, the 3D image of the structure, chosen from the menu, is incorporated into the scene. In the Figures 1 and 3, we employed a fiducial marker to facilitate the mapping of the abstract to the concrete experience for inexperienced user with the use of AR.

Besides these four views, a fifth view is being developed and improved: the cut plane view. Figure 4 shows two examples of cut plane views of segmented head structures.

[5] http://www.vuforia.com

[6] https://unity.com

[7] https://www.blender.org

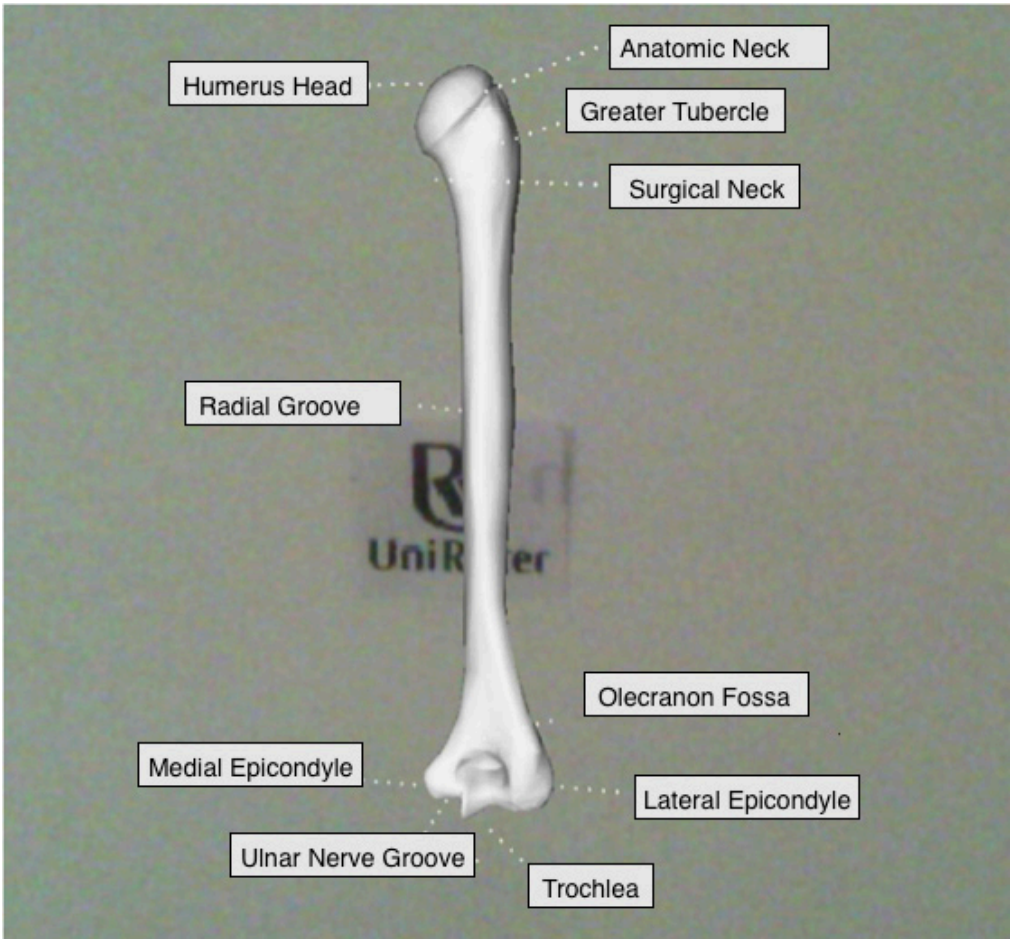


Figure 2. Image generated by AR with 3D model of the humerus bone with details.

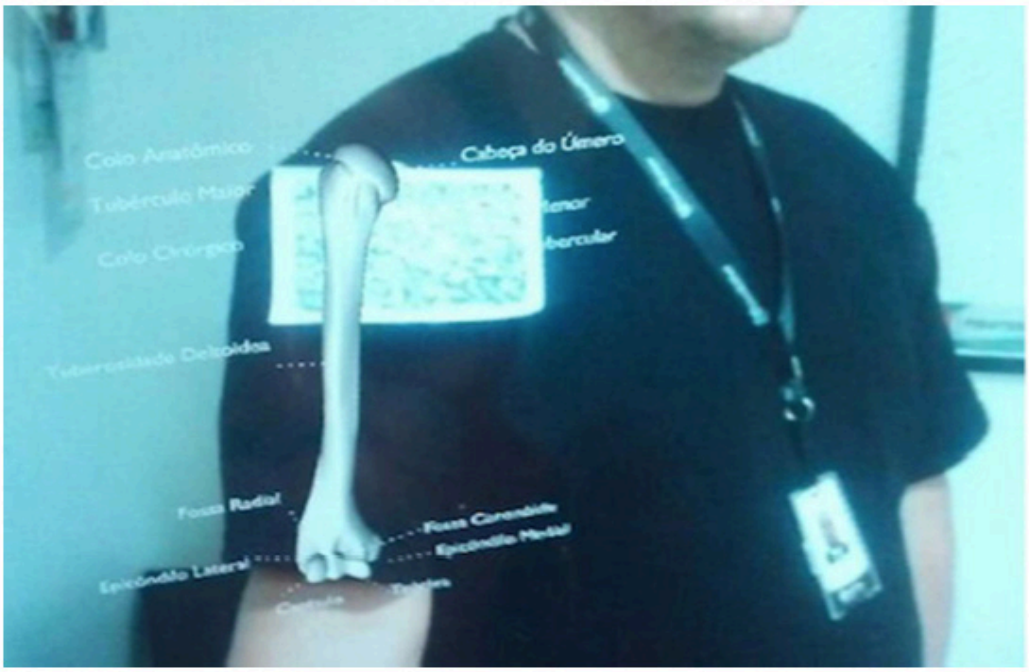
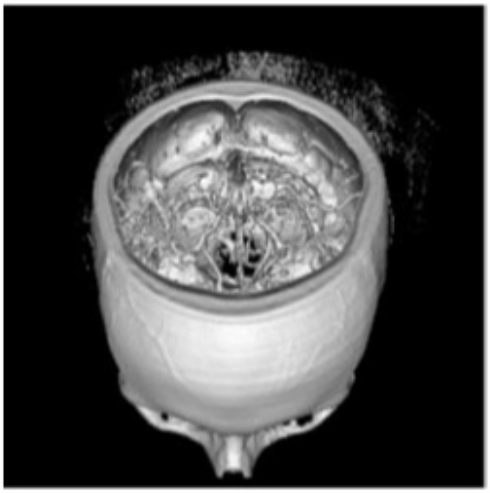
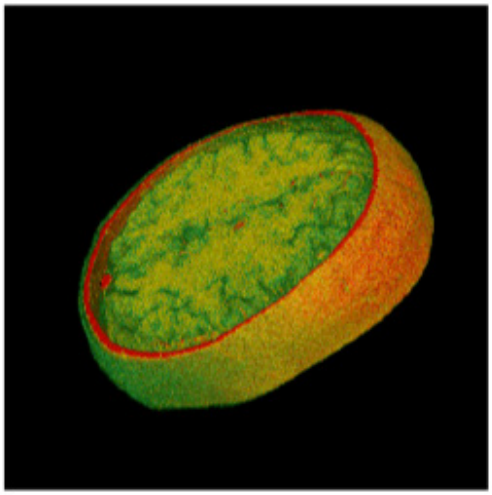


Image generated by RA with 3D model of the humerus bone on real scene with detailing.





The interaction with the anatomic objects in the scene occurs by three basics geometrics transformations: translation, scale and rotation.

4. Validation and Results

The proposed objectives was validated by two evaluation experiment based on the human humerus bone (Figures 1, 2 and 3) and the cat liver, being the last one validated considering the 3D model texture and transparency in order to show the vascular structures (Figures 5 e 6).

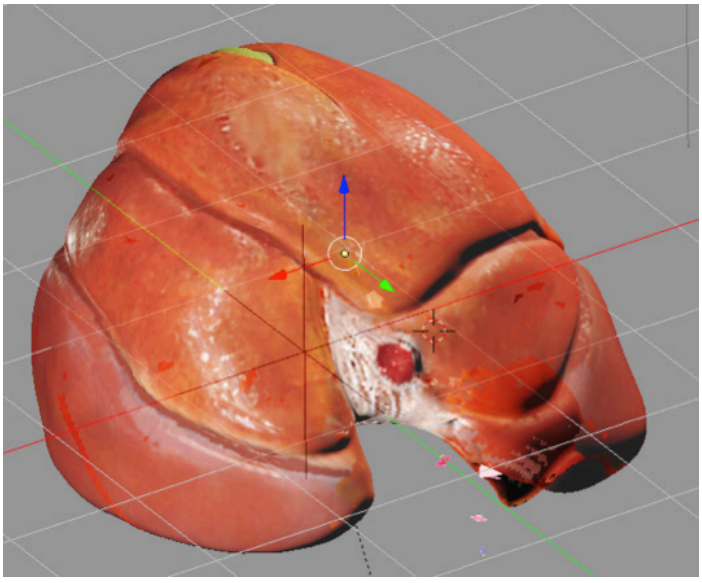
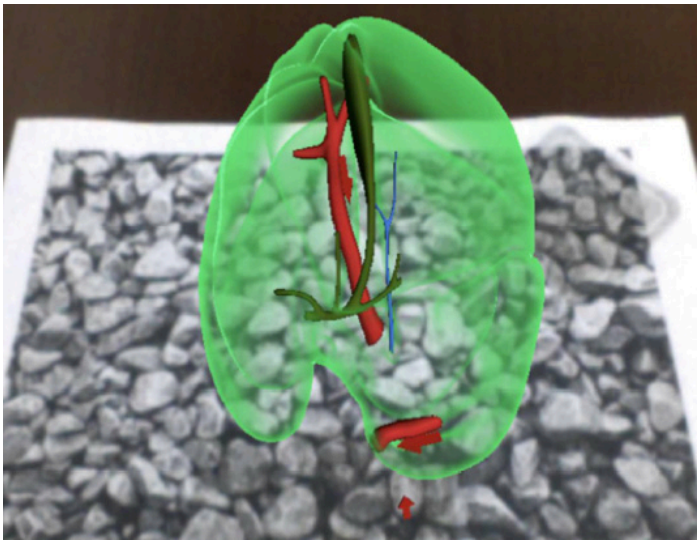


Figure 5. Texturized 3D model of a cat liver.

This experiment was based on tests of usability and has as hypothesis the following statement "The proposed application assists the teaching and the study of the human and/or animal anatomy".



In addition to the hypothesis, the evaluation criteria were:

• Efficiency: Does the application allow users to better visualize 3D anatomical structures on mobile devices with the visualization options of structures in a segmented and detailed way?

4.1 Case Study 1: Human Humerus Bone

4.1.1 Subjects

The application validation process had the participation of 5 (five) teachers and 23 (twenty-three) students of a Physiotherapy undergraduate course being 9 (nine) men and 14 (fourteen) women

between the ages of 22) And 45 (forty-five) years. Teachers have more than 20 (twenty) years of contact with the Physical Therapy area, and only 1 (one) of these had previously had experience with devices involving AR.

4.1.2 Procedure

The application was distributed to the subjects along with a manual explaining its operation. After testing the application free form, the group was submitted to a questionnaire composed of 5 (five) affirmations based on the likert scale composed by the assertions: "fully agree", "partially agree", "undecided", "partially disagree" and "fully disagree ". Users should choose only one alternative in each statement, representing the one that best matches their experience of using the application.

The questionnaire had the following statements:

1) The proposed application has an intuitive interface and simple handling and understanding.

2) The virtual visualization a segmented structure with individualized interaction helps in the understanding of its peculiarities in relation to the whole.

3) The possibility of AR visualization of a segmented part of the human or animal anatomy in mobile applications allows greater flexibility of user interaction during anatomy study.

4) The quality of the 3D rendered structure via AR provides a more immersive user experience when interacting with the virtual scene combined with the actual scene.

5) The visualization provided by the use of AR can aid the teaching and study of the anatomy more efficiently than traditional methods.

4.1.3 Results

From the validation experiment performed and discussed in the previous section, we set out to analyze the results and their discussion. In this sense, Figure 7 presents graphs with the distributions of the responses from each to each of the five statements described in the previous section.

As can be seen in Figure 7, the assertions ("fully agree" + "partially agree") make up 100% of the answers given to questions 1, 2, 3 and 4. Thus, it can be assumed that the proposed application is According to the evaluation criteria effectiveness and usability, presenting friendly interface and fulfilling its purpose by offering an intuitive interaction with the 3D representation of the humerus bone in this case study specifically.

Only affirmative 5 did not achieve full agreement - 30% of subjects were undecided on this question, probably because they needed more time to use the application in the classroom (the vast majority of participants in this experiment had never had contact with applications involving AR above).

Interquartile range (IQR) was also analyzed in order to summarize the data independently of the number of observations (Figure 8). It can be seen from the graph of Fig. 8 that the variability of the responses was almost nil with an IQR equal to 1.0 in statements 1, 2, 3 and 4 and with the value of 1.75 in affirmative 5, which confirms the convergence of users' opinions.

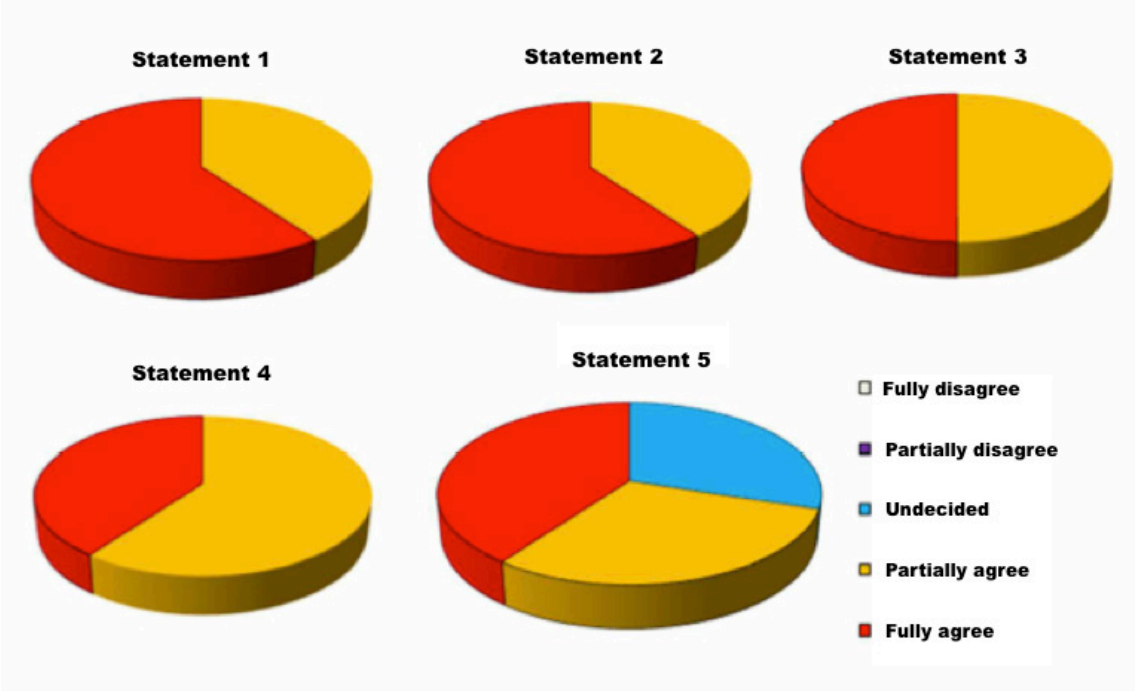
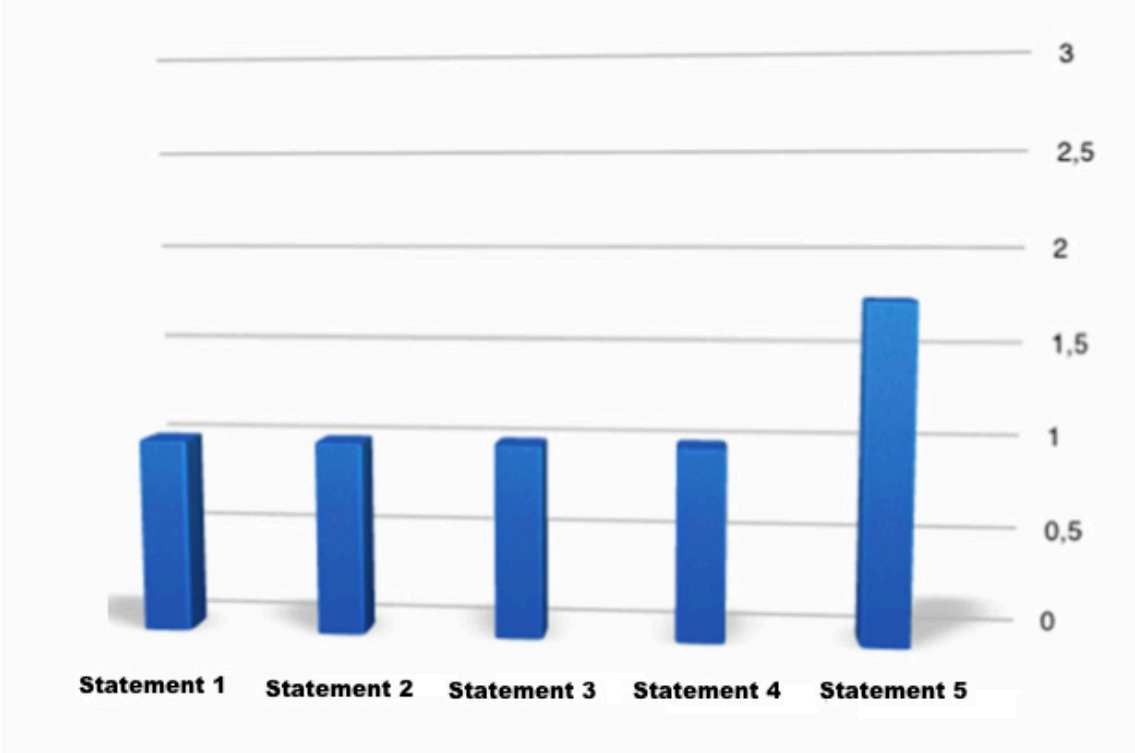


Figure 7. Graph with distribution of the users' answers to the questions proposed during the evaluation experiment for the case study 1.



Figre 8. Graph with the IQR values of the users' answers to the questions proposed during the evaluation experiment for the case study 1.

4.2 Case Study 2: Cat Liver

4.2.1 Subjects

4.2.2 Procedure

The subjects were taken to a computer lab to be introduced to the AR based visualizations developed and interact with them. Initially, the functionality of the visualization was explained, presenting screens and form of interaction.

Ad the first validation experiment, subjects were then asked to start the visualization and, at the end, respond to a post-test questionnaire composed of eleven questions and based on the likert scale with the following assertions: “fully agree", "partially agree" "undecided," "partially disagree" and "fully disagree". Subjects should choose only one alternative for each question in order to express their experience of using the application.

The questionnaire had the following statements:

1) The use of AR motivates me to participate in animal anatomy classes.

2) Virtual and interactive 3D models help the study and understanding of anatomical models.

3) The ability to view and interact with 3D models over real scenes via mobile devices helps learning as it is not limited to using a PC/notebook or paper atlas.

4) The interaction with the 3D model is intuitive.

5) The time elapsed between application initialization and the initial display of the 3D model is satisfactory.

6) The response time of the application to the interaction with it is satisfactory.

7) I realize that the use of AR can be applied to other content related to my undergraduate course.

8) I recommend and approve the use of AR in my undergraduate course.

4.2.3 Results

As can be seen in Figures 9 and 10 (IQR), most subjects agreed ("fully agree" and "partially agree") with the statements (around 80).

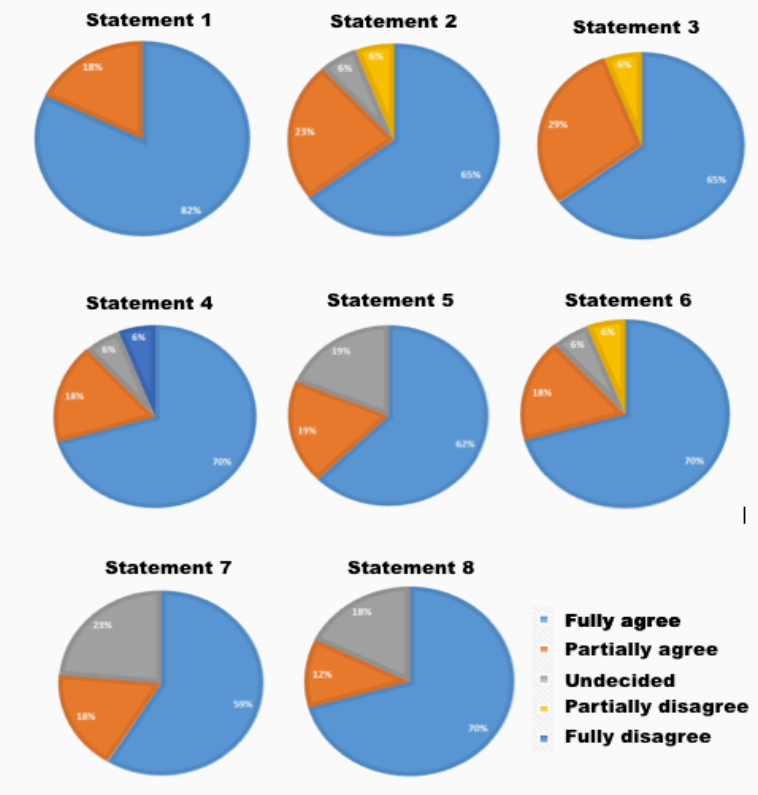
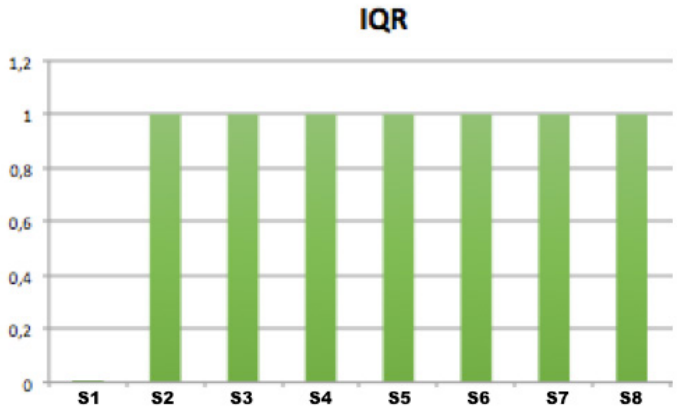


Figure 9. Graph with distribution of the users' answers to the questions proposed during the evaluation experiment for the case study 2.



Figre 10. Graph with the IQR values of the users' answers to the questions proposed during the evaluation experiment for the case study 2.

Based on these results, it is observed that, in addition to the proposed criteria, the hypothesis raised was also confirmed in this first moment, although it is felt the need to vary the sample of subjects and extend the validation time of the application in the classroom. In general, it was noted that the users approved the application and that it can be used in the teaching of human and animal anatomy in order to assist students' learning in an interactive and visual way.

However, new experiments, with different anatomical pieces, still have to be performed.

This work presented a study about the investigation of an application of AR aimed at assisting the teaching of anatomy having, as case studies, the visualization of the human humerus

bone and a cat's liver. During its development, it can be noted that the AR is a technology that brings a differential to methods commonly employed in teaching anatomy in undergraduate courses in the area of Health, since it integrates virtual and real objects in order to contribute to a more realistic and interactive visualization of structures in the human body as well as knowledge discovery.

As pointed out by the users, the usability combined with the effectiveness of the use of the AR visualization contributed to a positive evaluation of the same and signaling the will to use it continuously in the classroom. The interface was considered easy to handle and understood by users and the application made it possible to 3D visualization with adequate degree of realism, segmented form and individual interaction, meeting the main requirements raised by the interviewees in the initial phase of this study.

Thus, it is envisaged that AR can bring a new approach to teaching anatomy, both for students and teachers, through its integration into mobile devices. This is a valid alternative to traditional methods of studying anatomy employing corpses, 2D atlases on paper and mannequin models.

In future work, it is intended to extend the application to holographic visualizations besides the construction of simulations of interaction with anatomical parts in virtual environments with support to the use of RV glasses.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

We thank all the teachers and students who participated in this work, from the requirements gathering to the validation experiment.

References

[1] C. S. da Silva, G. Klein, D. M. Brandao, “Segmented and detailed visualization of anatomical structures for health education” in 35th International Conference of the Chilean Computer Science Society (SCCC), Chile, 2016.

[2] G. Nascimento, N. Robaina, B. Moreira, P. T. Mourão and A. Salgado. “A construção do corpo humano a partir de elementos químicos,” in Brazilian Symposium on Games and Digital Entertainment, Rio de Janeiro, RJ, 2009.

[3] R. Sousa, F.S.C. Moita, and A. Carvalho, Tecnologias digitais na educação, Editora da Universidade Federal da Paraíba, 2011.

[4] M. Prensky, Computer games and learning: Digital game-based learning, Mass: MIT Press, 2005.

[5] L. Johnson, S. Adams Becker, V. Estrada, A. Freeman, A, Horizon report 2014: Higher education edition, The New Media Consortium, 2014.

[6] T. G. Mendes, “Games e educação: diretrizes do projeto para jogos digitais voltados à aprendizagem”, Master Thesis, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2012.

[7] M. Mazzarotto, A. L. Battaiola, “Uma visão experimental dos jogos de computador na educação: a relação entre motivação e melhora do raciocínio no processo de aprendizagem” in Brazilian Symposium on Games and Digital Entertainment, Rio de Janeiro, 2009.

[8] S. K. Card, J. D. Mackinlay, B. Shneiderman, Readings in information visu- alization: using vision to think, USA: Morgan Kaufmann Publishers Inc., 1999.

[9] R. Spence, Information Visualization: design for interaction, Prentice Hall, 2007.

[10] M. Ward, G. Grinstein, D. Keim, Interactive Data Visualization: Foundations, CRC Press, 2015.

[11] C. Ware, Information Visualization: Perception for Design, Morgan Kaufmann, 2004.

[12] C. Ware, Visual Thinking: for Design, Morgan Kaufmann, 2008.

[13] R. Mazza, Introduction to Information Visualization, Springer, 2009.

[14] R. Azuma, “A survey of augmented reality” Presence: Teleoperators and Virtual Environments, 6(4), 55–385, 1997.

[15] C. H. Chien, C. H. Chien, T. S. Jeng, “An Interactive Augmented Reality System for Learning Anatomy Structure” in Proceedings of International MultiConference of Engineers and Computer Scientists, Hong Kong, 2010.

[16] R. F. Braz, Método didático aplicado ao ensino da anatomia humana “Anuário da produção acadêmica docente”, 3(4), 303–310, 2009.

[17] E. P. D. Silva, K. L. Santos, P. D. S. Barros, T. N. Silva, J. L. Souza, A. F. d. S. Mariano and M. B. A. Palma, “Utilização de cadáveres no ensino de Anatomia Humana: refletindo nossas práticas e buscando soluções”, in XIII Jornada de Ensino, Pesquisa e Extensãoo, Recife, 2013.

[18] E. Zhu, A. Hadadgar, I. Masiello, N. Zary, “Augmented reality in healthcare education: an integrative review,” PeerJ, 2(e469), 2014.

[19] W. Gomes and C. Kirner. “Desenvolvimento de Aplicações Educacionais na Medicina com Realidade Aumentada,” Bazar: Software e Conhecimento Livre, 1(1), 13-20, 2006.

[20] A. Safi, V. Castaneda, T. Lasser, N. Navab “Skin Lesions Classification with Optical Spectroscopy”, Springer, 2010.

[21] T. R. Coles, N.W. John, D. A. Gould, D. G. Caldwell, “Integrating Haptics with Augmented Reality in a Femoral Palpation and Needle Insertion Training Simulation” IEEE Transactions On Haptics, 4(3), 2011.

[22] J. Reitzin, “German Surgeon Successfully Uses iPad in Operating Room”, mHealthWatch, 2013.

[23] J. Bacca, S. Baldiris, R. Fabregat, S. Graf, & Kinshuk, “Augmented Reality Trends in Education: A Systematic Review of Research and Applications”, Educational Technology & Society, 17(4), 133–149, 2014.

[24] A. Jenny, “Augmented Reality in Dental and Medical Education, AR News, 2016.

[25] R. Parkes, N. Forrest, S. Baillie, “A Mixed Reality Simulator for Feline Abdominal Palpation Training in Veterinary Medicine”, Studies in Health Technology and Informatics, 142, 244-246, 2009.

[26] S. 1. Lee, J. Lee, A. Lee, N. Park, S. Lee, S. Song, A. Seo, H. Lee, J. I. Kim and K. Eom, “Augmented reality intravenous injection simulator based 3D medical imaging for veterinary medicine” Vet J. , 196(2),197-202, 2013.

[27] Veterinary Record, “Augmenting Reality in Anatomy”, News and Reports, 175:444, 2014.

[28] R. G. Thomas, N. W. John and J. M. Delieu. “Augmented reality for anatomical education” Journal of visual communication in medicine, 33(1), 6–15, 2010.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 479-484 (2017)

www.astesj.com

Special issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Intrusion detection in cloud computing based attack patterns and risk assessment

Ben Charhi Youssef*,[1], Mannane Nada[1], Bendriss Elmehdi[2], Regragui Boubker[2]

[1] *TIES Team, ENSIAS, Mohammed V University Rabat 1000, Morocco*

[2] *Professors ENSIAS, Mohammed V University Rabat 1000, Morocco*

ARTICLE INFO

Article history:
Received : 10 April, 2017
Accepted : 11 May, 2017
Online: 24 May, 2017

Keywords :
Intrusion detection
Cloud computing
Risk assessment
Attack pattern
IDS
IPS

ABSTRACT

This paper is an extension of work originally presented in SYSCO CONF.We extend our previous work by presenting the initial results of the implementation of intrusion detection based on risk assessment on cloud computing. The idea focuses on a novel approach for detecting cyber-attacks on the cloud environment by analyzing attacks pattern using risk assessment methodologies. The aim of our solution is to combine evidences obtained from Intrusion Detection Systems (IDS) deployed in a cloud with risk assessment related to each attack pattern. Our approach presents a new qualitative solution for analyzing each symptom, indicator and vulnerability analyzing impact and likelihood of distributed and multi-steps attacks directed to cloud environments. The implementation of this approach will reduce the number of false alerts and will improve the performance of the IDS.

1. Introduction

Cloud computing is a new emerging model in Information technology which can enable ubiquitous, convenient, on demand network access to a shared pool of configurable computing resources released with minimal management effort and can rapidly be provisioned. It represents a new opportunity for both customers and service providers, by improving IT efficiency, agility and reliability to reduce the cost of IT technologies. This technology allows customers to create an elastic environment with a multitude number of services, optimize resources and offer an alternative way for renting computing and storage infrastructure services.

Cloud computing services are the most vulnerable target for intruder's attacks due to their distributed environment and the accessibility of cloud services via the internet threating data and services security.

Detect and deter attacks can be difficult task for security administrators. Therefore, the use of IDS (Intrusion Detection system) can help both cloud providers and security administrator in order to monitor and analyze network traffic.

The goal of using such system is to detect and prevent attacks, using different algorithm of detection. Nevertheless, analyzing and monitoring symptoms produces a multitude alarms with a huge number of false ones impacting the effectiveness of such systems. The distributed and open structure of cloud layers makes the implementation of traditional IDS inefficient to be deployed in cloud environment.

In the aim of reducing the number of false alarms this paper proposes a correlation approach by analyzing risks related to each attack pattern. The approach consists of calculating risk related to each attack steps by analyzing symptoms, indicators and vulnerability to define the attack risk score then yield alert. This paper propose to include customers as part of security protection. It demonstrate how the implementation of risks can decrease the number of false alarms. Our work is extended by presenting the initial results of the implementation of the approach in cloud computing specially in SaaS layer.

The remainder of the paper is structured as follows: in section II we present the prior work done in this field. Section III presents the intrusion detection system and risk assessment methodology, in section IV, we describe our proposed detection process based on attacks patterns analysis by risk assessment. Section V presents the initial finding results of the implementation of our approach to detect an SQL injection attack in SaaS layer .The last section of this article will present a conclusion and future works.

[1] Corresponding Author: Ben Charhi Youssef, ENSIAS University Mohamed V Rabat 1000, MOROCCO | Email: bencharhi.youssef@gmail.com

2. Related Works

Intrusion detection system has an important role in the security and perseverance of active defense system against intruder hostile attacks. Through the past few years several papers have been presented with the purpose of detecting malicious threat against cloud with different approaches. Security researchers need to have a seamless mechanism to integrate and analyze various information generated by heterogeneous sources implemented in cloud environment with the aim to detect intrusion and reduce false positive alerts. [1]

Kleber, schulter et al. [2] have proposed an IDS service at cloud middleware layer with an audit system designed to cover attacks that Network IDS and Host IDS cannot detect. In [3] authors presented an approach of detection using five major classifiers to characterize the nature of an attack, classification by attack vector, attack target, operational impact, informational impact and by defense, which provide the network administrator with information of how to mitigate an attack.

With similar methodologies to those used in the AVOIDIT [4] taxonomy, S. A. and J. Hamilton [5] developed an ontology-based attack model to assess the security of an information system from the angle of an attacker. Goal of the assessment process is the evaluation of attack effects. Thus, the difference of system performance before and after an attack is calculated. The process consists of four phases. The first step consists in identifying system vulnerabilities by using automated vulnerability tools. Such tools assess computer system, applications or network regarding their vulnerabilities and generate sets of scan results. In the second phase, the developed ontology is used to determine which attacks might occur due to the identified vulnerabilities. By querying the ontology, the possible effects are obtained. Finally, in the last phase the attack effect is calculated. In [6] authors proposed a taxonomy with four dimensions that provides a classification covering network and computer attacks. Their taxonomy provides assistance in improving computer and network security as well as consistency in language with attack description. The first dimension being attack vector is used to classify the attack. The second dimension classifies the target of the attack. The third dimension consists of the vulnerability classification number, or criteria. The fourth and final dimension highlights the payload or effects involved.

Massimo Ficco in [7] proposes a hybrid and event correlation approach for intrusion detection in cloud computing. It consists of detecting intrusion symptoms by collecting diverse information at several cloud levels to perform complex event analysis presented in an ontology.

In [15] song et al propose a model which is based on behavioral and contextual analysis. Authors present two sets of attributes to be used for anomalies detection contextual attribute and behavioral attribute based on a probabilistic analysis.

Wang et al in [16] present an approach for TCP/IP Reassembly in Network Intrusion Detection and Prevention Systems to be used as DPI (Deep Packet Inspection) for network intrusion monitoring.

For risk assessment several methods exist. In [8] author proposes a method for a probabilistic model driven risk assessment with security requirements. The security requirements and their causal relationships are represented using MEBN (Multi-Entities Bayesian Networks) logic that constructs an explicit formal risk assessment model that supports evidence-driven arguments. In [9] author proposes security ontology for organizing knowledge on threats, assets and safeguards measures. This work constructs classification for each of these groups and creates a method for quantitative risk analysis, using its own framework.

In our previous work [10] we presented a new intrusion detection approach based on risk assessment for traditional IDS.

Although all of the previous works present useful taxonomy and ontology that can provide an informative baseline for cyber intrusions and attacks analysis, they lack the details needed to analyze all symptoms of attacks which can provide many false positive alerts in cloud environment. Moreover, this work proposes a detection approach based on risk assessment analysis related to attack pattern initially presented in [1], which means that the data owner and cloud provider will participate in attacks analysis.

For example the same attack against two different clients of the same services may have different impacts. However, IDS will detect both connections as malicious actions. That's could bring IDS to miss the impact analysis, and generate many false alarms.

Our solution addresses this issue in cloud services through the implication of risk assessment in the detection and analysis processes.

3. Intrusion detection and risk assessment

This section presents the Intrusion detection systems, the correlation process and the risk assessment methodology.

3.1. Intrusion detection systems

An IDS (Intrusion detection system) is used for monitoring the network or a host for suspicious or malicious activities. Next to an IDS there is also an Intrusion Prevention System (IPS). An IPS will do more than just monitoring the traffic t will also prevent malicious activities from taking place. IDS/IPS perform a diagnosis of system security to detect all suspicious actions and discover various security breaches based on different algorithms of detection. IDS is used to detect flaws and the different possible threats that attempt to compromise the confidentiality, integrity or availability of a resource with a capability to block malicious attacks using functionalities of prevention.

Providing an analysis of all network traffic with IDS will catch intruders and block multiple suspicious actions threating the security of IT systems. But will also lead to a great number of false positives and can identify authorized users as intruders, especially in cloud environment where all resources are shared between customers.

In order to withhold attacks and reduce the number of false alerts and improve efficiency of IDS, our solution defines both CSP (Cloud Services Provider) and the customers as part of the monitoring process by implementing a risk agent related to all cloud layers (IaaS, PaaS, SaaS).

3.2. Risk assessment

Risk assessment is the process of identifying security risks related to a system and determining their probability of occurrence, their impact, and the safeguards that would mitigate that impact. Risks can be defined as the potential that a given threat will exploit vulnerabilities of an asset or group of assets to cause loss or

damage to the assets. The main objective of risk assessment is to define appropriate controls for reducing or eliminating those risks

Based on the 27005[11] standard risk is evaluated by looking at the probability of successful attacks and the consequent severity of that attacks.

Generally there are four steps of risk assessment. The four steps are as follow:

- Threat Identification
- Vulnerability Identification
- Risk Determination
- Control Recommendation

The general equation of risk determination is:

$$Risk = Impact * Likelihood \quad (1)$$

Impact: (or consequence) refers to the extent to which a risk event might affect the enterprise expressed in the terms of: Confidentiality, Integrity and Availability

Likelihood: represents the possibility that a given event occurs.

This equation (1) will be implemented in our approach with the aim to encourage customers to assess the security risks related to their information and to simplify the analysis of all detected events for intrusion detection systems in cloud computing.

4. Attack pattern analysis in cloud with risk assessment

An attack pattern is an abstraction mechanism for describing how a type of observed attack is executed. Following the pattern paradigm, it also provides a description of the context where it is applicable and then, unlike typical patterns, it gives recommended methods of mitigating [10].

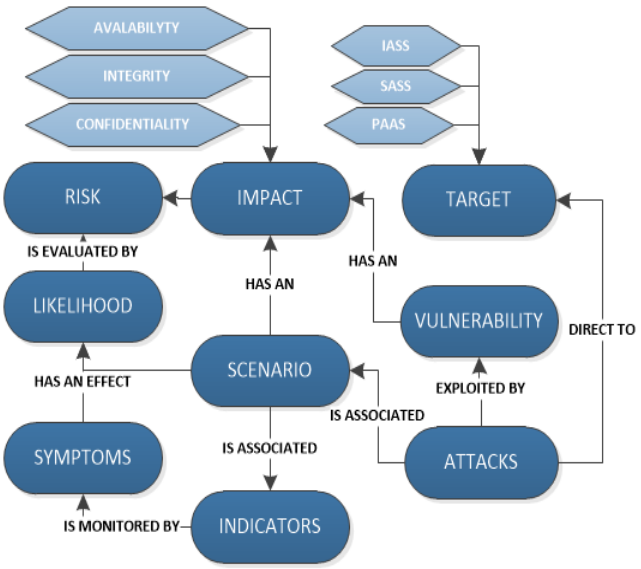


Figure 1: Attack ontology

Figure 1 shows a simplified view of our attack ontology designed to cloud environment.

When an attack occurs, it uses several paths (from reconnaissance to exploitation) by using scenarios defined in the process by indicators and symptoms, in the aims to gain unauthorized access to data. The analysis of impact simplifies the monitoring process and minimizes false-alarms.

Each attack scenario has an impact in either confidentiality, availability or/and integrity. The data owner on the risk-agent will define the impact of each indicator, symptom and vulnerability depending on the nature of the data.

The possibility that a given event will occur is defined as Likelihood. The risk score of the detected attack changes and simplifies the diagnostic of alarms (false positive or successful attack), in the same time symptoms and attack indicators scenario upturn the likelihood in the risk agent.

The analysis of attack pattern uses the risk agent to calculate the score of all indicators and symptoms then to define the nature of the packet. We propose a definition for risk (Re) as a product of the Probability (Pe) of a security compromise, a threat event, and its potential Impact or Consequence (Ie):

$$Re = Pe * Ie \quad (2)$$

The correlation is used to aggregate attack scenarios and symptoms produced by all the probes in the cloud environment.

Equation (2) represents the formula of risk assessment, the likelihood Pe, whereas (Ie) may be assigned a value on a numerical scale. Each indicator Is, Vulnerability Ve, Symptoms Sy have different impact. The value of Pe will be increased related to each detection of an attack pattern.

By using the formula of risk assessment of all suspicious actions we can get a risk score related to each attack:

$$Ra = 1/n \sum Ie * Pe \quad (3)$$

$$Ie = \sum (Is + Ve + Se + Sye) \quad (4)$$

The impact Ie and likelihood Pe of each symptom and attack will be defined by the data owner and cloud provider in a risk agent deployed in all cloud layers. The goal of such correlation is to get attacker's behavior and risk related to each action to by calculating impact and lure a potential attacker from critical system.

The value of Ra related to each attack will determine if the action is a successful attacks or a false positive depending of the sensitivity of targeted data related to risk values.

IDPSs serve four essential security functions: they monitor, analyze detect and respond to all suspicious activities as depicted in the functional layer.

The IDS/IPS detects intrusion by analyzing the collected data in all cloud layers based on knowledge base and the attack pattern ontology, calculates the risk of the attack then decides if it represents a suspicious action or a false-positive.

A software Agent is related with each probe defining the impact and likelihood of each detected symptoms. The risk server will determine the impact of each attack pattern then the risk of the attack. Implementing, such methodology can minimize the number of false-positive alarms and increase the effectiveness of Intrusion Detection System.

5. Implementation of approach

In this chapter we present the initial finding while implementing the discussed idea in cloud environment in the aims to decrease the number of false-alarms and to improve the reaction in the detection intrusion system in cloud.

5.1. Architecture of implementation

IDS implementation in cloud computing requires an efficient, scalable and virtualization-based approach. In cloud computing, user data and application is hosted on cloud service provider's remote servers and both cloud user and CSP (Cloud Services Provider) has a limited control over data and resources.

In the implementation of the proposed approach we have used OPENSTACK [12] as cloud platform with a collection of open source software project that developers and cloud computing technologist can use to setup and run their cloud compute and storage infrastructure.

SURICATA [17] was used as intrusion detection system which is a free and open source, fast and robust threat detection engine.

The simulation setup is a Cloud with different layers. We have chosen a fixed value in risk agent. The main goal is to analyze all actions based on the impacts and likelihoods of vulnerabilities and detected symptoms to evaluate the efficiency of our works.

The architecture of implementation and the used correlation process is presented in Figure 2:

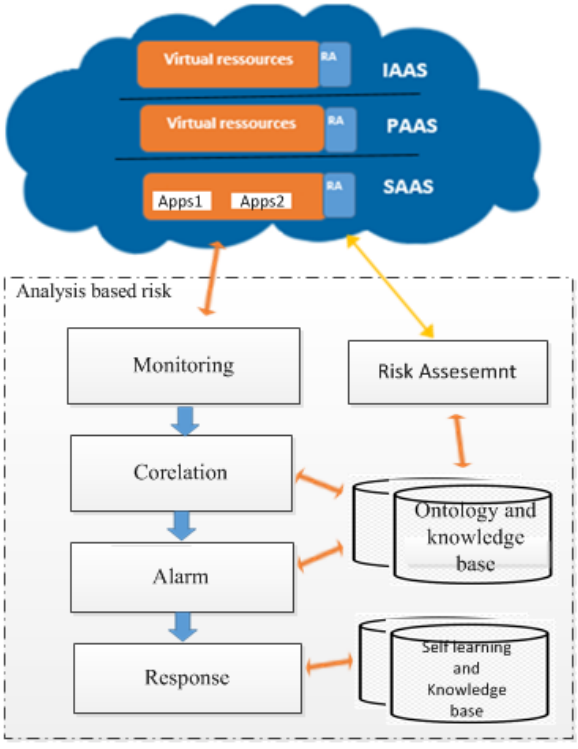


Figure 2: Architecture of implementation

5.2. Detection of SQL injection in SaaS

An attack based on SQL injection can be defined as an attack which consists of insertion of a SQL query via the input data from [13].

In this attack malicious user will use several steps to gain an access to a SaaS service and a successful SQL injection exploit can threat the confidentiality, integrity and availability of database. However the same attack in two application in a SaaS service could have different impact.

Based on the SQL injection attack pattern defined in [14] the scenario can be presented as below:

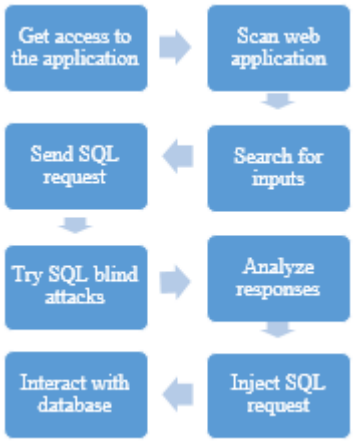


Figure 3: SQL injection attack pattern

In the scenario of implementation each cloud client must define a risk related to this attack pattern

- **Vulnerabilities**: Input validation ,Inputs are not properly validated by the application
- **Scenario** : Try different approaches for adding logic to SQL queries
- **Indicators** : Scan a website application, analyze responses positive or negative
- **Symptoms**: Use automated scan tools, try "Blind SQL Injection" techniques to extract information about the database, Send different requests

The Table below shows a simplified use of risk assessment used to estimate an SQL injection attack impact against two SaaS applications with a fixed values:

Attack pattern	Likelihood in app1	Likelihood in app2	Impact in app1	Impact in app2	Ra app 1	Ra app 2
Vulnerability	1	3	3	1	3	3
Scenarios	3	2	4	3	12	6
Indicators	2	2	4	3	8	6
Symptoms	2	3	4	1	8	4
Attack risks					7,75	4,75

Table 1: Attack analysis

Attack Pattern against SaaS services is composed of 4 attributes: Vulnerability, Scenarios, Indicators of attack and symptoms. Each attribute define an impact and has an initial value of likelihood determined by customers in risk agent.

Ra app 1 present the risk of each attribute, while attack risks is calculated using the equation (3). Table1 demonstrate how the same attack against two customers in the same Cloud service should have different impact.

The graphs below show results, before and after the use of risk analysis, obtained while exploiting the SQL injection vulnerability directed to two customers data in the same SaaS service with the use a static risk agent related to this attack pattern as presented in *table 1*.

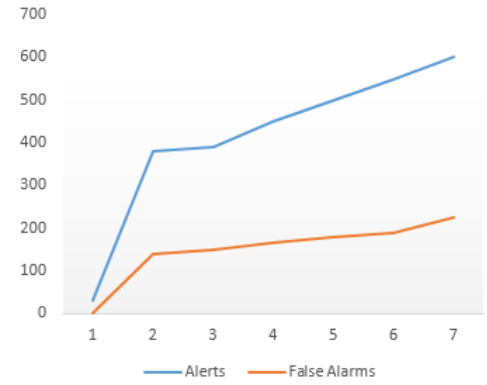


Figure 4: Number of false alarms before implementation

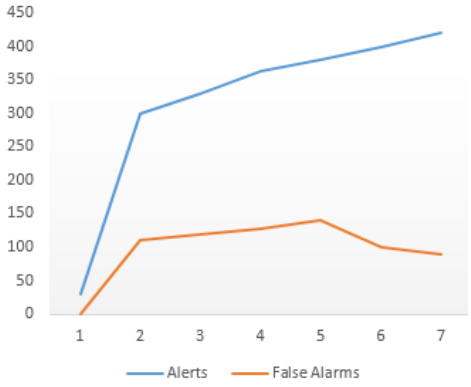


Figure 5: Number of false alarms after implementation

5.3. Discussion

The specific and complex characteristics of cloud computing environment make the implementation of intrusion detection system more difficult with the multitude of alarms and the huge number of false alarms. Therefore, there is a need of new approach of detection and analysis of malicious activities for both cloud consumers and CSP to check the effectiveness of the current security controls that protect data stored in cloud.

Figure 5 demonstrates how the implementation of risk assessment in cloud layer for two customers reduces the number of false positive alarms. With such an approach, the cloud consumers can check the effectiveness of the current security controls that protect an organization's assets and the service providers can maximize and win the trust of their cloud consumers if the level of risk is not high. Also the cloud consumers can perform the risk assessment to be aware of the risks and vulnerabilities present in the current cloud computing.

6. Conclusion and future works

Based on the observation that consumers must be involved in data protection and attack deterrence in cloud computing, this paper provides a novel approach in the disposition of intrusion detection system. We suggest a model to recognize and analyze malicious actions by using a risk assessment related to attack pattern.

The same attack should have different impact on different customer's data. This paper classifies attacks by target, symptoms, scenarios, impact and likelihood defining the attack pattern than estimate impact and risk related to all malicious activities against cloud services. Both CSP and consumer will participate in the correlation process. This classification will help to provide a methodology of analysis in depth of all suspicious actions in the aim to minimize the number of false alarms and increases the efficiency of intrusion detection system in cloud services.

The initial implementation for securing a SaaS service and detecting an SQL injection attack against two customers data's has demonstrated the efficiency of our approach in both decreasing the huge number of false alarms and increase the effectiveness of the IDS system.

Future works will be focused on the implementation of our approach to detect sophisticated attacks in all cloud layers with the implementation of distributing risk agent.

References

[1] S. Mumtaz, L. Khan, "Performance of Grid-Integrated Photovoltaic/Fuel Cell/ Electrolyzer/Battery Hybrid Power System" in 2nd International Conference on Power Generation Systems and Renewable Energy Technologies, Islamabad Pakistan, 2015.

[2] Youssef, Ben Charhi, Bendriss E. "Intrusion detection in cloud computing based attacks patterns and risk assessment." Systems of Collaboration (SysCo), International Conference on. IEEE, 2016.

[3] Kleber, schulter, "Intrusion Detection for Grid and Cloud Computing", in IEEE Journal: IT Professional, 19 July 2010.

[4] HANSMAN, Simon HUNT, Ray. "A taxonomy of network and computer attacks. Computers & Security", vol. 24, no 1, p. 31-43,2005.

[5] SIMMONS, Chris, ELLIS, Charles, SHIVA, Sajjan, et al. AVOIDIT: "A cyber attack taxonomy", IEEE, 2009.

[6] Hafez Amer, S., & hamilton Jr, J. A. "Intrusion Detection Systems (IDS) Taxonomy-A Short Review" in STN 13-2 Defensive Cyber Security: Policies and Procedures 2, 23, 2010.

[7] Patel, Ahmed, et al. "An intrusion detection and prevention system in cloud computing: A systematic review." Journal of network and computer applications 36.1, 25-41, 2013.

[8] Ficco, M., Tasquier, L., & Aversa, R. (2013, October). Intrusion detection in cloud computing. In P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC), Eighth International Conference on (pp. 276-283). IEEE, 2013.

[9] Z. Xuan,N. Wuwong , et al., "Information security risk management framework for the Cloud computing environments," in 2010 IEEE 10th International Conference on Computer and Information Technology (CIT), pp. 1328-1333, 2010.

[10] Ekelhart, A., Fenz, S., Klemen, M., & Weippl, E "Security ontologies: Improving quantitative risk analysis" in System Sciences, 2007. HICSS 2007. 40th Annual Hawaii International Conference on (pp. 156a-156a). IEEE,2007.

[11] Bencharhi, Y., Bendriss, E., & Regragui, B."Intrusion Detection System Based on Risk Assessment in Cloud Environment" in International Journal of cloud computing JCC, ICACON (4ICACON),2016.

[12] ISO/IEC 27005, Information Technology-Security techniques -Information security risk management. International Organization for Standardization, Geneva,2008

[13] Yadav, S. "Comparative study on open source software for cloud computing platform: Eucalyptus, openstack and opennebula" in international Journal Of Engineering And Science, 3(10), 51-54,2013.

[14] OWASP sql injection definition https://www.owasp.org/index.php/SQL_Injection

[15] Khatri, Jaimin K., and Girish Khilari. "Advancement in Virtualization Based Intrusion Detection System in Cloud Environment." International Journal of. Science, Engineering and Technology Research (IJSETR) 4.5,1510-1514, 2015.

[16] LIANG, Zhenkai, YIN, Heng, et SONG, Dawn. HookFinder: Identifying and understanding malware hooking behaviors. Department of Electrical and Computing Engineering, p. 41,2008.

[17] Wang, X., & Cronin, B "TCP/IP Reassembly in Network Intrusion Detection and Prevention Systems". International Journal of Information Security and Privacy (IJISP), 8(3), 63-76,2014.

[18] Hoque, M. S., Mukit, M., Bikas, M., & Naser, A. "An implementation of intrusion detection system using genetic algorithm". International Journal of Network Security & Its App lications (IJNSA), Vol.4, No.2, 2012.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 485-491 (2017)

www.astesj.com

Special issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Optimal Sizing and Control Strategy of renewable hybrid systems PV-Diesel Generator-Battery: application to the case of Djanet city of Algeria

Adel Yahiaoui*, [1], Khelifa Benmansour[1], Mohamed Tadjine[2]

[1]*Research Laboratory of Electrical Engineering and Automatics, Electrical Engineering Department, Yahia Fares University, Medea, Algeria*

[2]*Control Laboratory Processes, National Polytechnic School, 10 Avenue Hassen Badi, BP .182, El-Harrach Algiers, Algeria*

ARTICLE INFO

Article history:
Received: 26 March, 2017
Accepted: 11 May, 2017
Online: 24 May, 2017

Keywords:
Optimal size
Photovoltaic systems
Partial discharges
Urban areas
Hybrid power systems

ABSTRACT

A method for optimal sizing of hybrid system consisting of a Photovoltaic (PV) panel, diesel generator, Battery banks and load is considered in this paper. To this end a novel approach is proposed. More precisely a methodology for the design and simulation of the behavior of Hybrid system PV-Diesel-Battery banks to electrify an isolated rural site in southern Algeria Illizi (Djanet). This methodology is based on the concept of the loss power supply probability. Sizing and simulation are performed using MATLAB. The technique developed in this study is to determine the number of photovoltaic panels, diesel generators and batteries needed to cover the energy deficit and respond to the growing rural resident energy demand. The obtained results demonstrate the superior capabilities of this proposed method.

1. Introduction

PV system size and performance strongly depend on meteorological variables such as solar energy, wind speed and ambient temperature and therefore, to optimize a PV system, extensive studies related to the metrological variables have to be done [1].

The performance of a PV module strongly depends on the sun light conditions. Standard sunlight conditions on a clear day are assumed to be 1000 W of solar energy per square meter and it is sometimes called ‘‘one sun’’ or a ‘‘peak sun’’. Less than one sun will reduce the current output of a PV module by a proportional amount [2,3].

The generation of electricity through a hybrid system combining several renewable energy sources is of great interest to developing countries, Such as the Maghreb countries. These countries have many isolated regions far from traditional electricity distribution networks. Thus the extension of these networks would be of an exorbitant financial cost. To solve this problem, exploiting the renewable energy potential of these countries must be a priority.

Hybrid PV/diesel systems have greater reliability for electricity production than using PV system alone because diesel engine production is independent of atmospheric conditions. The PV/ diesel system will provide greater flexibility, higher efficiency and lower costs for the same energy quantity produced. In addition, PV/diesel system as compared to only diesel system provides a reduction in the operation costs and air pollutants emitted to the atmosphere [4]. Currently, some research works are carried out focusing on optimal sizing of hybrid PV/diesel generator systems so that the number of PV modules, storage battery capacity and diesel generator capacity can be optimally selected. Hybrid PV/diesel systems can be classified into two categories: series topology and parallel topology. In the series topology, the diesel generator is connected in series with the inverter and therefore it is not able to supply the load directly while the load demand is covered by the storage battery. The disadvantages of this topology that such systems require large inverter size and low overall efficiency due to the series connection, as well as when the PV system is connected to the diesel generator it leads to a limited control of the diesel generator. However, these disadvantages can be avoided in the parallel topology whereas the diesel generator is in parallel with the inverter and can supply the load and the battery at the same time [4].

Rural electrification in the southern countries is based in most cases on solar homes systems with limited use. This gives an impression that photovoltaic systems aimed the poor. Diesel generators generate more electricity but suffer from the maintenance costs and the provision of oil [5].

*Corresponding Author: Adel Yahiaoui, City of Almond, Bougara, Blida, Algeria, +213 795 517 302 | Email: yahiaoui.industry@gmail.com

The hybrid combination of photovoltaic-battery-diesel systems requires technical knowledge, economic and preliminary studies for the design, optimizing the choice of a suitable load profile, assessing consumption and simulation performance before installation and implementation of the systems. Each aforementioned item requires data and computer tools to accomplish it.

Many papers are discussed on design of hybrid systems with the different components. In 2006 Eftichios Koutroulis et al presented a methodology for optimal sizing of stand-alone PV/WG systems using genetic algorithms [6]. They applied design approach of a power generation system, which supplies a residential household. In 2007 Yang developed a novel optimization sizing model for hybrid solar-wind power generation system [7]. In 2016 R. Hosseinalizadeh et al presented an economic sizing of a hybrid (PV–WT–FC) renewable energy system (HRES) for stand-alone usages by an optimization-simulation model: Case study of Iran [8]. In 2015 B. Guinot developed a Techno-economic study of a PV-hydrogen-battery hybrid system for off-grid power supply: Impact of performances' ageing on optimal system sizing and competitiveness [9]. Recently, Leong Kit Gan et al developed a hybrid wind–photovoltaic–diesel–battery system sizing tool development using empirical approach, life-cycle cost and performance analysis: A case study in Scotland [10].

An overall power management strategy is designed for the system to coordinate the power flows among the different energy sources. Simulation studies have been carried out to verify the system performance using practical load profile and real weather data in south Algeria. It is to mention that there are many similar regions around the world with this typical situation that can be expanded.

In this paper, a methodology of optimal sizing of hybrid power generation system is proposed. The main objective of this work is to determine the optimal number of PV panels, Battery banks and diesel generators. For this purpose a new sizing algorithm, called loss power supply probability (LPSP) is utilized. The obtained results show that the proposed methodology finds optimal sizing of hybrid power generation system with a quality energy services for the rural population through of a reliable network. The simulated result illustrates the good applicability and performance of the proposed method.

The rest of the paper is organized as follows: Section 2 presents the city of our study. Section 3 presents the modeling of the system considered. Section 4 discuses about the LPSP algorithm. Section 5 and 6 describes how the problem can be solved by using the proposed method and results. Conclusions are presented in Section 7.

2. The case of Djanet city

2.1. The selected site

The site selected for this study is a rural village isolated Saharan named "Djanet" present in figure 1. This village is located in the province of Illizi and consists of a small number of homes.

2.2. Meteorological data

The performances of photovoltaic modules are directly correlated to climatic data that is in principle the solar irradiation and the ambient temperature. These data are taken through the acquisition system measured to step time is 10 minutes. For each day of the year the kit provides for the acquisition of ambient temperature in degrees Celsius (°C), and solar incident in Watts per square meter illumination (W / m2). The annual assessment of the site's climatic characteristics is shown in Figure 2 and Figure 3.

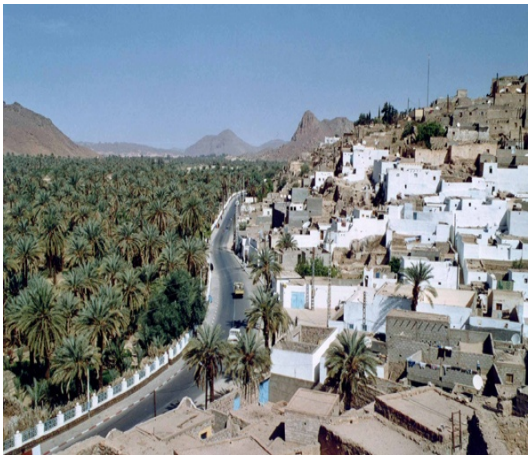


Figure 1: Geographical location of Djanet city in Algeria

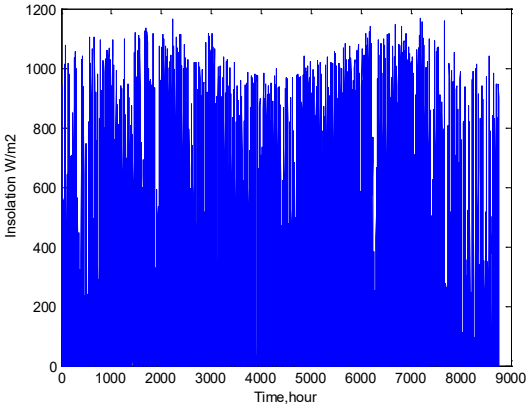


Figure 2: Annual insolation data for Djanet city

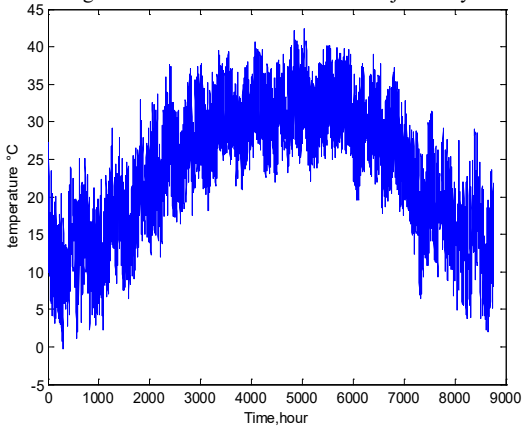


Figure 3: Annual temperature data for Djanet city

2.3. The load profile

The initial data in the implementation of a renewable energy production system as for any other energy system is the demand, which will be determined in relation to the and / or the loads to be supplied. This demand must be estimated as precisely as possible, both from a point of view of the powers called and from its temporal distribution, although it's often random nature makes this task very difficult.

The hybrid system considered is intended for supplying a load for domestic use. For our case study (Djanet city) in which there are ten dwellings, we consider a constant load profile throughout the year.

Figure 4 shows the distribution over the hours of the day of consumption of the conventional appliances used (lighting, refrigeration, TV, radio, ventilation); on which the difference between the quantities of energy consumed from one hour to another is seen. Hence, the importance of choosing economical appliances for isolated sites, to minimize installation costs.

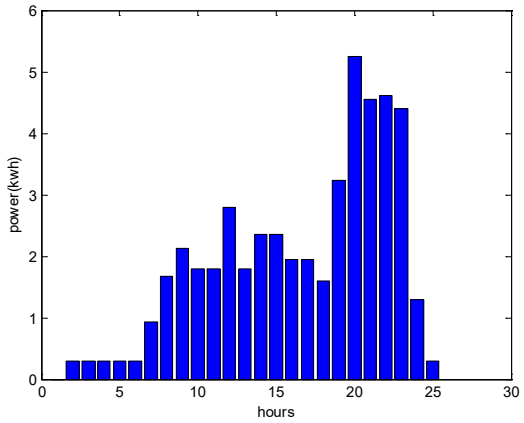


Figure 4: Curves of hourly load profiles.

3. Modeling of considered system

3.1. Photovoltaic system

Choosing a suitable model for simulation is very essential. The simple model of PV panel as a function of the area, the input variables of this model are solar insolation shown in Figure 1 and its efficiency. Detail equation of PV model can be described as follows [10,11,12]:

$$P_{pv}(t) = n.A_p.N_{pv}.G(t) \quad (1)$$

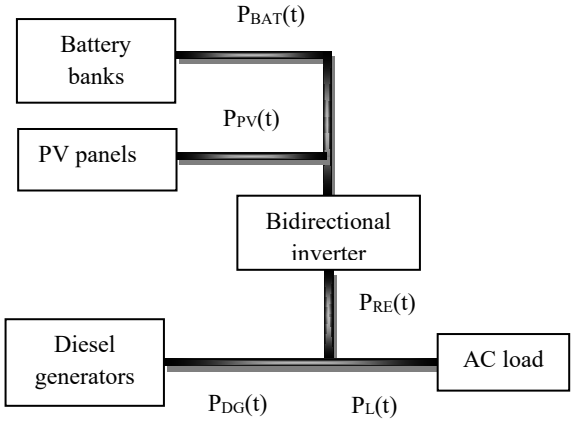


Figure 5: Configuration of the system

Where; n energy conversion efficiency, (%). A_p is the area of single PV panels, (m^2). N_{pv} is the number of PV panels. G (t) is the insolation data, (W/m^2).

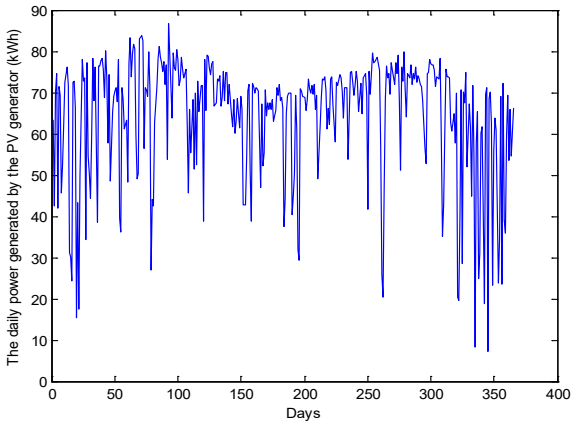


Figure 6: The daily power generated by PV modules

3.2. Battery Banks

Here, power from battery banks is required whenever The PV panels or Diesel generators (DG_s) are fail to supply load. Power from PV system through the inverter is hourly measured as follow:

$$P_{RE}(t) = \eta_{inv}.P_{pv}(t) \quad (2)$$

Where, η_{inv} is the inverter efficiency. Principles of charging and discharging scenarios are dependent on the state of P_{RE} and the load power demand at hour t ($P_L(t)$).

1. If the value of P_{RE} (t) > P_L(t) the remaining power will be used to charge battery banks by the following equation [13]:

$$C_{Bat}(t) = C_{Bat}(t-1) + \left(P_{pv}(t) - \frac{P_L(t)}{\eta_{inv}} \right) \eta_{Bat} \quad (3)$$

2. If the value of the P_{RE} (t) < P_L(t) insufficiency power will be supplied by diesel generators and/or the battery banks according to the following dispatch strategy:

- If the amount of power from battery banks is enough to handle the remaining power, the strategy allows the discharging process of battery banks with following equation [14]:

$$C_{Bat}(t) = C_{Bat}(t-1) + \left(P_{pv}(t) - \frac{P_L(t)}{\eta_{inv}} \right) \quad (4)$$

Where, C_{Bat} (t) is available battery banks capacity at hour t, C_{Bat} (t-1) is available battery banks at previous hour and η_{Bat} is yield of battery charging.

The value of C_B (t) could not be lower than minimum allowable energy level remained in battery banks, and it could not be higher than maximum allowable energy level during charging operations as follow [8,13]:

$$C_{Bmin} \leq C_B(t) \leq C_{Bmax} \quad (5)$$

- Whenever battery banks are fail to supply the remained power or the strategy does not allow for the battery banks to discharge, then diesel generator is started and battery banks will be neither charged.

$$P_{DG}(t) = P_L(t) - P_{RE}(t) \quad (6)$$

Where, $P_{DG}(t)$ is the output power from diesel generator unit.

3.3. Diesel generator

The fuel consumption of diesel generator unit is related with the rated power and its generated power. The fuel cost is calculated for a year as follows [15]:

$$AFC = C_f \sum_{t=1}^{Tend} F(t) \quad (7)$$

Where, F(t) is the hourly fuel consumption (liter /hour), based on load characteristic of the diesel generators. This parameter is calculated as follows [14]:

$$F(t) = (0.246 \times P_{DG}(t) + 0.08415 \times P_R) \quad (8)$$

Where, P_R is rated power of diesel generators, P_{DG} (t) is the power generated by diesel generators and C_f is the fuel cost per liter in US$/l.

4. Description of the LPSP method

A balanced system, the amount of energy supplied by the solar modules must at least compensate the amount of energy, depending on the place of installation, the season, and the particular conditions of consumption and use.

The sizing method presented consists in determining the optimum number of batteries and PV modules according to an optimization criterion, namely reliability, which is based on the concept of the LPSP (Loss of Power Supply Probability), It will allow us to dimension a realistic electricity production system supplying ten individual habitats for the Djanet site during a year.

Dimensioning a PV system means determining all the elements of the PV chain according to the loads such as the sunshine and the load profile. There is no fixed scale between consumption and generator size. There are several methods of dimensioning a PV system: by the hourly usability function, by the probability of energy loss, and by iterative methods ... etc.

LPSP is a probability technique introduced by AbouZahr and Ramakumar in 1990 the field of wind and photovoltaic generation systems [16, 17]. The probability of energy loss (loss of power supply probability) is widely used in the design and optimization of autonomous systems (stand-alone) wind-photovoltaic [18, 19, 20]. It is based on the concept of the LPSP, which is the probability that the system is experiencing a shortage at some time during a period of analysis, that is to say a state in which the system would not be able to provide energy to the load. From the information on the variability of resources and its correlation with the load, LPSP can be calculated. This method is particularly useful in assessing the behavior of an energy storage.

The 'LPSP' is defined as the fraction of the unsatisfied demand on that requested by the load. It expresses the rate of non-satisfaction of the load. Thus, this probability is defined as the sum of all unfulfilled demands 'LPS' on the total energy demanded by the load during a period of operation (For this study, T = 1 year), as follows [14] :

$$LPSP = \frac{\sum_{t=0}^{T} T_i}{T_p} \quad (9)$$

Where,

T_p is the period of study.

T_i is the duration corresponding to a power produced less than the demand.

5. Application of the LPSP to the PV installation

The first step before carrying out any dimensioning is to determine when we need electricity and to define our consumption which should not be underestimated or overestimated. This step requires a lot of reflection because an error at this stage will distort our results. So, at the outset, we choose the energy consumption profile, as well as the choice of an architecture of the system.

With the data from the first step, you will be able to know the number of photovoltaic modules needed to cover the electricity needs. To do this, it is necessary to calculate the energy that the modules must produce each day, and to know the sunshine of the region where the installation is located. The amount of electricity produced by the photovoltaic modules depends directly on this sunshine and therefore on the region.

Always the first step will also allow us to calculate the number of battery. The stored energy depends directly on the periodicity of your consumption.

The next step is to dimension the load regulator. This step does not involve any calculation, but there are a wide variety of regulators that differ according to their options, functions and power. It will therefore take time to choose the regulator best suited to your situation.

Another step is to choose the converter if this device is necessary, that is to say if our load is operating in AC mode. Indeed, the installation can already provide direct current 12, 24 or 48VDC.

6. Development of the proposed approach

The system configuration is shown in Figure 5; meanwhile, the flowchart of system operation is shown in Figure 7. The load demand is supplied by DGs, PV, and battery banks. The bidirectional inverter is required to supply power from battery banks to the load demand or feed surplus power from PV panels and DGs to the battery banks. Total power from battery banks, DGs and PV panels has to satisfy the load demand. The concept of operational strategy could be explained clearly as follow:

- If the load demand ($P_L(t)/ \eta_{inv}$) is less than the total power produced from photovoltaic panels ($P_{pv}(t)$), and the battery C_B is fully charged, then the banks of the battery are charged as amount as the capacity of the inverter and the excess power are dumped.

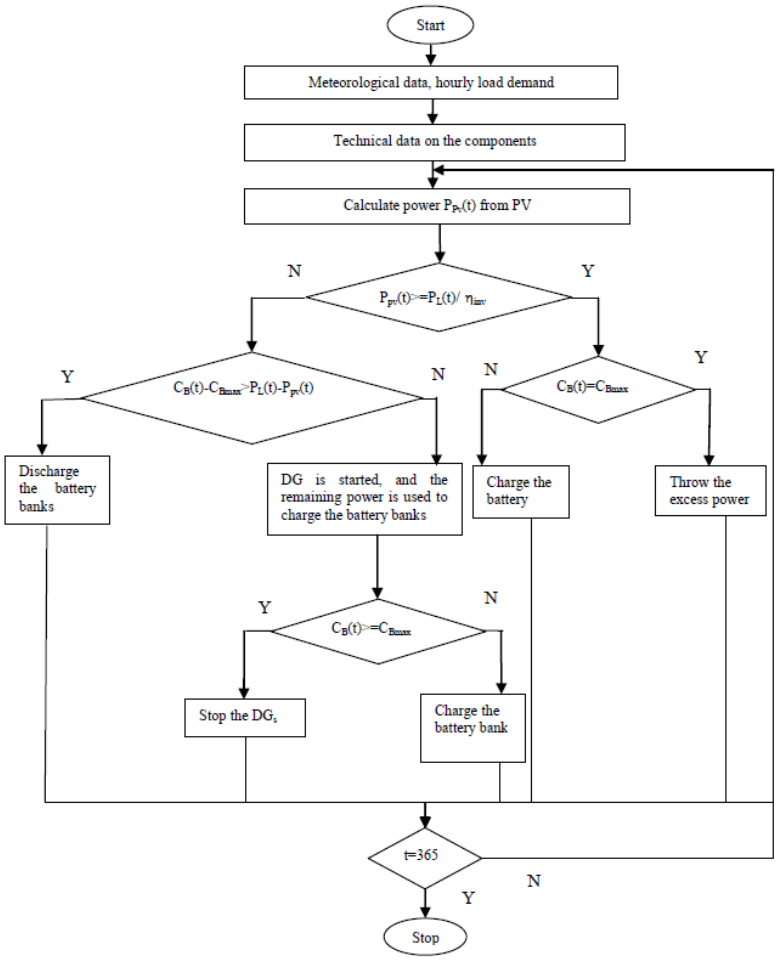


Figure 7: Flowchart of the algorithm for simulating hybrid system

- If the $C_B(t)$ is less than C_{Bmax} the excess power is used to charge the banks of the battery.
- If the load demand (P_L (t)/ η_{inv}) is greater than $P_{pv}(t)$ and C_B (t) of the banks of the battery is greater than C_{Bmin}, then the power of the insufficiency will be provided by the banks of the battery. Otherwise, if $C_B(t)$ of the battery banks equal C_{Bmin}, diesel generators are started to meet the demand for charging the banks of the battery.
- However, if the power of the inadequacy is greater than the capacity of the inverter, the battery banks are discharged in quantity that the capacity of the inverter, and then the diesel generators are started to meet the load demand.

7. Application and results

The main objective of this work is to determine the optimal configuration of the hybrid photovoltaic, diesel generator with storage to satisfy the power demanded by the load.

Several methods of designing hybrid energy systems have been carried out by researchers. The method we are going to use is based on a technical-economic analysis. We will first apply the LPSP

Table 1: Number of PV modules in terms of LPSP

LPSP	0.01	0.02	0.03
Number of PV modules	221	219	216
Number of batteries	120	105	90
Number of diesel generators	3	3	2

Table 2 : Monthly energy balance

	Energy produced by PV (kWh)	Average storage capacity (kWh)	Fuel consumption (Liter)	The operating hours of DG
January	1300	88.6046	6,33	1,00
February	1420	87.8545	0,00	0,00
Marche	1620	97.2668	0,00	0,00
April	1550	54.4687	88,62	14,00
May	1650	54.4379	82,29	13,00
June	1440	48.3240	151,92	24,00
July	1440	49.8778	177,24	28,00
August	1640	51.8210	88,62	14,00
September	1550	52.6139	101,28	16,00
October	1690	96.4810	0,00	0,00
November	1350	87.1510	0,00	0,00
December	1170	85.6003	0,00	0,00

Table 3 : Annual energy balance

Nature	Value
Annual PV Energy produced	17820 kWh
Average annual storage capacity	69,90 kWh
Annual fuel consumption	696,30 Liters
Annual energy produced by diesel generator	2200 kWh
Annual excess energy	4030,8 kWh

(Loss of Power Supply Probability) method. Then, using the optimal configuration that satisfies the demand obtained, consisting of the number of photovoltaic panels, number of batteries and diesel generators, we use the approach proposed in the previous section to simulate the hybrid system.

In our system, we chose LPSP = 0.03 which gives a number of PV modules equal to 216 Depending on weather conditions (light and temperature) and load profile, the number of photovoltaic modules is given in Table 1.

Table 4 : Daily production percentage

Nature	Percentage
Photovoltaic	94.4627 %
Diesel Generator	5.5373 %
Annual excess energy	1.5229%

Table 1 show that by increasing the value of the LPSP, the number of PV panels, battery banks and diesel generators is decreased. In other words, the variation of components depends on the LPSP.

It is thus possible to conclude from various simulations that the availability of photovoltaic panels plays a key role in the optimal system design and in the precise design of the energy production of the system.

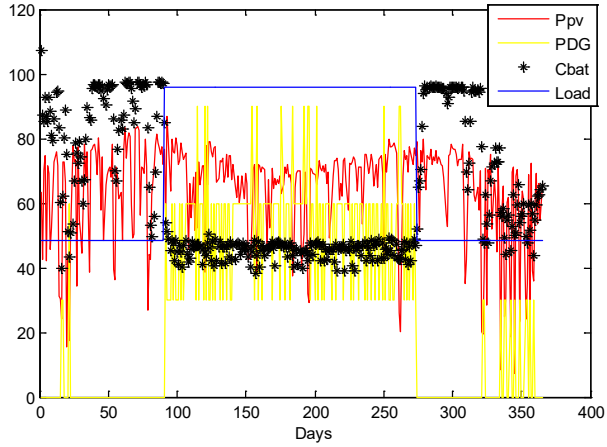


Figure 8. The energy management of hybrid system

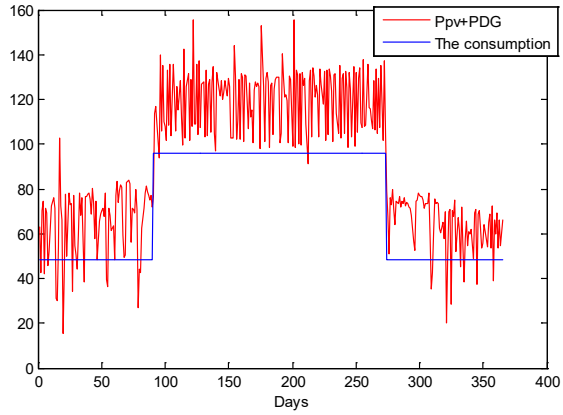


Figure 9: The daily power generated by PV panels and DG$_s$ compared with the consumption

The simulation results presented in this paper provides monthly, hourly and daily results, which allows us to make monthly and annual balance sheets. Figure 8 shows a monthly variation in the production of photovoltaic energy, diesel generators and average storage capacity respectively, we see that the two production pikes are for the months of April and September (iterations 93 and 279), and production is minimal in the month of December (in the iteration 345), these results affect the ability of storage, the relationship between the two parameters is proportional.

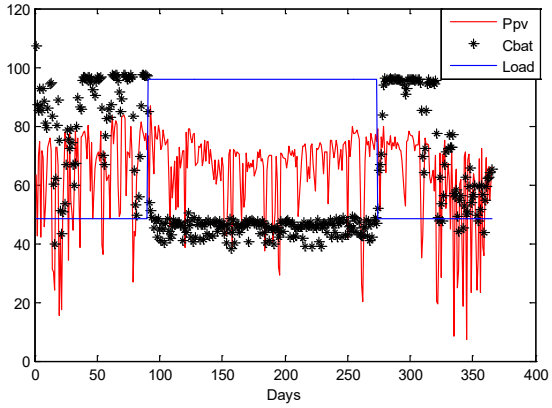


Figure 10: energy management of only PV system

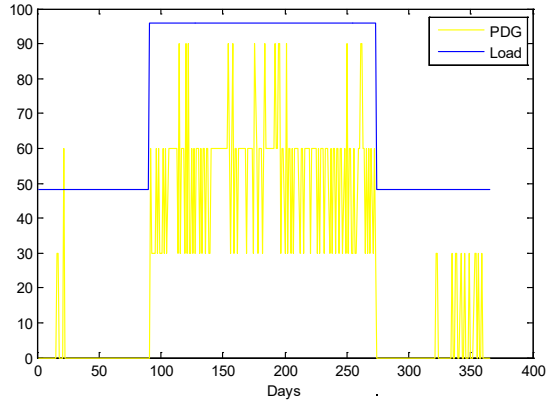


Figure 11: energy management of only DG system

The energy produced by the diesel generator, as shown in Table 3 vary from one month to another, they are highest in the months of April, May, June (in the iterations 92 to 272), and they are minimal in the months of January, February, March, April, May, October, November and December (in the iterations 1 to 91 and 273 to 365). Note that the results are synchronized by sunshine, temperature and load, we also find that the percentage of photovoltaic is important (94.4627%) by contribution to 5.5373% of the energy produced by the diesel generator.

The excess energy is low in the order of 1.5229%. The energy produced by the system during 12 months compared with the consumption is shown in Figure 9. it can be observed that the consumption is minimum in the months of January, February, March, October, November and December (in the iterations 1 to 88 and 274 to 365), which is causing a decrease in the total energy produced, and it increases proportionally with consumption in the rest iterations (months of April, may, June, July, August and September).

The simulation results of only PV system, only DG system are presented in Figures 10 and 11 respectively. From these results it can be concluded that the system is able to fully satisfy the load demand.

The design procedure resulted in determination of best configuration of the system.

The stand-alone hybrid PV-DG system with battery bank seems to be a motivational techno-economic solution to meet the energy demand of consumer.

8. Conclusion

This paper presents a new method based on loss power supply probability (LPSP) to solve the optimal sizing of a hybrid PV/Diesel/Battery system. The optimal number of PV panels, Diesel generators and batteries is determined using LPSP to supply the load. The main objective of this work is to examine the feasibility and performance to electrify rural community by the hybrid solution, then we realized that the operation of the system depends not only on the input and output variables which are the sunshine, the temperature and the load to be supplied but also the interdependence between different devices. An improperly sized system component can compromise which was verified by simulation results the proper functioning of the entire system. Thus, we can conclude that the combination of energy sources (PV-Diesel-Battery) is required to be able to cover the energy deficit and meet the growing energy demand of rural inhabitant. We believe that the technical and economic simulation results obtained were satisfactory and demonstrated the effectiveness of the studied hybrid system.

A successful hybrid system requires an energy saving approach and a rigorous design and installation of components that meet the need.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The research work presented in this paper was jointly carried out by the Research Laboratory in Electrical Engineering and Automatics (LREA) at the University of Yahia Fares in Medea-Algeria and the Control Laboratory Processes (LCP) at the National Polytechnic School of Algiers-Algeria. We would like to thank all those involved in our project.

References

[1] T. A Khatib, “Review of designing, installing and evaluating standalone photovoltaic power systems” Journal of Appl. Sciences.,;10:1212–28, 2010.
[2] R. Mukund, “Wind and solar energy” CRC Press LLC., 1999.
[3] The German solar energy society, “Planning and installing photovoltaic systems”, 2005.
[4] M. Ashari, V. Nayar, “An optimum dispatch strategy using set points for a photovoltaic (PV)–diesel battery hybrid power system” Sol. Ener., 66:1–9, 1999.
[5] A. Yahiaoui, K. Benmansour, M. Tadjine, “A new sizing algorithm of renewable hybrid systems PV-Diesel Generator-Battery: application to the case of Djanet city of Algeria” 8th International Conference on Modelling, Identification and Control (ICMIC), Algeria, 2016.
[6] E. Koutroulis et al, “Methodology for optimal sizing of stand-alone photovoltaic/wind generator systems using genetic algorithms” Sol. Ener., Volume 80, Issue 9, Pages 1072-1088, 2006.
[7] H. Yang et al, “a novel optimization sizing model for hybrid solar-wind power generation system” Sol. Ener., Volume 81, Issue 1, Pages 76-84, 2007.
[8] R. Hosseinalizadeh et al, “Economic sizing of a hybrid (PV–WT–FC) renewable energy system (HRES) for stand-alone usages by an optimization-simulation model: Case study of Iran” Renew. Sust. Ener. Rev., Volume 54, Pages 139-150, 2016.
[9] B. Guinot et al, “Techno-economic study of a PV-hydrogen-battery hybrid system for off-grid power supply: Impact of performances' ageing on optimal system sizing and competitiveness” Inter. Journal of Hydro. Ener., Volume 40, Issue 1, Pages 623-632, 2015.
[10] T. Senjyu, D. Hayashi, A. Yona, N.Urasaki and T. Funabashi, “optimal configuration of power generating systems in isolated island with renowable energy” Renew. Ener., vol.32, PP. 1917-1933, 2007.
[11] B. Nelson, M. H. Nehrir, C. Wang, “unit sizing cost analysis of stand-alone hybrid wind/PV/fuel cell power generation systems” Renew. Ener, vol. 31, PP. 1641-1656, 2006.
[12] A. Yahiaoui, K. Benmansour, M. Tadjine, “Control, analysis and optimization of hybrid PV-Diesel-Battery systems for isolated rural city in Algeria” Sol. Ener, vol. 137: 1–10, 2016.
[13] H. Suryoatmojo, A. A. Elbaset, F. A. Pamuji, D.C. Riawan, Nursalim, M. Abdillah, “optimal sizing and control strategy of hybrid PV-Diesel-Battery systems for isolated island” ADCONP 2014 Hiroshima.
[14] B. S. Borrowy, Z. M. Salameh, “Methodology for optimally sizing the combination Battery Bank and PV array in a Wind/PV hybrid system” IEEE Trans. Ener. Conv, Vol. 12, N°1, pp. 73-78, 1997.
[15] H. Suryoatmojo, A. A. Elbaset, T. Hiyama, “Economic and reliability evaluation of Wind-Diesel-Battery system for isolated island considering CO_2 emission” IEEJ Trans. PE, vol. 129(8), 2009.
[16] I. Abouzahr, R. Ramakumar, “Loss of power supply probability of stand-alone wind electric conversion systems” IEEE Trans. Ener. Conv, vol. 5, n° 3, pp. 445-452, 1990.
[17] I. Abouzahr, R. Ramakumar, “Loss of power supply probability of stand-alone photovoltaic systems”, IEEE Trans. Ener. Conv, vol. 6, n° 1, pp. 1-11, 1991.
[18] F. M. A. Ghali, M. M. Abd El Aziz, F. A. Syam, “Simulation and analysis of hybrid systems using probabilistic techniques” Proceedings of IEEE Power Conversion Conference, Nagaoka, pp. 831-835, 1997.
[19] F. Giraud, Z. M Saameh, “Steady-state performance of a grid-connected rooftop hybrid wind-photovoltaic power system with battery storage” IEEE Trans. Ener. Conv, ISBN 0885-8969, vol. 16, n° 1, pp. 1-7, 2001.
[20] Enrico Fabrizio, “modeling of multi energy systems in buildings”, PhD Thesis, National Institute of Applied Sciences in Lyon, in July 2008.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 492-497 (2017)

www.astesj.com

Special issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

RFID Antenna Near-field Characterization Using a New 3D Magnetic-Field Probe

Kassem Jomaa[*,1], Fabien Ndagijimana[1], Hussam Ayad[2], Majida Fadlallah[2], Jalal Jomaah[2]

[1]*IMEP-LaHC Laboratory, Grenoble-Alpes University, Grenoble, France*

[2]*Physics Department - Faculty of Sciences, Lebanese University, Beirut, Lebanon*

ARTICLE INFO

Article history:
Received : 05 April, 2017
Accepted: 24 April, 2017
Online: 24 May, 2017

Keywords :
Near-field Measurements
Magnetic-field Probe
Exposure
Electromagnetic fields
RFID

ABSTRACT

In this paper the design of a new 3D magnetic field (H-field) probe with a near-field scanning system is presented, then the near electromagnetic fields radiated by a Library RFID system is characterized. The proposed system is developed in order to determine the magnetic near-field emitted by electronic devices. The designed isotropic H-field probe consists of three orthogonal and identical loops each of diameter of 6 mm having 3 turns. The antenna factor of the designed probe is presented for a frequency range from 10 MHz to 1 GHz. The designed probe is tested and validated using a standard passive circuit as a device under test. An RFID reader antenna is also designed and simulated on HFSS (high frequency structural simulator) and the radiated magnetic field, obtained by simulations, is then compared to the real measured one above the fabricated circuit. The obtained levels are checked if they satisfy the European and ICNIRP Electromagnetic Fields Guidelines.

1. Introduction

The everyday use of devices and systems emitting radio frequency electromagnetic fields is continuously increasing. Some of these devices are operating in the vicinity of human body, in which people are in the near-field region of the radiating source, and they are exposed to electromagnetic fields. Radio Frequency Identification (RFID) systems are used today in wide variety aspects of everyday life. There are wide areas of RFID applications such as: cashless payments, transportation ticketing systems, residential door keys, antitheft protection systems, library systems [1] etc. RFID technologies are intended to recognize or identify remotely placed label on an object, animal or person. The system consists of the reader or RFID antenna and of a label (or tag) composed of a RFID chip and an antenna (e.g. pasted into a book in a library). The label is activated by a modulated electromagnetic wave processed by the reader with an antenna. It returns a signal to reader in one of the frequency bands allocated to the RFID applications: 125 to 150 kHz; 13.56 MHz; 433 MHz; 860 to 960 MHz; 2.45 GHz. The distances covered with systems can reach up to several meters [2].

Due to the development of such applications, there must be a study on the relationship between the exposure to electromagnetic fields and the health of humans. Near-field region is an important part that should be analyzed in this study. Even though there are different methods for field mapping, passive detection is always an economical way of measuring vector electromagnetic fields. Various researches have been done in the development and characterization of electric and magnetic probes. In [3] the results of a probe with a CPW fed transmission line in the frequency range of 0.1 to 0.3GHz is presented whereas in [4] a resonance suppressed magnetic field probe which operates in the 1-7GHz frequency band is presented. All these probes measure only one component of the electric or magnetic field at a time. Also, the larger size of these probes reduces the performance by reducing the spatial resolution as the size of the probe is inversely proportional to the spatial resolution of the probe [5].

It is necessary to have all the 3 components of magnetic or electric fields for accurate prediction of radiated emissions [6]. But it requires a large scanning time in order to complete one surface of a printed circuit board during the near field measurements. In [7] three dimensional (3D) near field scanner for IC chip level measurements is proposed. It consists of 3D near field scanner with a magnetic probe head. Two separate magnetic fields probe (to measure normal and tangential component of magnetic field) for measuring three dimensional fields in high frequency planar circuits is proposed in [8]. Still, these methods don't reduce the scanning time because each component has to be measured

[*]Corresponding Author: Kassem Jomaa, IMEP-LaHC Laboratory, Grenoble-Alpes University, Grenoble, France Email: kassem.jomaa@imep.grenoble-inp.fr

separately. In order to minimize the scanning time we propose a 3D scanning system consisting of a near field scanner and a 3D magnetic field probe.

In this work a new 3D magnetic field probe which measures the three components of the magnetic field simultaneously is designed and validated. It is used to measure the three magnetic field components of the radiated emissions from a device under test (DUT), over a broad range of frequencies. The near-field exposure assessment of the radiated field from RFID reader antennas used in Libraries is then studied. An RFID rectangular loop antenna operating at 13 MHz is designed and fabricated. The designed antenna is simulated on HFSS using finite element method (FEM) then the simulations are compared to the real measurements of the magnetic field near the antenna using the designed probe.

2. Probe Design

2.1. Magnetic field Sensors Theory

Electromagnetic field is a combination of two types of quantities: electric and magnetic fields. In the near-field regions they should be both measured since they are not related to each other as in the far-field region. So we have six field components three for each field, each one should be measured separately. In this work we will focus in the magnetic field measurement. Magnetic field passing through a loop generates a voltage across its terminals according to Faraday's law, which states that the induced voltage is proportional to the rate of change of magnetic flux through a circuit loop:

$$V_{emf} = -\frac{d\phi}{dt} = -\frac{d}{dt}\iint \vec{B}.d\vec{S} \tag{1}$$

Each single loop is capable to measure one component of magnetic field. For total magnetic field distribution three orthogonal components $\boldsymbol{H_x}$, $\boldsymbol{H_y}$, and $\boldsymbol{H_z}$ must be measured. The overall magnetic field $\boldsymbol{H_t}$ can be determined from the root sum square of these quantities:

$$H_t = \sqrt{H_x^{\,2} + H_y^{\,2} + H_z^{\,2}} \tag{2}$$

2.2. Design Probe

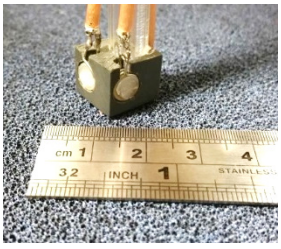


Fig. 1. 3D H-field Probe

The designed 3D ***H***-field probe consists of three orthogonal identical loops combined in a 1×1.2×1.3cm^3 plastic cube having their centers at the same plane (Fig. 1). Each loop has 3 turns with radius of 3mm. The terminals of each loop are connected to the channels of the Teledyne LeCroy oscilloscope (which also measures the frequency domain as a spectrum analyzer) through a coaxial cable. The difference between the loop centers was taken into account in post processing. The whole loop structure is attached to the robot arm in the scanning system.

2.3. Probe Calibration

Calibration is a process done to find the probe performance factor which is the antenna factor (AF). It is like a transfer function between the measured voltage across the loop and the magnetic field in the scanned geometry. The calibration of electromagnetic field probes needs to generate a reference field which intensity and spatial configuration are known with good accuracy inside a certain volume where the calibrated probe is placed.

Several probe calibration techniques are available nowadays. Transverse Electromagnetic (TEM) cells [9], waveguides, and also a 50 ohm simple microstrip line can be used.

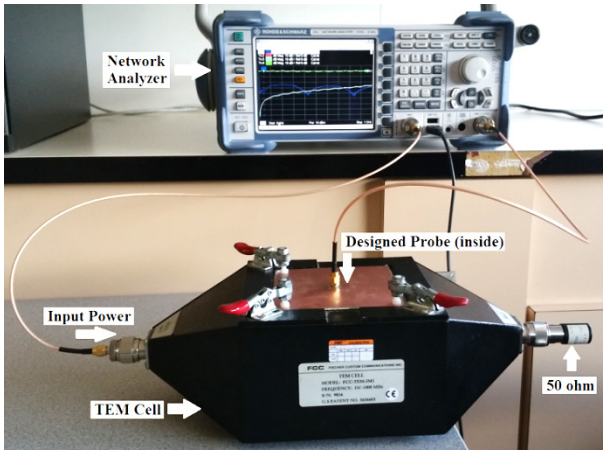


Fig. 2. Calibration Setup

In this work FCC-TEM-JM1 cell was used to produce calculable electric and magnetic field strengths. It operates in the transverse electromagnetic mode, so that both the ***E*** and ***H***-field components generated between the septum and outer conductor have the characteristics of a wave propagating in free space. The field strength can be calculated from the dimensions of the cell, its impedance at the measurement plane and the input power as shown in the equations below:

$$E[V/\mathrm{m}] = \frac{\sqrt{2.Z_c.P_{in}[\mathrm{watt}]}}{h[m]} \tag{3}$$

$$H[A/m] = \frac{E[V/m]}{\eta[\Omega]} \tag{4}$$

Where Z_c is the input impedance of the measuring instrument in ohms

P_{in} is the input power in watts

h is the distance in meters between the septum and the outer conductor of the TEM cell

η is the intrinsic impedance of the free space in ohms

$$\eta = \sqrt{\frac{\mu_0}{\varepsilon_0}} \tag{5}$$

The antenna factor then can be calculated by placing the probe inside the TEM cell and measuring the induced voltage $V_{ind.}$ across the loop terminals, according to the following equation:

$$AF[\mathrm{dB}\frac{A/m}{V}] = 20\log\left(\frac{H[\mathrm{A/m}]}{V_{ind.}[V]}\right) \tag{6}$$

Where $V_{ind.}$ is the induced voltage in the loop (in volts) and it is given by:

$$V_{ind.} = \sqrt{2.Z_c.|S_{21}|^2.P_{in}[watt]}.e^{j.\arg(S_{21})} \tag{7}$$

The AF of each loop in the proposed probe is shown in Fig.3 for a frequency range of [10MHz – 1GHz] along with the AF factor of a single loop.

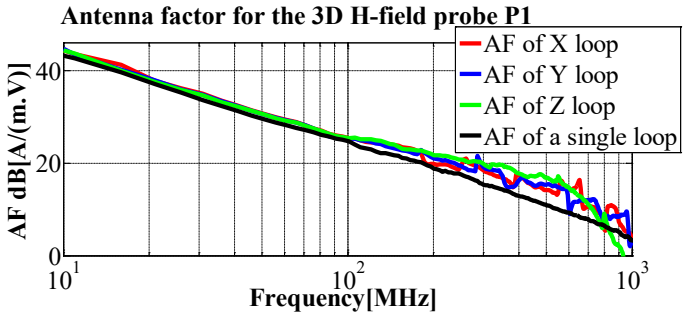


Fig. 3. Measured antenna factor versus frequency for the designed probe

The antenna factor varies from 44 dB(A/m.V) for low frequencies and up to 5 dB(A/m.V) at 1 GHz. Compared to the single loop AF, the antenna factor of each loop in the 3D probe has fluctuations above 200MHz . These are due to the coupling between the individual loops in the 3D probe. This factor is important to know the level of the field above the DUT when using the probe in the near-field measurement system.

3. Scanning System Validation

The 3D scanning system consists of a scanning robot, Teledyne LeCroy WaveRunner 640Zi oscilloscope, PC and the designed 3D H-field probe (Fig.4). The scanning robot is connected with the oscilloscope to the PC. The PC is programmed to control the movement of the robot and the data collection from the spectrum analyzer. The plastic arm of the robot holds the probe above the DUT. The movement of the robot arm can be precisely adjusted to 0.1mm/step by a 3D movable controller system. The output of the probe is connected to the 3 channels of the oscilloscope using coaxial cables. Since there is no preamplifier, the noise floor of the spectrum analyzer limits the measurements to frequencies above 10MHz.

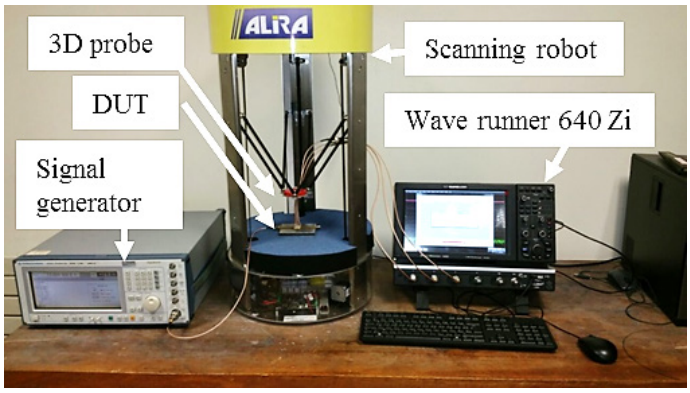


Fig. 4. Measurement System Setup

To validate the test bench, the magnetic near-field of a simple passive circuit used as DUT is measured and then the experimental results are compared to theoretical ones. The electromagnetic field components, depends on the geometrical parameters (Fig. 6) and the injected power of the DUT. A standard circuit that is consisting of a cylindrical conductor above a ground plane is used as the DUT. The DUT is terminated by 50 ohm load at one end and excited by a power of 16dBm by a signal generator at the other end (Fig. 5).

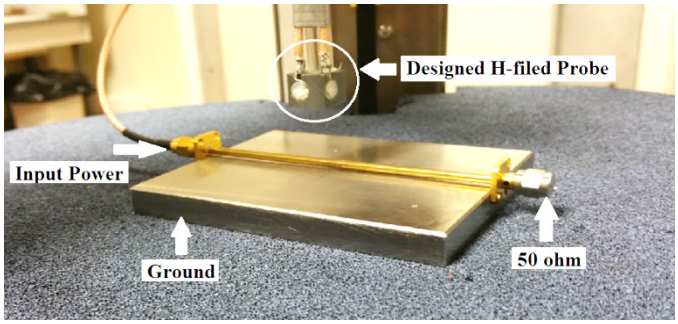


Fig. 5. Standard Validation Circuit

The magnetic field was computed using the expressions of the transverse electric and magnetic field components that are obtained by electrostatic calculation with the use of image theory [10]:

$$E_y = 8K\frac{yzn}{(y^2+(z+n)^2)(y^2+(\mathrm{z-n})^2)} \tag{8}$$

$$E_z = 4K\frac{n(y^2-z^2+n^2)}{(y^2+(z+n)^2)(y^2+(\mathrm{z-n})^2)} \tag{9}$$

$$H_y = -\frac{1}{\eta}E_z \tag{10}$$

$$H_z = \frac{1}{\eta}E_y \tag{11}$$

Where, $n = \sqrt{h^2-a^2}$ and $K = \dfrac{\sqrt{2PZ_c}}{\ln\left(\dfrac{h+n}{h-n}\right)}$

The propagating waves along the cylindrical conductor in the above circuit (Fig. 5) are characterized by a quasi-TEM mode. So we will only measure the transverse components of the magnetic field ($\boldsymbol{H_y}$ and $\boldsymbol{H_z}$).

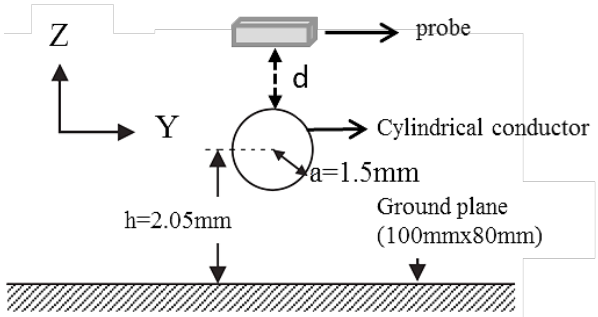


Fig. 6. Geometrical dimensions for the validation circuit

The ***z-axis*** is normal to the scanned plane, and the probe moves in the ***x*** and ***y*** directions measuring the magnetic field radiation from the DUT. Measurements were done at four different highs (4, 6, 8, and 10mm), and the scan was along ***y-axis*** with a step of 1mm. The computed theoretical field values are compared with the measured ones and they are presented in Fig. 7 and Fig. 8.

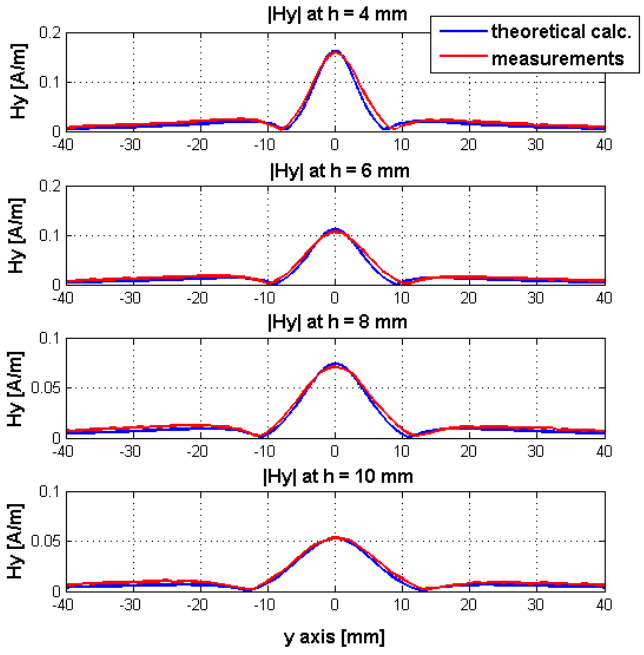


Fig. 7. Tangential H-field component H_y

Fig. 7 shows the tangential magnetic field component $\boldsymbol{H_y}$, while Fig. 8 shows the normal one $\boldsymbol{H_z}$. The comparison shows that the field measurements are in good agreement with the theoretical calculations for the 3D probe. Also good spatial resolution is obtained for the probes in measuring the magnetic field components.

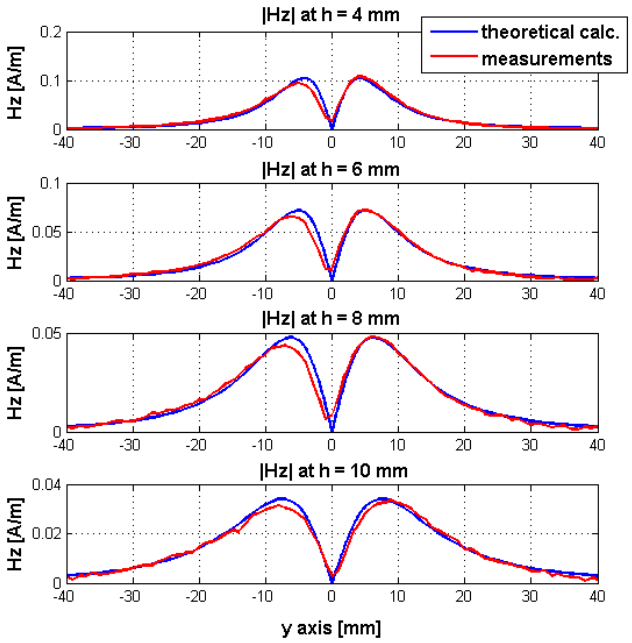


Fig. 8. Normal H-field component H_z

4. RFID System Characterization

4.1. RFID Loop Antennas

Most of the RFID readers are designed with loop antennas. Loop antennas feature simplicity, low cost and versatility. They are made of a loop of a conductor that may have various shapes: circular, triangular, square, elliptical, etc. They are widely used in communication links up to the microwave bands (up to 3GHz). They are also used as electromagnetic field probes.

Loop antennas are usually classified as electrically small and electrically large antennas. Electrically small loop antennas are those who have their circumference less than tenth of the wavelength ($C < 0.1\lambda$), while in electrically large antennas the circumference is near the wavelength ($C \approx \lambda$). Small loops are equivalent to a magnetic dipole and they have a far-field pattern very similar to that of a small electric dipole normal to the plane of the loop [11].

4.2. Designed RFID Antenna

In this paper, a rectangular small loop antenna operating at 13 MHz is designed. It is the typical RFID reader antenna used in library RFID systems. First the RFID reader antenna was designed and simulated using HFSS; it consists of 1-turn rectangular loop and a matching network. The rectangular loop has the dimensions of 140mm×140mm with a line width of 5mm.

The loop antenna has inductive input impedance; hence a parallel RC circuit should be used as a matching circuit. An impedance matching circuit was designed using surface mounted device (SMD) lumped elements to match the antenna to the 50Ω feed line at the operating frequency. Fig. 9 shows the corresponding circuit of the designed loop antenna, and Fig. 10 shows the 3D far field pattern where the loop lies in ***xy*** plane.

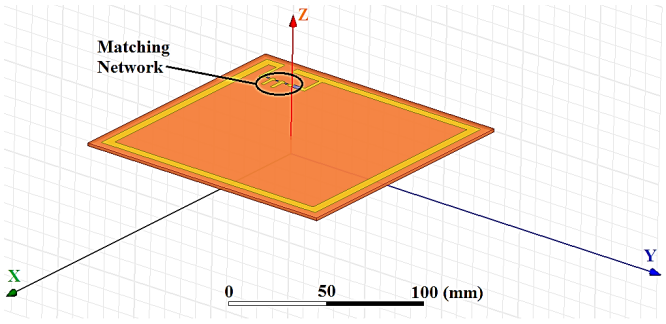


Fig. 9: Simulated RFID rectangular loop antenna

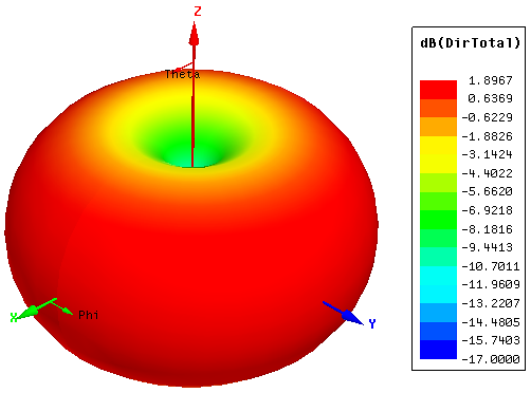


Fig. 10: 3D far field pattern of the designed antenna

The antenna prototype is then fabricated on a dielectric substrate FR4 with the dimensions of 150mm×150mm and 1.6mm thickness as shown in Fig. 11. The dimensions of the square loop are 140mm×140mm, and the width of the copper trace is 5mm.

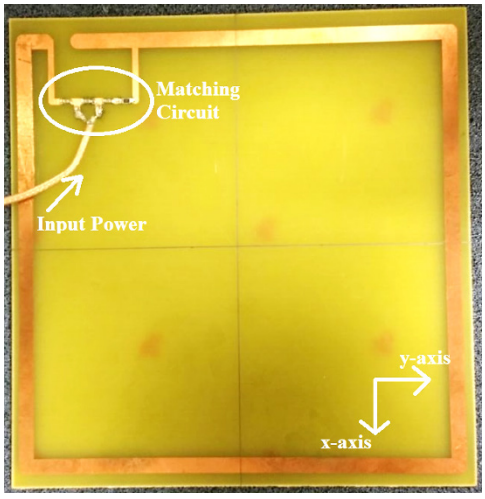


Fig. 11: Fabricated RFID loop antenna

The reflection coefficient measurement in Fig. 12 shows that the antenna resonant frequency is exactly at 13MHz. The measured reflection coefficient of the antenna is in good agreement with the simulation.

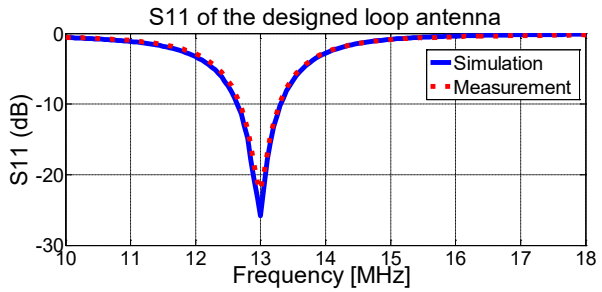


Fig. 12: Reflection coefficient of the designed RFID loop antenna

4.3. Characterization of the Designed RFID Antenna

The RFID antenna was excited with a ***CW*** signal having a source power of 10dBm, which is a typical power supply value for the RFID readers used in libraries applications [12]. The field mapping was done along ***y-axis*** and at two different heights (h=1cm & h=5cm) above the RFID reader loop antenna. These heights were chosen since they are the typical heights for the human exposure to library RFID systems. Then the measured magnetic field is compared with the simulated values and compared with ICNIRP guidelines for exposure limitations [13].

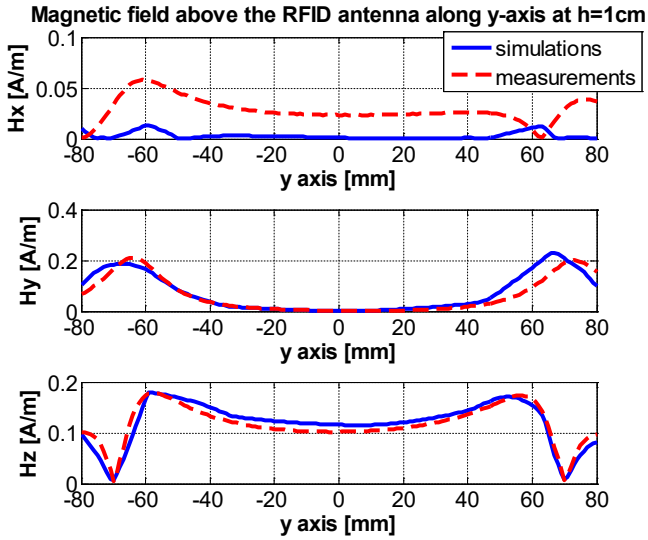


Fig. 13: Magnetic field strength above the RFID reader antenna along y-axis at height h=1cm

The magnetic field strength of the three components ($\boldsymbol{H_x}$, $\boldsymbol{H_y}$, and $\boldsymbol{H_z}$) at the two different heights above the RFID reader antenna is shown in figures Fig. 13 and Fig. 14. The simulation values obtained with HFSS are in good agreement with the measured ones, except the $\boldsymbol{H_x}$ component. This mismatch is due to the low level of $\boldsymbol{H_x}$ field if compared to the other components, so the coupling effect on the ***x*** component loop in the 3D probe is too high in this case. At h=1cm the $\boldsymbol{H_y}$ component has a maximum level of 0.2A/m while $\boldsymbol{H_z}$ has a maximum level of 0.18A/m where $\boldsymbol{H_x}$ has a level below 0.06A/m. At h=5cm the $\boldsymbol{H_y}$ component has a maximum level of 0.04A/m while $\boldsymbol{H_z}$ has a maximum level of 0.07A/m where $\boldsymbol{H_x}$ has a level below 0.02A/m.

At our operating frequency the reference level in the Guidelines is 0.2μT, which is equal to 0.16A/m in free space. This limit value refers to full body exposure for 6 minutes period. Considering the RFID reader used in library systems, there is a partial body exposure and the exposure lasts just for few seconds. So the obtained values at h=1cm which are little bit above the allowed ones (0.04A/m for $\boldsymbol{H_y}$ and 0.02A/m for $\boldsymbol{H_z}$) in the Guidelines are acceptable for this application. Moreover, the RFID reader antenna in Library systems is placed under a table so there is about 2cm between the antenna and the table surface.

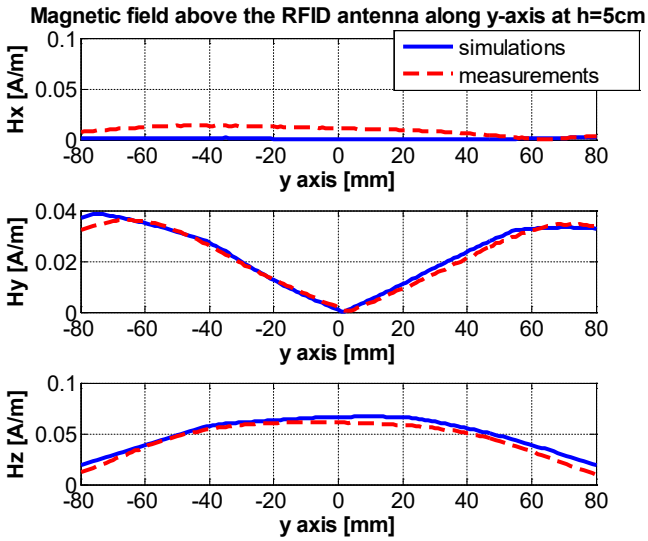


Fig. 14: Magnetic field strength above the RFID reader antenna along y-axis at height h=5cm

5. Conclusion

In this paper we have presented a near-field measurement system with a new 3D magnetic field probe design. The system consists of a 3D magnetic field probe, 3D scanning robot, and a digital oscilloscope. The designed probe was calibrated using a TEM cell, and the antenna factor between 10 MHz and 1 GHz was presented. The system was validated using a wire over a ground plane loaded with 50 ohm impedance. Good results were obtained for the used probe in measuring the magnetic field in the near region. Then an RFID reader loop antenna, used in Library RFID systems, was simulated and designed. The magnetic field strength near the designed antenna is then measured with the designed probe and analyzed. The simulations obtained with HFSS and the measurement results were compared to the action levels stated in ICNIRP Electromagnetic Fields Guidelines.

Acknowledgment

This research work was done at PHELINE, a test lab common with CSTB and Grenoble-Alpes University. It is specialized in the study of the electromagnetic fields phenomenon and its interactions with the environment. This work was partially funded by ANSES - French Agency for Food, Environmental and Occupational Health & Safety.

References

[1] G. Viziteu, B. Florean, G.M. Moraru, and A. Pintea, "An Overview of RFID Technology Used In Library", International Conference and Exposition on Electrical and Power Engineering (ICEPE), 2012, pp. 108-111.

[2] Klaus Finkenzeller, "RFID HANDBOOK Fndamentals and Applications in Contactless Smart Cards, Radio Frequency Identification and Near-field Communication, Third Edition", 2010.

[3] Osofsky and S. E. Schwarz, "Design and performance of a non-contacting probe for measurements on high-frequency planar circuits," IEEE Trans. Microw. Theory Tech., vol. 40, no. 8, pp. 1701– 1708, Aug. 1992.

[4] Jung-Min Kim, Woo-Tae Kim, and Jong-Gwan Yook, "Resonance-Suppressed Magnetic Field Probe for EM Field-Mapping System" IEEE Trans. Microw. Theory Tech.,, vol. 53, no. 9, September 2005.

[5] S. Jarrix *et*.al "Probe Characterization for Electromagnetic Near-Field Studies" IEEE Transactions On Instrumentation And Measurement, Vol. 59, No. 2, pp.292-300, February 2010.

[6] H.Weng, D.G.Beetner, R.E.DuBroff, "Prediction of Radiated Emissions Using Near Field Measurements" IEEE Transactions On Electromagnetic. Compatibility, vol.53, no.4, pp. 891-899, November 2011.

[7] Muroga et.al "3D magnetic –near-field scanner for IC chip-level noise coupling measurements," IEEE Transactions On Magnetics, vol. 49, no. 7, 4P. july 2013.

[8] Y. Gao and I. Wolff, "A new miniature magnetic field probe for measuring 3-dimensional fields in planar high frequency circuits," IEEE Trans. Microw. Theory Tech., vol. 44, no. 6, pp. 911–918, Jun. 1996.

[9] J.P. Karst, C.Groh, H. Garbe, "Calculable field generation using TEM cells applied to the calibration of a novel E-filed probe", IEEE Trans. Electromagnetic compatibility, vol. 44, no.1, pp. 59-71, February 2002.

[10] D. Baudry, A. Louis, and B. Mazari, "Characterization of the open-ended coaxial probe used for near-field measurements in EMC applications," Prog. Electromagn. Res., vol. PIER 60, pp. 311–333, 2006.

[11] Balanis, C. A., "Antenna Theory Analysis and Design, 2nd Ed.", John Wiley & Sons, New York, 1997.

[12] Klaus Finkenzeller, "RFID HANDBOOK Fundamentals and Applications in Contactless Smart Cards, Radio Frequency Identification and Near-field Communication, Third Edition", 2010.

[13] International Commission on Non-Ionizing Radiation Protection, "ICNIRP guidelines for limiting exposure to time varying electric, magnetic and electromagnetic fields (up to 300GHz)", Health Physics, Vol. 74, No. 4, pp. 494-522, 1998.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 498-504 (2017)

www.astesj.com

Special issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Design, Fabrication and Testing of a Dual-Range XY Micro-Motion Stage Driven by Voice Coil Actuators

Xavier Herpe*, Matthew Dunnigan, Xianwen Kong

School of Engineering and Physical Sciences, Heriot-Watt University, Edinburgh, EH14 4AS, UK

ARTICLE INFO

Article history:
Received: 23 February 2017
Accepted: 14 April, 2017
Online: 24 May, 2017

Keywords:
Compliant Mechanism
Dual-Range Manipulation
XY Motion Stage
Voice-Coil Actuator

ABSTRACT

Nowadays, compliant micro-motion stages offer high positioning accuracy and repeatability. When assembled with conventional stages, the resulting dual-range positioning systems allow for nanometer scale accuracy within a centimeter scale working area. However, such systems invoke a high cost and large footprint. This paper presents a low-cost, compact compliant XY micro-motion stage where dual-range manipulation has been implemented in an effort to maintain the workspace area and increase the positioning accuracy. The method presented in this work employs the use of a large range/low resolution linear encoder to measure the position of the stage (coarse positioning), before using short range/high resolution capacitive sensors to correct the positioning error (fine positioning). This solution achieves a positioning accuracy close to 1μm within a workspace of ±2.2mm×2.2mm. The compliant stage has been machined from Aluminum 7075-T6 using Wire Electric Discharge Machining (WEDM). This material is well known for its large reversible strain and its suitability for compliant mechanisms.

1. Introduction

Due to advantages such as compactness, cost reduction and enhanced performance, compliant XY motion stages are promising alternatives to conventional linear stages. They have a wide range of applications, such as fibre alignment, semi-conductor positioning, scanners for an Atomic Force Microscope (AFM) and micro-assembly. Their inclusion in micro-motion applications has allowed for accuracy and repeatability values in the nanometre scale. Compliant stages have been reported to have no backlash; no friction; no noise emission and no need for lubrication. However, they also have several disadvantages such as non-linear behaviour and limited working area. A multitude of compliant XY motion stages have been reported in the literature [1-10], with a range of motion from $105 \times 105\mu m^2$ to $25 \times 25mm^2$ and a positioning accuracy from 10μm down to a few nanometres. Combining high accuracy and large workspace implies expensive metrology such as ultra-high resolution/large range sensors and actuators. The aim of this work is to use the concept of dual-range manipulation to achieve a positioning accuracy of a few micrometres within a workspace of a few millimetres, while keeping the cost as low as possible.

The concept of dual-range manipulation is to use two sets of actuators and/or sensors within the same system. The first set allows for positioning in a large workspace but has a low positioning accuracy. The second set covers a smaller workspace, just enough to compensate the positioning error of the coarse mechanism and has a high positioning accuracy.

The main advantage of dual-range manipulation is the cost reduction. It also avoids the need for complex calibration, tight manufacturing and assembly tolerances.

Most conventional dual-range micro-motion stages simply consist of two stages serially connected, simplifying their design and assembly. Some stages can achieve a very high absolute positioning accuracy over a large range of motion by using ultra-high resolution/large range sensors such as laser interferometers. However, this implies a high cost and often a large foot print. For instance, the stage designed in [11] consists of a piezo (PZT) actuated stage mounted on the top of a DC motor driven stage, resulting in a $300\times300mm^2$ workspace and a positioning accuracy of ±10nm. A 3-DOF stage was designed in [12] using linear motors for the coarse positioning and voice coil actuators (VCAs) for the fine positioning. The system has a workspace of $500\times500mm^2$ and can achieve nanometre scale accuracy, with a repeatability of 50nm. The positioning stage developed in [13] combines a 2-DOF coarse positioning stage driven by linear motors and a 6-DOF fine positioning stage driven by VCAs and magnetic bearings. The stage has a $300\times300mm^2$ workspace with an accuracy of 10nm and 15nm along the X and Y axes, respectively. The 3-DOF stage from [14] combines two linear motors and four PZT actuators, giving it a working range of $200\times200mm^2$ working range and an accuracy of 13nm. The stage designed in [15] also combines VCAs and PZT actuators. It can achieve an accuracy of ±20nm within a

*Corresponding Author: Xavier Herpe, School of Engineering and Physical Sciences, Heriot-Watt University, Edinburgh, UK | Email: xh28@hw.ac.uk

workspace of 25×25mm^2. Another 1-DOF stage was developed in [16] using a linear motor for the coarse motion and a VCA for the fine motion. The positioning accuracy achieved was 10nm within the working range of 300mm.

The high cost, high accuracy and large workspace of these stages are not justified for miniaturised product assembly applications. In addition, stacking two stages on the top of each other increases the moving mass and therefore slows the dynamic response

Other stages use high resolution/short range sensors such as capacitive sensors for the fine positioning. This solution involves cheaper equipment but allows for the same positioning accuracy. However, positioning is relative to the current coarse position of the stage. For instance, the 2-DOF stage developed in [17] consists of a PZT actuated stage mounted on top of a DC motor and a lead screw driven stage. It can reach sub-micrometre accuracy using capacitive sensors and has a workspace of 200×100mm^2. The 1-DOF stage designed in [18] uses a VCA for the coarse positioning and a PZT actuator for the fine positioning. The high resolution linear encoder used allows for a positioning accuracy of 20nm within a workspace of 80mm.

In some cases, the dual-range manipulation is carried out using a single compliant stage in which two sets of actuators and/or sensors are integrated. For instance, the 1-DOF stage developed in [19] consists of a PZT actuator directly mounted on a VCA with a positioning accuracy in the nanometer scale. For each direction of motion, the XY motion stage designed in [20] uses one PZT actuator connected between the base and a lever mechanism to amplify the motion and a second one connected between the lever mechanism and the stage for the fine positioning. The positioning accuracy achieved is 10nm but the workspace is only 119.7×121.4µm^2. Another compliant dual-range XY motion stage designed in [21] uses only two VCAs and two sets of strain gauges, one for the coarse positioning and one for the fine positioning. The resolution of the coarse motion is 8.52µm for a range of -1.94mm to 2.47mm; while it is 1.63µm for the fine motion in the range of -0.22mm to 0.31mm. The advantage of this solution is that once the stage has been calibrated, both coarse and fine positioning are absolute. However, the fine positioning mode cannot be used for the whole workspace.

This paper presents the design and testing of a dual-range micro-motion XY motion stage based on our previous work from [22, 23]. A novel coarse/fine positioning concept is used to achieve low cost high accuracy positioning. Firstly, the force-displacement characteristics of the stage are defined using Finite Element Analysis (FEA). The fabricated stage is then evaluated and the results are compared with FEA. Finally, the dual-range positioning system including the compliant stage is designed, fabricated and test results are presented. Unlike most of the available dual-range stages, this system uses the same actuator for both coarse and fine positioning in a parallel setup, resulting in a reduced cost, a reduced footprint and a faster response. This concept can be used for systems with larger workspace and with different types of sensor and actuator, so that the right cost/accuracy balance can be achieved.

Table 1. Initial geometrical parameters for the XY motion stage

Parameters	E (MPa)	σ_{max} (MPa)	P (kg/m3)	N
Values	71700	505	2810	0.33

2. Compliant XY Motion Stage Design

The structure of the compliant XY motion stage presented in this paper is based upon the stage designed in [22] and implemented in [23]. The dimensions of the beams are chosen to be 45mm in length to increase the range of motion, 6 mm in height to maintain a high stiffness along the Z-axis and 0.5mm in thickness to limit the force input requirement. As with most of the compliant stages reported in the literature, this stage is made of aluminium 7075-T6. This material is well known for its large reversible strain. In order to reduce the stress concentration, corners of 0.5mm radius have been added at both ends of the beams.

2.1. Force-displacement analysis

In order to size the actuators and define the travel range of the stage, FEA is carried out using ABAQUS.

Firstly, a buckling analysis is carried out, where the buckling point of the beams will determine the travel range of the stage. For short beams, the force required to cause buckling is often much higher than the force required to reach the material's yield strength. However, for longer or thinner beams, buckling can occur before the yield strength is reached. In the present case, buckling occurs at the inner beams (Fig. 1) when the stress at the centre increases suddenly and the difference between the input and the output displacement of the stage increases significantly. From Fig. 2, the buckling point is estimated to occur when the input displacement is around 4.05mm, corresponding to an input force of 351.3N.

To continue, a stress/strain analysis is carried out. An input displacement is gradually applied on one side of the stage until the yield strength of the material is reached. The output displacement, the maximum stress and the reaction force are recorded. In this case, the yield strength of 505MPa is attained when the input displacement is 3.24mm, corresponding to a reaction force of 208.1N. The yield strength of Aluminium 7075-T6 is therefore the main limit to the travel range of the stage. To ensure a long fatigue life and to limit the force input requirement, the travel range of the final design is limited to ±2.2mm × 2.2mm. This corresponds to an input force of approximately 70N.

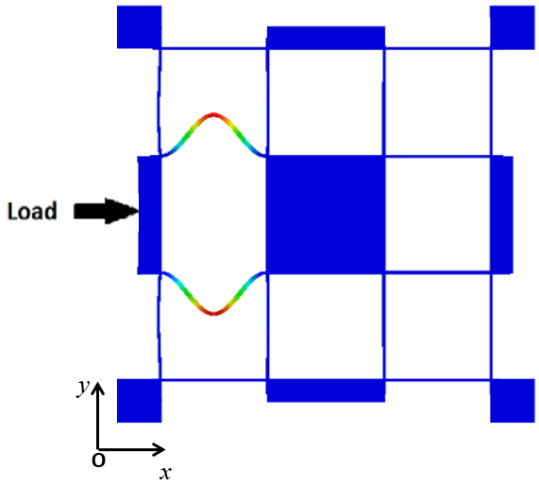


Fig. 1. Buckling point when a load is applied along X

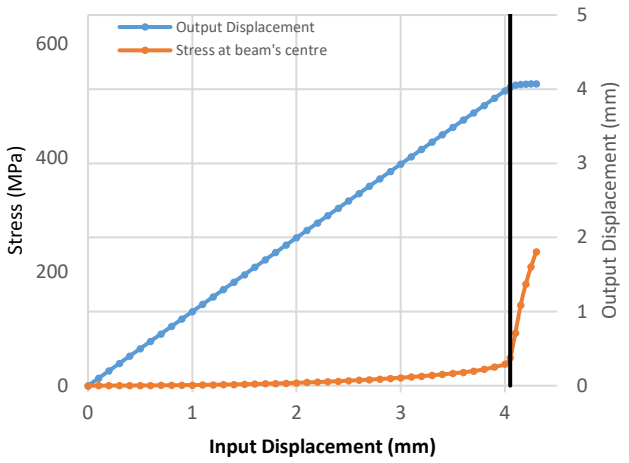


Fig. 2. Stress and output displacement response to a large input displacement

The output displacement is also compared with the input displacement. The axial deformation of the inner parallelogram beams causes the output displacement to be slightly different from the input displacement. The results show that for an input displacement between 0 and 2.2mm, the difference between the input and the output displacement can reach 5µm.

2.2. Coupling analysis

It is desirable for a compliant stage to have a minimal cross-coupling between the axes of motion. An analysis is therefore carried out to evaluate the parasitic displacement along the Y direction when a load is applied in the X direction. An input displacement of 2.2mm is first applied along the Y direction, as presented in Fig. 3. The maximum parasitic displacement is 65µm (Fig. 4), corresponding to a coupling of 3%. An input displacement of 2.2mm is then gradually applied along the X direction. The results show that the maximum displacement error along Y is 67µm (Fig. 5).

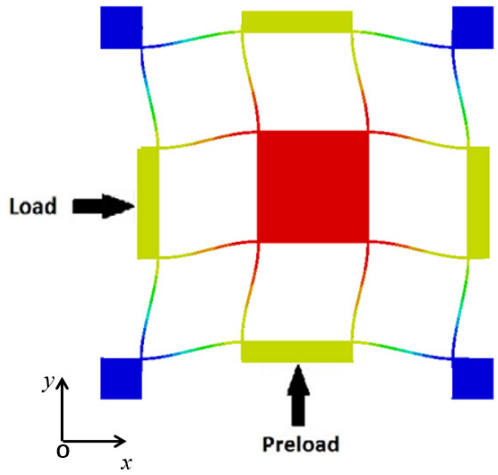


Fig. 3. Deformed XY motion stage with input displacement applied along X and Y directions

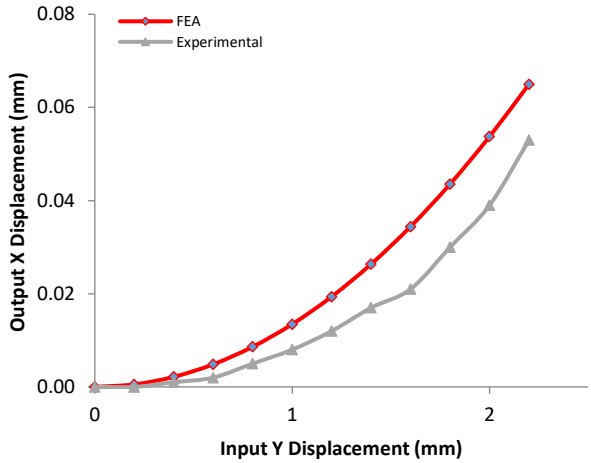


Fig. 4. Measurements of the parasitic displacement for single direction loading

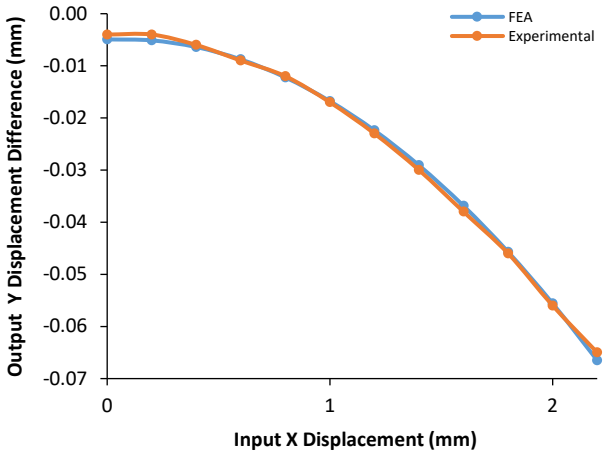


Fig. 5. Coupling error between X and Y

2.3. Dynamic analysis

In order to limit the effect of external vibrations, the first natural frequency of the compliant stage must be kept as high as possible. A high ratio between the first two natural frequencies and the third natural frequency is also desirable as it guarantees that the stiffness along the two directions of motion is much lower than the stiffness along all other directions, thus avoiding unwanted motion. The stiffness ratio of compliant stages usually lies between 2 and 5 [4, 5, 10, 24-26]. A dynamic analysis of the stage is carried out with ABAQUS using the Lanczos Eigen solver (Fig. 6). The first two modes correspond to simultaneous vibrations along both the X and Y-axes, which occurs at 55.8Hz. The third mode, corresponding to a rotation around the Z-axis occurs at 249.5Hz. Finally, the fourth mode, corresponding to vibrations along the Z axis occurs at a frequency of 313.3Hz. The ratio between the first two natural frequencies and the third one is greater than 4.

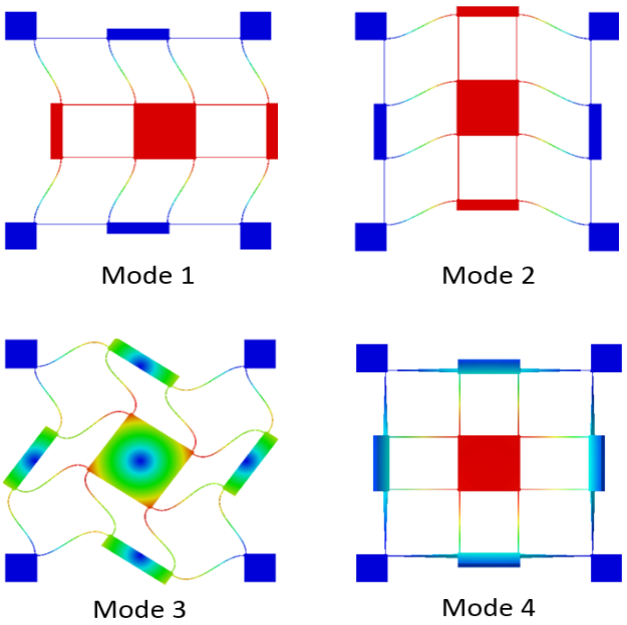


Fig. 6. Modal analysis results from FEA

3. Compliant XY Micro-motion Stage Fabrication and Testing

Most of the compliant stages reported in the literature are fabricated using Wire Electric Discharge Machining (WEDM). This technique allows for tight tolerances and a smooth surface finish. However, it is costly. Laser cutting and CNC machining were also considered but these processes are not suitable for machining 0.5mm thick beams. The first one would melt the beams and the second one would bend the beams because of the cutting force applied. The stage was fabricated using (Fig. 7). Corners of 0.5mm radius have been added to reduce the stress concentration. The stage is driven by two VCAs from Moticont (LVCM-051-064-02) with a 12.7mm stroke, 68.2N continuous force and a force constant of 21.6N/A. The moving coils are guided using high accuracy preloaded miniatures guides from SKF (LZMHS12-37T2P1). The VCAs are controlled by two Ingenia Pluto Drives, allowing for a position command resolution of 1.07µm over ±2.2mm along both directions using a 12-bit Analogue to Digital converter (ADC). The displacement of the actuators is measured using two linear encoders from Renishaw (ATOM4T0-150) with a resolution of 100nm.

3.1. Force-displacement test

A test is carried out to establish the relationship between the input displacement and the reaction force. The input displacement along the X-axis it maintained at 0mm while it is varied from 0mm to 2mm along the Y-axis. The linear encoders are used to measure the position of the VCAs. As the servo drives include a built-in current sensor, the force is derived from the current using the VCA's force constant. The results are shown in Fig. 8. The force error of the first few readings is due to the internal friction in the bearings and can be neglected. The maximum force error, occurring at 2mm, is 11.1% when compared with FEA.

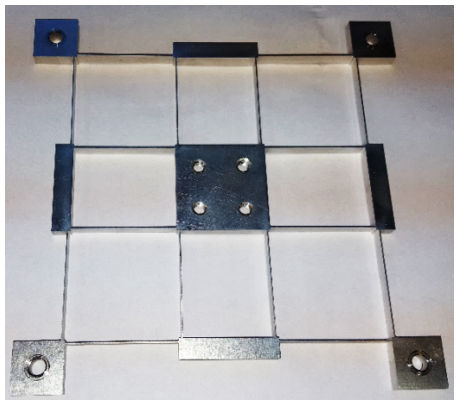


Fig. 7. Fabricated compliant XY micro-motion stage

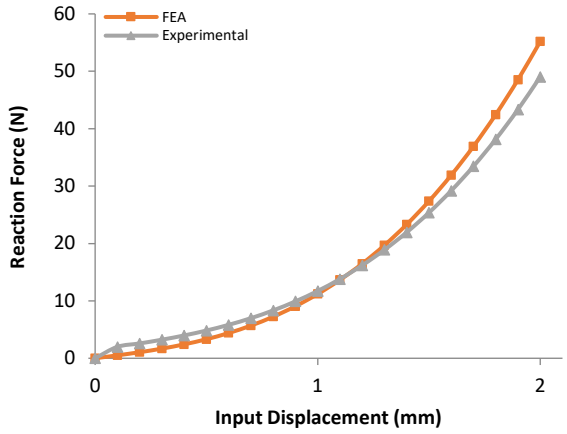


Fig. 8. Stiffness along a single direction

3.2. Frequency response analysis

In order to obtain the frequency response of the XY motion stage, a testing rig is set up to allow the stage to vibrate freely. An impulse is used to generate free vibrations which are measured along the X, Y and Z axes using a 3-dimensional accelerometer (ICP-T356A16), placed at the centre of the stage. The sensitivity of the accelerometer is 100mV/G and its output signal is processed by a Dual Channel Accelerometer Amplifier (FE-376-IPF) and recorded using a National Instruments data acquisition card (6008) with a sampling rate of 10kHz.

Labview is used to obtain the frequency domain response of each direction of motion using the Fast Fourier Transform (FFT). The testing rig setup is presented in Fig. 9. A peak in amplitude in each spectrum corresponds to the natural frequency along the corresponding direction of motion. The results are then compared with FEA, taking into account the mass of the accelerometer. The recorded resonant frequencies are 47Hz along the X and Y-axes and 267Hz along the Z-axis. When compared with FEA, the corresponding errors are 15.8% along the X and Y-axes and 14.8% along the Z-axis respectively. This error can be due to the mounts applying a small preload in the beams when being screwed.

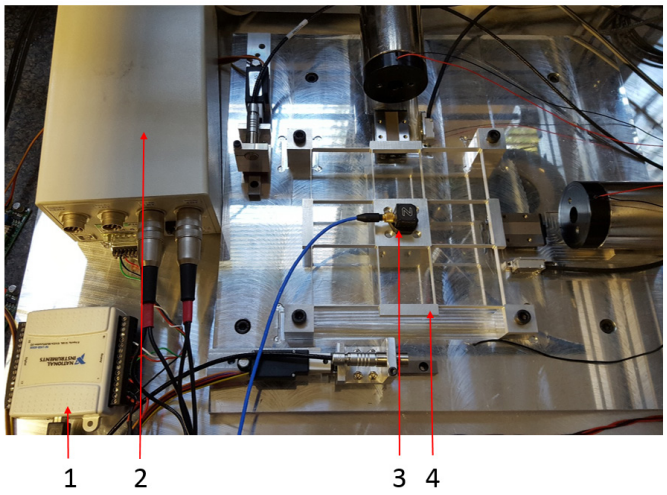


Fig. 9. Experimental vibration test setup. 1) Data acquisition card, 2) Amplifier, 3) 3D accelerometer, 4) Compliant XY motion stage

4. Dual-Range Positioning XY Micro-motion Stage

The assembled stage is presented in Fig. 10. It consists of a platform directly fixed to the centre of the compliant stage. Two small plates are mounted on the platform and are used as a target to measure the output displacement of the XY motion stage. The total moving mass is approximately 1kg.

4.1. Input Position Control

The input position control of the XY motion stage controls the linear displacement of the VCAs with a Proportion-Integral-Derivative (PID) loop using the two linear encoders as feedback. The output displacement (i.e. displacement of the moving platform) is measured with two capacitive sensor heads (CS1) from Micro-Epsilon with DT6110 controllers, giving a resolution of 100nm and a reading accuracy of 0.5µm over a range of 1mm. These sensors are mounted on a carrier and are moved along a guiding rail using two Firgelli L12-P micro linear actuators with a 10mm stroke and ±0.1mm repeatability.

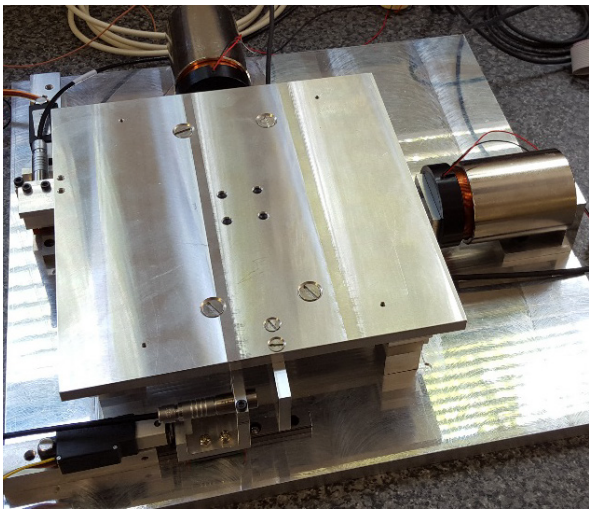


Fig. 10. Fabricated XY motion stage

The input displacements along the X and the Y directions are controlled with LABVIEW 2014 through the analogue outputs of a Labjack U6 Pro data acquisition card.

To begin with, a single direction displacement test is carried out over a range of 0mm to 2.2mm along Y. The results show that the difference between the input and the output displacement is 4µm and that the parasitic translation along the transverse direction is 53µm (Fig. 4), corresponding to a coupling of 2.4%.

A bi-directional loading test is then carried out. The input displacement of 2.2mm previously applied along the Y direction is maintained and an input displacement of 2.2mm is gradually applied along the X direction. The results are presented in Fig. 5. The first observation made is that the results are in accordance with FEA, where the maximum error is less than 3µm for 2.2mm input displacement. The second observation is that the maximum recorded parasitic displacement is 50µm.

4.2. Dual-range positioning

The observations made from the input position control test show that there is a coupling between the two axes of motion that cannot be predicted when moving the stage along both directions simultaneously. There is also a difference between the input and the output displacement of up to 14µm. To compensate for this positioning error, a set of high resolution sensors is added to the system in order to directly measure the output displacement of the stage. These sensors are mounted on a carrier and are moved along a guiding rail using two Firgelli L12-P micro linear actuators with a 10mm stroke and ±0.1mm repeatability. The dual-range positioning procedure for each direction of motion is as follow:

i. The stage is moved to its desired position in open-loop mode (coarse positioning)
ii. The high resolution sensors are moved close to the moving platform and the relative position is recorded and used as the initial position for the next step
iii. The positioning error is measured using an external system
iv. The stage is moved by the positioning error measured in closed-loop mode using the feedback from the high resolution sensors (fine positioning)

A second PID loop is added to control the output displacement of the stage using the readings from the high resolution sensors. The simplified plant model is presented in Fig. 11. The gains are tuned using the trial-and-error method and will be refined at a later stage if necessary.

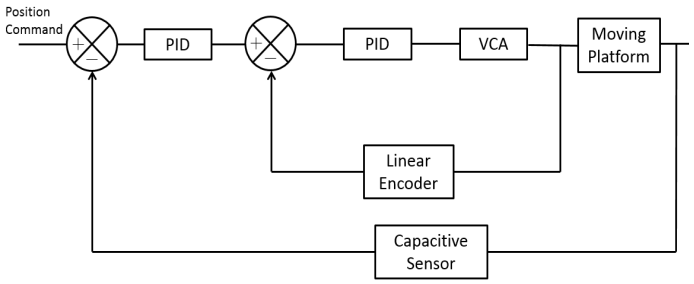


Fig. 11. Fine positioning control loop for the XY micro-motion stage

The external system used to measure the positioning error depends on the system in which the XY motion stage is integrated. This choice is up to the user. For instance, a high resolution vision system can be used to measure the position of the moving platform on which a calibration grid is placed. Alternatively, a coordinate measurement machine (CMM) was used in [23] to measure the positioning error.

Sensor selection

The choice of high resolution sensor is critical, however it involves a trade-off between the cost of the system and the measuring accuracy . Table 2 presents a qualitative analysis of the four types of sensors considered for this application based on quotations from RENISHAW®, KEYENCE Ltd. and MICRO-EPSILON Ltd. The most suitable solution in terms of this trade-off between cost and accuracy is the capacitive sensor. Therefore, the sensors used to measure the open-loop behaviour of the compliant stage are reused for the closed-loop control of the fine positioning, along with the linear actuators. The dual-range setup of the XY motion stage is presented in Fig. 12.

Testing

Since the capacitive sensors directly measure the position of the stage, they can be used to estimate the relative positioning accuracy. It is assumed that due to the nonlinear stiffness of the compliant stage, the force disturbance is higher for large displacements. Therefore, a test is carried out for an input displacement of 0mm along both directions and then for 2.2mm along both directions. The sensors' readings are recorded for a period of 20 seconds in fine positioning mode and the accuracy is estimated from the maximum amplitude of oscillation once the steady state is reached. Fig. 13 shows the steady state response along both directions for 0mm loading. It can be clearly seen that the maximum amplitude of oscillation is smaller than ±0.5µm, which is beyond the reading accuracy of the capacitive sensors. Fig. 14 shows the steady-state response for 2.2mm bidirectional loading. The amplitude of oscillation increases and reaches ±2µm. This oscillation is due to the nonlinearities arising from the compliant stage. Additionally, the resolution of the analogue to digital converter (ADC) used to read the position command sent by the data acquisition card is only 12-bits. This corresponds to an incremental position resolution of 1.12µm. Controllers with a higher resolution would allow for a higher positioning accuracy.

5. Conclusion

A compliant XY micro-motion stage made from Aluminium 7075-T6 was characterised and fabricated, with a travel range of ±2.2mm×2.2mm and a coupling of 2.4%. The simulation results in terms of working area and stiffness are in accordance with FEA. The maximum displacement error of the micro-motion stage measured in coarse positioning mode was 67µm. The concept of dual-range manipulation was therefore introduced using capacitive sensors for the fine positioning. The relative positioning accuracy was reduced to less than ±2µm in the worst-case scenario. This error can be reduced by using higher resolution motor controllers and sensors. Therefore, there is a great potential for compact, low-cost and high accuracy micro-motion using dual-range positioning. The stage will later be integrated in a hybrid miniaturised product assembly system.

Table 2. Sensors' rating

Sensor	Accuracy	Sensing Range	Cost	Environmental Sensitivity
Capacitive	+++	-	+	+
Interferometer	+++	+	--	+
Eddy current	+	--	-	+++
LVDT(1)	--	+++	+++	+++

(1) Linear Variable Differential Transformer

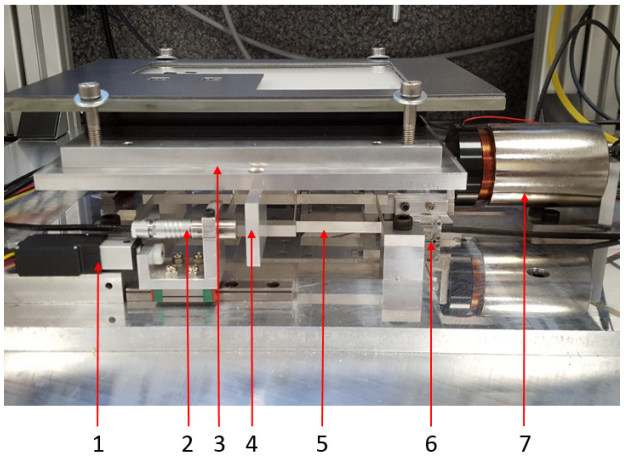


Fig. 12. Side view of the XY motion stage. 1) Linear actuator, 2) Capacitive sensor head, 3) Moving Platform, 4) Target, 5) Compliant XY stage, 6) Linear encoder, 7) VCA

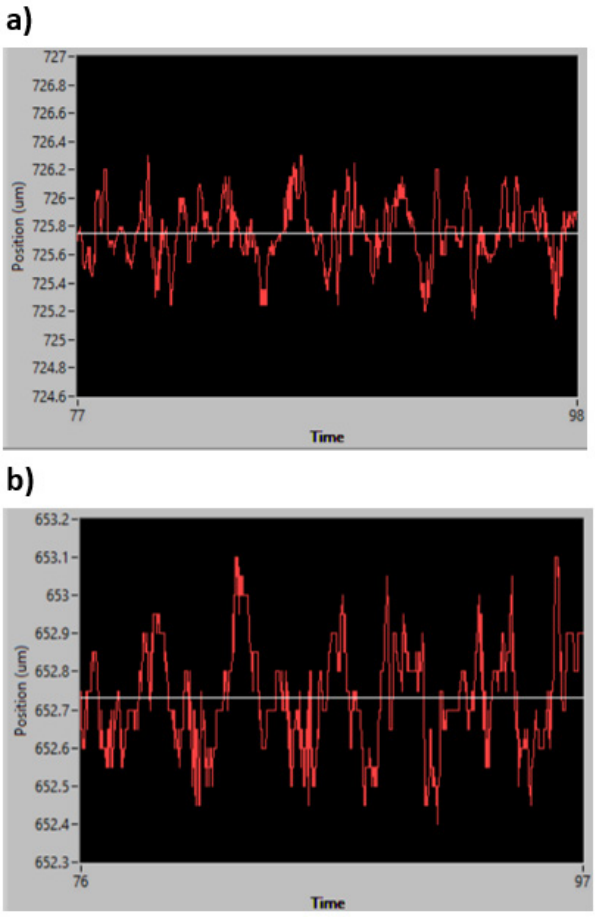


Fig. 13. Capacitive sensors readings in fine positioning mode at position 0mm along a) the X-axis, and b) the Y-axis

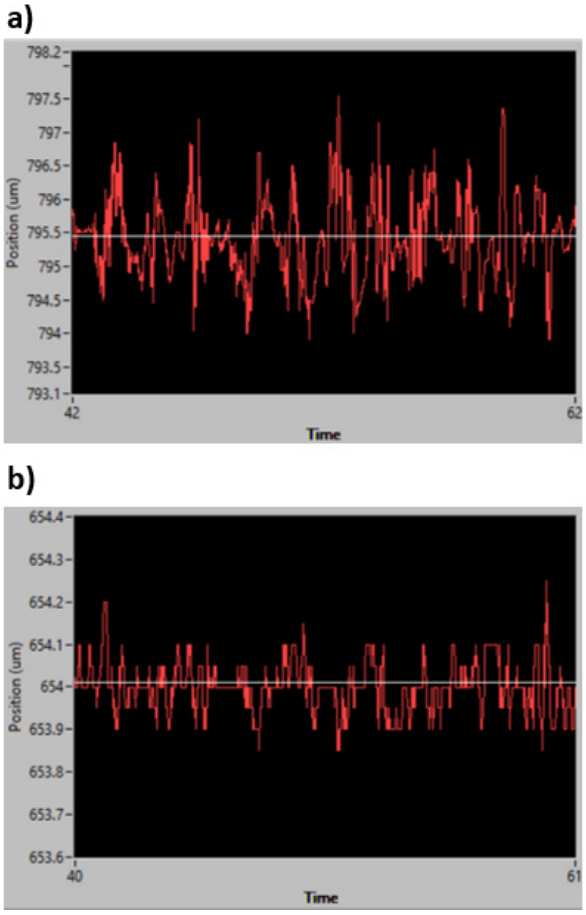


Fig. 14. Capacitive sensors readings in fine positioning mode at position 2.2mm along a) the X-axis, and b) the Y-axis

Acknowledgment

The authors would like to thank the financial support from the Engineering and Physical Sciences Research Council (EPSRC), UK under Grant No. EP/K018345/1.

References

[1] F. Wang, C. Liang, Y. Tian, X. Zhao, and D. Zhang, "A Flexure-Based Kinematically Decoupled Micropositioning Stage with a Centimeter Range Dedicated to Micro/nano Manufacturing," *IEEE/ASME Transactions on Mechatronics,* vol. PP, pp. 1-1, 2015.

[2] D. Jie, L. Sun, Y. Liu, Y. Zhu, and H. Cai, "Design and Simulation of a Macro-micro Dual-drive High Acceleration Precision XY-stage for IC Bonding Technology," in *Proceeding of the 6th International Conference on Electronic Packaging Technology*, 2005, pp. 161-165.

[3] Y. Tian, B. Shirinzadeh, and D. Zhang, "Design and dynamics of a 3-DOF flexure-based parallel mechanism for micro/nano manipulation," *Microelectronic Engineering,* vol. 87, pp. 230-241, 2010.

[4] G. Hao and J. Yu, "A Completely Kinematostatically Decoupled XY Compliant Parallel Manipulator through New Topology Structure," in *Proceedings of the IFToMM Workshop on Fundamental Issues and Future Research Directions for Parallel Mechanisms and Manipulators*, 2014.

[5] S. Xiao and Y. Li, "Design and analysis of a novel flexure-based XY micro-positioning stage driven by electromagnetic actuators," in *Proceeding of the International Conference on Fluid Power and Mechatronics*, 2011, pp. 953-958.

[6] Y. Li, J. Huang, and H. Tang, "A Compliant Parallel XY Micromotion Stage With Complete Kinematic Decoupling," *IEEE Transactions on Automation Science and Engineering,* vol. 9, pp. 538-553, 2012.

[7] J. Yu, Y. Xie, Z. Li, and G. Hao, "Design and Experimental Testing of an Improved Large-Range Decoupled XY Compliant Parallel Micromanipulator," *Journal of Mechanisms and Robotics,* vol. 7, pp. 044503-044503, 2015.

[8] G. Hao and X. Kong, "A Novel Large-Range XY Compliant Parallel Manipulator With Enhanced Out-of-Plane Stiffness," *Journal of Mechanical Design,* vol. 134, pp. 061009-061009, 2012.

[9] Q. Xu, "New Flexure Parallel-Kinematic Micropositioning System With Large Workspace," *IEEE Transactions on Robotics,* vol. 28, pp. 478-491, 2012.

[10] G. Hao, Q. Meng, and Y. Li, "Design of Large-range XY Compliant Parallel Manipulators Based on Parasitic Motion Compensation," in *Proceedings of the ASME International Design Engineering Technical Conferences & Computers and Information in Engineering Conference (IDETC/CIE)*, 2013

[11] C.-H. Liu, W.-Y. Jywe, Y.-R. Jeng, T.-H. Hsu, and Y.-t. Li, "Design and control of a long-traveling nano-positioning stage," *Precision Engineering,* vol. 34, pp. 497-506, 2010.

[12] Y. Park, "Precision motion control of a three degrees-of-freedom hybrid stage with dual actuators," *IET Control Theory & Applications,* vol. 2, pp. 392-401, 2008.

[13] Y. M. Choi and D. G. Gweon, "A High-Precision Dual-Servo Stage Using Halbach Linear Active Magnetic Bearings," *IEEE/ASME Transactions on Mechatronics,* vol. 16, pp. 925-931, 2011.

[14] S. Jing-Chung, W. Chia-Hung, C. Bor-Yu, and J. Wen-Yuh, "Control of a Long-stroke Precision Scanning Stage," in *Proceeding of the Mediterranean Conference of Control and Automation (MED)*, 2014, pp. 311-315.

[15] S. Jing-Chung, J. Wen-Yue, L. Qun-Zhong, and W. Chia-Hung, "Control of a High Precision Positioning Stage," in *Proceeding of the 7th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, 2012, pp. 931-935.

[16] Y. Song, J. Wang, K. Yang, W. Yin, and Y. Zhu, "A dual-stage control system for high-speed, ultra-precise linear motion," *Int J Adv Manuf Technol,* vol. 48, pp. 633-643, 2009.

[17] S. Kwon, W. K. Chung, and Y. Youm, "On the Coarse/fine Dual-stage Manipulators with Robust Perturbation Compensator," in *Proceeding of the IEEE International Conference on Robotics and Automation (ICRA)*, 2001, pp. 121-126 vol.1.

[18] L. Zhang, Z. Long, J. Cai, and J. Fang, "Design of a linear macro–micro actuation stage considering vibration isolation," *Advances in Mechanical Engineering,* vol. 7, pp. 1 - 13, 2015.

[19] F. Jiwen, L. Zhili, Z. Lufan, and N. Longsheng, "Driving Process and Control Analysis in Macro-micro Dual Stage," in *Proceeding of the IEEE International Conference on Information and Automation (ICIA)*, 2013, pp. 37-42.

[20] H. Tang and Y. Li, "Design, Analysis, and Test of a Novel 2-DOF Nanopositioning System Driven by DualMode," *IEEE Transactions on Robotics,* vol. 29, pp. 650-662, 2013.

[21] Q. Xu, "Design and Testing of a Novel XY Micropositioning Stage with Dual Ranges and Resolutions," in *Proceeding of the IEEE International Conference on Robotics and Automation (ICRA)*, 2014, pp. 2351-2356.

[22] X. Herpe, R. Walker, K. Xianwen, and M. Dunnigan, "Analysis and Characterisation of a Kinematically Decoupled Compliant XY Stage," in *Proceeding of the 21st International Conference on Automation and Computing (ICAC)*, 2015, pp. 1-6.

[23] X. Herpe, R. Walker, M. Dunnigan, and X. Kong, "Design, Fabrication and Testing of a Hybrid Micro-motion XY stage Driven by Voice Coil Actuators," in *Proceeding of the International Conference for Students on Applied Engineering (ICSAE)*, 2016, pp. 153-158.

[24] L.-J. Lai, G.-Y. Gu, and L.-M. Zhu, "Design and control of a decoupled two degree of freedom translational parallel micro-positioning stage," *Review of Scientific Instruments,* vol. 83, p. 045105, 2012.

[25] Y. Li and Z. Wu, "Design, analysis and simulation of a novel 3-DOF translational micromanipulator based on the PRB model," *Mechanism and Machine Theory,* vol. 100, pp. 235-258, 6// 2016.

[26] K.-B. Choi and D.-H. Kim, "Monolithic parallel linear compliant mechanism for two axes ultraprecision linear motion," *Review of Scientific Instruments,* vol. 77, p. 065106, 2006.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 505-512 (2017)

www.astesj.com

Special issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Self-Organizing Map based Feature Learning in Bio-Signal Processing

Marwa Farouk Ibrahim Ibrahim*, Adel Ali Al-Jumaily

[1]School of Mechanical and Electrical Engineering, Faculty of Engineering and IT, University of Technology Sydney, NSW 2007, Australia

ARTICLE INFO

Article history:
Received: 02 April, 2017
Accepted: 11 May, 2017
Online: 24 May, 2017

Keywords:
unlabeled data
unsupervised learning
feature learning
deep learning
bio-signal processing
self-organizing map
support vector machine
extreme learning machine
discriminate analysis
classifier fusion

ABSTRACT

Feature extraction is playing a significant role in bio-signal processing. Feature identification and selection has two approaches. The standard method is engineering handcraft which is based on user experience and application area. While the other approach is feature learning that based on making the system identify and select the best features suit the application. The idea behind feature learning is to avoid dealing with any feature extraction or reduction algorithms and to train the suggested model on learning features from input bio-signal by itself. In this paper, Self-Organizing Map (SOM) will be implemented as a feature learning technique to learn the model extract the features from the input data. Deep learning approach will be proposed by deploying SOM to learn features. In the proposed model, the raw data will be read then represented by using different signal representation as Spectrogram, Wavelet and Wavelet Packet.
The newly represented data will be fed to self-organizing map layer to generate features, and finally, the performance of the suggested scheme will be evaluated by applying different classifiers such as Support Vector Machine, Extreme Learning Machine, Evolutionally Extreme Learning Machine and Discriminate Analysis Classification. Analysis of Variance (ANOVA) and confidence interval for different classifiers will be calculated. As an improving step for the results, classifier fusion layer will be implemented to select the most accurate result for both training and testing set. Classifier fusion layer led to a promising training and testing accuracies.

1. Introduction

This paper is an extension of work originally presented in 2016 8th Cairo International Biomedical Engineering Conference (CIBEC) [1]. However, we will implement Self-Organizing Map as a feature learning technique after representing raw data by using different data representation techniques. Supervised learning is considered as highly limited technique despite being applied to various applications. Still, most of the applications require engineering handcraft extraction by using different algorithms. This means that the primary target is, to represent the data by using good feature representation. Whenever feature representation is good, classification error should be expected to be less.

However, engineering handcraft representation is still exhausting and time-consuming in addition to that, it depends on researcher experience. Many suggested feature learning techniques may be applied to enhance feature representation automatically and save time and effort. The chief governor on the performance of used feature learning technique is the classification error. The most approach for implementing feature learning belongs to deep learning concept. Deep learning started in 1986 when Rina Dechter presented the basics for first and second order deep learning [2]. Later, deep learning is considered as a branch of machine learning where, it depends on multilayer by implementing the algorithm, in each layer, to represent data. Each layer feature output should be the input to the next cascaded layer [3]. Each layer in deep learning is implemented by using a hidden layer of artificial neural networks [4], where features are being learned by using the artificial neurone as shown in fig 1 where a very simplified flow chart of deep learning steps. Deep learning learns features in a hierarchal concept. The learning is being processed layer by layer, and extracted notations are being learned from lower level layers [5]. Deep learning can be used in supervised learning by translating the data into useful representations and remove any redundancy in

*Marwa Farouk Ibrahim, 15 Broadway, Ultimo, NSW 2007, Australia, 0061405565739 & marwafaroukibrahim.ibrahim@student.uts.edu.au

representations or features. The significant advantage of deep learning is that it can be used in unsupervised learning, where data is unlabelled. Unlabelled data is more challenging than labelled one. Neural history compressors[6] and deep belief networks[7] are considered as an example for implementing deep learning for unsupervised learning.

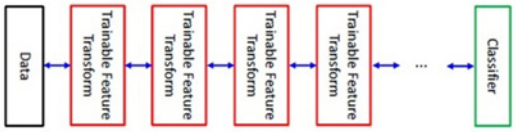


Fig.1. Deep Learning simplified steps

This paper is structured by giving a brief on previous work that was published in finger movement classification and deep learning implementation for the biomedical signal. Very simplified review on the self-organizing map and three bio signal representations (Spectrogram, Wavelet and Wavelet Packet) will be represented where, Discriminate Analysis, Support vector machine, Evolutional Extreme Learning Machine, and Extreme learning machine will be used as an evaluation step. Analysis of Variance (ANOVA) and confidence interval for different classifiers will be calculated. Finally, classifier fusion will be added to improve both training and testing accuracies.

2. Previous Work

In this paper, we will suggest a deep learning model that will be able to learn features by itself without utilising any feature extraction or reduction technique. The proposed model will be skilful in classifying between ten finger movements. Finger movement classification was presented previously in many types of research as in [8] where the researchers proposed a feature projection technique by implementing Fuzzy Neighbourhood Preserving Analysis (FNPA) with QR-decomposition. The primary purpose of Fuzzy Neighbourhood Preserving Analysis (FNPA) is to decrease the distance, to the maximum extent, between samples of the same class and increase it between samples of different classes. The researchers' purpose was to classify ten finger movements, and they verified average accuracy 91% by using only two channel electrodes. After that, other authors proposed a finger movement classification system. Where, they employed Spectral Regression Discriminant Analysis (SRDA) for dimensionality reduction, Extreme Learning Machine (ELM) for classification and smoothed the results by using majority voting method. They achieved classification accuracy 98% from two electromyography channels only [9]. Then the authors used the same data set collected by using two channel electrodes, which was considered a challenging task, to hire Spectral Regression Discriminate Analysis (SRDA) for dimensionality reduction, kernel-based Extreme Learning Machine (ELM) for classification and smoothed the results by using majority voting method [10]. Moreover, they applied Particle Swarm Optimization (PSO) to optimise the kernel based ELM. Later, other investigators suggested a model classifying between 9 finger movements (classes) by using two electrodes [11]. They implemented seven-time domain features and evaluated the system by applying Artificial Neural Network (ANN) and k- nearest neighbours (KNN) as classifiers. Other investigators suggested using Nonnegative Matrix

Factorization (NMF) technique to select accurate and dependable features. They employed neural networks for classification and calculated the accuracy for both simple and complex flexion where they achieved 95% for simple flexion and 87% for complex one [12]. After that other researchers suggested collecting surface electromyography signal using eight electrodes to classify between nine hand motions. They hired two different neural networks for classification. The first neural network architecture verified 83.43% accuracy while the second one achieved 91.85% [13]. Another experiment was published in [14] where the authors employed time-domain descriptors (TDD) to measure the power spectrum characteristics for electromyography signal which in turn reduced the computationally expensive feature construction traditional techniques. This suggested method improved the error percentage by 8% in comparison with other methods [14]. Another trial was published for recognising finger movements by using microneedle with average accuracy 94.9% [15]. On the other hand, many researchers claimed for implementing a general model that should be able to learn features by itself without employing any feature extraction and reduction techniques which led to deep learning clue. Many published experiments were done in hiring deep learning for biomedical data. As for example but not limited to, in [16] the authors developed a system by using an ensemble of convolutional neural networks to classify medical images. Convolutional neural networks were implemented for learning features by tunning and for classification as well. Other researchers suggested employing convolutional neural networks to learn features and developing low-level features only [17]. The authors in [18] developed a deep learning system to classify between different plaque constituents in Carotid Ultrasound. Later, the texture of lung pattern was analysed by implementing deep learning technique after employing convolutional neural networks [19]. Consequently, we can consider deep learning as a feature learning system where the system will extract the best features suit the application or in another word the system will learn features without employing any feature extraction or feature reduction algorithm. We will propose a deep learning model by engaging self-organising map in learning features from the represented bio signal. The bio signal will be represented using either spectrogram, wavelet or wavelet packet. We will evaluate the behaviour of our suggested model by hiring different classifiers as support vector machine, discriminant analysis, extreme learning machine and evolutionally extreme learning machine. Evolutionally extreme learning machine will lead to higher accuracy results than other implemented classifiers. Analysis of variance and confidence interval for different classifiers will be calculated as well. Finally, and as a development stage, classifier fusion layer will be added to select best local classifier. Accordingly, our proposed model will learn features by itself without hiring any feature extraction or reduction algorithm and verify well-accepted accuracy values. Moreover, implementing analysis of variance and confidence interval is considered as an in-depth step to understand the nature of the relation between our employed classifiers and to know our trustful range of accuracy results.

3. Self-Organizing Map

Self-Organizing Map is a typical artificial neural network which has two main characteristics. The first one is that it follows unsupervised learning technique to generate a lower dimensional copy of input training set. The new dimension of the data is typically two dimensions, and the new training set is called map. While the second characteristic is that it applies error correction by using back propagation technique or gradient descent and at the same time, keeps the properties for the input by using neighbourhood function [7]. Sometimes Self-Organizing Map is called Kohonen map in referring to Teuvo Kohonen who first invented it in 1980 [8]. Self-Organizing Map may represent the nonlinear version of principal component analysis [9]. The new two-dimensional data is considered as the features that will be used for classification steps. SOM is used in unsupervised learning. The network is being trained to generate features from the input data at the output instead of producing classes at the output. The input to the self-organizing map is the represented data either by spectrogram or wavelet or wavelet packet while the output is considered as the features. The classifier uses features produced from the self-organizing map as an input while; its output is the class equivalent to input data. Fig.2 shows the procedures of signal processing by using self-organizing map.

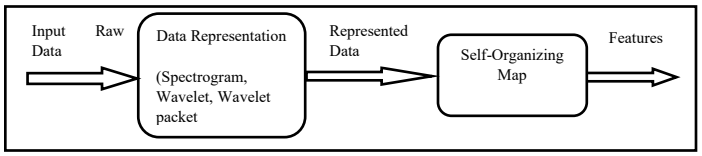


Fig.2. Procedures of SOM signal representation

Fig 3 shows the mapping of a point in the input space to its new point in the output space.

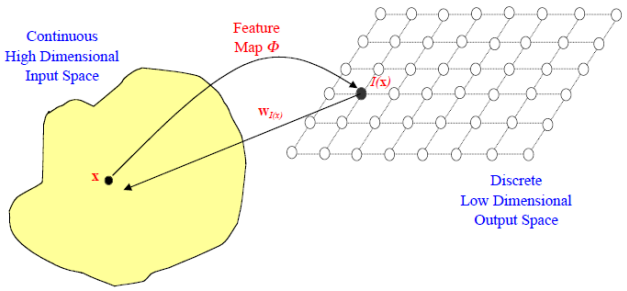


Fig.3. Mapping input space to output space in SOM

In this algorithm, the self-organizing map is being used to reduce the dimension of the data into lower and more useful representation. Consequently, self-organizing map enhances in extracting essential features or good representations of input data. So, the self-organizing map is in the context of feature learning where we have trained data. We will extract some features from training data then, apply the algorithm on the testing set and calculate the classification error.

4. Bio-Signal Processing

4.1. Bio Signal Representation

To improve the performance and to get lower classifier error, we suggest some signal representations can be done on the data before being introduced to the self-organizing map stage. The first data representation is to take the spectrogram of our input data. The spectrogram is to represent the spectrum of frequencies for our signal in a visual manner. Mathematically, Spectrogram can be calculated by taking the square of the magnitude of Short-Time Fourier Transform (STFT) or; it can be named short term Fourier transform. Short time Fourier transform is a Fourier-related transform, where it divides the long-time signal into shorter segments in time. These separated segments should be equal in length and then compute the frequency and phase for each segment separately. So, simply it is Fourier transform but for shorter segments than calculating it for the whole signal at one time[10]. The spectrogram output of the data is being trained by the self-organizing map to extract features then; these extracted features should be used in classification stage. Moreover, wavelet was applied as signal representation Wavelet is considered as small shifted and scaled partitions of the original signal. Fourier transform is representing the signal as a sinusoidal wave with different frequency while wavelet is representing the signal in the form of sharp changes. Wavelet allows us to capture better feature representation for abrupt changes in the signal. However, smooth representation by Fourier is useful in case of the smooth signal. We used Haar wavelet in our software. Fig.4 shows Fourier transform representation and wavelet representation. As a comparative study, we applied wavelet packet for the signal representation. We used six levels for sym10 at sampling frequency 2 kHz.

Fig.4. Fourier transforms representation and wavelet representation

4.2. Classifiers

We implemented four classification algorithms, where the first is Extreme Learning Machine (ELM), second is Discriminate Analysis (DA), third is Support Vector Machine (SVM), and the fourth is Self-Adaptive Evolutionary Extreme Learning Machine. We compared the accuracy for six different types of SVM (Linear SVM, Quad SVM, Cubic SVM, Fine Gauss SVM, Medium Gauss SVM and Coarse Gauss SVM) and picked up the most accurate result. Also, we compared Linear Discriminate Analysis (LDA) and Quadrature Discriminate Analysis (QDA) and selected the most accurate result. Fig.5 shows the suggested procedures for our implemented feature learning suggested model

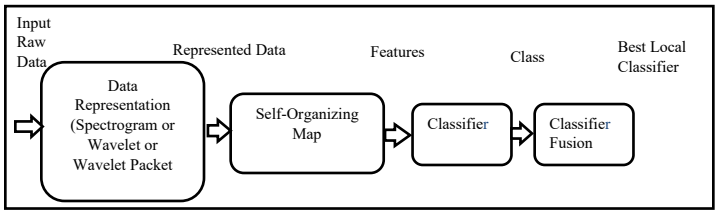


Fig.5. Procedures of Suggested Feature Learning Model

An extreme learning machine is being used in modelling complex systems either linear or nonlinear. The modelling algorithm uses gradient free fast convergence tool [20-23], for N observations $\{(x_i, t_i)\}^N$, where $x_i \in R^d$ and $t_i \in R^m$. ELM

chooses both the input weights and hidden biases in a random manner.

$$y_i = \sum_{j=1}^{L} \beta_j \, g(a_j . x_i + b_j) = t_i \quad , \; j = 1, \dots, N \qquad (1)$$

Where $a_j = [a_{j1}, a_{j2}, \dots, a_{jn}]^T \in R^d$ is the input weight, x_i is the input signal & $b_j = [a_{j1}, a_{j2}, \dots, a_{jn}]^T \in R \; (j = 1,2, \dots, L)$ is the bias. Those are the learning parameters of jth hidden node, $\beta_j = [\beta_{j1}, \beta_{j2}, \dots, \beta_{jn}]^T \in R^m$ is the output weight that relates output node and hidden node. Equation (1) can be simplified

$$H . \beta = T \qquad (2)$$

Where H is the hidden matrix and can be defined as

$$H = \begin{bmatrix} g(a_1 . x_1 + b_1) & \cdots & g(a_L . x_1 + b_L) \\ \vdots & \ddots & \vdots \\ g(a_1 . x_N + b_1) & \cdots & g(a_L . x_N + b_L) \end{bmatrix} \qquad (3)$$

So simply j th element of H represents the corresponding output vector of the jth hidden neuron with respect to ith input x_i , the values of H parameter is being selected in a random manner Subsequently, the estimate of the β_j^* , a_j^* , b_j^* parameters will be calculated using equation (4).

$$\|H(a_1^*, \dots, a_L^*, b_1^*, \dots, b_L^*)\beta^* - T\| = min_\beta \|H(a_1, \dots, a_L, b_1, \dots, b_L)\beta - T\| \qquad (4)$$

The cost function can be defined by the following equation

$$SSE = \sum_{i=1}^{N} \left[\sum_{j=1}^{L} \beta_j \, g(a_j . x_i + b_j) - t_i \right]^2 \qquad (5)$$

Where SSE is the sum square of error.

The least square solution of the $H . \beta = T$ can be determined as

$$\beta = H^{\dagger} T \qquad (6)$$

Where $H^{\dagger}$ is the Moore-Penrose generalized inverse of the hidden layer matrix H .

The self-adaptive evolutionary Extreme learning machine introduces the optimisation algorithm by applying differential evolutionary optimisation technique on objective function $f(\theta)$ by picking up a certain number of populations N_P and number of generation G until the targeted level of convergence obtained [24-28]. Where, the ith parameter can be defined as

$$\theta_{i,G} = \left[\theta_{i,G}{}^1, \theta_{i,G}{}^2, \dots, \theta_{i,G}{}^D \right] \; i = 1,2, \dots, N_P \qquad (7)$$

Where D is the described searching space.

The procedures of Differential Evolutionary can be defined as following

First, the cover parameter space should be initialized by using the following equation

$$\theta_{i,G} = \theta_{min} + rand(0,1) . (\theta_{max} - \theta_{min}) \qquad (8)$$

Where, θ_{min}& θ_{max} are the minimum and maximum boundaries respectively. The values of θ_{min}& θ_{max} are predetermined.

Second, generation of the new mutual vector by calculating difference vectors of randomly picked population vector. These can be calculated by one of four strategies

Strategy 1:

$$v_{i,G} = \theta_{r_1^i,G} + F . (\theta_{r_2^i,G} + \theta_{r_3^i,G}) \qquad (9)$$

Strategy 2:

$$v_{i,G} = \theta_{r_1^i,G} + F . \left(\theta_{best,G} - \theta_{r_1^i,G}\right) + F . \left(\theta_{r_2^i,G} - \theta_{r_3^i,G}\right) + F . (\theta_{r_4^i,G} - \theta_{r_5^i,G}) \qquad (10)$$

Strategy 3:

$$v_{i,G} = \theta_{r_1^i,G} + F . \left(\theta_{r_2^i,G} - \theta_{r_3^i,G}\right) + F . (\theta_{r_4^i,G} - \theta_{\mathrm{r_5^i,G}}) \qquad (11)$$

Strategy 4:

$$v_{i,G} = \theta_{i,G} + F . \left(\theta_{r_1^i,G} - \theta_{i,G}\right) + F . \left(\theta_{r_2^i,G} - \theta_{r_3^i,G}\right) \qquad (12)$$

Where r_1^i, r_2^i, r_3^i, r_4^i and r_5^i are mutually exclusive integers numbers which are randomly selected. $0 \leq F \leq 2$ is implemented to adjust the scaling difference, and the ruling parameter is randomly created as well within the range $0 \leq K \leq 1$.

Thirdly, the crossover is used to increase the varieties of the confused parameters vector with respect to mutant vector $v_{i,G} = [v_{i,G}{}^1, v_{i,G}{}^2, \dots, v_{i,G}{}^D]$ at generation G a trail vector $u_{i,G} = [u^1, u_{i,G}{}^2, \dots, u_{i,G}{}^D]$ is generated according to the following equation.

$$u_{i,G} = \begin{cases} v_{i,G}{}^j \; if \; (rand \; j \leq CR \; OR \; (j = jrand) \\ \theta_{i,G}{}^j \; Otherwise \end{cases} \qquad (13)$$

Where CR is the crossover rate to adjust the friction of the parameter values copied, rand j is jth evaluation of a uniform random number created in [0, 1] and jrand is randomly selected an integer from [1, D].

Fourthly, Selection is conducted using fitness function for each objective and corresponding trail vector, and the one at a lower value of fitness function is kept as population for the next generation. Then, second and fourth steps are repeated until the objective met or maximum iteration reached.

To overcome the restrictions of mutually selecting trail vector generation strategies and its associated ruling parameters, Self-Adaptive Evolutionary Extreme Learning Machine (SAEELM) is implemented to optimise the ELM output weight matrix using the following procedures

Initialize the hidden layer matrix

$$\theta_{i,G} = g\left(a^T{}_{1,(k,G)}, \dots, a^T{}_{L,(k,G)}, b^T{}_{1,(k,j)}, \dots, b^T{}_{L,(k,G)} \right) \qquad (14)$$

Where $j = 1,2, \dots, L$ are randomly generated, G generation, and $k = 1,2, \dots, N_P$ number of population, and then calculate the output matrix by using the following formula

$$\beta_{k,G} = H_{k,G}{}^{\dagger} T \quad (15)$$

$$RMSE_{k,G} = \sqrt{\frac{\sum_{i=1}^{N} \left\| \sum_{j=1}^{L} \beta_j \, g\left(a_{j,(k,G)} \cdot x_i + b_{j,(k,G)}\right) - t_i \right\|^2}{m.N}} \quad (16)$$

Where, $H_{k,G}$ and $\theta_{k,G+1}$ can be defined by using the following equations

$$H_{k,G} = \begin{bmatrix} g\left(a_{1,(k,G)} . x_1 + b_{1,(k,G)}\right) & \cdots & g\left(a_{L,(k,G)} . x_1 + b_{L,(k,G)}\right) \\ \vdots & \ddots & \vdots \\ g\left(a_{1,(k,G)} . x_N + b_{1,(k,G)}\right) & \cdots & g\left(a_{L,(k,G)} . x_N + b_{L,(k,G)}\right) \end{bmatrix} \quad (17)$$

$$\theta_{k,G+1} = \begin{cases} u_{k,G+1} \; if \; RMSE_{\theta k,G} - RMSE_{uk,G+1} > \varepsilon . RMSE_{\theta k,G}, \\ u_{k,G+1} \; if \; \left| RMSE_{\theta k,G} - RMSE_{u\,k,G+1} \right| < \varepsilon . RMSE_{\theta k,G} \\ \qquad and \; \left\| \beta_{uk,G+1} \right\| < \left\| \beta_{\theta k,G} \right\| \\ \theta_{k,G} \quad elsewhere \end{cases} \quad (18)$$

All trail vectors are $u_{k,G+1}$ created the $(G+1)^{th}$ generation are evaluated by implementing equation 16 with small tolerance ε.

The Analysis of Variance (ANOVA) was calculated for different classifiers. Where, we grouped average testing accuracies for three signal representation techniques (Spectrogram, Wavelet and Wavelet Packet) that result in P value equalled to 0.0613. The P value indicated that there was no practical difference between any of these four classifiers as P value was larger than 0.05 although its value was not that far from 0.05.

Moreover, the Confidence interval for each classifier was calculated with confidence score equals 60%. Generally speaking, confidence interval has two limits one is a higher limit, and the other is the lower one. We are confident in our result by level 60 % as long as testing accuracy, for each classifier, within our interval.

5. Implementation

In this section the data acquisition that we followed will be displayed and explained in more details and results will be discussed.

5.1. Data Acquisition

We used FlexComp Infiniti™ to collect the surface Electromyography signal. MyoScan™ T9503M Sensors were used to obtain two EMG. Those two sensors were put on the subject forearm as shown in Fig.6.

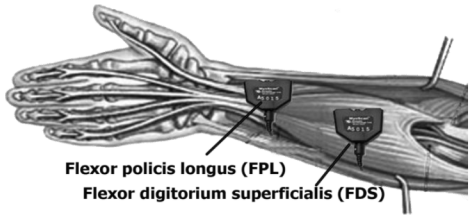


Fig.6. Placement of the electrodes

We have nine subjects where each subject was asked to do one of ten different finger movements for five seconds then take rest for other five seconds. Each movement was reiterated for six times. Then, the subject was asked to perform the same scenario for another finger movement class till finishing the whole ten finger movement classes. The signal was amplified by gain 1000 and sampled at the rate of 2000 sample per second.

The main target is to use collected EMG signal and feature learning technique and classify the finger movement for the subjects into ten classes. Fig.7 shows the targeted ten classes

Three-folded validations for data was used where the training set was 2/3 of the whole data, and the residual 1/3 was assigned for testing set.

The bio signal was filtered to remove any noise that might be superimposed on data. So, the implemented filter was to ensure that our processed signal is free from noise.

The average training and testing accuracy were calculated by simulating the system for each subject separately then summing the training or testing accuracy for all participants and dividing the result by the number of our subjects.

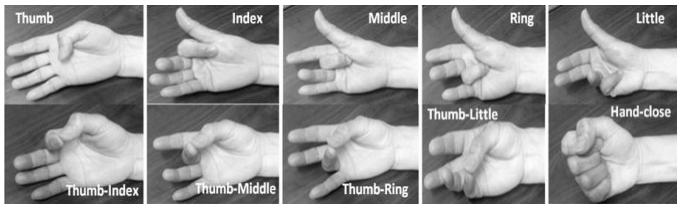


Fig.7. Ten different finger movements

5.2. Results

Table I shows training and testing accuracies for different signal representations and different classifiers

TABLE I. TESTING AND TRAINING ACCURACY

Signal Representation	Average Training Accuracy	Average Testing Accuracy	Classification Algorithm
Spectrogram	99.41%	81.54%	Support Vector Machine
Spectrogram	98.26%	83.29%	Extreme Learning Machine
Spectrogram	96.71%	91.11%	Evolutionally Extreme Learning Machine
Spectrogram	96.95%	83.13%	Discriminate Analysis
Wavelet	99.84%	85.11%	Support Vector Machine
Wavelet	98.03%	84.91%	Extreme Learning Machine
Wavelet	99.35%	93.14%	Evolutionally Extreme Learning Machine
Wavelet	98.61%	86.77%	Discriminate Analysis
Wavelet Packet	98.42%	90.47%	Support Vector Machine
Wavelet Packet	97.83%	87.96%	Extreme Learning Machine
Wavelet Packet	99.50%	94.44%	Evolutionally Extreme Learning Machine
Wavelet Packet	96.30%	89.98%	Discriminate Analysis

We went through different experiments by altering self-organizing map configuration until we reached the current configuration where the size of the two-dimensional map was 10x10, the layer topology function employed was 'hextop' and the neurone distance function hired was 'linkdist'. This configuration was to compromise between simulation time and accuracy level. These parameters were optimum regarding the accuracy level and the time consumed in simulation in which, increasing size of our map would recall higher simulation time without a tangible impact on accuracy level. The bio signal was filtered before signal representation stage and neural networks phase to guarantee a substantial barrier against noise level that might be superimposed on our collected bio signal.

The same set of data was introduced to a pattern recognition system, as shown in Fig.8, where we applied the traditional feature extraction and feature reduction techniques. The extracted features were slope sign change, zero crossing, waveform length, skewness, mean absolute value, root mean square and autoregressive. Also, we applied linear discriminant analysis as feature reduction method. The training accuracy was 96.60% and testing accuracy were 88.92%.

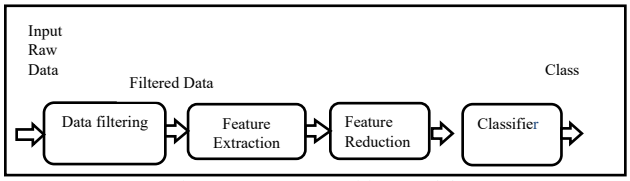


Fig.8. Pattern Recognition Model

It is evident from the results that are shown in Table I that using wavelet packet as a signal representation generated a higher testing accuracy than both wavelet and Spectrogram. Moreover, wavelet had a higher testing accuracy than Spectrogram and had a lower testing accuracy than wavelet packet. Support vector machine showed impressive classification ability for wavelet packet than its ability in spectrogram signal representation. Support vector machine testing accuracy for wavelet packet signal representation was very close to testing accuracy for discriminate analysis classifier but still higher. Moreover, support vector machine showed moderate classification ability for wavelet signal representation and poor classification ability for spectrogram signal representation. On the other side, Extreme learning machine excelled the other classification techniques for spectrogram signal representation. However, testing accuracy for spectrogram signal representation was very close to its value resultant from using Discriminate Analysis classifier. Discriminate analysis showed higher testing accuracy for wavelet signal representation than both extreme learning machine and support vector machine for the same data set and wavelet signal representation while, testing accuracy for wavelet signal representation by using support vector machine was higher than testing accuracy resultant from extreme learning machine for same wavelet signal representation but, close and lower than testing accuracy for discriminate analysis. The testing accuracies, for three signal representations, were moved to another higher level by implementing Evolutionally Extreme Learning Machine as a classifier and this is due to the optimisation algorithm that is being followed by this technique. The only drawback for evolutionally extreme learning machine was its simulation time as it was considered longer than the time consumed during simulation of an extreme learning machine. However, the high accuracy levels gained from evolutionally extreme learning machine was very encouraging to implement it.

We calculated the ANOVA value for four different classifiers. The P value resulted in 0.0613, which is slightly less than 0.05. Although the P value is slightly smaller than 0.05 However, it is yet indicating that there is no sensible difference between those four implemented classifiers

Table II shows the confidence interval for four different performed classifiers

TABLE II. CONFIDENCE INTERVAL for various CLASSIFIERS

Classifier	Confidence Interval High	Confidence Interval Low	Confidence Interval
Extreme Learning Machine	86.54%	84.24%	2.3%
Discriminate Analysis	88.29%	84.96%	3.32%
Support Vector Machine	87.89%	83.53%	4.36%
Evolutionally Extreme Learning Machine	93.71%	92.08%	1.63%

Fig.9. Shows the clustered column chart which is representing our confidence interval for our used classifier. The used confidence score is 60%. We are confident in our result by level 60 % as long as testing accuracy, for each classifier, within our interval. It is apparent from these results that evolutionally extreme learning machine has the narrowest range which is 1.63%. Support vector machine has the widest interval, which impacts the gap between higher and lower confidence interval. Moreover, the second widest interval is Discriminate Analysis which results in 3.32% and finally, extreme learning machine generates the second narrowest interval that is 2.3%.

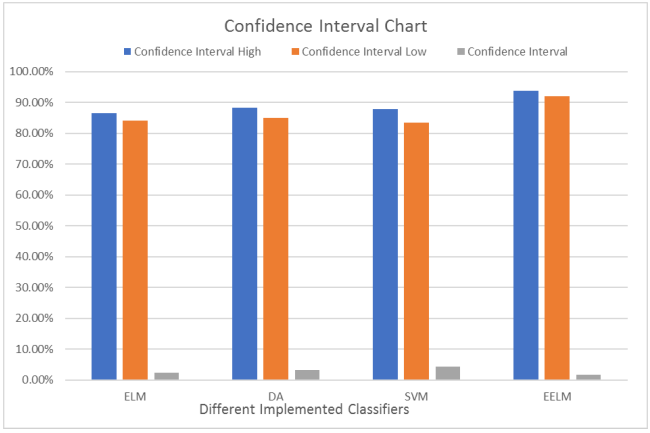


Fig.9. Confidence Interval

Classifier fusion layer was added where; it selected the best local classifier by using dynamic classifier selection algorithm. Both training and testing accuracies were improved as shown in Table III.

TABLE III: CLASSIFIER FUSION TESTING AND TRAINING ACCURACY

Signal Representation	Average Training Accuracy	Average Testing Accuracy	Classification Algorithm
Spectrogram	98.42%	93.07%	Classifier Fusion
Wavelet	99.35%	95.25%	Classifier Fusion
Wavelet Packet	99.73%	96.60%	Classifier Fusion

The testing accuracies were improved by adding classifier fusion where wavelet packet testing accuracy exceeded 96% while, wavelet testing accuracy became higher than 95%, and finally, spectrogram became more than 93%. In comparing Table III results with Table I results we can conclude that classifier fusion improved results. Fig.10 shows the number of learned feature for each data representation.

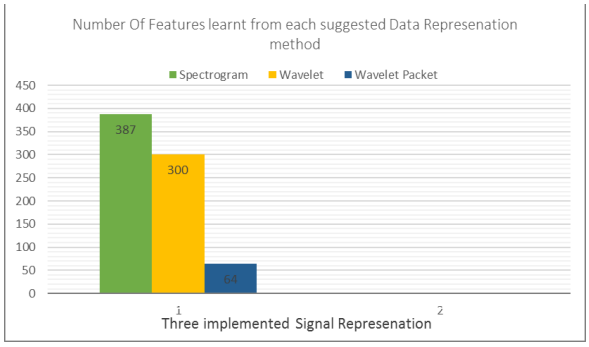


Fig.10. Number of learned features for each data representation

6. Conclusion

The self-organizing map was time-saving and easy feature learning technique, where it reduced dimensions of data without too much loss of the information and observations that should be associated with data. The self-organizing map was considered one of the types of neural networks which use dimension reduction (typically two dimensions) and back propagation for error correction. Although the simulation time was not long, we got a good testing accuracies.

Spectrogram, wavelet and wavelet packet helped in better representing raw data. Support vector machine showed good classification ability in Wavelet Packet while discriminate analysis showed an impressive result in wavelet signal representation and good ability for classification for Spectrogram although Extreme learning machine excelled in Spectrogram testing accuracy than using discriminate analysis.

Evolutionally extreme learning machine was considered a very real improvement for our accuracies percentages. This was due to the optimisation algorithm that existed in the model. As all testing accuracies, for evolutionally extreme learning machine, exceeded 90%. The only drawback for evolutionally extreme learning machine was its simulation time which was relatively longer than that consumed during implementing extreme learning machine. On the other hand, the accuracy values achieved by employing evolutionally extreme learning machine as a classifier was higher than those reached by hiring extreme learning machine in the classification stage.

ANOVA value was 0.0613 as calculated above. The value was not yet to be less than 0.05. However, it was very close to 0.05. This meant that there was no practical difference between classifiers. Despite this, we believed that using evolutionally extreme learning machine pushed ANOVA value to be as close as 0.05.

The confidence interval was calculated by confidence score 60%. And as shown above, the narrowest interval was for an evolutionally extreme learning machine. While the widest range for support vector machine. These intervals gave us an indication of how far we were confident in our resulting testing accuracy. And if the testing accuracy was in our range, so we would be confident in our result by 60%.

Classifier Fusion layer enhanced in increasing testing accuracies for the three used signal representations. Classifier fusion algorithm followed was selecting best local classifier. The only drawback for classifier fusion was increasing simulation time as it became relatively high than before. However, simulation time consumed by the self-organizing map as feature learning was short.

The average simulation time was around 20 seconds which was considered as a short time. However, it was very far from real time hands applications (150m seconds to 300 m seconds). As a future work we are planning to mix all features for three suggested data representation and select top useful features by applying indexing technique then use classifier fusion to combine between reaching real simulation time in real hands applications (150m seconds to 300 m seconds) and verifying high testing accuracies rate. Also, as a future work, we are planning to introduce raw data to our model without any signal representation to save simulation time and build a practical system.

References

[1] M. F. I. I. A. A. Al-Jumaily, "PCA indexing based feature learning and feature selection," presented at the 2016 8th Cairo International Biomedical Engineering Conference (CIBEC), Cairo, Egypt, 2017.

[2] R. Dechter, "Learning While Searching in Constraint-Satisfaction-Problems," presented at the Proceedings of the 5th National Conference on Artificial Intelligence, Philadelphia, 1986.

[3] L. Y. Deng, D, "Deep Learning: Methods and Applications," *Foundations and Trends in Signal Processing,* vol. 7, pp. 3-4, 2014.

[4] Y. Bengio, "Learning Deep Architectures for AI," *Foundations and Trends in Machine Learning,* vol. 2, no. 1, pp. 1-127, 2009.

[5] Y. C. Bengio, A.; Vincent, P., "Representation Learning: A Review and New Perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 35, no. 8, pp. 1798–1828, 2013.

[6] J. Schmidhuber., "Learning complex, extended sequences using the principle of history compression," *Neural Computation,* vol. 4, pp. 234–242, 1992.

[7] G. E. Hinton, "Deep belief networks," *Scholarpedia,* vol. 4, no. 5, p. 5947, 2009.

[8] S. K. Rami N. Khushaba, Dikai Liu, Gamini Dissanayake, "Electromyogram (EMG) Based Fingers Movement Recognition Using Neighborhood Preserving Analysis with QR-Decomposition," presented at the 2011 Seventh International Conference on Intelligent Sensors, Sensor Networks and Information Processing, Adelaide, SA, Australia, 2011.

[9] R. N. K. a. A. A.-J. Khairul Anam, "Two-Channel Surface Electromyography for Individual and Combined Finger Movements," presented at the 35th Annual International Conference of the IEEE EMBS, Osaka, Japan, 2013.

[10] K. A. a. A. Al-Jumaily, "Swarm-based Extreme Learning Machine for Finger Movement Recognition," presented at the 2014 Middle East Conference on Biomedical Engineering (MECBME), Doha, Qatar, 2014.

[11] M. H. P. C. B. V. Rao, "EMG signal based finger movement recognition for prosthetic hand control," presented at the Communication, Control and Intelligent Systems (CCIS), 2015, Mathura, India, 2015.

[12] G. R. N. a. H. T. Nguyen, "Nonnegative Matrix Factorization for the Identification of EMG Finger Movements: Evaluation Using Matrix Analysis," *IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS,* vol. 19, no. 2, pp. 478-485, 2015.

[13] S. H. Rocio Alba-Flores, and A. Shawn Mirzakani, "Performance Analysis of Two ANN Based Classifiers for EMG Signals to Identify Hand Motions," presented at the SoutheastCon, 2016, Norfolk, VA, USA, 2016.

[14] A. H. A.-T. Rami N. Khushaba, Ahmed Al-Ani, and Adel Al-Jumaily, "A Framework of Temporal-Spatial Descriptors based Feature Extraction for Improved Myoelectric Pattern Recognition," *IEEE Transactions on Neural Systems and Rehabilitation Engineering,* vol. PP, no. 99, 2016.

[15] D. S. K. Minjae Kim, and Wan Kyun Chung, "Microneedle-Based High-Density Surface EMG Interface with High Selectivity for Finger Movement Recognition," presented at the 2016 IEEE International Conference on Robotics and Automation (ICRA), Stockholm, Sweden, 2016.

[16] J. K. shnil Kumar, David Lyndon, Michael Fulham, and Dagan Feng, "An Ensemble of Fine-Tuned Convolutional Neural Networks for Medical Image Classification," *IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS,* vol. 21, no. 1, pp. 31-40, 2016.

[17] Y. Z. Ruikai Zhang, Tony Wing Chung Mak, Ruoxi Yu, Sunny H. Wong, James Y. W. Lau, and Carmen C. Y. Poon, "Automatic Detection and Classification of Colorectal Polyps by Transferring Low-Level CNN Features From Nonmedical Domain," *IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS,* vol. 21, no. 1, pp. 41 - 47 2016.

[18] A. G. Karim Lekadir, An` angels Betriu, Maria del Mar Vila, Laura Igual, Daniel L. Rubin, Elvira Fernandez, Petia Radeva, and Sandy Napel, "A Convolutional Neural Network for Automatic Characterization of Plaque Composition in Carotid Ultrasound," *IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS,* vol. 21, no. 1, pp. 48 - 55, 2016.

[19] M. A. Stergios Christodoulidis, Lukas Ebner, Andreas Christe, and Stavroula Mougiakakou, "Multisource Transfer Learning With Convolutional Neural Networks for Lung Pattern Analysis," *IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS,* vol. 21, no. 1, pp. 76 - 84 2016.

[20] Z. Q.-Y. Huang G-B, Siew C-K, "Extreme learning machine: a new learning scheme of feedforward neural networks," *In Proceedings of the international joint conference on neural networks, Budapest, Hungary,* pp. 985–990, 2004.

[21] Z. Q.-Y. Huang G-B, Siew C-K, "Extreme learning machine: theory and applications.," *Neurocomputing,* vol. 70, no. 1-3, pp. 489–501, 2006.

[22] J. Tang, C. Deng, and G. B. Huang, "Extreme Learning Machine for Multilayer Perceptron," *IEEE Transactions on Neural Networks and Learning Systems,* vol. 27, no. 4, pp. 809-821, 2016.

[23] Y. Miche, A. Sorjamaa, P. Bas, O. Simula, C. Jutten, and A. Lendasse, "OP-ELM: Optimally Pruned Extreme Learning Machine," *IEEE Transactions on Neural Networks,* vol. 21, no. 1, pp. 158-162, 2010.

[24] J. Cao, Z. Lin, and G.-B. Huang, "Self-Adaptive Evolutionary Extreme Learning Machine," *Springer Science+Business Media, LLC. 2012,* 18 July 2012 2012.

[25] L. D. S. Pacifico and T. B. Ludermir, "Evolutionary extreme learning machine based on particle swarm optimisation and clustering strategies," in *Neural Networks (IJCNN), The 2013 International Joint Conference on*, 2013, pp. 1-6.

[26] N. Zhang, Y. Qu, and A. Deng, "Evolutionary Extreme Learning Machine Based Weighted Nearest-Neighbor Equality Classification," in *2015 7th International Conference on Intelligent Human-Machine Systems and Cybernetics*, 2015, vol. 2, pp. 274-279.

[27] K. F. Man, K. S. Tang, and S. Kwong, "Genetic algorithms: concepts and applications [in engineering design]," *IEEE Transactions on Industrial Electronics,* vol. 43, no. 5, pp. 519-534, 1996.

[28] S. Jie, Y. Yongping, W. Peng, Y. Liu, and H. Shuang, "Genetic algorithm-piecewise support vector machine model for short-term wind power prediction," in *Intelligent Control and Automation (WCICA), 2010 8th World Congress on*, 2010, pp. 2254-2258.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 2, No. 3, 513-519 (2017)
www.astesj.com
Special Issue on Recent Advances in Engineering Systems

ASTES Journal
ISSN: 2415-6698

A delay-dependent distributed SMC for stabilization of a networked robotic system exposed to external disturbances

Fatma Abdelhedi [*1], Nabil Derbel[2],

[1]*Digital research center of Sfax, technopole of Sfax*, 3021, *Tunisia*

[2]*University of Sfax, Electrical Engineer, Sfax Engineering School*, 3000, *Tunisia*

ARTICLEINFO

Article history:
Received: 05 April, 2017
Accepted: 11 May, 2017
Online: 01 June, 2017

Keywords :
Sliding Mode Control(SMC)
Robotic systems
Synchronization
Communication time-delay
Multi-agent systems
Stability measurement errors
Disturbances

ABSTRACT

In this paper, a delay-dependent controller based on the sliding mode concept is proposed to stabilize a networked robotic system in a decentralized synchronization scheme. In addition to being affected by communication time-delays between agents, and seen that external disturbances obviously affect any physical and dynamic system, an unsettling action resulting from measurement errors affects the position and the velocity state vectors of agents. Then, it has been proved that the synchronizing algorithm enables cooperative agents, acting in a disturbed environment, to efficiently accomplish a shared task and to compensate delayed communication data. The proposed controller has been implemented in a cooperative robotic system, where the controller proves its robustness face to disturbances.

1 Introduction

The need for robust robotic controllers, that respond to greater productivity and faster manufacturing devices, has been a hot topic in the last few years. This need obviously imposes high pressure on researchers to acquire secure and effective robots, especially with the increasing humanitarian and social exigency, which require interactions between robots and humans. Thence, this paper presents an extension of a research work originally presented in the 8th International Conference on Modelling, Identification and Control [1]. Recently, several similar synchronization studies have been investigated in several complex mechanical systems, where flexibility and accuracy have been highly recommended [2], [3], [4], [5], [6]. There exist numerous samples of sophisticated technologies dedicated to modern life, such as material transporting [7], [8], advanced agriculture aggregation [9], building tasks [10], etc. Indeed, in light of the accelerated pace of development, there has become a growing demand for cooperative multi-agent robotic systems to serve contemporary industries and manufacturing applications [11], [12]. In this context, it has been proved that the flexibility and the maneuverability guaranteed by a cooperative robotic system are far beyond the working capacity of one single manipulator. In the other hand, the time-delay phenomenon obviously appears in nature [13], especially in several control systems (aircraft, chemical or process control systems) either in the control input, in the state or in measurements. In contrast to the ordinary differential equations, delayed systems are infinite dimensional in nature, and time-delayed data transmission presents in the most of cases, a source of instability. As an illustration, a state estimator is designed in [14] to a class of uncertain discrete-time Markovian jump neural networks, with time-varying delays, in order to estimate the network states through available output measurements. Indeed, the study presented in [15] seeks to resolve the problem of data packet dropout and transmission delays resulted by communication channels in the networked control system. So, great efforts have been invested in the synthesis of uncertain networked systems with time-delay, seen that the stability issue as well as performances of the system control with delay are both of theoretical and practical importance. However, in several robotic applications, the uncertainty may arise from many real-world sources such as errors of the robot modeling, unpredictable robot actuation, unpredictable movements in the environment, etc [16], [17]. As well, while robots are executing their assigned global task, some assumptions under which controllers were built may be invalidated: loads or robots may fail, the environment may change, etc. The cumulative effect of

[*]Fatma Abdelhedi, Sfax–Tunisia, fatma.abdelhedi@live.com

https://dx.doi.org/10.25046/aj020366

such uncertainty sources can be hard to manage during the task execution. Thus, to ensure a fluid response to the distributed multi-robot system, tradeoffs between multi-robot flexibility, time delay adjustment, uncertainty compensation, and task execution efficiency should be considered. To achieve it, we present the stability analysis of a robust control concept which combines the SMC design with a particular class of the graph theory, aiming to stabilize a networked robotic system.

2 Introduction to the SMC

While considering the SMC concept, the state trajectory evolution is divided into two parts [18]:

- from the initial state to the intersection with the sliding surface, this part is called "the reaching phase",
- from the intersection with the sliding surface to the origin, this part is called " the sliding phase".

During the reaching phase, the attractiveness condition of the sliding surface $\dot{S}(x)S(x) < 0$ is verified. This condition is global but it does not guarantee a finite sliding time. To ensure a finite sliding time, the condition becomes:

$$\dot{S}(x)S(x) < -\varepsilon_1 - \varepsilon_2|S| \;\; if \;\; S(x) \neq 0 \; and \; \varepsilon > 0 \qquad (1)$$

This sliding time can be imposed by choosing the sliding surface of the form:

$$\dot{S}(x) = \; -\mu S(x) - \mu\Omega \,\mathrm{sign}\,[S(x)] \qquad (2)$$

This expression, which has been adopted in this study, is a solution of a differential equation [19]. During the sliding phase ($S(x) = 0$ and $\dot{S}(x)S(x) < 0$), the closed loop system has the same behavior than $S(x) = 0$.
In general, for a system of order n which is described in the state space by:

$$\begin{cases} \dot{x}_i = x_{i+1} \;\; i = 1, ..., n-1 \\ \dot{x}_n = f(x,t) + g(x,t)u(t) \end{cases} \qquad (3)$$

the linear sliding surface according to the state has the form:

$$S(x) = Cx = \sum_{i=1}^{n} c_i x_i \qquad (4)$$

In the sliding phase, the looped system will have the same behavior as the system is described by:

$$\begin{cases} \dot{x}_i = x_{i+1} \;\; ; i = 1, ..., n-2 \\ \dot{x}_{n-1} = -\sum\limits_{i=1}^{n-1} c_i x_i \;\; ; c_n = 1 \end{cases} \qquad (5)$$

This is a linear system of order $(n-1)$ [20]. Thus, when the sliding regime is reached (after an interval t_g of time), the operating point is going to stay on the surface whose equation is $S(x) = 0$. Therefore, the looped system has an insensitivity to parameter variations of the system to be controlled.

3 Problem formulation and Mutual SMC synchronization concept

Let us consider p manipulator robots described by the following differential equation:

$$M_i(\ddot{q}_i) + C_i(q_i, \dot{q}_i)\dot{q}_i + G_i(q_i) \;\; = \;\; \tau_i \qquad (6)$$

for $i = 1 \cdots p$. Each robot has n degrees of freedom, for which:

- $q_i \in \mathbb{R}^n$ denotes the vector of measured displacements and i is the number of cooperative robots,
- $\dot{q}_i \in \mathbb{R}^n$ is the measured velocity vector,
- $\ddot{q}_i \in \mathbb{R}^n$ is the vector of articulatory acceleration,
- $M_i(q_i) \in \mathbb{R}^{n \times n}$ is the inertia matrix, which is symmetric uniformly bounded and positive definite,
- $C_i(q_i, \dot{q}_i)\dot{q}_i \in \mathbb{R}^n$ is the vector expressing the Coriolis and the centrifugal forces,
- $G_i(q_i) \in \mathbb{R}^n$ is the vector of gravitational forces,
- $\tau_i \in \mathbb{R}^n$ denotes the vector of external torques and forces applied at each joint.

The state vector is presented as: $x_i = \begin{pmatrix} q_i \\ \dot{q}_i \end{pmatrix}$, and its control vector is $u_i = \tau_i$. To achieve a coordinated control motion, a cross coupling synchronizing approach (that is derived from the graph theory [21]) has been applied to the networked system. Thus, the cross-coupled SMC concept is adopted in this work to control the whole multi-agent system, which in turn becomes considered as a single generalized system. The typical procedure for implementing such a control module is to first build a global error model in real time [3], based on the information feedbacks derived from all system members. In order to achieve this goal, we define the tracking error as follows:

$$\zeta_{1i}(t) = q_i(t) - q_d(t) \qquad (7)$$

where:

- $q_d(t) \in \mathbb{R}^n$ denotes the desired position of the robot.

The 'cross-coupling' concept consists in presenting a synchronization error with reference to the calculation of differences between position errors. This error is presented as follows [3]:

$$\zeta_{2i}(t) \;\; = \;\; \sum_{j \neq i}^{p} \Lambda_{ij}\,[q_i(t) - q_j(t-\tau)] \qquad (8)$$

where $\Lambda_{ij} = \Lambda_{ji}$ are symmetric positive-definite matrices which reveal an idea about the communication quality between the ith and j^{th} agent. Finally, the

global error concerning each agent of the system is given by:

$$\varepsilon_i = \zeta_{1i} + \int_0^t \zeta_{2i}(\mu) d\mu \tag{9}$$

Let's consider the sliding surface:

$$s_i = \dot{\varepsilon}_i + \lambda_i \varepsilon_i$$

and let's define the control law:

$$u_i = u_{eq} + \Delta u \tag{10}$$

Hence, equality $\dot{s}_i = 0$ yields the following equivalent control expression:

$$u_{eq} = M_i\Big[\ddot{q}_{id} - \dot{\zeta}_{2i} - \lambda_i(\zeta_{2i} - \dot{\zeta}_{1i})\Big] + C_i(q_i, \dot{q}_i)\dot{q}_i + G_i \tag{11}$$

and the discontinuous control part becomes:

$$\Delta u = -M_i^{-1}\, K_i \,\text{sign}\,(s_i) \tag{12}$$

where K_i is a definite positive diagonal matrix.

Applying the control law (10) yields the expression of the differential with respect to time of vector s_i:

$$\dot{s}_i = -K_i \,\text{sign}\,(s_i) \tag{13}$$

The advantage of the adopted decentralized architecture is that it allows to each robot to have local information about relative positions of its neighbors. So each agent can communicate and share information with robots placed just nearby, according to the utilized cross coupling approach.

4 Stability analysis

To prove the stability of one agent in the system, we consider a Lyapunov function:

$$V_i \;=\; \frac{1}{2}\, s_i^T K_i^{-1} s_i \;> 0 \tag{14}$$

Its differentiation with respect to time gives:

$$\dot{V}_i \;=\; -s_i^T \,\text{sign}\,(s_i) \;< 0 \tag{15}$$

By considering the whole multi-agent system, the chosen Lyapunov candidate function becomes:

$$V = \sum_i^p V_i > 0$$

and after its differentiation with respect to time we obtain:

$$\dot{V} = -\sum_i^p s_i^T \,\text{sign}\,(s_i) < 0$$

This proves the stability of the overall system.

5 Convergence to the desired trajectory

It is clear that $s_i(t)$ goes to zero after the reaching phase ($t > t_0$). In the sliding phase, we have $s_i = 0$. That is for $t > t_0$:

$$\varepsilon_i(t) \;=\; \varepsilon_i(t_0) e^{-\lambda(t-t_0)} \tag{16}$$

$$\dot{\varepsilon}_i(t) \;=\; -\lambda \varepsilon_i(t_0) e^{-\lambda(t-t_0)} \tag{17}$$

It is obvious that:

$$\lim_{t \longrightarrow +\infty} \varepsilon_i(t) \;=\; 0 \tag{18}$$

$$\lim_{t \longrightarrow +\infty} \dot{\varepsilon}_i(t) \;=\; 0 \tag{19}$$

Then, we can write that for $t > t_1 = t_0 + \frac{5}{\lambda}$:

$$\varepsilon_i(t) \simeq 0\,, \;\; \dot{\varepsilon}_i(t) \simeq 0 \tag{20}$$

Let's assume that the following quantities are uniformly bounded for $t > t_1$:

$$\|q_d(t) - q_d(t-\tau)\| \;\leq\; m_1 \tag{21}$$

$$\|\dot{q}_d(t) - \dot{q}_d(t-\tau)\| \;\leq\; m_2 \tag{22}$$

$$\left\| \int_{t-\tau}^{t} q_d(\varsigma) - q_d(\varsigma - \tau) d\varsigma \right\| \;\leq\; m_3 \tag{23}$$

Let's write, for $t > t_1$, $\dot{\varepsilon} = 0$:

$$\begin{aligned} &\dot{q}_i(t) - \dot{q}_d(t) + \left(\sum_{j \neq i} \Lambda_{ij}\right)[q_i(t) - q_d(t)] \\ &- \sum_{j \neq i} \Lambda_{ij}\,[q_j(t-\tau) - q_d(t-\tau)] \\ &+ \sum_{j \neq i} \Lambda_{ij}\,[q_d(t) - q_d(t-\tau)] \;=\; 0 \end{aligned} \tag{24}$$

Denoting:

$$d_i(t) \;=\; \sum_{j \neq i} \Lambda_{ij}[q_d(t) - q_d(t-\tau)] \tag{25}$$

we can write that:

$$\dot{\zeta}_{1i}(t) + \left(\sum_{j \neq i} \Lambda_{ij}\right)\zeta_{1i}(t) - \sum_{j \neq i} \Lambda_{ij}\zeta_{1j}(t-\tau) + d_i(t) = 0 \tag{26}$$

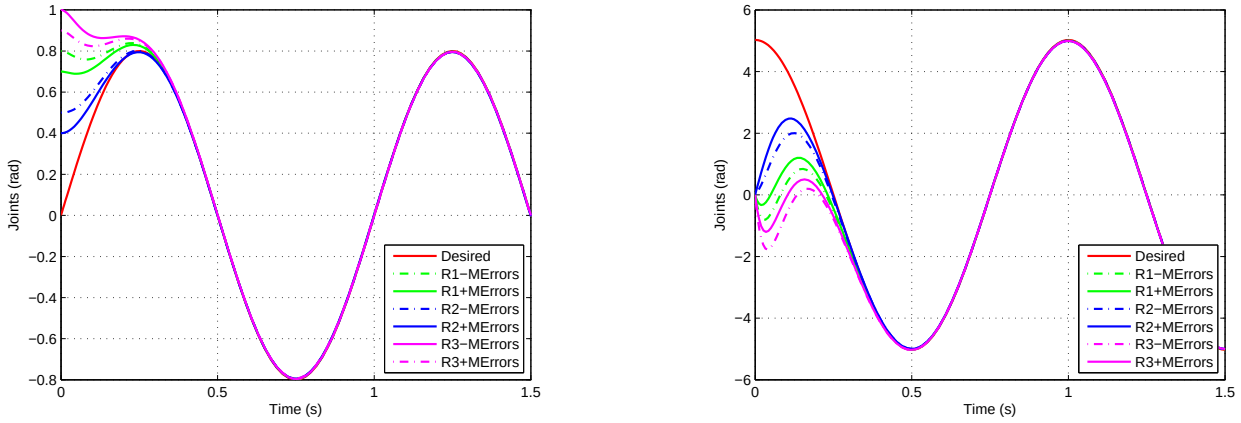
Let's define: $\zeta_1 = [\zeta_{1i}^T\; \zeta_{2i}^T \cdots\;]^T$ and $d = [d_1^T\;\; d_2^T \;\cdots\;]^T$. Then, we have:

$$\dot{\zeta}_1(t) = A\zeta_1(t) + B\zeta_1(t-\tau) + d(t) \tag{27}$$

where A is a block-diagonal matrix, for which each bloc-diagonal term is:

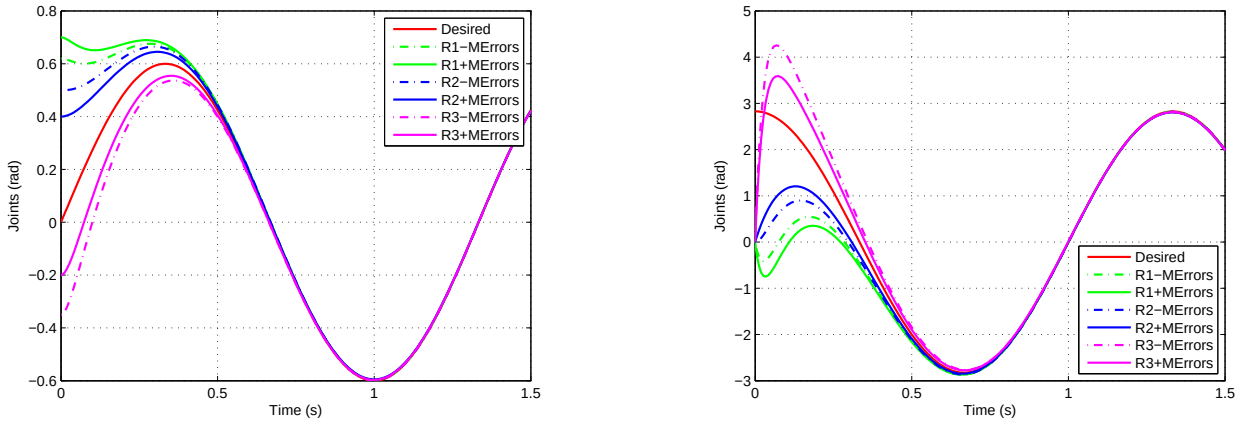
$$A_i = -\sum_{j \neq i}^{p} \Lambda_{ij} \tag{28}$$

which is a symmetric definite negative matrix. This shows that all eigenvalues of matrix A_i have strictly negative real parts. Consequently, A is a Hurwitz matrix.



(a) (b)

Figure 1: The 1st joints SMC:(a) Trajectory tracking and position synchronization,(b) Velocity synchronization.



(a) (b)

Figure 2: The 2nd joints SMC:(a) Trajectory tracking and position synchronization,(b) Velocity synchronization.

Table 1: Joints initiation and parameters

Articulation	Mass	Length
q_1	8 kg	0.4 m
q_2	6 kg	0.3 m
q_3	0.5 kg	0.3 m

Using properties of matrices A and B, and using the fact that $d(t)$ is uniformly bounded, we can deduce that $\zeta_1(t)$ is uniformly bounded [22, 23, 24]. This shows that $\zeta_{2i}(t)$ and:

$$\int_0^t \zeta_{2i}(t)dt = \varepsilon_i - \zeta_{1i}(t) \tag{29}$$

are uniformly bounded. This yields that:

$$\lim_{t \longrightarrow +\infty} \zeta_{2i}(t) = 0 \tag{30}$$

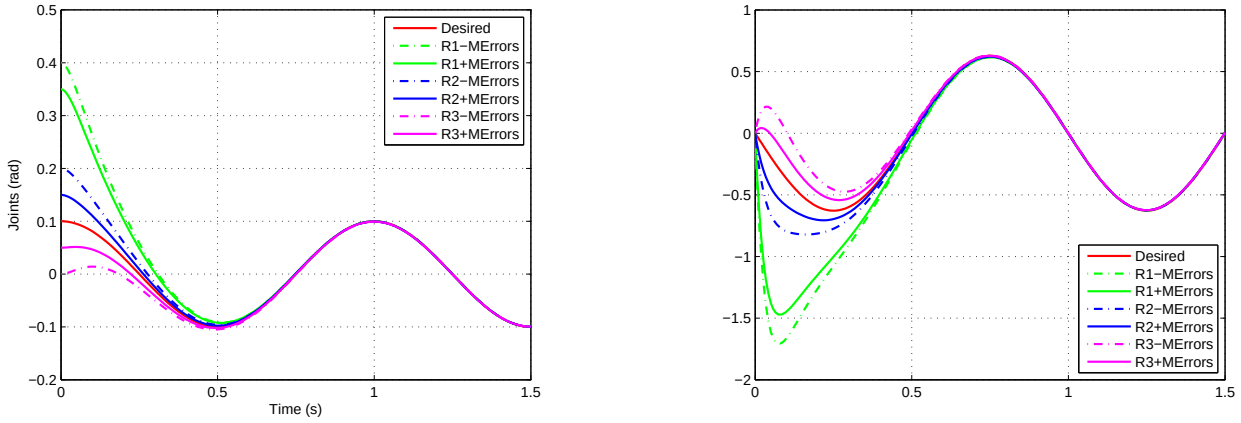
In consequence, we have, for $t > t_1$:

$$\varepsilon(t) = \dot{\zeta}_{1i} + \zeta_{2i} \simeq 0 \tag{31}$$

Finally, we can easily deduce that for $t > t_1$:
$\zeta_1(t) \simeq 0$. This proves that positions of al robots converge to their desired trajectory.

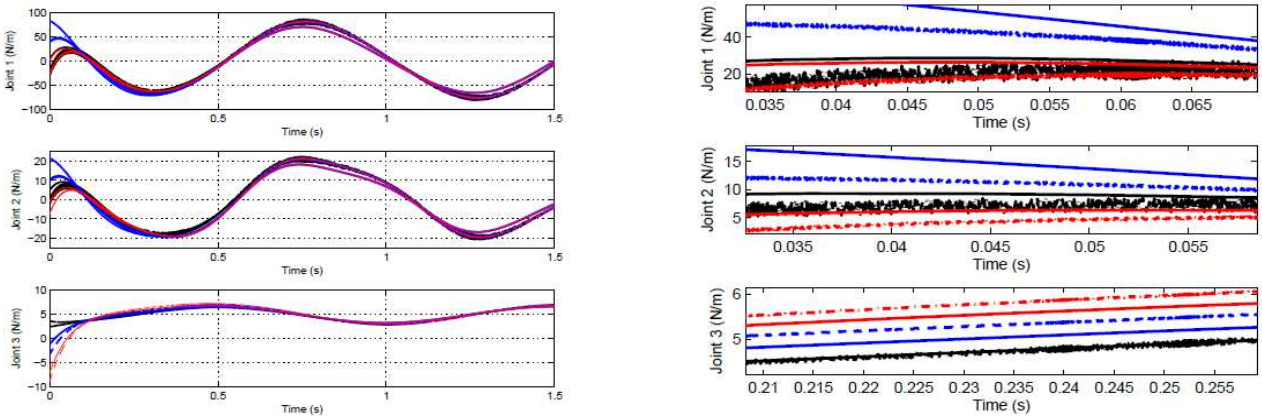
6 Application

Consider the SMC law applied to the networked robotic system for a trajectory tracking control task. Here, the cooperation is defined as a group of robots that work together to accomplish a required task, where resources are shared between all agents, and the action performed by each robot takes into account actions performed by others. The required behavior of the proposed synchronizing SMC is to handle measurement errors appearing in the controller of one robot in the system, the communication time delay lag between cooperative robot movements, and at the same time to retain all properties and system performances. So that, the designed controller must have a sufficient authority which is able to dominate the undesirable disturbances affecting one robot, and which



(a) (b)

Figure 3: The 3rd joints SMC:(a) Trajectory tracking and position synchronization,(b) Velocity synchronization.



(a) (b)

Figure 4: (a) Applied torque behaviors: The black lines present behaviors of the robot affected by the measurement errors, (b) Enlarged scales: Dotted lines- In the presence of external disturbances. Continuous lines- Without external disturbances,

have been transmitted to other neighbors due to the feedback control sharing imposed by the cross coupling design. Dynamic parameters of the manipulator robot model are mentioned in Table I. We note that:

- R_i: designate the cooperative robots, $i = 1..3$.
- R-MErrors: presents the robot behavior without measurement errors (in the presence of a constant communication time delay z = 0.4s).
- R+MErros: presents the robot behavior in the presence of measurement errors affecting the position vector of the first robot R1 (in the presence of a constant communication time delay z = 0.4s).

Discussion: Simulation results present a comparison between behaviors of the time-delayed networked system acting with and without presence of measurement errors. The robust synchronization of the networked system, realized by the synchronizing SMC has been proved when the time delay rate is valued around 0.4 s, even in the presence of external disturbances. In fact, position and velocity behaviors of the multi-agent system are almost similar in both cases as shown in Figs.1, 2 and 3, despite the addition of measurement errors (at the position vector of the first robot) whose impact reaches 20% of its right value. So the global system addresses imposed perturbations and different positions of cooperative robots rapidly attain their desired trajectories. At the same time, the synchronization effect is clearly presented despite the existence of communication time delay between robots. Besides, the generated velocity curves also reflect the insensitivity of the proposed control approach to different perturbations applied to the system, as demonstrated in Figs.1(b), 2(b) and 3(b). This proves the robustness of the synchronizing SMC in achieving adequate motion control and meeting the efficiency requirement throughout a middle affected by a constant communication time-delay and relatively high valued external disturbances. Concerning the control input behavior, we note that the signum function (which is responsible to generate the chattering phenomenon) has been replaced by an approximating function (Hyperbolic tangent) in order to better present the disturbance effect on the system. This is justified by the fact that the high frequency sig-

nals resulted from the chattering phenomenon cover the measurement error impact, whose amplitude is much smaller in comparison with the chattering one. So in the absence of measurement errors, torques behave smoothly as shown in Fig.4(a) and obviously in Fig.4(b). However, it is clear that input torques are influenced by the presence of disturbances, where slight oscillating signals disrupt input torques of the affected robot, and lower disturbing signals appear in the torque behaviors of other agents not affected. This shows that each robot influences other teammates under the synchronizing control effect. So, feedbacks derived from unaffected robots help the affected agent to overcome external disturbances which he meets.

7 Conclusion

In this paper, we have developed an analytical approach to define adequate synchronizing stability conditions in the presence of constant communication delays, and illustrated them for a decentralized architecture of a class of networked robotic systems. This analysis shows that the overall system synchronization, even in a middle disturbed by measurement errors, can be successfully guaranteed by transmitting feedbacks between cooperative system agents. Obtained comparison between simulation results shows that the networked system can easily reach desired behaviors and move smoothly when the considered communication time delay is valued around 0.4 s, and the rate of measurement errors attain 20presence of low disturbing frequencies in the control torques of agents proves that the proposed controller addresses disturbances and guarantees a robust synchronized motion task.Further researches will involve consideration of the system convergence rate with respect to the time-varying delay functions, to the SMC constraints, and the limit rate of time delays which can be supported and compensated by such a variable structure controller.

8 References

1. F. Abdelhedi, A. Jribi, & N. Derbel, Robust SMC for synchronization of networked time delay robotic systems. 8th International Conference on Modelling, Identification and Control (ICMIC), 2016.

2. M. Fathallah, F., Abdelhedi, & N. Derbel, A surveillance camera algorithm based on the sliding mode approach with measurement's disturbance. The 13th International Multi-Conference on Systems, Signals and Devices (SSD), 783–786, 2016.

3. D. Sun, Synchronization and control of multiagent systems, vol. 41. CRC Press, 2010.

4. F. Abdelhedi, Y. Bouteraa and N. Derbel, Second Order Sliding Mode Based Synchronization Control for Cooperative Robot Manipulators, Advances and Applications in Nonlinear Control Systems Springer International Publishing, 2016.

5. F. A. Bouaziz, Y. Bouteraa, and N. Derbel, Control energy comparison between 1st and 2nd order sliding mode approach with application to a SCARA robot, In 13th International Multi-Conference on Systems, Signals and Devices, 2016.

6. P. Gratia, A. Magomedov, T. Malinauskas, M. Daskeviciene, A. Abate, S.Ahmad, ... & M. K. Nazeeruddin, A Methoxydiphenylamine Substituted Carbazole Twin Derivative: An Efficient Hole Transporting Material for Perovskite Solar Cells. Angewandte Chemie International Edition, **54**(39), 11409–11413, 2015.

7. Y. K. Wang, Z. C. Yuan, G. Z. Shi, Y. X. Li, Q. Li, F. Hui, ... & L. S. Liao, DopantFree Spiro Triphenylamine/Fluorene as Hole Transporting Material for Perovskite Solar Cells with Enhanced Efficiency and Stability. Advanced Functional Materials, **26**(9), 1375–1381, 2016.

8. A. Z. Abbasi, N. Islam, & Z. A. Shaikh, A review of wireless sensors and networks' applications in agriculture. Computer Standards & Interfaces, **36**(2), 263–270, 2014.

9. C. Laschi, & M. Cianchetti, Soft robotics: new perspectives for robot bodyware and control. Frontiers in bioengineering and biotechnology, **2**(3), 2014.

10. A. Jribi, F. Abdelhedi, Y. Bouteraa, and N. Derbel, Kinematics and a Comparison Between Two SM Control Strategies for a 5DOF Serial Robot for Tele-Echography. In Applications of Sliding Mode Control 157–174. Springer Singapore, 2017.

11. S. J. Chung, U. Ahsun, J. J. E. Slotine, D. W. Miller, Application of synchronization to cooperative control and formation flight of spacecraft, In proceedings of the AIAA Guidance, Navigation and Control Conference and Exhibit. Hilton Head, South Carolina, 2007.

12. M. Wu, Y. He, and J. H. She, Stability analysis and robust control of time-delay systems, 43–47. Berlin: Springer, 2010.

13. P. Shi, Y. Zhang, & R. K. Agarwal, Stochastic finite-time state estimation for discrete time-delay neural networks with Markovian jumps. Neurocomputing, **151**, 168–174, 2015.

14. M. Yu, L. Wang, T. Chu, & F. Hao, Stabilization of networked control systems with data packet dropout and transmission delays: continuous-time case. European Journal of Control, **11**(1), 40–49, 2005.

15. R. Olfati Saber, Flocking for multi agent dynamic systems: algorithms and theory. IEEE Transactions on Automatic Control, **51**(3):401–420, 2006.

16. B. Yang, & H. Fang, Forced consensus in networks of double integrator systems with delayed input. Automatica, **46**(3), 629–632, 2010.

17. A. S. Nouri, F. A. Bouaziz, and N. Derbel, On the Sliding Control, In Applications of Sliding Mode Control Springer Singapore, 2017.

18. F. Abdelhedi, N. Derbel, Adaptive Second order sliding mode control based cross coupling concept for synchronization of robotic systems, Advances in Systems, Signals and Devices, Issues on Systems, Analysis and Automatic Control, 2016.

19. P. Lopez & A.S. Nouri, Thorie lmentaire et pratique de la commande par les rgimes glissants, Mathmatiques et applications **55**, SMAI, Springer – Verlag, 2006.

20. M. C. Golumbic, Algorithmic graph theory and perfect graphs **57**, Elsevier, 2004.

21. N.N. Krasovskii, Stability and motion : applications of Lyapunov's second method to differential systems and equations with delay. Stanford University Press, Stanford, 1963.

22. Gu. K. Discretized lmi set in the stability problem of linear uncertain time-delay systems, International Journal of Control, **68**:923–934, November, 1997.

23. Gu. K. An integral inequality in the stability problem of time-delay systems, In Proceedings of the 39th IEEE Conference on Decision and Control, **3**:2805–2810, 2000.

24. Y. Miladi Siala, Commande du comportement chaotique d'un moteur pas pas. Master report, Ecole Nationale d'Ingnieurs de Sfax-Tunisia, June, 2012.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 2, No. 3, 520-531 (2017)
www.astesj.com
Special Issue on Recent Advances in Engineering Systems

ASTES Journal
ISSN: 2415-6698

Modelization of cognition, activity and motivation as indicators for Interactive Learning Environment

Asmaa Darouich*, Faddoul Khoukhi, Khadija Douzi

Computer Laboratory of Mohammedia (LIM), Faculty of Sciences and Technology Mohammedia (FSTM), Hassan II University of Casablanca, 20650, Mohammedia, Morocco.

ARTICLEINFO

Article history:
Received: 05 April, 2017
Accepted: 12 May, 2017
Online: 01 June, 2017

Keywords :
Cognition
Activity
Motivation
Navigation traces

ABSTRACT

In Interactive Learning Environment (ILE), the cognitive activity and behavior of learners are the center of the researchers' concerns. The improvement of learning through combining these axes as a structure of indicators for well-designed learning environment, encloses the measurement of the educational activity as a part of the learning process. In this paper, we propose a mathematical modeling approach based on learners actions to estimate the cognitive activity, learning behavior and motivation, in accordance with a proposed course content structure. This Cognitive indicator includes the study of knowledge, memory and reasoning. While, activity indicator aims to study effort, resistance and intensity. The results recovered on a sample of students with different levels of education, assume that the proposed approach presents a relation among all these indicators which is relatively reliable in the term of cognitive system.

1 Introduction and Purpose

In Interactive Learning Environment (ILE) the evaluation of learners is based on interactions analysis when accessing e-learning systems.
Several studies [1, 2, 3] look into calculation and deduction of indicators from analysis of interaction traces, in order to extract meaningful information about the learner activity and behavior, this allows tutors to observe, assess and regulate the activities of learners .
To some extent, the choice of data to analyse mainly depends on the objectives of observation, supposedly with individual or collective character, using quantitative or qualitative value, knowledge's study. The use of extracted knowledge from traces has enabled the development of several sectors such machine learning, artificial intelligence, e-learning and e-commerce. A concept that turns more and more to realization in real time [4, 5]. In fact, exploring traces as a background information identifies learners' navigation paths and reconstruct their activities. [6] demonstrate the relevance of navigation in systems oriented toward the elderly, indicating that linear navigation is more suitable for these users, through comparing two designs of the same email Web application with linear and hypertextual navigation styles. However, the use of traces is a common approach by focusing on the processing on digital traces; just a few works involves another kind of traces in this aspect of the research. [7] propose a system that evaluates the activity of learners using log files data, based on four indicators; classification indicator of learners based on their activity (weekly), indicator of activity of each learner throughout the course (weekly), learner classification indicator according to their performance (for each event) and performance indicator of each learner.
The active involvement of learners in their learning process can be reinforced by a meta-cognitive regulation strategy by alternating learning with repeated tests of their knowledge [8]. [9] propose that the computer support provided to the tutor enable him/her to follow up individually a collective activity in which the learner is engaged. In complicated learning conditions, learners have an increased commitment to learn with increased cognitive effort that would paradoxically result in more sustained attention [10, 11]. The voluntary efforts

*Corresponding Asmaa Darouich, asmaa.darouich@gmail.com

https://dx.doi.org/10.25046/aj020367

of the learner result in the acquisition of knowledge, however the quality of learning returns with added value, which positively influences cognitive performance.

Effort, or as [12] call, motivational excitement, occurs only if a number of conditions are met. First, there should be expectation that some behavior will lead to some desirable outcome values (task incentive). Secondly, the behavior must be difficult but considered to be at its capacity and justified by the potential gain. A concept that is very close to mental effort is the mental load [13].
The effort is expected to be associated with cognitive processes reflected by the Electroencephalography measurements. Where cerebral activity in default mode [14, 15] and cortical inhibition [16, 17] suggests that this measure reflects different levels of effort, with a high alpha for low effort levels.
The human cognitive system is a complex processing device capable of acquiring, conserving, processing and transmitting information. Cognition improves gradually from birth through the interactions of the child with his/her environment. Cognitive development appears punctuated by acquisitions that are necessary for the establishment of subsequent cognitive processes. Actually, cognitive science has identified four key factors for learning: attention, active learner engagement, feedback and consolidation [18]. These four factors are based on the functioning characteristics of the cognitive system and are deeply rooted at an early age.

On the basis of the [19] model, the cognitive architecture underlying the learning phenomena consists of working memory and long-term memory. In neuroscience, memory is defined as the ability to acquire, retain and restore information. Stimulated by more than a century of neuropsychological study, researchers have come to distinguish these forms of memory. Maintaining a certain amount of memory in short-term memory would eventually allow it to be transferred into long-term memory for more sustainable storage. This process would be facilitated by mental work of repetited information , hence the expression "working memory" [20]. The working memory can be defined as the temporary maintenance and the manipulation of information during the realization of cognitive spots. However, long-term memory can be seen as an adaptive function of the human being to his cultural, social and emotional environment [21].
Knowledge is an element of our memory that allows us to recognize the world around us, to interpret, to understand our environment, but also to act on it. Knowledge is mainly evaluated in the form of a simple reminder or through comprehension activities.
In the didactic of learning environments, knowledge given to be study always comes from an institutional demand and what individuals have to build are knowledge that will enable them to analyze a situation or to allow an action.
When designing a pedagogical scenario, knowledge derived from the psychology of learning, [22] presents human knowledge by categorizing it into six different formats, these formats corresponding to particular learning processes. Contrary to the proposal of [23], which updated the taxonomy of Bloom, these authors do not establish any hierarchy between the different knowledge formats and even consider that some knowledge does not exist according to certain formats.
Based on the idea that cognition groups together the various mental processes of perception, memorization and reasoning, [24] made an important contribution when it differentiates four types of cognitive reasoning: deductive reasoning, inductive reasoning, abductive reasoning and reasoning by analogy. However, reasoning takes the form of a cognitive process to produce a conclusion based on premises.
The studies conducted in the cognitive psychology help to structure behavioral models. However, the analysis of interactions provides a large amount of data approach that identify the motivational aspect as a point that cannot be detached from cognitive and activity aspect of learner. The number of definitions of the motivation term is almost a large topic. This happens because motivation represents a hypothetical construct and not a material entity, which means that we use observable manifestations to infer a state of motivation.
Motivation is not a skill, but a state of mind influenced by internal and external factors to the individual. [25] defines it as "the orientation of the conduct, the efforts deployed and finally the persistence of this effort over time". In a psychological perspective, [26] emphasizes the characteristics of the environment in which behavior is manifested: motivation is defined as the set of "physiological and psychological processes responsible for initiation, maintenance and cessation behavior and the appetitive or aversive value conferred on the elements of the environment on which this behavior is exercised ". The work environment is likely to influence motivation, and also the work environment of the worker and, more specifically, the social beliefs of the employee [27]. The work of [28] on the theory of objectives setting establishes that an individual is motivated when he is working towards a clear, challenging and meaningful objective for which he will have a feedback. The proponents of the theory of self-determination, on the other hand, consider that the individual is motivated by the needs of social belonging, competence and autonomy, which would be the most decisive need [29]. The theory [30] relates motivation to three variables: valence (value attributed to the result), instrumentality (relation between performance and retribution) and expectation (results) which combined, can induce

motivation. On the other hand, analysing resistance to learn is directly observable at a large scale as some students fail assessments on a course. [31] suggests, that this resistance manifests through a reduction in motivation, suggested by lateness or absence from sessions, or a reduction in participation in group exercises and discussions. The individual learning resistance is linked to his/her defense mechanisms as a set of actions that purpose to the reduction or elimination of any change in relation to the integrity and stability of his/her biopsychic [32].

This paper is an extension of work originally reported in Proceedings of the 11th International Conference on Intelligent Systems: Theories and Applications (SITA). Thereby, we expose firstly the methodology of structuring a course content on which we base our modeling indicators approaches [33]. Secondly we modeled cognition [34], activity and motivation [35] indicators also approached in [36], by addressing their structure that use navigation traces on a learning session, likewise as a concept of procedures integrated in the adaptive learning content system we named S-CAMO. Developing those approaches complete studies that use simpler types of indicators [1, 2, 3], which is not sufficient to reflect the learner profile in a general way. The last section expose the statistical validation of the proposed indicators modeling.

2 Educational collections for course structure

Conceiving a flexible learning content toward learner profile with his different needs requires and includes structured components.
We propose an educational collections Ct that offer the possibility of being redesigned under different personalization in proportion to the same degree of subordination of the collection to form a semantic model of dynamic content.
Following the conceptualization of a structured educational content ("As shown in Algorithm 1"), educational collections can be one of four types; A course Cr_i "(2)", a chapter Ch_j "(3)", a unit U_k "(4)" or an educational component Cpt_h "(5)" as a pedagogical concept Cp_h or a pedagogical activity Ap_h.

2.1 Pedagogical objective

The pedagogical objective O^{Ct} "(1)" of a collection is conceptualized from the objective to learn dimension Ld and the junction of the cognitive Cd with the knowledge Kd dimensions ("As shown in Figure 1").

$$O^{Ct} = \{Ld, Cd, Kd\} \quad (1)$$

Algorithm 1: Structuring pedagogical content

1: **for** Create

$$Cr_i \in [Cr_1 \cdots Cr_l] \quad (2)$$

do
2: Assigning the Objective O^{Cr_i}
3: **for** Create

$$Ch_j \in [Ch_1 \cdots Ch_s] \quad (3)$$

do
4: Assigning the Objective O^{Ch_j}
5: **for** Create

$$U_k \in [U_1 \cdots U_c] \quad (4)$$

do
6: Assigning the Objective O^{U_k}
7: **for** Create

$$Cpt_h \in [Cpt_1 \cdots Cpt_{n+m}] \quad (5)$$

do
8: Assigning the Objective O^{Cpt_h}
9: **end for**
10: **end for**
11: **end for**
12: **end for**
13: Pedagogical content with a dynamic and modular structure
14: $O^{Cr_i} = \{O^{Ch_j}\}_{j\in[1\cdots s]} = \{O^{U_k}\}_{k\in[1\cdots c]} = \{O^{Cpt_h}\}_{h\in[1\cdots n+m]}$

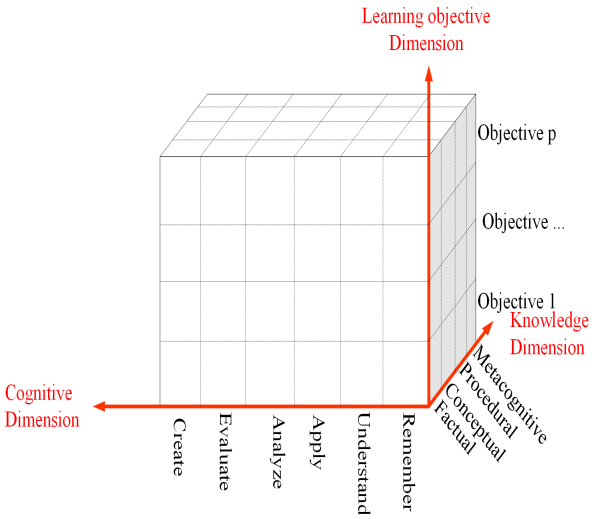


Figure 1: Pedagogical objective

Tutors create educational collections with taking into account the progress of the objectives to acquire throughout the learning process ("As shown in Algorithm 1"). By means of joining to each collection a learning objective, however, for each one is assigned one of the 24 levels ("As shown in Table 1") arranged by the junction of the knowledge type and the cognitive process identified in the revised Bloom Taxonomy [23]. Each level is represented by the coefficient $Pd(Ct)$ ("As shown in Table 1") which we propose is equal to the product of the value of the

knowledge type {1..4} and the value of the cognitive process {1..6} compared to all levels of the taxonomy.

Table 1: Coefficient of learning objectives

$Pd(Ct)$	(1)Factual	(2)Conceptual	(3)procedural	(4)Metacognitive
(1)Remember	1/24	2/24	3/24	4/24
(2)Understand	2/24	4/24	6/24	8/24
(3)Apply	3/24	6/24	9/24	12/24
(4)Analyze	4/24	8/24	12/24	16/24
(5)Evaluate	5/24	10/24	15/24	20/24
(6)Create	6/24	12/24	18/24	24/24

3 Modeling Cognitive, activity and motivation indicators

In the framework of developing Cognitive, Activity and Motivation aspects of learners, deduced from navigation traces and consulting concepts presented in course content. As a fundament of an adaptive learning content, we detail these indicators that are updated each time the learner progresses and pass his unit test.

$$\left(C^{U_k}, A^{U_k}, M^{U_k}\right) \quad (6)$$

3.1 Navigation traces

The navigation traces are collected in the form of a web browsing data, the flow of these interactions data reflects the interaction of a learner ("As shown in Figure 3").
A trace $Tr^k(t_i)$ in a learning session with an index k regroups all interactions data captured at every time of consultation t_i, it is characterized by;

t: Duration of learning a unit.

N: Total number of traces visited when learning a unit.

$f_{Tr^k(t_i)}$: Frequency of consultation of an interaction trace.

$d_{Tr^k(t_i)}$: Duration of an interaction trace.

3.2 Cognitive indicator

The cognitive activity indicator "(7)" for a learner at a given unit is a component defined by a $3-tuple$ of explanatory variables related to that unit.

$$C^{U_k} = (D_{knowledge_{U_k}}, D_{memorizing_{U_k}}, D_{reasoning_{U_k}}) \quad (7)$$

$D_{knowledge_{U_k}}$: The knowledge degree is calculated after the learner accessed and proceeded to a unit as well as having passed this unit test T^{U_k} ("As shown in Figure 2). Indeed, a unit test is a transit from one unit to another in order to ensure the assimilation of the pedagogical objectives and knowledge presented to the learner related to each pedagogical component of this unit ("As shown in Algorithm 2"). However, the calculation of knowledge acquisitions follows a linear process in the learning activity.

Algorithm 2: Processing the level of knowledge $\{D_{knowledge_{U_k}}\}$

for each $U_k \in [U_1 \cdots U_c]$ **do**
  for each $[Cpt_1 \cdots Cpt_{n+m}]$ **do**
    for v' questions $Q_i^{Cpt_h}$ associated with the pedagogical component Cpt_h. Each acquisition Aq_{Cpt_h} of this pedagogical component Cpt_h is linked to a vector $[1..v'] \subset [1..v]$ of questions $Q_i^{Cpt_h}$. **do**
      if $\sum_1^{v'} R(Q_i^{Cpt_h}) \geq \frac{v'}{2}$, with $R(Q_i^{Cpt})$ as answers given by a learner **then**
        $Aq_{Cpt_h} = 1$, Cpt_h is acquired.
      else if
        then $Aq_{Cpt_h} = 0$, Cpt_h is non acquired.
      end if
    end for
  end for
  Calculate $D_{knowledge_{U_k}} = \frac{\sum_1^{n+m} Aq_{Cpt_h}}{n+m}$, $D_{knowledge_{U_k}} \in [0,1]$.
  if $\sum_1^{n+m} Aq_{Cpt_h} \geq \frac{(n+m)}{2}$. **then**
    T^{U_k} is assessed validated with $state_{D_{knowledge_{U_k}}} = acquired$ and learner is redirected to unit U_{k+1}.
  else if
    then T^{U_k} is assessed not validated with $state_{D_{knowledge_{U_k}}} = non-acquired$ and the learner is redirected to the same unit U_k.
  end if
end for
Save the learner degree and state of knowledge $\{D_{knowledge_{U_k}}, state_{D_{knowledge_{U_k}}}\}$.

$D_{memorizing_{U_k}}$: The memorization degree is the quantification of memorized knowledge in a unit U_k. This degree represents the sum of the acquisitions multiplied by the coefficient of objective of a pedagogical component to which they correspond, divided by the number of acquisitions to be acquired $(n+m)$ for each unit ("As shown in Algorithm 3").

Algorithm 3: Processing the memorization degree $\{D_{memorizing_{U_k}}\}$

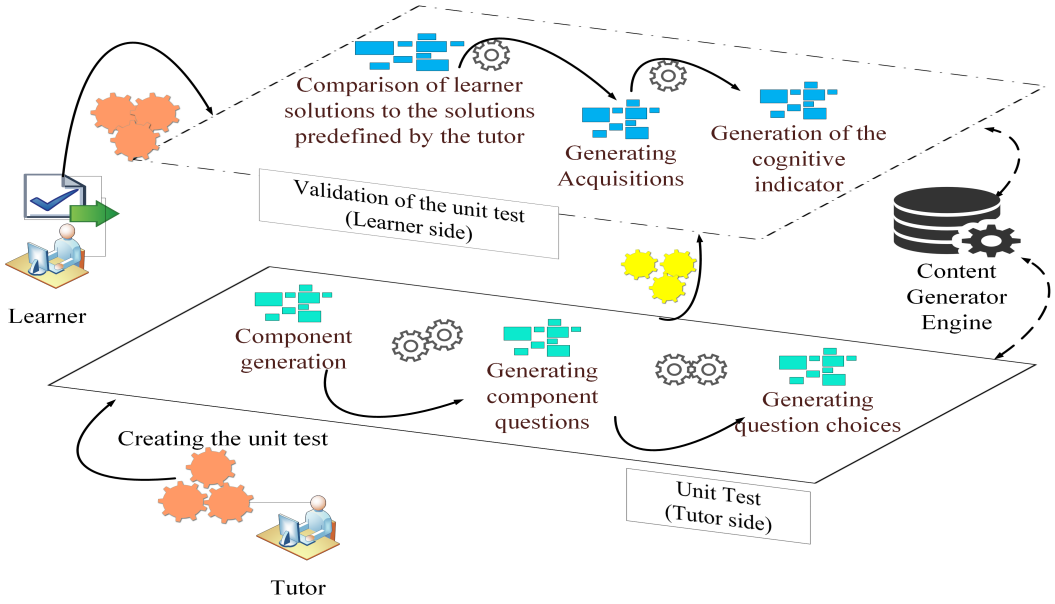


Figure 2: Generation of the cognitive indicator

1: **for** each $U_k \in [U_1 \cdots U_c]$ **do**
2: **for** each $[Cpt_1 \cdots Cpt_{n+m}]$ **do**
3: **for** All acquisitions Aq_{Cpt_h} of a pedagogical component Cpt_h **do**
4: **Calculate** $D_{memorizing_{U_k}} = \frac{\sum_{h=1}^{n+m} Aq_{Cpt_h} \times Pd(Cpt_h)}{\sum_{h=1}^{n+m} Pd(Cpt_h)}$
5: **end for**
6: **end for**
7: **end for**
8: Save the degree of memorization $\{D_{memorizing_{U_k}}\}$

$D_{reasoning_{U_k}}$**:** The reasoning degree refers to the ability of an effective knowledge understanding, proposed on a unit U_k, it represents the sum of all the given answers $R(Q_i^{Cpt})$ multiplied by their logical reasoning coefficient divided by the total number v of questions present in the unit test T^U (”As shown in Algorithm 4”).

Algorithm 4: Processing the reasoning degree $\{D_{reasoning_{U_k}}\}$

1: **for** each $U_k \in [U_1 \cdots U_c]$ **do**
2: **for** each $[Cpt_1 \cdots Cpt_{n+m}]$ **do**
3: **for** each $\{Q_i^{Cpt_h}\}_{i\in[1..v]}$, where $R(Q_i^{Cpt_h})$ is the learner's answer to the question $Q_i^{Cpt_h}$ **do**
4: **if** $\exists$ dependence between $Q_i^{Cpt_h}$ and $Q_{i-1}^{Cpt_h}$ **then**

$$P(Q_i^{Cpt_h} \mid Q_{i-1}^{Cpt_h}) = \begin{cases} 0, R(Q_i^{Cpt_h}) = 0 \,\&\, R(Q_{i-1}^{Cpt_h}) = 0 \\ \frac{1}{2}, R(Q_i^{Cpt_h}) = 1 \,\&\, R(Q_{i-1}^{Cpt_h}) = 0 \\ 1, R(Q_i^{Cpt_h}) = 1 \,\&\, R(Q_{i-1}^{Cpt_h}) = 1 \end{cases} \quad (8)$$

5: **else if** $\nexists$ dependence between $Q_i^{Cpt_h}$ and $Q_{i-1}^{Cpt_h}$ **then**
6:

$$P(Q_i^{Cpt_h} \mid Q_{i-1}^{Cpt_h}) = \begin{cases} 0, & R(Q_i^{Cpt_h}) = 0 \\ 1, & R(Q_i^{Cpt_h}) = 1 \end{cases} \quad (9)$$

7: **end if**
8: **Calculate** $D_{reasoning_{U_k}} = \frac{R(Q_1^{Cpt_h}) * P(Q_1^{Cpt_h}) + \sum_{i=2}^{v} R(Q_i^{Cpt_h}) * P(Q_i^{Cpt_h} | Q_{i-1}^{Cpt_h})}{v}$
where

$$P(Q_1^{Cpt}) = \begin{cases} 0, & R(Q_1^{Cpt_h}) = 0 \\ 1, & R(Q_1^{Cpt_h}) = 1 \end{cases} \quad (10)$$

9: **end for**
10: **end for**
11: **end for**
12: Save the reasoning degree $\{D_{reasoning_{U_k}}\}$.

3.3 Activity indicator

The activity indicator ”(11)” measures the degree to maneuver learning on a given unit. It is a component defined by $3-tuple$ of explanatory variables linked to this unit (”As shown in Figure 3”).

$$A^{U_k} = (Res_{U_k}, Int_{U_k}, Efr_{U_k}) \quad (11)$$

Res_{U_k}**:** Resistance to learn expresses the difficulty of learning all the pedagogical objectives presented in a unit.

$$Res_{U_k} = \frac{\sum_{j\in[1..N]} f_{Tr^k(t_i)}(j)}{N} * d^2_{Tr^k(t_i)}(j) \quad (12)$$

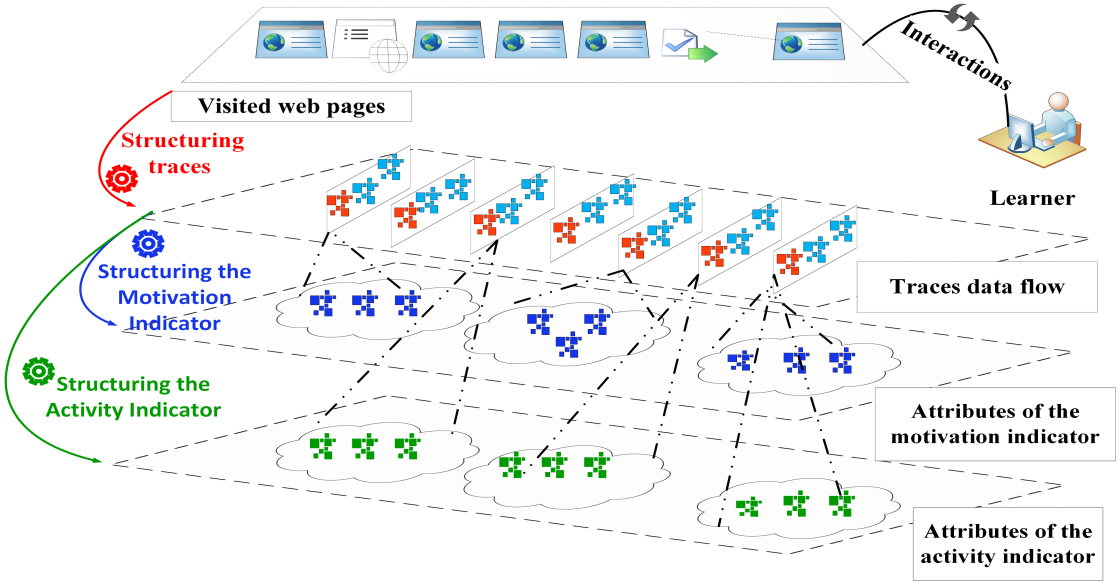


Figure 3: Structuring Activity and Motivation Indicators

Int_{U_k}: Learning intensity expresses the intensity to learn all the pedagogical objectives presented in a unit.

$$Int_{U_k} = N/t \tag{13}$$

Efr_{U_k}: Learning effort expresses the effort to learn all the pedagogical objectives presented in a unit.

$$Efr_{U_k} = Pd(U_k) * Res_{U_k} * {Int_{U_k}}^2 * t \tag{14}$$

3.4 Motivation indicator

The motivation indicator "(15)" measures the degree of motivation in learning all the pedagogical objectives presented in a unit ("As shown in Figure 3).

$$M^{U_k} = \sum_{j \in [1..N]} Pd(U_k) * \frac{f_{Tr(t_i)}(j)}{N} * \frac{d_{Tr(t_i)}(j)}{t} \tag{15}$$

4 Method

The perplexity of the proposed indicators modelization, depends on the reliability of the proposed scale. In fact, we implemented all procedures of our approach in an adaptive learning content system we named S-CAMO.

4.1 Learning environment

The system S-CAMO comprises the module "web navigation tracker" which is an extension that retrieves the data of navigation on the Internet between the time of the connection and the disconnection to the system. This extension will appear as an icon in the browser's menu bar, where the learner will have to enter his username and password and then log in. If the learner exists in the database, then the profile (Learner name, learner identifier) will persist in the local storage of "Google Chrome". However, just after installing this extension, authentication will remain active as long as the learner does not choose to disconnect manually. This extension detects and traces access to any web page in any open tab of the browser "Google Chrome", as well as verifies the learner's adhesion to each learning session opened in the platform, through the validity of its domain name. The extension sends the navigation or interaction data to the web server.

4.1.1 Interception of navigation data

For each successful authentication of a learner, the web service provided by the platform retrieves his identifier ("As shown in Figure 4). However, whenever the learner changes web page (url), the browser adds a new function that will be delegated to the extension level. The process of the new function includes:

1. The retrieval of the information of each navigation data which are url, the url parameters (parameter of the request), the title of the web page;

2. Testing and verifying the authentication at the platform, in order to send the information associated with this new intercepted and formulated navigation data.

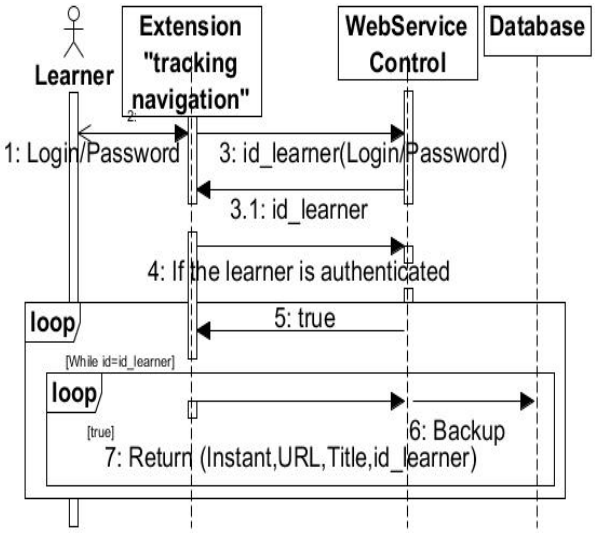


Figure 4: Process of intercepting a navigation trace by "web navigation tracker" module

4.2 Participants

The sample of learners on which the test was conducted is equal to 50 students, organized into 3 approximately equal groups. The group G_1 and G_2 contained 17 learners, while G_3 contained 16 ones. These learners were heterogeneous in relation to their level of education and had at least an associate degree.
Learners had a laptop, nevertheless, those who could not bring one, had at their disposal laptops or desktop laboratory. For the relevance of the study, we made sure that the module "web navigation tracker" was installed before the test.

4.3 Materials

We create a course ("As shown in Table 2") while respecting the learning content structuring described above, also to each educational collection was assigned a pedagogical objective.
All students followed the same course and performed the presented tests unit .

5 Task analysis results

This study recovers values of all indicators' variables according to the proposed approach, by studying the learners traits.
We have obtained the number of learners who have acquired or not the knowledge concepts presented at level of the chapter Ch_1 structure ("As shown in Table 3).

Table 3: Tests unit of chapter Ch_1 results.

$state_{D_{knowledge_{U_k}}}$	U_1	U_2	U_3
$acquired$	34	48	40
$non-acquired$	16	2	10

Table 4: Abbreviation of indicators variables

abbreviation	variables
$D_{knowledge_{U_k}}$	v_1
$D_{memorizing_{U_k}}$	v_2
$D_{reasoning_{U_k}}$	v_3
Res_{U_k}	v_4
Int_{U_k}	v_5
Efr_{U_k}	v_6
M^{U_k}	v_7

The three groups present approximately the same characteristics with respect to the educational components of the studied chapter. In fact, all groups reflect a strong correlation among the variables of the cognitive indicator relatively to all the tested units. All groups have no resistance to learn in relation to all units except for group G_1 who presents moderately a resistance to learn unit U_1. All groups have a low intensity to learn relatively all units, however more is the intensity to learn less there is a resistance. The effort to acquire knowledge and to memorize for the group G_1 remains average relatively to unit U_1 while it is low for groups G_2 and G_3, however there is no effort to reason for all groups for all units. All groups show moderate resistance linked to the effort to learn the unit U_2, however, the group G_1 has an effort related to the resistance at unit U_1. Unlike group G_2 and G_3 where resistance is not related to effort. As well as group G_3 has a resitance related to the effort at unit U_3 while group G_2 and G_3 where resistance is not related to effort. Intensity to learn is moderately low compared to all units. Motivation is weakly related to the variables of the cognitive indicator, it is moderately linked to the resistance to learn and weakly related to the intensity of work. All groups have a moderately weak relationship between motivation and effort .

6 Conclusions and discussion

The many aspects profile of learners serve on adapting their learning environment. However, the lack of holistic measures in learning environment make us propose a compilation of indicators by developing a theoretical conception that include the cognition, activity and motivation in a learning process.The study involve the measurement of the proposed indicators' structures, it reflects the behavior and trend of learners on processing information in learning interactions. In this context, we develop

Table 2: Structure of Ch_1.

Ct	Cd	Kd	$Pd(Ct)$	Ld
Cr_1	Apply	Procedural	9/24	Learning HTML and CSS
Ch_1	Understand	Conceptual	4/24	Learning The basics of HTML and CSS
U_1	Understand	Factual	2/24	Introduction
Cpt_1	Remember	Factual	1/24	Definitions and roles
Cpt_2	Remember	Factual	1/24	Versions
Cpt_3	Understand	Factual	2/24	Local work
Cpt_4	Understand	Factual	2/24	The text editor
U_2	Understand	Factual	2/24	The basics of HTML
Cpt_1	Understand	Factual	2/24	Proficiency of Elements, tags, attributes
Cpt_2	Understand	Factual	2/24	Defining Structure
Cpt_3	Understand	Factual	2/24	Display
Cpt_4	Understand	Factual	2/24	Indentation and comments
Cpt_5	Understand	Factual	2/24	Titles and paragraphs
Cpt_6	Understand	Factual	2/24	Spaces and line breaks
Cpt_7	Understand	Factual	2/24	Strong, em and mark
Cpt_8	Understand	Factual	2/24	Listes
Cpt_9	Understand	Factual	2/24	Internal and external links
Cpt_{10}	Understand	Factual	2/24	Other Links
Cpt_{11}	Understand	Factual	2/24	Images
Cpt_{12}	Understand	Factual	2/24	Validation
U_3	Understand	Conceptual	4/24	The basics of CSS
Cpt_1	Understand	Factual	2/24	Selectors, Properties
Cpt_2	Understand	Factual	2/24	Emplacement
Cpt_3	Understand	Factual	2/24	Comments
Cpt_4	Understand	Factual	2/24	Single selectors
Cpt_5	Understand	Factual	2/24	Id and class
Cpt_6	Understand	Factual	2/24	Inheritance
Cpt_7	Understand	Conceptual	4/24	Block vs inline
Cpt_8	Understand	Conceptual	4/24	Div and span

Table 5: Correlation matrix (Pearson) for group G_1 toward the unit U_1.

Var	v_1	v_2	v_3	v_4	v_5	v_6	v_7
v_1	**1**	*	*	*	*	*	*
v_2	**0,9763**	**1**	*	*	*	*	*
v_3	**0,8089**	**0,7444**	**1**	*	*	*	*
v_4	**0,5319**	**0,5314**	0,4427	**1**	*	*	*
v_5	-0,3943	-0,3942	-0,4222	**-0,6701**	**1**	*	*
v_6	0,4221	0,4335	0,2749	**0,7679**	-0,0678	**1**	*
v_7	0,1746	0,2400	-0,0120	**0,6594**	-0,4728	0,4132	**1**

Table 6: Correlation matrix (Pearson) for group G_2 toward the unit U_1.

Var	v_1	v_2	v_3	v_4	v_5	v_6	v_7
v_1	**1**	*	*	*	*	*	*
v_2	**0,9417**	**1**	*	*	*	*	*
v_3	**0,7276**	**0,5491**	**1**	*	*	*	*
v_4	0,1653	0,1883	-0,0480	**1**	*	*	*
v_5	-0,1236	-0,1468	-0,0236	**-0,7593**	**1**	*	*
v_6	-0,0340	-0,0934	-0,0436	0,3459	0,2720	**1**	*
v_7	-0,0707	0,0112	-0,1221	**0,5028**	**-0,5128**	-0,1491	**1**

Table 7: Correlation matrix (Pearson) for group G_3 toward the unit U_1.

Var	v_1	v_2	v_3	v_4	v_5	v_6	v_7
v_1	**1**	*	*	*	*	*	*
v_2	**0,9465**	**1**	*	*	*	*	*
v_3	**0,7899**	**0,6318**	**1**	*	*	*	*
v_4	-0,1103	-0,1054	0,0172	**1**	*	*	*
v_5	-0,2860	-0,2256	-0,1790	**-0,6800**	**1**	*	*
v_6	**-0,5034**	-0,3807	-0,2676	0,1603	**0,5861**	**1**	*
v_7	-0,1322	-0,1718	0,1119	**0,6086**	-0,1038	0,3820	**1**

Table 8: Correlation matrix (Pearson) for group G_1 toward the unit U_2.

Var	v_1	v_2	v_3	v_4	v_5	v_6	v_7
v_1	**1**	*	*	*	*	*	*
v_2	**1,0000**	**1**	*	*	*	*	*
v_3	**0,7787**	**0,7787**	**1**	*	*	*	*
v_4	0,2433	0,2433	0,0330	**1**	*	*	*
v_5	0,0835	0,0835	0,0522	**-0,7769**	**1**	*	*
v_6	**0,5768**	**0,5768**	0,2773	0,3029	0,2851	**1**	*
v_7	0,1338	0,1338	0,0446	**0,6863**	-0,4735	0,1091	**1**

Table 9: Correlation matrix (Pearson) for group G_2 toward the unit U_2.

Var	v_1	v_2	v_3	v_4	v_5	v_6	v_7
v_1	**1**	*	*	*	*	*	*
v_2	**1,0000**	**1**	*	*	*	*	*
v_3	**0,8845**	**0,8845**	**1**	*	*	*	*
v_4	-0,1224	-0,1224	-0,3195	**1**	*	*	*
v_5	-0,2007	-0,2007	-0,1233	-0,3503	**1**	*	*
v_6	-0,2327	-0,2327	-0,3430	0,4431	**0,6647**	**1**	*
v_7	0,0947	0,0947	0,2639	0,0348	-0,4126	-0,4577	**1**

Table 10: Correlation matrix (Pearson) for group G_3 toward the unit U_2.

Var	v_1	v_2	v_3	v_4	v_5	v_6	v_7
v_1	**1**	*	*	*	*	*	*
v_2	**1,0000**	**1**	*	*	*	*	*
v_3	**0,8585**	**0,8585**	**1**	*	*	*	*
v_4	-0,0124	-0,0124	-0,1452	**1**	*	*	*
v_5	-0,1733	-0,1733	0,0944	**-0,7590**	**1**	*	*
v_6	-0,3695	-0,3695	-0,1906	0,3853	0,2777	**1**	*
v_7	0,1702	0,1702	0,0438	0,3602	-0,1726	0,1085	**1**

Table 11: Correlation matrix (Pearson) for group G_1 toward the unit U_3.

Var	v_1	v_2	v_3	v_4	v_5	v_6	v_7
v_1	**1**	*	*	*	*	*	*
v_2	**0,9461**	**1**	*	*	*	*	*
v_3	**0,9024**	**0,8295**	**1**	*	*	*	*
v_4	-0,0795	-0,1861	-0,0150	**1**	*	*	*
v_5	-0,0930	0,0175	-0,1375	**-0,5699**	**1**	*	*
v_6	-0,1737	-0,1584	-0,1821	0,4723	0,4199	**1**	*
v_7	-0,0322	-0,1200	0,0994	0,3538	0,2557	**0,4968**	**1**

Table 12: Correlation matrix (Pearson) for group G_2 toward the unit U_3.

Var	v_1	v_2	v_3	v_4	v_5	v_6	v_7
v_1	**1**	*	*	*	*	*	*
v_2	**0,9201**	**1**	*	*	*	*	*
v_3	**0,6952**	**0,5768**	**1**	*	*	*	*
v_4	0,1357	0,1553	0,0472	**1**	*	*	*
v_5	-0,2563	-0,2060	-0,1317	**-0,6814**	**1**	*	*
v_6	-0,1320	-0,0371	-0,0406	0,2194	**0,5400**	**1**	*
v_7	0,1781	0,2682	0,1773	0,3929	-0,2815	-0,0695	**1**

Table 13: Correlation matrix (Pearson) for group G_3 toward the unit U_3.

Var	v_1	v_2	v_3	v_4	v_5	v_6	v_7
v_1	**1**	*	*	*	*	*	*
v_2	**0,9317**	**1**	*	*	*	*	*
v_3	**0,7830**	**0,6418**	**1**	*	*	*	*
v_4	0,1757	0,1825	0,1453	**1**	*	*	*
v_5	-0,1821	-0,2657	-0,1122	**-0,8190**	**1**	*	*
v_6	-0,0332	-0,0881	-0,0297	**0,7285**	-0,2264	**1**	*
v_7	0,1746	0,1418	0,0468	**0,6077**	-0,4340	**0,5587**	**1**

an educational content structure which was already implemented and which mobilizes the achievement of the specific estimation of the updated indicators after each pedagogical unit. The results show specifically a high relation among the cognition attributes and a moderatly interrelated linkage among all the studied variables. We assume that the scale we have presented in this paper prove the efficiency of the proposed modelization.

References

[1] L. S. Settouti, "Systmes Base de Traces Modlises: Modles et Langages pour l'exploitation des traces d'Interactions," Ph.D Thesis, 2011.

[2] N. Bousbia, "Analyse des traces de navigation des apprenants dans un environnement de formation dans une perspective de dtection automatique des styles dapprentissage," Ph.D Thesis, 2011.

[3] P. T. N. Diem, "Spcification et conception de services d'analyse de l'utilisation d'un environnement informatique pour lapprentissage humain," Ph.D Thesis, tel-00689025, version 1 - 19 Apr 2012.

[4] L. France, J.-M. Heraud, J.-C. Marty, T. Carron, and J. Heili, "Monitoring virtual classroom: Visualization techniques to observe student activities in an e-learning system," In ICALT, 716-720. IEEE Computer Society, 2006.

[5] E.W. Black, K. Dawson, and J. Priem, "Data for free: Using LMS activity logs to measure community in online courses," The Internet and Higher Education, 11(2), 65-70, 2008.

[6] D. Castilla, A. Garcia-Palacios, I. Miralles, J. Breton-Lopez, E. Parra, S. Rodriguez-Berges, C. Botella, "Effect of Web navigation style in elderly users," Computers in Human Behavior, 55(Part B), 909-920, 2016.

[7] Angel A.Juan, T. Daradoumis, S. Caball, F. Xhafa, "SAMOS Project: A data analysis model based on control charts to monitor online learning processes," Barcelona, Spain, July 2nd, 2008.

[8] Jeffrey D. Karpicke and Henry L. Roediger III, "The Critical Importance of Retrieval for Learning," Science, American Association for the Advancement of Science, 319(5865), 966-968, 2008.

[9] C. Laperrousaz, "Le suivi individuel d'apprenants engags dans une activit collective distance : tacsi : un environnement informatique support aux activits du tuteur," Ph.D Thesis, 1 vol. (295 p.), 2006.

[10] Franklin M. Zaromb, Jeffrey D. Karpicke and Roediger III, Henry L., "Comprehension as a basis for metacognitive judgments: Effects of effort after meaning on recall and metacognition," Journal of Experimental Psychology: Learning, Memory, and Cognition, American Psychological Association, 36(2), 552-557, 2010.

[11] Bjork E. L. and Bjork, R. A., "Making things hard on yourself, but in a good way:Creating desirable difficulties to enhance learning,"

Chapter 5, New York: Worth Publishers, M. A. Gernsbacher, R. W. Pew, L. M. Hough & J. R. Pomerantz (Eds.), Psychology and the real world: Essays illustrating fundamental contributions to society, 56-64, 2011.

[12] Jack W. Brehm and Elizabeth A. Self, "THE INTENSITY OF MOTIVATION," Ann. Rev. Psychol., Department of Psychology, University of Kansas, Lawrence, Kansas 66045, Annual Reviews Inc., 40, 109-31, 1989.

[13] A. W. K. Gaillard and C. J. E. Wientjes, "Mental load and work stress as two types of energy mobilization," Work & Stress, 8(2), 141-152, 1994.

[14] M. Izzetoglu and S. C. Bunce and K. Izzetoglu and B. Onaral and a. K. Pourrezaei, "Functional brain imaging using near-infrared technology," IEEE Engineering in Medicine and Biology Magazine, 26(4), 38-46, 2007.

[15] Laufs, H. and Krakow, K. and Sterzer, P. and Eger, E. and Beyerle, A. and Salek-Haddadi, A. and Kleinschmidt, A., "Electroencephalographic signatures of attentional and cognitive default modes in spontaneous brain activity fluctuations at rest," Proceedings of the National Academy of Sciences of the United States of America, 100(19), 1105311058, 2003.

[16] Brouwer A.-M. and Hogervorst M. A. and Herman P. and Kooi F., "Are You Really Looking? Finding the Answer Through Fixation Patterns and EEG," Proceedings of the 5th International Conference on Foundations of Augmented Cognition. Neuroergonomics and Operational Neuroscience: Held As Part of HCI International 2009, FAC '09, 329-338, 2009.

[17] Van Dijk H. and Schoffelen J.-M. and Oostenveld R. and Jensen O., "Prestimulus Oscillatory Activity in the Alpha Band Predicts Visual Discrimination Ability," Journal of Neuroscience, Society for Neuroscience, 28(8), 1816-1823, 2008.

[18] Dehaene S., "Les quatre piliers de l'apprentissage, ou ce que nous disent les neurosciences," Ecole des Neurosciences de Paris Ile-de-France / ENP, November, 2013.

[19] Atkinson R.C. and Shiffrin R.M., "Human memory: A proposed system and its control processes," K.W. Spence & J.T. Spence (Eds.), The psychology of learning and motivation: Advances in research and theory, New York: Academic Press, 2, 742-775, 1968.

[20] A. Baddeley, "The episodic buffer: a new component of working memory?, " Trends in Cognitive Sciences, 4(11), 417-423, 2000.

[21] J. R. Anderson, "Learning and Memory: An Integrated Approach," Journal of Developmental & Behavioral Pediatrics, New York, NY, John Wiley and Sons, Inc, 17(2), 125-126, 1996.

[22] M. Musial and F. Pradre and A. Tricot, "Prendre en compte les apprentissages lors de la conception d'un scnario pdagogique," Recherche et formation, 68, 15-30, 2011.

[23] D. R. Krathwohl, "A Revision of Bloom's Taxonomy: An Overview, Theory Into Practice," Routledge, 41(4), 212-218, 2002.

[24] L. Festinger, "A Theory of Cognitive Dissonance," Stanford University Press, 291, 1957.

[25] C. Louche, "Introduction la psychologie du travail et des organisations. Concepts de base et applications," L'orientation scolaire et professionnelle, 37(2), 289-290, 2008.

[26] J.-M. COQUERY, " Motivation dans Grand dictionnaire de la psychologie," BLOCH, H., CHEMANA, R. et GALLO, A. (dirs), Paris, Larousse, 480-482, 1991.

[27] T. Wisner, "Motivation au travail dans les institutions publiques et l'effet des facteurs culturels : une illustration dans le contexte hatien," Ph.D Thesis, 2016.

[28] Edwin A. Locke, "Toward a theory of task motivation and incentives," Organizational Behavior and Human Performance, 3(2), 157 - 189, 1968.

[29] E. L Deci and R. M Ryan, "Intrinsic Motivation and Self-Determination in Human Behavior," Contemporary Sociology, New York: Plenum Pres, 3(2), January 1985.

[30] Vroom V.H, "Work and motivation," Wiley Work and motivation, Oxford, England, 1964.

[31] J. Atherton, "Resistance to learning: a discussion based on participants in in-service professional training programmes," Journal of Vocational Education & Training, 51(1), 77-90, 1999.

[32] V. A. Negrii, "Psychological Problems of Learning Resistance," Procedia - Social and Behavioral Sciences, Volume 86, Pages 116-121, 2013.

[33] A. Darouich, F. Khoukhi and K. Douzi, "A dynamic learning content pattern for adaptive learning environment," 2015 10th International Conference on Intelligent Systems: Theories and Applications (SITA), 1-6, 2015.

[34] A. Darouich, F. Khoukhi and K. Douzi, "A cognitive indicator activity for dynamic learning content pattern in adaptive learning environment," 2016 11th International Conference on Intelligent Systems: Theories and Applications (SITA), 1-6, 2016.

[35] A. Darouich, F. Khoukhi and K. Douzi, "Mining fuzzy motivation indicator in learning environment through Human Computer Interaction," 2013 Science and Information Conference, London, 712-720, 2013.

[36] F. Khoukhi and M. E. Ammari, "The adaptive learning system with CAM indicators," 2012 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT), Sousse, 2012, pp. 114-120.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 2, No. 3, 532-539 (2017)
www.astesj.com
Special Issue on Recent Advances in Engineering Systems

ASTES Journal
ISSN: 2415-6698

Homemade array of surface coils implementation for small animal magnetic resonance imaging

Fernando Yepes-Calderon[*,1], Olivier Beuf[2]

[1]*Children's Hospital Los Angeles, Neurosurgery Division; 4650 Sunset Blvd, Los Angeles, CA 90027, USA*

[2]*Universite de Lyon, CREATIS INSA-Lyon; Villeurbanne 69100, France*

ARTICLEINFO

Article history:
Received: 06 April, 2017
Accepted: 12 May, 2017
Online: 01 June, 2017

Keywords:
MRI
Small Animal
Receptor antennas
Mutual induction
Parallelization

ABSTRACT

Small animal modeling is an exciting research field where human pathogenic frameworks can be replicated in a controlled environment. Accurate Imaging is in high demand when modeling abnormalities and, magnetic resonance imaging plays a vital role due to its demonstrated lowest intrusion when compared with other imaging methods. However, the required high-resolution yields low-quality images and often, critical events are masked. In this manuscript, we improve the images in small animal frameworks by modifying the reception coils to parallelize the readings and by dealing with the mutual induction produced at the dimensions required for the studied subjects.

1 Introduction

Magnetic Resonance Imaging (MRI) has become a ubiquitous tool in imaging research and clinical radiology due to its capacity to provide non-invasive, in-vivo images, both in humans and in animal models. Small animals are widely used in disease research as their size make them particularly easy to manipulate; however, MR imaging is also significantly more challenging in smaller structures. The trade-off between the spatial resolution and signal-to-noise ratio (SNR) becomes exceptionally important when scanning small volumes.

Despite improvements in high field magnet technology at 7T and beyond, several factors reduce image quality so that larger fields may not mean better images. In particular, T1 lengthening, T2 shortening, and chemical shifts artifacts account for lowering the image quality well below theoretical expectations in high field magnets [1]. In contrast, parallel acquisitions improve image quality while avoiding all of these issues [2]. Simultaneous acquisitions require single receptors to be split into two or more simple loops while maintaining the field of view (FOV). This dimensional reduction is, initially, convenient because smaller receptors provide higher sensitivity and field stability [2], [3]. Nevertheless, when two or more coils share a common space, their mutual inductance (MI) [4] deteriorates the receptor characteristics of the coils to the point where images can not be produced.

Here we present strategies to increase the quality of MRI images in surface coils and to minimize the deteriorating effects of MI. The first involves a shared capacitor placed between adjacent loops that counteracts the rising MI at Larmor frequency (F_r). A second strategy, accomplished with pre-amplifiers, consists in diminishing the electrons' flow among the receptors, so the factor $L\frac{di}{dt}$ is reduced; therefore, MI declines as well. This last strategy is intended to deal with high order MI, which is likely to be an issue when reading parallelization involves more than two channels.

We also determine the effect of implementing these MI reducing strategies both, independently and combined by analyzing the SNR, the field stability and the sensitivity of the coils for the images produced.

2 State-of-the-art technology

The concept of reducing the coils' area to produce more efficient receptors, which was introduced by Roemer in [2] is both, attractive and challenging. It has numerous technical advantages, such as the possibility of creating adaptable devices that can fit almost any volume, while maintaining modularity [5]. A reasonable attempt at using multiple surface coils was presented by the Petals Project [6]. In that work, the receptor device consisted of a crown of resonant loops for hydrogen nuclei magnetized at 1.5T. The authors

[*]Fernando Yepes-Calderon, 1300 N. Vermont Av, Los Angeles-CA, USA & fernandoyepesc@gmail.com

https://dx.doi.org/10.25046/aj020368

elegantly avoided near (first order) and next to near (higher order) MI by placing a gap between receptors. One drawback of this approach is that images of the brain presented not only progressive in-depth deterioration, but also fuzzy regions between the "petals." In spite of these issues, the Petals Project work showed the validity of the multiple coil approach and provided an important step towards applicability, while demonstrating the flexibility of the device.

The reception capabilities of the coil circuitry are not only dependent on its electrical length, but also on its geometrical configuration. For example, in [7], planar devices were build for which proper decoupling was enforced at the cost of reduced flexibility. In humans, [5] designed an interfacing box that allowed to connect up to 16 coils with different shapes, generating images with good contrast, though with an FOV of $200cm^2$. Notwithstanding these high dimensions (a surface coil of $200cm^2$ is too big for scanning a rat or mice), the circuitry is stable due to the larger values of its reactive components. The dynamically adaptable tuning-and-matching box is feasible due to the relatively big size of humans, but its implementation with smaller devices has various drawbacks associated with the electrical length of connectors and their dependence on frequency. The problem of making coils invisible to each other in an array disposition is well depicted in [8], where the authors proposed a geometrical method by placing a butterfly coil in parallel with a loop component. Through this configuration, a 90° combination of signals isolated the array from second and superior order MI while critical overlapping accounted for first order MI. This implementation was accomplished by arranging commercial coils to suit the quadrature configuration and presents and alternative to the low impedance preamplifiers in Roemer's small coil strategy [2]. Yet in [7], the authors stated one more motivation to implement smaller receptors. Apart from offering better field stability and more in deep tissue penetration, its implementation, due to the inherent FOV reduction and the necessity to recover it, forces the use of parallel imaging (PI), which in turn reduces acquisition time. Following this initiative, several other authors looked into the so-called array coil. In particular, in [9] and [10], human array coils were designed for PI implementation, which allowed the use of standard reconstruction methods such as Simultaneous Acquisition of Spatial Harmonics (SMASH) [11], Sensitivity Encoding (SENSE) [12], Generalized Auto-Calibrating Partially Parallel Acquisitions (GRAPPA) [13], [14] and their improvements or variations [15]. The idea is that a complete positioning axis can go from being encoded in a voxel-by-voxel fashion to being wholly or partially encoded in one TR depending on the length of the array and its relative dimensions about the subject being scanned. Consequently, the acquisition time is dramatically reduced.

For small animal imaging, the reduced size of the subjects imposes new challenges. For example, circuits may be too long or of unsuitable geometries. Also, the dynamic components are smaller, making the circuitry less stable. In spite of these issues, array coils have been used for specific applications like in [16], where rodents are scanned at ultra-high magnetic fields. In this case, the images are read with a three-channeled device, but each channel has been built to detect a different element's nuclei, and therefore, their resonance frequency is different. Thus, MI is minimized by the selectiveness of the receptors. In one more small animal implementation, a novel imaging method capable of detecting redox-status, oxygenation and free radicals has also been introduced as a functional MRI strategy. This has been called Electron Parametric Resonance (EPR) [17], and is implemented through a 4-channel coil array. Here, the receptors are activated sequentially; thus, there is no need of MI decoupling methods.

The work presented here is a step toward creating operative MRI surface coils. Our model consists of a single loop coil that is replicated to build a two elements receptor. Once the coils share the same space, the MI affects the reading. Two strategies for avoiding MI are implemented. The first consists of a shared capacitor that counteracts the MI by resonating at the working frequency with the inductive effect; therefore, turning the induction circuit in a purely resistive device. The second strategy consists of adding a low impedance pre-amplifier to the parallel channels, so the read signals are transported as voltage instead of current; consequently, the induction principle of Faraday is minimized.

The design suits the small animal dimensions. At these sizes and working frequencies, one faces instability of the electronic components, high influence of geometry and the fact that even a simple cable can become any active component – a capacitor or an inductor – therefore making of the resonance condition, a moving target.

3 Materials and Methods

All experiments presented here were performed using a Biospec 4.7T MRI Bruker system, setup to enable the surface coil mode and PI. This system is usually sold with volumetric coils as the original acquisition device; Although, these antennas may be changed to suit animal dimensions. The software can be adapted to either use volumetric coils that have the emitter and the receptor on the same device or to use two different devices; thus making of the surface coil, the receptor device. Under the surface coil scheme, the emitter can be any of the individual coil sets provided by the manufacturer.

The new devices are Printed Card Board (PCB) loops designed using the Eagle software [18]. The associated circuitry was created following the procedure depicted in [19]. The devices were simulated and tested individually in a network analyzer (Agilent - E5062A-275) before coupling them to the MRI scanner. All coils were matched at 50Ω and tuned at

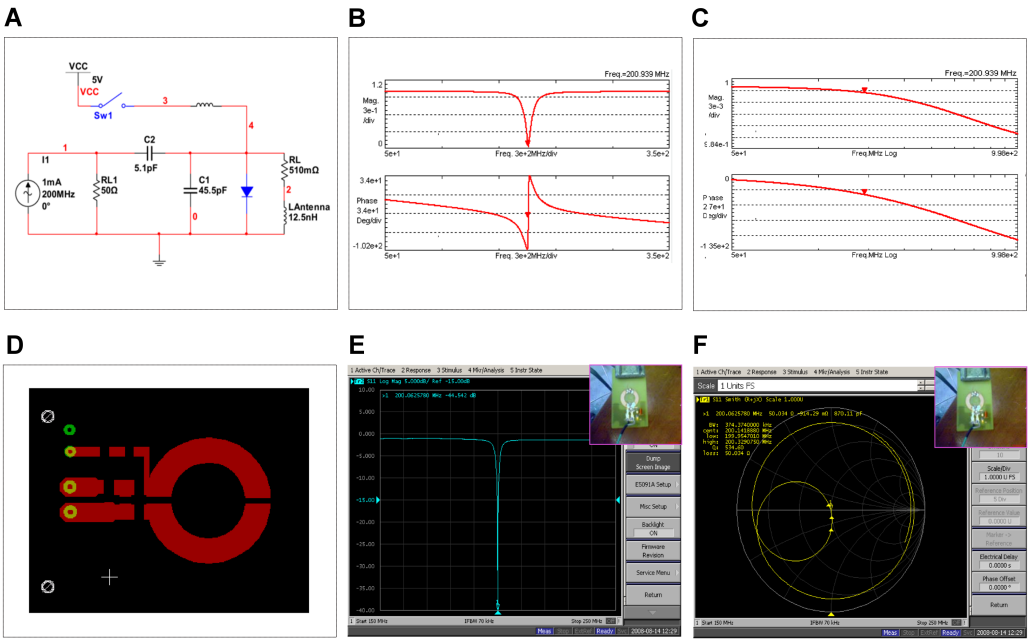


Figure 1: MRI single surface coil fabrication. **A**. Schematic of the circuit. **B**. Simulation of magnitude and phase responses of the circuit in A. **C**. The output of the circuit in magnitude and phase when the decoupling pin diode (line 4 in A) is on, providing protection to the coil against MRI RF pulsation. **D**. PCB of the real single surface coil. **E**. The amplitude response of the designed circuit read in the network analyzer around frequency F_r. **F**. Zero Index of signal reflexion in the coil's terminals depicted in a Smith diagram. The PCB in D, when mounted, is used to create the displays in panels E and F (see upper right corner of these two panels).

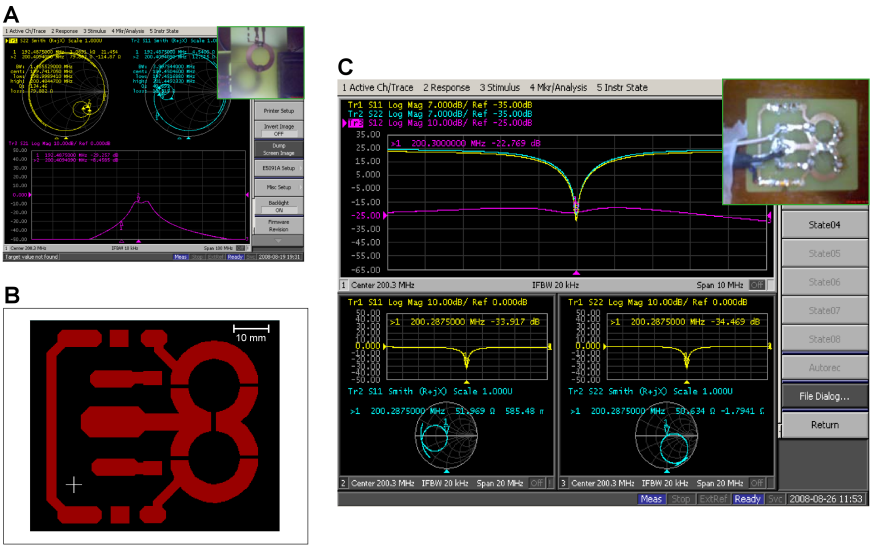


Figure 2: Two loops MRI receptor. **A**. Reduced emission/reception capabilities due to MI (see loops in the upper right panel). **B** PCB of the two-elements coil. **C**. Proof-of-concept: two coils magnetically separated by a shared capacitor. Here, the loops behave as if they were single devices, similar to the scheme in figure 1-E. The Smith diagrams show perfect matching and tuning in both loops

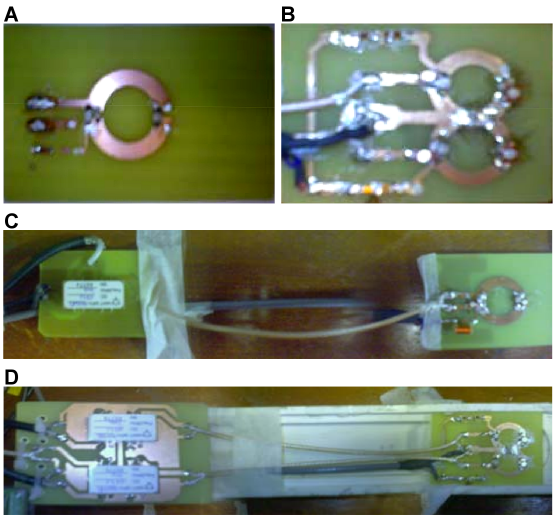


Figure 3: **A**.Single loop device (SD) **B**. Two elements device (DD), using a shared capacitor for decoupling. **C**. Single loop device with pre-amplifier (SDWP) connected trough a $\frac{\lambda}{4}$ cable. **D**. Two elements device with pre-amplifier (DDWP), using a shared capacitor and connected trough a $\frac{\lambda}{4}$ cable.

(f_r) which corresponds to 200.3 MHz, as dictated by hydrogen spins magnetized at 4.7T. The different devices that we tested consist of:

- **SD** → A single-loop coil, $15mm$ in diameter.
- **DD** → A two-elements coil with shared capacitor, $2x15mm$ in diameter.
- **SDWP** → A single-loop coil with low impedance amplifier, $15mm$ in diameter.
- **DDWP** → A two-elements coil with shared capacitor and pre-amplifiers, $2x15mm$ in diameter.

Figures 1 and 2 show the circuit design and simulation, as well as evidence of accurate tuning and matched properties for the SD and DD devices.

3.1 Reducing di/dt. The low-impedance pre-amplifier

The role of the low-impedance pre-amplifier is to reduce current flow trough the coils' terminals; therefore, the MI is also lowered [20]. The objective of this strategy is to obtain a circuit that has a high impedance on the coil side and, simultaneously, a low impedance at the preamplifier's terminals at frequency F_r. In order to accomplish this, the $\lambda/4$ concept described in [21] is used. Briefly, the $\lambda/4$ length happens to be the distance at which a coaxial cable like the one used for transmission inside the scanner, becomes the desired device (high impedance on one side, a low impedance in the opposite). Using the network analyzer and the Smith diagram, this length was found to be $21cms$.

The $\lambda/4$ cable is then plugged-in to the terminals of the two devices already created. Figure 3 shows the final setup of the receptors.

The images are created on phantoms using T1 contrast and multi-slice-multi-echo sequence (MSME-pvm) [22]. The repetition and echo times (TR and TE) are 261.4 and 10.7 ms, respectively, with a flip angle of 180°; isometric in-plane resolution of $0.136mm$; slice thickness of $2mm$; $35x35mm^2$ FOV; $256x265$ matrix, 8 slices for single loop coils, and 16 for the two-elements ones. The acquisition time is kept at 3.06 minutes for all experiments. The phantom for single loops coils consists of a circular syringe 15mm in diameter, while the two-elements devices scanned a syringe of 30 mm in diameter. Both phantoms contain a mix of water, salt (4.5%) and gadolinium. We performed six acquisitions per device. The five most centered slices in the volume of interest of each experiment are used to compute statistical comparisons (see Table 1).

3.2 Measures of coils performance and comparisons mechanisms

Measures of SNR, field stability, and visual depth were gathered from the acquired images. The SNR is obtained in the time-domain as the ratio between the signal - high-intensity pixels arranged in a circular fashion - and the standard deviation of the intensities in the background. The field stability was characterized by the horizontal profile while visual depth was obtained from the vertical profiles, where amplitude and signal decay are separately analyzed. All measures were done in axial views. A Shapiro-Wilk test was run over the data to determine the normality of the distribution. T-tests were performed on all the normally distributed data. However, the horizontal and vertical (amplitude and decay) profiles resulted in not normally distributed data, so we used the Mann-Whitney U test for statistical comparisons. Table 1

summarizes the factors we analyzed, the statistical tests, the comparisons we used and the questions that we addressed with each test. Statistical significance is assessed using a threshold of $p = 0.05$. All the statistics mentioned above were computed with SPSS.

4 Results

This section presents the results of applying the measurements described in Subsection 3.2. Color coding and labeling are provided; moreover, the colors are coherent throughout all sections for a fast association between the result and the device under test. The leftmost panel in each figure shows the regions where the data were gathered.

4.1 SNR comparisons

For each device, the SNR was measured by taking as signal a disk in the high-intensity region at the center of the image and, as noise, the standard deviation of intensities in the rest of the FOV.

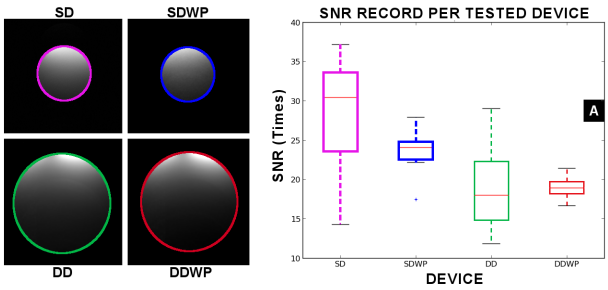


Figure 4: Panel A. Single loop devices exhibit a higher image quality as measured by the SNR, nevertheless it is in this group (the single devices) where the pre-amplifier generates the largest gap in quality when comparing with its counterpart (SDWP). Remarkably, in both, single and double device groups, the use of the pre-amplifier reduces the noise dispersion. In the double-loop devices, the pre-amplifier does not modify the SNR.

No statistical differences in SNR were found (t(14)=1.47, p=0.16) between SD ($\overline{x} = 28.54 \pm 7.95$) and SDWP ($\overline{x} = 23.44 \pm 3.48$). For the DD devices, the average for DD and DDWP were respectively 19.05 ± 6.35 and 18.99 ± 1.64. They were not statistically different from each other (t(10)=0.024, p=0.98).

4.2 Field stability

A Gaussian filter is applied to the original profile aiming to recover its low-frequency component, which is in turn, a measure of the field stability. The cutoff frequency is automatically extracted from the properties of the given profile by truncating its Fourier spectrum at 90% of the spectral energy content. Posteriorly, a homemade program iteratively looks for a polynomial that fits the filtered signal. The algorithm stops when the correlation coefficient reaches the value of 0.995, then, the index of the fitted polynomial is saved for each profile. In SD devices, the fitted polynomial found from 3 to 7 constraints, with an average of 5 for the two included devices. In DD devices, there is a natural increment of the tortuosity due to the boundary created by the two channels. The asymmetry in the profile was a constant in all the tested devices; we hypothesize this is due to a difference in the copper deposition in the construction process.

4.3 Visual depth

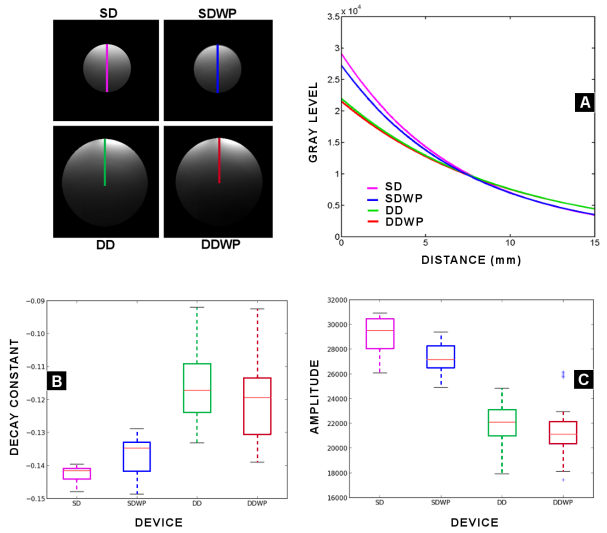


Figure 6: Comparison of vertical profiles. The vertical profile can be compared among all tested devices since the composition in the phantoms is the same for all devices regardless of their size. Panel A shows the decay while penetrating the phantom in all devices. Panels B and C show the data distribution in decay and amplitude respectively for each tested device.

In the single loop device, the use of the pre-amplifier reduces the dispersion in the decay constant. Nonetheless, the mean value for this decay is faster when the pre-amplifier is used. Both versions of the two elements loop allow the system to decay more slowly compared with their single loop counterparts. The amplitude in SD and DD devices can not be compared due to the loss of quality factor evidenced when looking at **Figures 1-E** and **2-C**. Additionally, the amplitude and the decay constant can be separated for this specific analysis since the peak amplitude of the signal depends solely on construction specifications while the signal decay is a function of recruited spins. Considering that the phantom contains the same solution for all tested devices, the signal decay can be use for comparing all built devices, but an amplitude comparison will be only fair if done intra-group.

Table 1: **Questions addressed and statistical tests performed**

Factor	Test	Comparisons	Addressed question
SNR	T-test	(SD-SDWP)	Is the quality of the images affected by addition of the pre-amplifier?
SNR	T-test	(DD-DDWP)	Is the quality of the images affected when the pre-amp and the shared capacitor are in place?
Horizontal	MWU	(SD-SDWP)	Is the field homogeneity affected by the pre-amp?
Horizontal	MWU	(DD-DDWP)	Is the field homogeneity affected by the pre-amp if another MI avoiding strategy is used?
Amplitude	MWU	(SD-SDWP)	Is the amplitude of the acquisitions affected by the use of the pre-amp?
Amplitude	MWU	(DD-DDWP)	Is the amplitude of the acquisitions affected when the pre-amp and the shared capacitor are in place?
Decay	MWU	(SD-SDWP)	Is the use of the pre-amp affecting the visual depth?
Decay	MWU	(DD-DDWP)	Is the visual depth affected when pre-amp and the shared capacitor are in place?
Decay	MWU	(SD-DD)	Is the visual depth of the coils increased when the inner diameter of the device decreases?
Decay	MWU	(SDWP-DDWP)	Is visual decay affected when two MI avoiding strategies are implemented together?

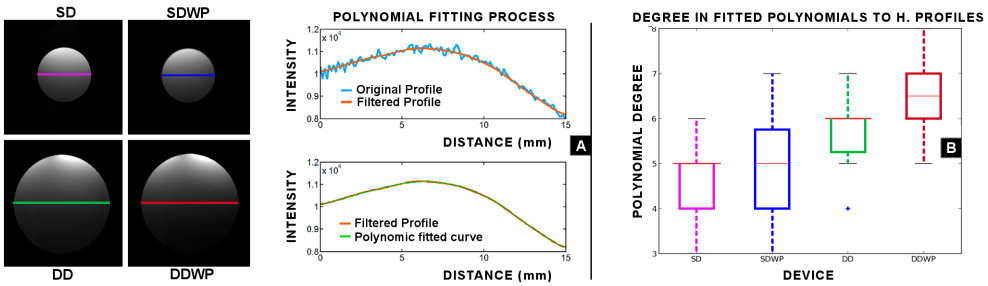


Figure 5: Panel A. Summary of the procedure for the horizontal profiles. Panel B. Distribution in tortuosity of the horizontal profiles in each device, measured as the degree of the first polynomial adjusted at 0.995 index of correlation

Table 2: **Results of statistical tests**. * experiments where statistical differences are found

Factor	Test	Comparisons	Average $\overline{x}\pm$ Standard deviation	p-value	Statistic values
SNR	T-test	(SD-SDWP)	$(28.54 \pm 7.95),(23.44 \pm 3.48)$	0.16	t=1.47 df=14
SNR	T-test	(DD-DDWP)	$(19.05 \pm 6.35),(18.99 \pm 1.64)$	0.97	t=0.03 df=9
Horizontal	MWU	(SD-SDWP)	$(4.46 \pm 0.86),(4.76 \pm 1.33)$	0.41	U= 396.50, Z=0.83
Horizontal	MWU	(DD-DDWP)	$(5.63 \pm 0.85),(6.5 \pm 0.97)$	0.0013*	U=247.50, Z=3.22
Amplitude	MWU	(SD-SDWP)	$(29096.36 \pm 1586.54),(27222 \pm 1315.57)$	0.001*	U=731.00, Z=-4.15
Amplitude	MWU	(DD-DDWP)	$(21971.4 \pm 1665.32),(21526.8 \pm 2375.99)$	0.1421	U=550.00, Z=-1.48
Decay	MWU	(SD-SDWP)	$(-14.24 \pm 0.002),(-0.1366 \pm 0.005)$	0.0002*	U=193.50, Z=-3.79
Decay	MWU	(DD-DDWP)	$(-0.115 \pm 0.010),(-0.119 \pm 0.012)$	0.224	U=533.00, Z=1.23
Decay	MWU	(SD-DD)	$(29096.36 \pm 1586.54),(21971.4 \pm 1665.32)$	0.001*	U=0.00, Z=6.65
Decay	MWU	(SDWP-DDWP)	$(27222 \pm 1315.57),(21526.8 \pm 2375.99)$	0.001*	U=109.50, Z=5.03

The built devices are capable of MRI imaging, that should be and indubitable statement at this point due to the vast evidence presented with the Smith diagrams in **Figures 1** and **2**. However, we show in **Figure 7**, proof of the images that may be acquired with the sort of surface receptors designed and built along the development of the experiments that support the statements in this manuscript.

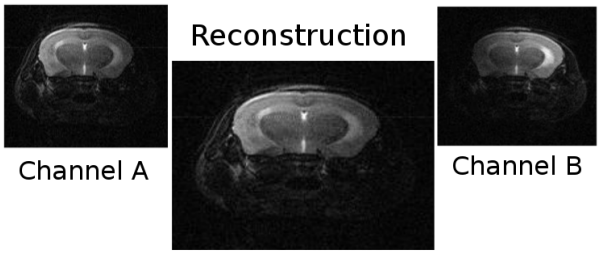


Figure 7: Image of mice's brain using a DDWP device

5 Discussion

Our hypothesis is that the DD devices behave similarly to the SD devices when MI is diminished or avoided, and this claim is validated in those experiments with no statistical differences. Hence, we focus the discussion on those trials that resulted in statistical differences, and on their practical implications. Refer to the entries labeled with * in Table 2.

A significant difference (U= 247.50, Z=3.22, p=0.001) is found between DD ($\overline{x} = 5.63 \pm 0.85$) and DDWP ($\overline{x} = 6.5 \pm 0.97$). The mean rank of DD was 32.25, and that of DDWP was 23.75. These results have a low practical impact since the horizontal dynamics can be controlled by software reconstruction. In this case, we used a geometrically driven algorithm to merge the two channels, but customized setups can be implemented to optimize the smoothness of the boundary [1]. Nevertheless, the high degree of the fitted polynomials, even in the single loop devices, suggest a lack of uniformity in the copper deposition of our devices. This is observable in figure 5-A, which belongs to the SD device. A perfectly build coil within a flat B_0 field should have a profile that can be modeled with a two-degree polynomial.

In the amplitude analysis, there was a significant difference ($p \leq 0.05$) between SD ($\overline{x} = 29096.36 \pm 1586.54$) and SDWP ($\overline{x} = 27222 \pm 1315.57$) (U= 731.00, p=0.0013). The mean rank of SD was 21.13, and the mean rank of SDWP was 39.87, suggesting that the use of the pre-amplifier negatively affected the reading capabilities of the device; However, the practical implications of this finding are null, since it affects the single loop devices where the pre-amplifier does not introduce any added value. The pre-amplifier was connected to a single loop device with the purpose of creating a fair counterpart to compare the outcomes of the DDWP device. More importantly, the pre-amplifier in DD devices does not result in statistical differences, a fact of high relevance towards applicability.

In the decay analysis, there is a significant difference ($U = 193.50, p < 0.001$) between SD ($\overline{x} = -14.24 \pm 0.002$) and SDWP ($\overline{x} = -0.1366 \pm 0.005$). The mean rank of SD decay is 21.95, and the mean rank of SDWP decay is 39.05. There is no significant difference ($U = 533.00, p = 0.22$) between DD ($\overline{x} = -0.115 \pm 0.010$) and DDWP ($\overline{x} = -0.119 \pm 0.012$). The mean rank of DD decay is 33.27, and that of DDWP decay is 27.73. These results imply that visual depth is affected by the pre-amplifier in the group of devices that are not affected by the MI. In contrast, the pre-amplifiers do not have a significant influence on the signal in DD devices, where MI exists.

There was also a significant difference ($U = 0.00, p < 0.00$) between SD ($\overline{x} = -14.24 \pm 0.002$) and DD ($\overline{x} = -0.115 \pm 0.010$). The mean rank of SD decay was 45.50, and the mean rank of DD decay was 15.50. The reduction of the diameter in the coils improves the reading capabilities of these devices. There was a significant difference ($p \leq 0.05$) between SSWP ($\overline{x} = -0.1366 \pm 0.005$) and DDWP ($\overline{x} = -0.119 \pm 0.012$); $U = 109.50, p < 0.001$. The mean rank of SDWP decay was 41.85, and the mean rank of DDWP decay was 19.15, which suggest that both strategies can be used together without affecting the quality of the images.

6 Conclusions

The small animal modeling field would be greatly benefited if accuracy in the visual monitoring methods is improved. An imaging setup with better outcomes would also yield more robust results, and the derived conclusions would be less questionable. In the end, animal experimentation is the only means that humanity has to test and study the progression of diseases and to anticipate the outcomes of procedures in living subjects. In this manuscript, we demonstrated that a better MRI receptor device, suitable to work at small animal dimensions is feasible. To this purpose, we implemented MI avoiding strategies and proved that they could be used together without significantly affecting the signal quality while enabling a more efficient use of the scanning time.

References

[1] O. Beuf, F. Jaillon, and H. Saint-Jalmes, "Small-animal MRI: Signal-to-noise ratio comparison at 7 and 1.5 T with multiple-animal acquisition strategies," *Magn Reson Mater Phy*, vol. 19, pp. 202 – 208, 2006.

[2] P. Roemer, W. Endelstein, C. Hayes, S. Souza, and O. Mueller, "The NMR Phased Array," *Magenetic Resonance In Medicine*, no. 16, pp. 192–225, 1990.

[3] D. Gareis, T. Wichmann, T. Lanz, G. Melkus, M. Horn, and P. M. Jakob, "Mouse MRI using phased-array coils," *NMR Biomed.*, vol. 20, pp. 326 – 334, 2007.

[4] K. Sahay and S. Pathak, *Basic Concepts Of Electrical Engineering*. New Age International, 2006.

[5] N. D. Zanche, J. A. Massner, C. Leussler, and K. P. Pruessmann, "Modular design of receiver coil arrays," *NMR Biomed.*, vol. 5, 2007.

[6] S. S. Hidalgo, R. Rojas, and F. A. Barrios, "A petal resonator volume coil for MR neuroimaging," *Revista Mexicana de Fisica*, vol. 52, no. 3, pp. 272 – 277, 2006.

[7] R. F. Lee, C. R. Westgate, R. G. Weiss, D. C. Newman, and P. A. Bottomley, "Planar strip array. PSA," *Magnetic Resonance in Medicine*, vol. 45, no. 1, pp. 673 – 683, 2001.

[8] C. C. Guclu, E. Boskamp, T. Zheng, R. Becerra, and L. Blawat *Journal of Magnetic Resonance Imaging*, vol. 19, pp. 255 – 258, 2004.

[9] B. Keil and L. L. Wald, "Massively parallel MRI detector arrays," *Journal of Magnetic Resonance*, vol. 229, pp. 75 – 89, 2013.

[10] C. Qian, G. Zabow, and A. Koretsky, "Engineering novel detectors and sensors for MRI," *Journal of Magnetic Resonance*, vol. 229, pp. 67 – 74, 2013.

[11] D. Sodickson and W. Manning, "Simultaneous acquisition of spatial harmonics (SMASH): fast imaging with radiofrequency coil arrays.," *Magnetic Resonance in Medicine*, vol. 38, pp. 591–603, 1997.

[12] K. P. Pruessmann, M. Weiger, M. B. Scheidegger, and P. Boesiger, "SENSE: Sensitivity Encoding for Fast MRI," *Magnetic Resonance in Medicine*, vol. 42, no. 3030, pp. 952 – 962, 1999.

[13] M. A. Griswold, P. M. Jakob, R. M. Heidemann, M. Nittka, V. Jellus, J. Wang, B. Kiefer, and A. Haase, "Generalized autocalibrating partially parallel acquisitions. GRAPPA," *Science*, vol. 47, pp. 1202 – 1210, 2002.

[14] P. Qu, G. X. Shen, C. Wang, B. Wu, and J. Yuan, "Tailored utilization of acquired k-space points for GRAPPA reconstruction," *Journal of Magnetic Resonance*, vol. 174, pp. 60 – 67, 2005.

[15] M. Blaimer, F. Breuer, M. Mueller, R. M. Heidemann, M. A. Griswold, and P. M. Jakob, "SMASH, SENSE, PILS, GRAPPA How to Choose the Optimal Method," *Magn Reson Imaging*, vol. 15, no. 4, pp. 1 – 10, 2004.

[16] D. Gareis, T. Neuberger, V. C. Behr, P. M. Jakob, C. Faber, and M. A. Griswold, "Transmit-receive coil-arrays at 17.6T, configurations for 1H, 23Na, and 31P MRI," *Magn Reson Engineering*, vol. 29, pp. 20 – 27, 2006.

[17] K.-I. Yamada, R. Murugesan, N. Devasahayam, J. A. Cook, J. B. Mitchell, S. Subramanian, and M. C. Krishna, "Evaluation and Comparison of Pulsed and Continuous Wave Radiofrequency Electron Paramagnetic Resonance Techniques for in Vivo Detection and Imaging of Free Radicals," *Journal of Magnetic Resonance*, vol. 154, pp. 287 – 297, 2002.

[18] Cadsoftusa, "Eagle pcb software," 2008.

[19] R. M. Arnold, "Transmission Line Impedance Matching Using the Smith Chart," *IEEE Transactions in Microwave Theory Technology*, vol. 22, no. 11, pp. 977–978, 1974.

[20] A. Reykowski, S. M. Wright, and J. R. Porter, "Design of Matching Networks for Low Noise Preamplifiers," *Magnetic Resonance in Medicine*, no. 33, pp. 848–852, 1995.

[21] I. Bahl, *Lumped Elements for RF and Microwave Circuits*. Artech House, 2003.

[22] F. Hennel, "Experiment Programming with ParaVision on BioSpec and PharmaScan Systems," pp. 28–32, 2002.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 2, No. 3, 540-544 (2017)
www.astesj.com
Special Issue on Recent Advances in Engineering Systems

ASTES Journal
ISSN: 2415-6698

An Encryption Key for Secure Authentication: The Dynamic Solution

Zubayr Khalid[*,1], Pritam Paul[1], Khabbab Zakaria[2], Himadri Nath Saha[1]

[1]*Institute of Engineering and Management, Computer Science and Engineering, 700091, India*

[2]*Jadavpur University, Department of Power Engineering, 700098, India*

ARTICLEINFO

Article history:
Received: 06 April, 2017
Accepted: 15 May, 2017
Online: 01 June, 2017

Keywords :
AES
Cryptography
Dynamic Password
EYE
Feistel Block Cipher

ABSTRACT

In modern day technology, the Information Society is at risk. Passwords are a multi-user computer systems usual first line of defence against intrusion. A password may be textual with any combination of alphanumeric characters or biometric or 3-D. But no authentication protocol is fully secured against todays hackers as all of them are Static in type. Dynamic authentication protocol is still a theoretical concept. In this paper, we are focusing on a concept of authentication technique which is actually dynamic in genre, i.e. the password here will change in t time (where t is as small as possible). This technique comprises of both hardware and software part. In this paper, we have covered the idea of generating an efficient algorithm that can work as the final in the Dynamic Password Authentication system. We have used standard deviation within statistics to generalize the possible password which is further secured by Feistel Block Cipher and Advanced Encryption Standard technique (AES), leading and following the said mathematics respectively. In order to allow the system to create variable password in the least time interval possible, we must make sure our process is not much complex.

1 Introduction

This paper is an extension of the research work named Secure Authentication with Dynamic Password presented in the IEEE IEMCON, , Vancouver, Canada, 2016. [1]

When it is the matter with network security, we have to be updated always. As for example, today we simply cannot afford any security system that includes a dilapidated protocol (as for example Caesar Shift Cipher). Security protocols are used to protect several different medium using passwords, which supposed to be known only to the users of the system. Textual password authentication system is the most widely used protocol, though it is not much secure against Shoulder Surfing Attack, Brute Force Attack [2]. Smartcard concept, 3-d password concept are not totally protected either [3], [4]. Biometric system may be the most powerful authentication system ever, but it is much costly [2]. Dynamic password authentication system is the latest system that has come in the field. In this work we have discussed about the cryptographic way to generate the dynamic password for the authentication protocol described in the said conference. [1]

The authentication protocol described in the conference is a combination of both the hardware and software. Here we are imagining a chip which will be used as the key to unlock different ids. According to the main work, we have taken a random textual password e.g. Password. The ASCII values of each letter acts as equivalence to mass where we have used Einstein's Theory on bending Space-Time Continuum because of Mass. Now we have placed the textual password (here Password) on 3-d box which is described by the X-Y-Z coordinates.[1]

Later we have used the concept of the eye. The EYE (technically a sensor) is the only object inside the 3-D box.

The EYE will actually take a random position inside the 3-D box (i.e. the X, Y, Z coordinates of it will be selected randomly). Now the EYE will trace the endings of the shifted parts in the plane by joining them.

[*]Zubayr Khalid, Computer Science and Engineering, Institute of Engineering and Management, +917596872617 & write2zubayr@gmail.com

https://dx.doi.org/10.25046/aj020369

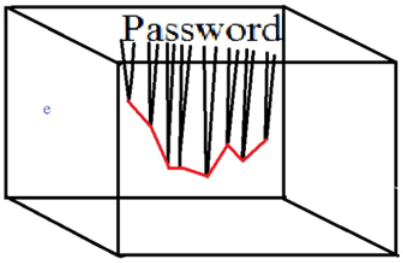


Fig.1: Traced path according as the position of the eye

If we manage to let the eye to change its position in t time (t is as small as possible), then during each time interval we will have a different path as shown.

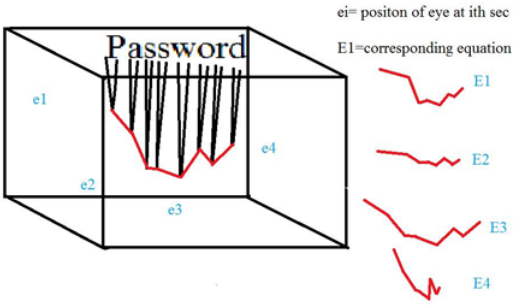


Fig.2: Traced path by the eye

Since at a certain time, the received traced path is the key of the real password, by changing the position of the EYE, and hence by getting different traced path in each time interval t, we gain the dynamically changeable password.

In this paper we have tried to make an algorithm that will be able to convert the traced path (technically a geometric figure) to a combination of binary digits (0 and 1). This expression gained at a certain time will be the password at that time interval. We will try to add some cryptographic concept in the algorithm to make a final defence in the system. The mathematical idea used to convert the geometric figure an expression of standard deviation (in decimal) will be both followed by and lead two different cryptographic concepts the first one is general Feistel Block Cipher concept,[5], [6] and the later one is AES (Advanced Encryption Standard) algorithm.

2 Security with Feistel Block Cipher Model

We will move further assuming that the textual password chosen by the user was Password. Before going to the generation of the password from the traced path, we need to find the coordinate of the points of the path, which we will do in this process. We will follow the general Feistel Block Cipher concept, i.e. at first, we will divide the plaintext into 2 equal parts say L and R. The plaintext will be the ASCII representation of the textual password. Hence we have the plaintext ASCII representation of Password 8097115115119111114100. [5],[6]

Now the string will be divided into 2 halves based on the length of it. If the length of the string is found even, the L, R will be of equal length. If the length of the string is found odd, the length of L will be 1 greater than that of R.

Here we have the length 22. Hence L will be 80971151151 and R will be 19111114100. Each digit in L and R are added and the results will be stored in R‘ and L‘ respectively. The addition of digits in L is 39 and the addition of digits in R is 20. Hence R‘ will store 39 and L‘ will store 20.

We will count (R‘ modulo 2) and (L‘ modulo 2). To find the axis of the traced path, we will follow a concept

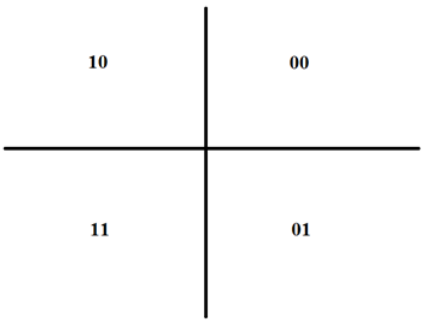
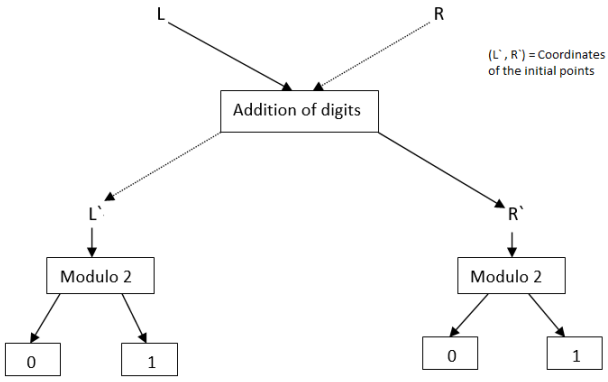


Fig.3: X-Z Coordinate Plane

As for example if (L‘ modulo 2) is 1 and (R‘ modulo 2) is 1 then we will consider that the traced path will be in 3rdquadrant and proceed to the mathematics (hence 0 represents positive quadrant and 1 represents negative quadrant). The values of L‘ and R‘ will be the x-coordinate and z-coordinate of the initial point in the traced path respectively.



L` mod 2	R` mod 2	quadrant
0	0	first
0	1	forth
1	0	second
1	1	third

Fig.4: Feistel Block Cipher Model to get quadrant and initial coordinate

In our example, (L‘ modulo 2) is 0 and (R‘ modulo 2) is 1. Hence the traced expression will be in the 4thquadrant. The initial point will be (20, -39). It

should be mentioned that the graph, starting from a specific quadrant, may end in a different quadrant.

A case to be noted that in this concept we havent used Feistel Block Cipher completely. Here we havent used any key and there is no functioning it with L. Instead, we have changed both L and R to R‘ and L‘ respectively. The swapping of L and R to R‘ and L‘ is done only once instead of having several rounds.

3 Derivation of the Traced Path

Suppose there are n letters in the password. So, we have n corresponding extreme points in the x-z graph. Here, we connect each 2 consecutive extreme points to get one straight line. Thus, in total we get (n-1) of such straight lines, each having a definite equation f(z,x)=0. The equation of each straight line is provided by the system.

Our job is to find variance (similar to what is used in case of problems with large number of bodies)[7], and thus standard deviation for the distance of each part from x axis (z) for each of (n-1) Straight lines.

$$Variance = \sigma^2 = < z^2 > - < z >^2 \quad (1)$$

$< z >= \int_a^b z\rho(x)dx$ where (a, z_1), (b, z_2) are the two extremities of a Straight lines where $b > a$. Here $\rho(x)$ is the probability density such that $\rho(x)dx$ defines the probability of an individual being in the interval of (x,x+dx).

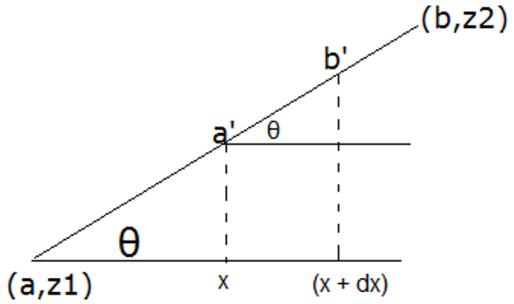


Fig.5: Derivation of traced path

Here, length of $a'b' = dxSec\theta$
Length of $ab = xSec\theta$
So,

$$\rho(x)dx = \frac{dx}{x} \quad (2)$$

Therefore,

$$< z >= \int_a^b \frac{zdx}{x} \quad (3)$$

$$< z^2 >= \int_a^b \frac{z^2 dx}{x} \quad (4)$$

Performing the integrations, we can have σ^2
Now, standard deviation= σ
Thus, we will have (n-1) no. of standard deviations for (n-1) of Straight lines.

Let, these standard deviations are σ_1, σ_2,..... $\sigma_{(n-1)}$

Let, $\sigma_k = \Sigma_{i=1}^{n-1} \frac{\sigma_i}{n-1}$, is the standard deviation.

Standard Deviation of those standard deviations is:

$$\sigma_p = \frac{\sqrt{\Sigma_{i=1}^{n-1}(\sigma_k - \sigma_i)^2}}{n-1} \quad (5)$$

σ_p is the password at that moment.

Come to think about it, for a password with n letters, there are $(26+10+30)^n = 66^n$ [26⟶ English letters, 10⟶numbers from 0 to 9, 30⟶special characters] no. of different possibilities, and thus 66^n no. of different combinations of Straight lines in z-x graph. Each of the point in z-x graph can arbitrarily be on each of the 4 coordinates. So, there are 4^npossibilities for each of 66^n different possibilities. Finally, the final standard deviation σ_p depends on the position of average of different σ_i. There are (n-1) of such partitions.

So, probability of a random password, matching the real password be:

$$Probability = \frac{1}{(66\times 4)^n \times (n-1)} = \frac{1}{264^n \times (n-1)} \quad (6)$$

For a 8 digit password, this probability is $\sim 6 \times 10^{-21}$

4 Final Security with AES Algorithm

This process is the final process in the whole system. Here we will use Advanced Encryption Standard (AES) to get the encrypted cipher text [8]. The cipher text will be the actual password in the dynamic password authentication system.

Reasons behind selecting AES encryption for our process:

(i) AES is one of the most efficient and strongest encryption systems.

(ii) Here we are not much concern of any decryption of the encrypted text (i.e. the password). Hence theres no necessity of choosing asymmetric encryption system like RSA. [9]

(iii) Comparing Twofish encryption system and AES encryption system, Twofish slightly leads in terms of security. But AES is faster than Twofish. Since our system gets high security due to the changeable password, we will allow it to change in the minimum time interval possible. Hence we are choosing AES for our authentication system. [10]

According to AES system, we will have a plaintext and key of 128 bits and there will be total 10 rounds- 9 regular rounds and 1 final round.

We will get the plaintext from the standard deviation of the previous mathematics.

Let the output from the mathematics is 6.4467. To have a plaintext of 128 bit (i.e. 16 bytes), the output is multiplied with pi (round 16 bytes decimal). Here

the output will be:20.25290535989732. Each digit excluding the point (2, 0, 2, 5, 2, 9, 0, 5, 3, 5, 9, 8, 9, 7, 3, 2) is considered and the corresponding binary digit in 8 bit is calculated (i.e. 00000010, 00000000, 00000010, 00000101, etc). Hence we have 128 bit plaintext.

To get the key, the textual password (here Password) is converted to binary. To get a 128 bit (16 byte) key the textual password will have to be of 16 letters. If the textual password is less than 16 letters then there will be a repetition of letters until it reaches 16 letters. If there is more than 16 letters then there will be excise of letters from right to left until it reaches 16 letters. At the end there will be text to binary conversion.

As for example, for the textual password Password since there is 8 letters, after updating we have PasswordPassword. Hence the key will be: 01010000 01100001 01110011 01110011 01110111 01101111 01110010 01100100 01010000 01100001 01110011 01110011 01110111 01101111 01110010 01100100.

Now we will move for AES algorithm.

The algorithm our system will follow can be described in the figure: [11]

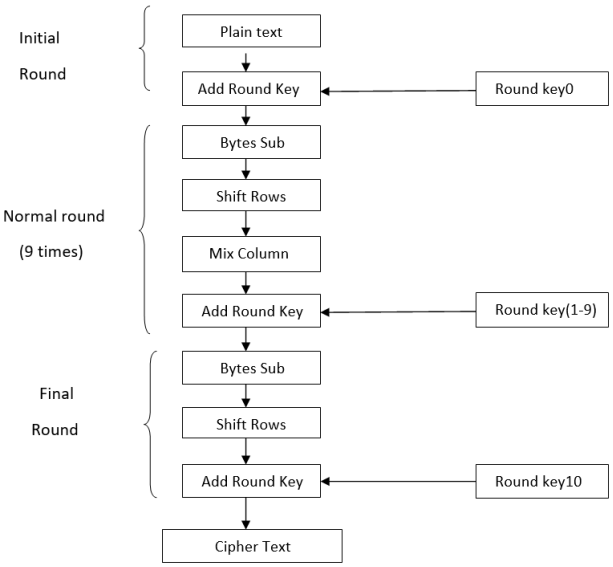


Fig.5: AES Algorithm

The cipher text we get after whole 10 rounds is the actual password of the system. It is to be noted that we havent done anything with decryption since we are not concerned with the decoding of password to original text.Since the password consists of 128 different characters each having 2 possibilities (0,1), the number of different possible passwords is 2^128.

5 Conclusion

In the main work we have already clarified that the concept is secured against most of the common security threats like Brute Force Attack, Timing Attack, Key Logger and Shoulder Surfing Attack. Moreover, it is shown that the Space and Time Complexity of the concept is less than some other common authentication techniques. The whole idea has been developed to make a secure yet simple concept, creating a user-friendly platform. The time interval t between two different traced paths will depend on the time required for the execution of the two encryption algorithms (Feistel Block Cipher and Advanced Encryption Standard technique), along with the completion of the standard deviation. Less is the t, greater is the possibility of having 2 different traced paths in a time interval T, hence greater is the security. The algorithms used in generating the binary password from the graphical traced path are recognised as some secure encryption techniques. Hence they can perform as the final defence as well. At the end we should mention that since we have worked with a completely theoretical concept, it should go through an experimental procedure to detect the flaws and disadvantages.

References

1. Z. Khalid, P. Paul, S. P. Chattopadhyay, A. N. Biswas, Secure Authentication with Dynamic Password, in 2nd IEEE IEMCON and 7th IEEE Annual Information Technology, Electronics and Mobile Communication Conference, Vancouver Canada, 2016.
2. C. Kuo , S. Romanosky , L. F. Cranor , Human Selection of Mnemonic Phrase-based Passwords, [online] available- *https* : *//cups.cs.cmu.edu/soups/*2006/
3. By the editors of Time-Life Books, Computer Security (Understanding Computers),Time-Life Books , 1990.
4. F. A. Alsulaiman, A. El Saddak, A Novel 3D Graphical Password Schema, in Virtual Environments, Human-Computer Interfaces and Measurement Systems, Proceedings of 2006 IEEE International, La Coruna Spain, 2006
5. W. Stallings, Cryptography and Network Security (Fourth Edition), Pearson Education Inc., 2006

   $proceedings/p67_kuo.pdf$, 2006
6. C. De Canniere, A. Biryukov, B. Preneel, An Introduction to Block Cipher Cryptanalysis, Proceeding of The IEEE, Vol. 94, No. 2, 2006
7. D. J. Griffiths, Introduction to Quantum Mechanics, Prentice Hall, Inc. USA, 1995.
8. D. Selent , Advanced Encryption Standard, Rivier Academic Journal, Volume 6, Number 2, 2010
9. P. K. Arya, M. S. Aswal, V. Kumar, Comparative Study of Asymmetric Key Cryptographic Algorithms, International Journal of Computer Science & Communication Networks, Volume 5, Issue 1, 2015

10. S. A. M. Rizvi, S. Z. Hussain, N. Wadhwa, Performance Analysis of AES and Twofish Encryption Schemein International Conference on Communication Systems and Network Technologies, Katra, Jammu, India, 2011.

11. H. Lee, K. Lee, Y. Shin, AES Implementation and Performance Evaluation on 8-bit Microcontrollers, (IJCSIS) International Journal of Computer Science and Information Security, Vol. 6, No. 1, 2009

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 2, No. 3, 545-552 (2017)
www.astesj.com
Special Issue on Recent Advances in Engineering Systems

ASTES Journal
ISSN: 2415-6698

Multi-Domain Virtual Network Embedding with Coordinated Link Mapping

Shuopeng Li [*], Mohand Yazid Saidi, Ken Chen

L2TI, Institut Galiléee Université Paris 13 Sorbonne Paris Cité Villetaneuse, France

ARTICLEINFO

Article history:
Received: 03 April, 2017
Accepted: 12 May, 2017
Online: 01 June, 2017

Keywords :
Network Virtualization
Virtual network embedding
Multi-domain

ABSTRACT

Network Virtualization, which allows the co-existence of various logical networks on shared physical infrastructure, has become popular in recent years. The optimal mapping of virtual resource to physical resource is a major issue in network virtualization. This problem, called virtual network embedding (VNE), has been well explored in the context of one physical domain, which is in practice operated by a single infrastructure provider (InP). However, the needs of virtual network (VN) is rapidly growing, and quite a number of VNs have to be established across multi-domain. For multi-domain VNE, infrastructure providers can no longer just solve their own single domain VNE problem, but have to cooperate to build the whole VN. Therefore, new challenge arises for the multi-domain VNE, compared to traditional single domain VNE. The existing investigations on this problem mainly focus on decomposing a VN to sub VN for each domain, but little attention has been paid to the joint relation between intra-domain and inter-domain (peering) links. In this paper, we propose a multi-domain link mapping method which combines the intra and peering link mapping so as to optimize the overall resource utilization. Our approach is easy to be deployed since it is based on current Internet architecture. Evaluation shows that our approach brings improvements related to existing methods.

1 Introduction

Network virtualization [1] is regarded as a solution to overcome some weakness of traditional network architecture. It makes easy to support various separated logical networks running over shared physical network. In virtualized network architecture, the service provider (SP) creates and manages virtual networks (VN) for end users,while infrastructure provider (InP) deploys the substrate network (SN) equipment and offers the necessary physical resources.

An important step of network virtualization is to establish VNs above SNs. This is referred as *virtual network embedding* (VNE). The VNE problem aims to find a mapping from the VN to SN in a way that objectives (*e.g.* cost) are optimized and constraints (*e.g.* bandwidth) are satisfied.

Large networks in current Internet architecture are organized by autonomous system (AS). An AS is one or several physical networks controlled by a single authority. In this article, we use the vocabulary of *"domain"* to denote the whole substrate network under the control of a single InP.

As VNs are getting more and more deployed, VNs in multi-domain will be more and more considered by potential VN users. Establishment of multi-domain VN is more difficult than the one on single domain for at least two reasons:

- First, a single domain VNE problem is mainly solved by linear programming (LP). If we had a complete vision of all the domains, a multi-domain VNE could be considered as a single domain VNE with a very large domain, so computationally harder to solve.
- More importantly, for various reasons (technical, commercial, etc.), the acquisition of full information in multi-domain is costly and often not possible. Only limited information is exchanged between InPs via protocols like BGP, so single domain approach cannot be re-used.

To address these challenges, multi-domain VNE

[*]Shuopeng Li, L2TI, Institut Galiléee Université Paris 13 Sorbonne Paris Cité Villetaneuse, France, li.shuopeng@univ-paris13.fr

https://dx.doi.org/10.25046/aj020370

frameworks are developed. Usually, there is a VN decomposition step followed by local sub VN mapping in each InP. Some authors introduced a broker-like additional actor, termed *Virtual Network Provider* (VNP) [2], between SP and InPs. The role of this VNP consists in assembling multi-domain information, decomposing VN and achieving the multi-domain VNE.

Existing solutions mainly focused on the decomposition of a multi-domain VN to each domain. One of the shortcomings in these frameworks is the lack of efficient link mapping method especially for the peering links which interconnect two domains.

In this paper, we propose an efficient framework of link mapping in multi-domain virtual network embedding context, which jointly consider the mapping of intra and peering links. In our approach, the peering links are mapped simultaneously along with intra domain links. Our approach is based on information usually disseminated by classical routing protocol (like BGP). Our simulation results prove that this solution results in better utilization of substrate resources.

The rest of this paper is organized as follows. Section 2 provides an overview of the related work. Section 3 presents our network model. Section 4 presents our multi-domain VNE solution. The evaluation results are shown in section 5. Section 6 concludes this paper.

2 Related Work

2.1 Single domain VNE

The problem of single domain VNE is NP-hard [3] [4]. A basic off-line approach is proposed in [5], which performs the embedding in 2 stages (node then link mapping). The Multi Commodity Flow (MCF) is introduced in [6] to embed the virtual links The method in [7] takes into account the virtual links in node mapping stage. It privileges such node mapping that reduce the length of substrate link path.

Since 2-stage VNE solutions are lack of cooperation, some solutions mapping nodes and links in the same stage have been proposed. An approach based on subgraph isomorphism detection is presented in [8]. An other model in [9] applies the Markov Random Walk to rank nodes and then embeds links and nodes by using back-tracking strategy based on breadth-first search. In order to meet the change of requirements over time, dynamic VNE is proposed in [10] .

2.2 Multi-domain information disclosure

Because of politic and efficiency reasons, InPs can't disclose their complete information to others, so it is critical to make clear the information disclosure policy.

The proxy VNP could get peering links location and resource information but intra-domain links cannot be assumed to be available to VNP [11].

In [12], three types of resource information in each domain are provided to VNP:

- Node: its location, available capacity and unit price.
- Peering link: its vertices, available capacity and unit price.
- Intra-domain link: a length-based price for connecting any two nodes in its domain.

Based on the information disclosure policy above, we will describe our network model in the next section.

2.3 Multi-domain VNE

Multi-domain VNE framework can be decomposed into three major components [13]:

(i) partitioning the VN request into each InP via multi-domain node mapping method,

(ii) establishing inter-domain connection (peering links) between InPs using inter-domain paths,

(iii) embedding each sub VN request in each InP using intra-domain algorithm.

Based on multi-domain information model introduced in previous section, some centralized multi-domain VNE solutions are proposed in [14][11][12][15].

Many of them mainly focus on the *first component*. In [14], the authors introduce the cost of mapping a virtual node to a domain and the cost of mapping a link between two substrate nodes. Their node mapping algorithm optimizes the total embedding cost. The approach in [12] adopts the node mapping method [7] on a full-mesh topology which complies with partial information disclosure policy.

The *second component* is not very well explored compared to the first component. Existing solutions use simple policies to establish peering links. In [14], each peering link is considered as a single VPN (Virtual Private Network) connection. In [12], the flow of peering links is unsplittable between two domains, while the intra-domain sub virtual links are splittable. The peering link path is determined by Dijkstra's algorithm on VNP layer. Since VNP layer topology is not modified over time because of cost efficiency, Dijkstra's algorithm based peering links have always the same path. This phenomenon will result in difficulty of later intra-domain mapping. In [11], a virtual node is first mapped to substrate peering node to determine the peering link and the InP it belongs to. This approach is suitable for traffic matrix based VN [16] but not topology based VN.

Since establishing peering links is a part of link mapping, the chosen peering nodes will probably influence the intra-domain paths. We believe that there exist some inter-dependencies between the 2nd and

3rd components. To this end, we propose a framework which maps peering links jointly with intra-domain links in each InP.

3 Network Model

We adopt the usual substrate and virtual network model [7]. In addition, we describe VNP layer information based on existing multi-domain information model.

3.1 Substrate Network

A domain InP_i is modelled as an undirected graph $G_i^S(N_i^S, L_i^S)$, where N_i^S is the set of substrate nodes in domain i, L_i^S is the set of internal substrate links. Each substrate node $n_i^s \in N_i^S$ is associated with a CPU capacity $cpu(n_i^s)$ and a geographic location $loc(n_i^s)$. Each substrate link $l_i^s \in L_i^S$ is associated with a bandwidth capacity $bw(l_i^s)$.

Assuming that the substrate network covers K domains, there are some peering nodes (border nodes) which have peering links with other domains. The peering nodes set is denoted by N_i^{SP} ($N_i^{SP} \subset N_i^S$). The peering links between InP_i (*i.e.* G_i^S) and InP_j (*i.e.* G_j^S) is denoted by L_{ij}^S. We denote by $P_i^S = \cup_{j=1}^K L_{ij}^S$ the set of all of the peering links of InP_i, and by P^S with $P^S = \bigcup_{i=1}^K P_i^S = \bigcup_{(i,j)\in(1...K)^2} L_{ij}^S$ the set of all of the peering links. The complete substrate network $G^S(N^S, L^S)$ is thus obtained as follows: $N^S = \bigcup_{i=1}^K N_i^S$, $L^S = (\bigcup_{i=1}^K L_i^S) \bigcup P^S$.

3.2 Virtual network

The virtual network is also modelled as an undirected graph $G^V(N^V, L^V)$, where N^V is the set of virtual nodes and L^V is the set of virtual links. Each virtual node $n^v \in N^V$ is associated with a CPU capacity demand $cpu(n^v)$, a geographic location $loc(n^v)$ and a distance constraint $dis(n^v)$ specifying how far a virtual node n^v can be placed from its $loc(n^v)$. Each $l^v \in L^V$ is associated with a bandwidth demand $bw(l^v)$. In addition, each virtual network G^V has a lifetime $t(G^V)$.

3.3 VNP layer model

VNP collects information provided by InPs. We assume that InPs provide exact information about their nodes, as well as the peering links. On the contrary, there is no exact information about the internal organisation of a domain. Similar to the existing solution [12], we assume that this information is given by InP for each couple of <node, peering node>, as if there was a *pseudo* direct link between these two nodes. Denote the set of these links by $L_i^P = \{l_{mn} / m \in N_i^S, n \in N_i^{SP}\}$, InP_i provides to VNP the set of linking cost C_i^P defined by

$$C_i^P = \{C(l_{mn}) / m \in N_i^S, n \in N_i^{SP}\}$$

where $C(l_{mn})$ represents a cost (distance, bandwidth, etc.) characterizing the link l_{mn}. This kind of information is actually what a routing protocol (BGP) reports to other AS.

Thus, the SN of an InP_i is *perceived* by VNP as a graph $G_i^P = (N_i^S, L_i^P)$. In this way, the whole substrate network that VNP perceives, referred as G^P, is defined as follows:

$$G^P = (\bigcup_i G_i^P) \bigcup P^S$$

i.e. the perceived vision for each domain and the exact vision of the inter-domain connections. With G^P, VNP can establish a kind of complete topology covering all the domains for achieving VN decomposition and link mapping.

4 Our proposition

To solve VNE in the context of multi-domain, we propose a novel algorithm that maps jointly intra and peering links.

We propose to handle each VN request with a 2-step process

- At the first step, VNP performs the node decomposition optimizing the node embedding.
- Subsequently, VNP performs a series of iterative downsizing VNE sub-solution, each of them optimizes both the intra and peering link mapping related to a domain.

The link mapping is determined, at each iteration, by the acting InP (called *mapper*). VNP is in charge of providing necessary information to the mapper. The generic work-flow of our algorithm is given by figure 1. The details are explained as below.

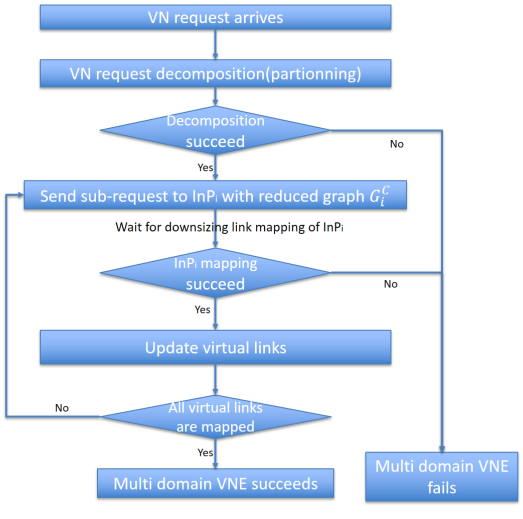


Figure 1: VNP workflow to embed a VN

4.1 Decomposition

Firstly, VNP decomposes the VN request with objective of minimizing the node mapping cost. In this stage, VNP associates each virtual node with a candidate set of substrate nodes that meet its $loc(n^v)$. VNP is free to use any multi-domain VN partitioning

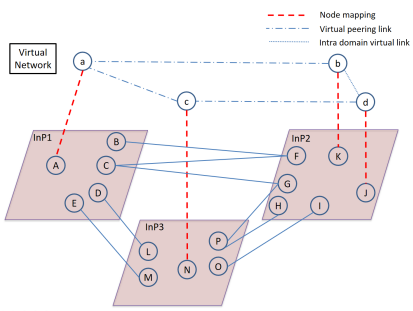


Figure 2: VN decomposition

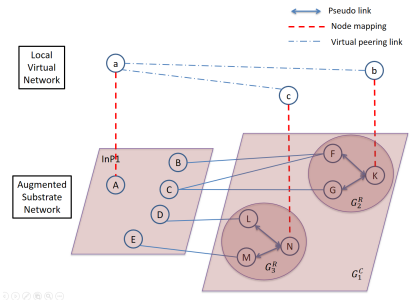


Figure 3: downsizing link mapping by InP_1

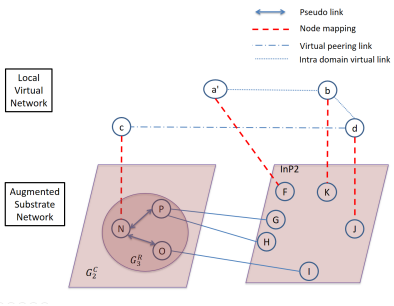


Figure 4: downsizing link mapping by InP_2

method (*e.g.* [14][12]). At the end of this stage, virtual nodes are embedded to different domains.

An example of VN decomposition is shown in figure 2. Three InPs are shown with their substrate nodes from A to P. They are connected via 2 or 3 peering links. Intra substrate links are not drawn. We suppose that a VN $\{a,b,c,d\}$ arrives. The VN decomposition step tells us that a, b, c and d are mapped to substrate nodes A, K, N and J, respectively. $\{a-b, a-c, c-d\}$ are virtual links which interconnect two different domains, while $\{b-d\}$ locates in only one domain.

4.2 An iterative downsizing VNE approach

Here we give a detailed presentation of the kernel of our proposal, which is formally given in algorithm 1.

4.2.1 Rationale

After VN decompostition step, since there is no domain who knows the complete information of any other one, embedding the virtual links which interconnect two different domains becomes an issue.

We notice that, VNP can build, for each InP_i, a reduced vision (denoted by G_i^R) from G_i^P. This vision contains all the peering links/nodes, as well as the substrate nodes on which a virtual node is embedded. Formally, $G_i^R = (N_i^R, L_i^R)$ where

$$N_i^R = N_i^{SP} \bigcup \{n_i^S \in N_i^S \,/\, \exists n^v \in N^V, M(n^v) = n_i^S\}$$

i.e., N_i^R is the union of all the substrate nodes supporting virtual nodes on domain i and all of its peering nodes. In a similar way, we define L_i^R as follows;

$$L_i^R = \{l_{mn} \in L_i^P / n \in N_i^R, m \in N_i^{SP}\}$$

i.e., L_i^R is the subset of L_i^P between N_i^R and N_i^{SP} containing only the links interconnecting a peering node and a node supporting a virtual node.

In order to achieve an efficient and pragmatic operation mode, we prefer that VNP plays its role of coordinator: It is VNP who decides which of the InP should have the privilege to map its peering links with others. It is also VNP who provides to the chosen InP (that we refer as *mapper*) the topology of the rest of the network according to its perception. In other words, the chosen InP (the mapper) extends its view to the rest of the network, by using the vision provided by VNP, the only one who has a kind of comprehensive view on all domains. In this way, the mapper obtains an augmented graph on which it will perform link mapping, including both its intra and peering links.

This process continues, domain after domain, until all of the virtual links are set. The selection criterion is the link utilization, the InP has most stringent link utilisation will be the first to map its peering links. The reason lies in that high link utilization denotes more constraints in the choice of path.

4.2.2 Building of the augmented graph

Let InP_i be the chosen mapper. Formally speaking, the vision of the other domains provided by VNP is $G_i^C = \bigcup_{j \neq i} G_j^R$, i.e. the reduced perceived vision of all the other domains. We only need to consider the case where all the domains are adjacent to the mapper. The case of a domain not adjacent to the mapper but to which the mapper has virtual links can be reduced to the adjacent case.

VNP communicates G_i^C to the mapper (InP_i) so that the latter can creates an augmented graph G_i^A, defined as follows:

$$G_i^A = G_i^S \cup P_i^S \cup G_i^C$$

This topology covers all of the accessible domains and can be used as a substrate graph on which the mapper performs VNE.

4.2.3 VN sub-request

VNP asks the mapper to perform a partial VNE, which concerns only the virtual links related to the mapper. We refer this partial VNE as a *sub-request* (L_i^{subv}). It is obtained from the current VN request by reducing it to virtual links related to the mapper.

4.2.4 An MCF-based link mapping

At this stage, the mapper gets an augmented vision of the whole substrate network, and a VNE sub request (L_i^{subv}), both from VNP. We have thus a classical VNE problem that we solve with the multi commodity flow

(MCF) based mapping algorithm (line 6 of algorithm 1).

At the end of this step, InP_i pre-allocates resources on the intra and peering links related to it and sends to VNP a virtual link update notification.

Let us illustrate it by our example. Assume in figure 2 that InP_1 is chosen as the 1st mapper. VNP builds the VNP-level graph vision $G_1^C = G_2^R \cup G_3^R$ (see figure 3) with G_2^R = ({*F*, *G*, *K*}, {*F-K*, *G-K*}) and G_3^R = ({*M*, *N*, *L*}, {*M-N*, *L-N*}). It builds also the sub-request L_1^{subv} = {*a-b*, *a-c*}, actually the virtual links *b-d* and *c-d* will be pruned since they haven't any extremity node supported by a substrate node in InP_1. VNP then sends G_1^C and L_1^{subv} to InP_1. The latter builds the augmented graph G_1^A which includes G_1^S (all the nodes and links in InP_1), the peering links (*B-F*, *C-F*, *C-G*, *E-M*), and G_C^1. InP_1 then applies the MCF-based algorithm to solve the embedding of L_1^{subv} on G_1^A.

4.3 Update and iteration

After each sub-request, the mapper (say InP_i) reports the results. In particular, it gives the results of the mapping of all of its inter-domain virtual links in the following manner.

Let $l^v(a,b)$ be the virtual link between a particular node $a \in InP_i$ and a particular node $b \in InP_j$, with $bw(l^v(a,b))$ as the required bandwidth. The MCF algorithm will map $l^v(a,b)$ into one or several paths. Denote by N^F the set of the peering nodes of InP_j through which a fraction of $l^v(a,b)$ is mapped. After the link mapping of InP_i, the set of virtual links $\{l^v(c,b)\}_{c\in N^F}$ is equivalent to the virtual link $l^v(a,b)$ with bandwidth demand:

$$\sum_{c\in N^F} bw(l^v(c,b) = bw(l^v(a,b))$$

It is to be pointed out that these links are totally inside inP_j and they replace $l^v(a,b)$.

As the mapping of InP_i is achieved, it will no longer appear as domain in the subsequent problem which contains only the remaining domains. However, the achieved mapping concerns only the links related to InP_i (intra as well as peering), the part of inter-domain virtual links on the other domains still has to be mapped. Each of such inter-domain virtual link related to the mapper can be transformed into the above described equivalent set which will be added to each concerned domain. For the sake of reading simplicity, we prefer to give an informal description here, instead of a formal one, which would generate some more heavyly-indexed notations.

In this way, we obtain a new VNE problem with:

- at the SN level, the retreat of InP_i and all the peering links related to it;
- at the VN side, the retreat of all the virtual links internal to InP_i and the replacement of all the inter-domain virtual links related to InP_i by their equivalent set which are added to corresponding domain.

This allows us to execute iteratively the downsizing mapping described in § 4.2. VNP repeats the process till its convergence which is *certain*, since the subset is reduced by at least one domain (the mapper) at each iteration.

Algorithm 1: Link mapping of InP_i as mapper

```
Input  : sub request virtual links L_i^subv
Input  : reduced perceived graph G_i^C
Output: virtual link update notification
1  begin
2      if L_i^subv = NULL then
3          return
4      end
5      create augmented substrate network
         G_i^A(N_i^A, L_i^A) ;
6      solve single domain VNE MCF problem;
7      foreach flow on substrate link l_mn do
8          if l_mn ∈ L_i^S ∪ L_ij^S then pre-allocate
             resource on link l_mn ;
9      end
10     send virtual link update notification;
11 end
```

In the example of figure 2 and 3, assume that InP_1 has chosen link *F-K* to map virtual link *a-b*. After sending its results to VNP, this latter deduces and creates a new virtual link *a'-b* with node mapping *a'* equal to *F*. This virtual link *a'-b* replaces virtual link *a-b*.

Now, the new problem (figure 4) contains only InP_2 and InP_3. Assuming that the InP_2 is chosen as mapper, the same process continues and our problem is eventually reduced to a single domain which is the last step of our algorithm.

4.4 Reject of virtual request

The resources are definitively allocated only if all the computation on different domains succeed. A COMMIT message is then sent by VNP to InPs so as to validate the resource reservation. Should a mapper report a failure, a DEALLOC message would be sent by VNP, which stops the process (VNE failure) and allows each domain to deallocate pre-allocated resources.

5 Reinforcement of our method

As mentioned at the end of 4.2.1, we choose the link utilization criterion to determine mapping sequence. This choice simplifies the algorithm, but may fail to get the optimal solution. In this section, we propose a reinforcement of our algorithm, which waives the constraint of sequence selection.

Fundamentally, domains are peers. From a domain's point of view, there are actually 2 "domains" : a single domain (itself) and a outside domain (others). Using our downsizing algorithm, a domain tries its best to map its intra and peering virtual links, but how does the outside domain (others) map the remaining virtual links? We notice that after the first

downsizing mapping, the problem is reduced to a multi-domain VNE on the outside domain (others) because the first mapper has mapped its own intra and peering virtual links. Following the downsizing logic, the problem will finally be reduced to a 2 domain VNE, on which a better solution can be easily found. Therefore, We first study the case of only 2 domains, and then we move to K domains VNE.

5.1 Two domain basic method

First, we consider the case of 2 domains. Assuming that the multi-domain VNE problem consists of 2 domains (denoted by InP_1 and InP_2), there are obviously 2 possible mapping sequences.

- First Solution $S_{1->2}$: InP_1 starts up the mapping processus as mapper and then InP_2 solves a single domain VNE.
- Second solution: $S_{2->1}$: InP_2 starts up the mapping processus as mapper and then InP_1 solves a single domain VNE.

These two solutions are sent to VNP, which compares the embedding cost of these solutions. The final solution of the 2 domain basic multi-domain VNE (denoted by $MDVNE(2)$) is the better one among the 2 solutions above:

$$S_{MDVNE(2)} = min\{S_{1->2}, S_{2->1}\}$$

Algorithm 2: $MDVNE(K)$

```
   Input  : VN request
   Output: embedding cost
   Output: mapping solution
1  begin
2      foreach InP_i do
3          link mapping of InP_i as mapper;
4          get embedding cost C(mapper);
5          if K > 2 then
6              solve MDVNE(K − 1) ;
7              get embedding cost C(others);
8          end
9          else
10             solve MDVNE(2);
11             get embedding cost C(others);
12         end
13         InP_i solution cost
             C(i) = C(mapper) + C(others) ;
14     end
15     return minimum C(i) and correspond
         solution;
16 end
```

5.2 Towards K domain solution

From the 2 domain basic method, K domain multi-domain VNE (denoted by $MDVNE(K)$) can be determined by a recursive algorithm. The detail of $MDVNE(K)$ is shown in Algorithm 2.

The multi-domain is fundamentally divided into 2 elements, a mapper and the others. The former is mapped using our downsizing method (line 3 in Algorithm 2), and the later is reduced to a K-1 domain problem (line 6 in Algorithm 2), until the basic 2 domain problem. The cost of the candidate is the sum of cost of the 2 elements (line 13 in Algorithm 2). The minimum cost candidate will be adopted as the mapping solution.

6 Performance Evaluation

We implemented a discrete event simulator to evaluate the performances of our method. The optimization problem is solved by IBM CPLEX library. Since we are basically interested by the link mapping, all the evaluated methods work with the same node decomposition by using the greedy algorithm of [6].

6.1 Evaluation Environment

The substrate networks are generated by GT-ITM tool [17]. 3 domains are generated. Each domain consists of 50 nodes and 100 links. The 3 domains are interconnected by 26 peering links. The CPU capacity of each node is chosen in [50,150]. The bandwidth capacity is selected in [50,100] for intra links and in [300,400] for peering links. The cost of the pseudo link between a border node and an intra node is chosen to be the inverse of the bandwidth capacity of the shortest path between these two nodes.

The virtual networks are also generated by GT-ITM tool. The virtual nodes of each VN follow a uniform distribution between 3 and 8. The virtual nodes are interconnected with probability 0.4. The CPU and bandwidth demands are uniformly chosen in [0,20]. The VN request arrival process is Poisson with arrival rate $\lambda \in (2...6)$ requests per 100 time units. The life time of each VN follows an exponential distribution with an average of 1000 time units. Each simulation lasts for 20000 time units.

6.2 Compared methods

We compare the following 3 methods :

(i) $lu-ciplm$: Link utilization coordinated intra and peering link mapping. This is our first method.

(ii) $r-ciplm$: Reinforced coordinated intra and peering link mapping. This is our second method.

(iii) *shen* [12]: This approach computes separately intra and peering links. The latter is determined according to Dijkstra's algorithm.

We used the following metrics for comparison:

- *VN request acceptance ratio*: the ratio of the accepted VN request over the total arrived VN requests;

- *Average link utilization*: the link utilization is the total allocated link resource over the total substrate resource. The allocated resource $\mathscr{U}$ is given by:

$$\mathscr{U} = \sum_{l^v \in L^V} \sum_{l^s \in L^S} bw(l^s, l^v)$$

where $bw(l^s, l^v)$ denotes the bandwidth committed on the substrate link l^s to embed the virtual link l^v.

- *Total revenue*: The revenue of a VN as the weighted sum of bandwidth and CPU:

$$\mathscr{R} = \beta \sum_{l^v \in L^V} bw(l^v) + \alpha \sum_{n^v \in N^V} cpu(n^v)$$

where α (resp. β) is the unit revenue for cpu (resp. bandwidth). .

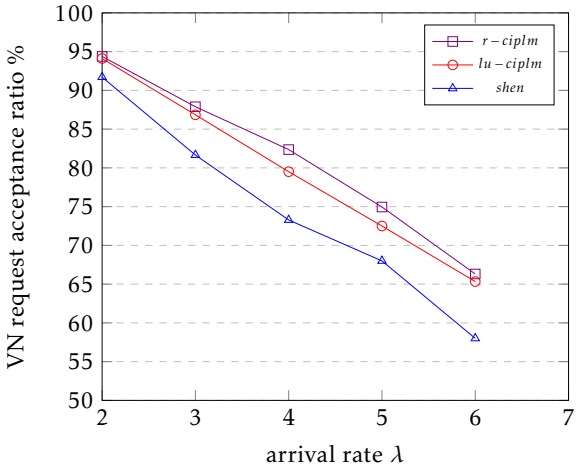


Figure 5: Acceptance ratio

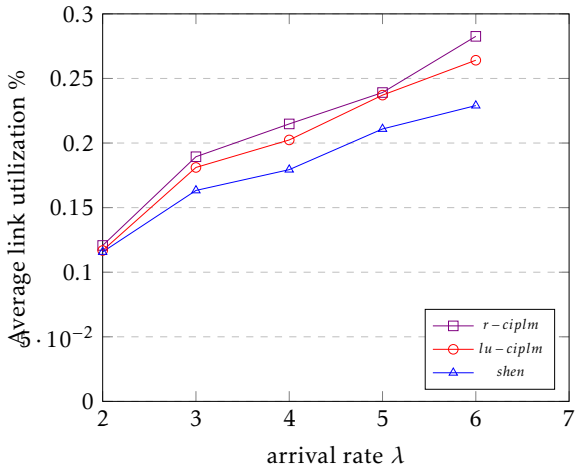


Figure 6: Average link utilization

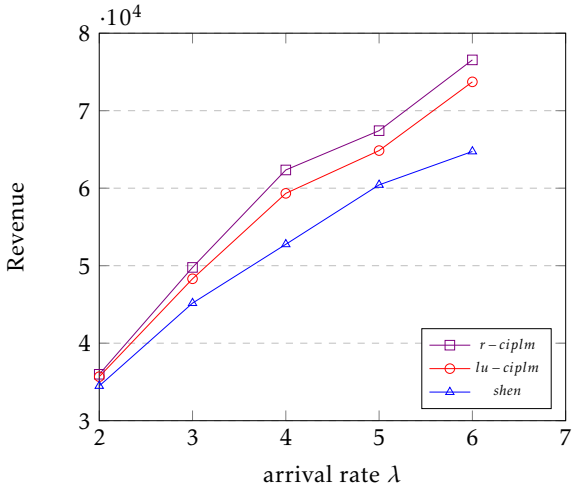


Figure 7: Total revenue

6.3 Result analysis

The simulation results are shown in Figure 5, 6 and 7. The VN request acceptance ratio is shown in Figure 5. The link utilization and revenue are shown in Figure 6 and Figure 7, respectively. We got the following observations:

- $r-cilpm$ is the best, followed by $lu-ciplm$, and then *shen* over all the three metrics.
- The difference between $lu-ciplm$ and *shen* is always significant.
- The difference between $r-ciplm$ and $lu-ciplm$ is not always significant.

To summarize:

- Our approaches are better than that of *shen*. Indeed, mapping jointly intra and peering links increases the efficiency. Our methods improve the performance. In these cases, traffic is splitted and sent to less loaded links, achieving in this way a better utilisation of the overall residual bandwidth.
- The out-performance of $r-ciplm$ is small compare to $lu-ciplm$. $r-ciplm$ ensures a cost-efficient mapping solution for every VN. The peering links are mapped jointly with the right domain, which leads to a better resource allocation. However, the peering links in $lu-cimplm$ are sometimes not perfectly mapped because link utilization does not always give the best mapping sequence.

7 Conclusion

Network virtualization attracts more and more attention in future network architecture, since it allows the (dynamic) building of a network suited to end-users need, without modifying the underlay infrastructures. Part of them will be built over several infrastructures run by different operators.

The virtual network embedding, which aims at establishing the optimal virtual networks on substrate networks, is a key issue in network virtualization. The fact of partial information makes the multi-domain VNE quite different from the single-domain VNE and this problem remains a challenge. Some multi-domain VNE solutions have been proposed in literature. Most of them focus more on VN decomposition into sub VN requests for each domain, so that the single-domain VNE can be applied subsequently. Few attention has been paid on the mapping of peering (inter-domain) links.

In this paper, we propose a novel multi-domain VNE algorithm which aims to optimize the peering link mapping. For this, we introduce a coordinator (VNP, VN Provider). The latter has the privilege to get a comprehensive vision of all of the domains as well as the peering links. It performs VN decomposition, then coordinates the optimized mapping of both intra and peering links, domain after domain, in an iterative (and converging) manner. The optimization is achieved by applying the MCF algorithm on an augmented graph related to each domain. Simulation shows that our approach optimizes the substrate resource utilization compared to existing method. Besides, our method is easy to deploy in current Internet architecture.

References

[1] T. Anderson, L. Peterson, S. Shenker, and J. Turner, "Overcoming the internet impasse through virtualization," *Computer*, no. 4, pp. 34–41, 2005.

[2] G. Schaffrath, C. Werle, P. Papadimitriou, A. Feldmann, R. Bless, A. Greenhalgh, A. Wundsam, M. Kind, O. Maennel, and L. Mathy, "Network virtualization architecture: proposal and initial prototype," in *Proceedings of the 1st ACM workshop on Virtualized infrastructure systems and architectures*, pp. 63–72, ACM, 2009.

[3] A. Belbekkouche, M. Hasan, and A. Karmouch, "Resource discovery and allocation in network virtualization," *Communications Surveys & Tutorials, IEEE*, vol. 14, no. 4, pp. 1114–1128, 2012.

[4] A. Fischer, J. F. Botero, M. Till Beck, H. De Meer, and X. Hesselbach, "Virtual network embedding: A survey," *Communications Surveys & Tutorials, IEEE*, vol. 15, no. 4, pp. 1888–1906, 2013.

[5] Y. Zhu and M. H. Ammar, "Algorithms for assigning substrate network resources to virtual network components," in *INFOCOM*, vol. 1200, pp. 1–12, 2006.

[6] M. Yu, Y. Yi, J. Rexford, and M. Chiang, "Rethinking virtual network embedding: substrate support for path splitting and migration," *ACM SIGCOMM Computer Communication Review*, vol. 38, no. 2, pp. 17–29, 2008.

[7] N. M. K. Chowdhury, M. R. Rahman, and R. Boutaba, "Virtual network embedding with coordinated node and link mapping," in *INFOCOM*, pp. 783–791, IEEE, 2009.

[8] J. Lischka and H. Karl, "A virtual network mapping algorithm based on subgraph isomorphism detection," in *Proceedings of the 1st ACM workshop on Virtualized infrastructure systems and architectures*, pp. 81–88, ACM, 2009.

[9] X. Cheng, S. Su, Z. Zhang, H. Wang, F. Yang, Y. Luo, and J. Wang, "Virtual network embedding through topology-aware node ranking," *ACM SIGCOMM Computer Communication Review*, vol. 41, no. 2, pp. 38–47, 2011.

[10] G. Sun, H. Yu, V. Anand, and L. Li, "A cost efficient framework and algorithm for embedding dynamic virtual network requests," *Future Generation Computer Systems*, vol. 29, no. 5, pp. 1265–1277, 2013.

[11] D. Dietrich, A. Rizk, and P. Papadimitriou, "Multi-domain virtual network embedding with limited information disclosure," in *IFIP Networking Conference, 2013*, pp. 1–9, IEEE, 2013.

[12] M. Shen, K. Xu, K. Yang, and H.-H. Chen, "Towards efficient virtual network embedding across multiple network domains," in *Quality of Service (IWQoS), 2014 IEEE 22nd International Symposium of*, pp. 61–70, IEEE, 2014.

[13] M. Chowdhury, F. Samuel, and R. Boutaba, "Polyvine: policy-based virtual network embedding across multiple domains," in *Proceedings of the second ACM SIGCOMM workshop on Virtualized infrastructure systems and architectures*, pp. 49–56, ACM, 2010.

[14] I. Houidi, W. Louati, W. B. Ameur, and D. Zeghlache, "Virtual network provisioning across multiple substrate networks," *Computer Networks*, vol. 55, no. 4, pp. 1011–1023, 2011.

[15] K. Guo, Y. Wang, X. Qiu, W. Li, and A. Xiao, "Particle swarm optimization based multi-domain virtual network embedding," in *Integrated Network Management (IM), 2015 IFIP/IEEE International Symposium on*, pp. 798–801, IEEE, 2015.

[16] C. Wang, S. Shanbhag, and T. Wolf, "Virtual network mapping with traffic matrices," in *Communications (ICC), 2012 IEEE International Conference on*, pp. 2717–2722, IEEE, 2012.

[17] GT-ITM *http://www.cc.gatech.edu/projects/gtitm/*.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 2, No. 3, 553-561 (2017)
www.astesj.com
Special Issue on Recent Advances in Engineering Systems

ASTES Journal
ISSN: 2415-6698

Semantic-less Breach Detection of Polymorphic Malware in Federated Cloud

Yahav Biran[*,1], George Collins[1], Borky John M[1], Joel Dubow[2]

[1]*Colorado State University, Department of Systems Engineering, Fort Collins, 80523, CO*

[2]*Fulcrum Co., Cyber Security, Centreville, 20120, VA*

ARTICLEINFO

Article history:
Received: 10 April, 2017
Accepted: 20 May, 2017
Online: 01 June, 2017

Keywords:
Anomaly detection
Polymorphic Malware
Cloud Federation

ABSTRACT

Cloud computing is one of the largest emerging utility services that is expected to grow enormously over the next decade. Many organizations are moving into hybrid cloud/hosted computing models. Single cloud service provider introduces cost and environmental challenges. Also, multi-cloud solution implemented by the Cloud tenant is suboptimal as it requires expensive adaptation costs. Cloud Federation is a useful structure for aggregating cloud based services under a single umbrella to share resources and responsibilities for the benefit of the member cloud service providers. An efficient security model is crucial for successful cloud business. However, with the advent of large scale and multi-tenant environments, the traditional perimeter boundaries along with traditional security practices are changing. Defining and securing asset and enclave boundaries is more challenging, and system perimeter boundaries are more susceptible to breach. This paper to describe security best practices for Cloud Federation. The paper also describes a tool and technique for detecting anomalous behavior in resource usage across the federation participants. This is a particularly serious issue because of the possibility of an attacker potentially gaining access to more than one CSP federation member. Specifically, this technique is developed for Cloud Federations since they have to deal with heterogeneous multi-platform environments with a diverse mixture of data and security log schema, and it has to do this in real time. A Semantic-less Breach detection system that implements a self-learning system was prototyped and resulted in up to 87% True-Positive rate with 93% True-Negative.

1 Introduction

As an increasing fraction of computing services move to the Cloud, there will be a proliferation of software characteristics, service models, and deployment options. Many organizations are moving into hybrid cloud/hosted computing models. A Cloud Service Provider (CSP) goal is to maximize its market share among the potential Service Providers (SP). Accommodating variable demands for computing resources requires an immense capacity, as it calls for providing for the maximum demand. In some cases, this drives them to underutilization of massive datacenter deployments. In other situations, the CSPs suffer over-utilization because of a miss in the market share, load and reliability projections. Both cases lead to sub-optimal utilization.

From the SPs perspective, they are most interested in availability and adaptability. The former refers to reliable service conditions that make its services available to the users it serves. The latter relates to the Vendor lock-in risk[1]. Single CSP provides a sub-optimal solution to the SP thus multi-cloud become an attractive solution. However, multi-cloud solution implemented by the SP requires expensive adaptations to the CSP's tools and service constructs that may vary among different CSPs.

Cloud Federation is a new paradigm that allows many CSPs to utilize computing resources optimally[2, 3]. Also, it allows SP to avoid the Vendor lock-in risk and provide service availability that can not be provided by a single CSP. No matter what the architecture, there is a need for to ensure the security and information assurance to users. Cloud Federation is an advantageous structure for aggregating cloud based services under a single umbrella to share resources and responsibilities for the benefit of the member cloud ser-

[*]Corresponding Author, ybiran@colostate.edu

https://dx.doi.org/10.25046/aj020371

vice providers. Federation is useful not only for sharing resources amongst cloud service providers but also for providing enclaves for interactions to perform domain-specific missions such as electrical grids and supply chains.

The Federation will need to assure that data transfers amongst the Federation's CSPs are secure. The Federation will, above all, need to detect any anomalous behavior occurring in transactions and resource sharing. In addition to the growing number of security tools, there is a need to log and identify security issues requiring attention early on in the process. In particular, breach detection in inter-cloud data transfer and communications is a particularly serious issue because of the possibility of an attacker potentially gaining access to more than one CSP federation member.

This paper to describe security best practices for Cloud Federation. The paper also describes a tool and technique for detecting anomalous behavior in resource usage across the federation participants. Specifically, this method is developed for Cloud Federations since they have to deal with a heterogeneous multi-platform environment with a diverse mixture of data and security log schema, and it has to do this in real time. This Semantic-less tool is described below after a description of the context of the issue.

The reminder of paper is organized as follows, Section II (Federated Cloud) discusses a new inter-cloud structure. It also describe some of the terminology used in this paper; Section III (Cyber Security Challenges in Federated Cloud) describes the core challenges of such inter-cloud system in a multi-layer model; Section IV (Semantic-less Breach Detection) discuss the tool we suggest for detecting the behavior of the anomalous system that runs in the Cloud Federation; Section V (Evaluation) discuss the breach-detection tool prototyped; Section VI (Analysis) analyze and present the prototype results; finally, Section VII (Conclusions) summarizes the main results and ongoing developments.

2 Federated Cloud

Organizations that serve and process a large number of simultaneous users and the user's often large quantities of data are termed "cloud computing services." These services enable convenient, on-demand, network access to a shared pool of configurable computing resources. The following section describes cloud federation and the reasons for its formation. It serves as background for the security design and breach detection mechanism we propose in a later section.

Cloud computing is occupying a rapidly growing market share of IT resources. IT related spending toward workload processing increased 32.8 percent in 2014, 29 percent in 2015, 34.1 percent during 2016, and expected to grow 42 percent during 2017[1]. Also, from environmental perspectives, grid energy powered by hydrocarbons are increasingly devoted to the growing power needs of cloud computing data centers. Carbon emission generated by data centers is growing from a 2011 level of 21.3 MtCO2e and expected to rise to 39.1 MtCO2e by 2020[2].

This makes cloud computing one of the fastest growing consumers' utility services, one that is projected to grow an order of magnitude in the next decade. [2] propose a new paradigm that will allow multiple cloud providers to utilize computing and energy resources optimally. From the customer's perspectives, the organization's total-cost-of-ownership is expected to shift from capital-expense-based, e.g., on-premise deployments to operational-expense-based, e.g., cloud service subscriptions. Thus, there is a need to create an efficient and transparent eco-system that allows the organization to match IT expenses with its planned cost structure.

Federated Clouds demonstrate a new paradigm that will allow multiple cloud providers to utilize optimally computing resources. It will do this by (1) lowering the data center's deployments per provider ratio, share and (2) scheduling available energy via aggregators and (3) lastly to employ, where appropriate, more renewable and carbon-free energies. It will quantify results and baseline the efforts based on work at the data centers within a single cloud provider. The federated cloud demonstrates the utility software container-based paradigm for achieving dense and elastic containerization technologies. It is centered on several discreet Linux Containers (LXC) managed by a Kubernetes resource management system [3] that acts as the governance engine. Moreover, the proposed solution scales out and optimizes cross-data centers and cross-regional deployments by computing and suggesting a cross-cloud provider's collaboration via a cloud aggregator. Furthermore, it suggests operating data centers employing maximum intermittent green energy sources[2]. Yet all of these advantages will be for naught if federation services cannot be supplied securely in the face of growing sophistication and quantities of cyber security threats. It is therefore essential that risks to data, computing resources and communications are managed such that the value of services provided exceeds the losses arising from cyber breaches.

2.1 Architecture

The Cloud Federation architecture is comprised of multiple CSPs, a Clouds-Coordinator, and a Clouds-Broker system. These are defined below.

Clouds-Coordinator. Acts as an information registry that stores the CSPs pricing offers and demand patterns. Clouds Coordinators periodically update the CSPs availability and offering prices. Also, a Clouds-Coordinator will help to employ, where appropriate, more renewable and carbon-free energies [3].

Clouds-Broker. Manages the membership of the CSP constituents. Both CSPs and SPs will use the Clouds-Broker to onboard the Cloud Federation. Also, Clouds-Broker will acts on behalf of the SP for resource allocation and provisioning requests. Clouds-Broker also provides a continuous ability to deploy SP software, configuration, and data to one or more CSP, so it provides the SoS IT agility goal.

[1] Gartner Says Worldwide Cloud Infrastructure-as-a-Service Spending to Grow 32.8 percent in 2015

[2] GeSI SMARTer2020 The Role of ICT in Driving a Sustainable Future. 2015. GeSI

[3] http://kubernetes.io

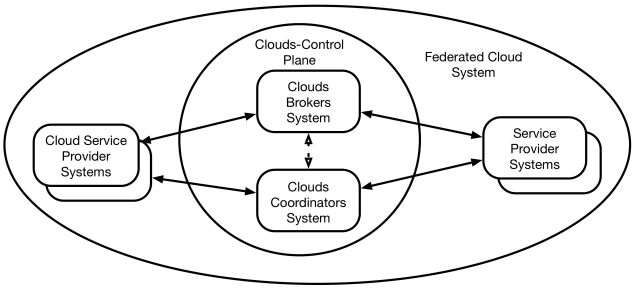


Figure 1: Proposed Cloud Federation Systems of System comprises of CSPs and SP that are managed by Clouds-Broker and Clouds-Coordinator

3 Cyber Security Challenges in Federated Cloud

The Cloud Federation has a global scale software and hardware infrastructure. We describe a progressive layers security model starting from the physical security of data centers, progressing to the hardware and software that underlies the infrastructure, and the constraints and processes to support the Cloud Federation operational security. The following section describes the Cloud Federation cyber security design throughout the data processing life cycle at a Cloud Federation. e.g., enables secure communication with tenants (SP) and its customers or control plane communication including CSP, Cloud-Brokers, and Clouds-Coordinator.

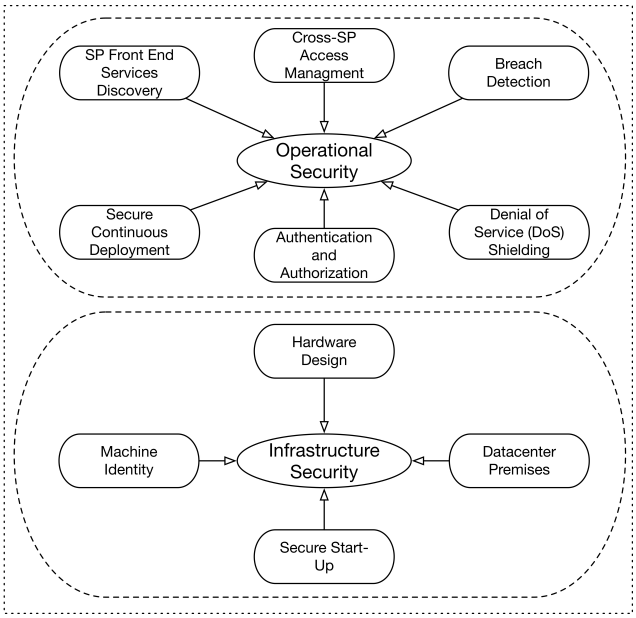


Figure 2: Cloud Federation Cyber Security Model includes two core layers, infrastructure security and operational security

Figure 2 describes the cyber security layers offered by the Cloud Federation. The following paragraph briefly describes the security elements corresponded with each layer[4]. Our extended cyber security model will emphasize the operational security with unique breach detection methodology. This was done since the operational security corresponds to the perimeter security of an enterprise system and the interface to the Federation members. Also, it will suggest a system for encryption of both inter microservices communication with emphasis on cross-CSP for tenants' workloads.

3.1 Infrastructure Security

The required baseline security level needed for cloud federation constituent's systems is referenced in Figure 1. It includes deployed facilities and computer systems managed by the CSPs or the Federation. The larger CSP's often exceed these baselines.

Datacenter Premises. CSPs design and build its data centers based on its expected computing capacities and service reliability manifested by their SLA and the redundancy levels of sub systems[4, 5]. The datacenter incorporates various components of physical security protections. Access to such facilities is governed by the CSP security operations. It uses technologies such as biometric identification, metal detection, metal detectors, and CCTV solutions[6].

Hardware Design. CSPs data centers run computing server machines fed by power distribution units and connected to a local network that is all connected to the edge of the wide network. The computing, digital storage, and networking equipment require a standard that ensures the required audit and validation of the security properties by components[7, 8], e.g., hardware security chip [9].

Machine Identity. confirms that any participating computing server in Cloud Federation can be authenticated to its CSP machine pool throughout a low-level management services[9]
Secure Start-Up. Ensures that CSPs servers are booting the correct software stack. Securing underlining components such as Linux boot loaders, OS system images and BIOS by cryptographic signatures can prevent an already compromised server from being continuously compromised by an ephemeral malware.

3.2 Operational Security

Operational security comprises the business flows between the SP with the cloud federation and the CSP it uses for processing workloads. The following section briefly discusses the required cybersecurity measures needed for SP and CSP business scenarios in a cloud federation.

Cross-SP Access Management: SP workloads are manifest in two workload types, (1) short-lived workloads. i.e., jobs that are terminate upon completion, and (2) long-lived workloads. i.e. services. The former workload might require connectivity to external services during its processing. The latter might expose serving endpoints to other services. e.g., short-lived jobs might require persistent storage to write its job results hence connecting to BigTable[5] storage server provisioned by other CSPs, which, in turn, require access management that uses credentials and certificates

[4]We extrapolate Google Cloud security model from https://cloud.google.com/security/security-design/
[5]https://cloud.google.com/bigtable/

stored within the Cloud Federation.

SP Front End Service Discovery: Long-lived workloads might expose public facing endpoints for serving other workloads or end-user requests. SP front-end services require publishing endpoints to allow other workloads within or external to the cloud Federation to discover their public facing entry point and this requires service discovery capabilities. Service discovery endpoints, and the actual service endpoints, are prone to risks such as Denial of Service attacks or intrusions originated by an attacker. We argue that current solutions offered by individual CSP's are sub-optimal because of the target scope of the intrusion. i.e., assuming an attack probability for a given CSP, running several CSPs reduces the risk by a factor of the number of CSPs. Later sections will formulate the risk function and show how cloud-federation minimizes those challenges by using the semantic-less breach detection system and show how most risks originate by crossing machine boundaries.

Secure Continues Deployment: Continuous Deployment (CD) is the function that allows cloud-native applications to get updated through an automated pipeline that is initiated by a new or updated code submission, compiled, tested through various quality gates until it is certified for deployment of the production systems and deployed seamlessly. Continues deployment enables cloud applications to innovate faster and safer no matter what number of machines are in the service pool. A secure continuous deployment service requires secured SP code and a configuration repository that authenticates to the target computing resource regardless of the CSP network segmentation. Traditional network segmentation, or fire walling, is a secondary security mechanism that is used for ingress and egress filtering at various points in the local network segment to prevent IP spoofing[10, 11]

Authentication and Authorization. In a federated cloud architecture, deployed workloads might require access to other services deployed by the federation. The canonical example will be an end user request service deployed in the Federation that triggers another micro-service within the SP architecture. Such cascading requests require multilayered authentication and authorization processes. i.e., a micro-service calls another micro-service and authenticate on behalf of the end user for audit trails supported by the end-user authentication token and the cascading micro-service tokens generated throughout the end user request. Figure 3 depicts the data flow during a call initiated by SP Micro-Service that runs in one of the federation's CSP denoted by CSP_i and CSP_j. A call initiated from SP_n that was provisioned in the federation as ms_n. The call destination runs on a different and sometimes the same SP. Let SP_m denote the destination SP. The call payload is encrypted by SP_n private key. The call arrived at an SP_m endpoint and checked for admission. SP_m admission control decrypts the call payload using SP_n public key that was submitted throughout the on-boarding process to the Cloud Federation. It is verified for authenticity and authorization of allowed call-sets. If admitted, SP_m calls and process the get_data() call and sends back the response to the originating SP, SP_n.

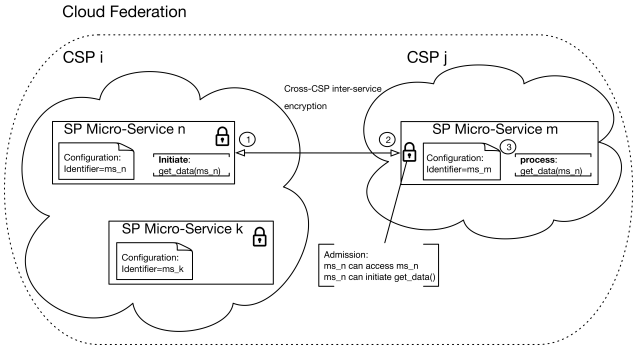


Figure 3: Authentication and authorization in Cloud Federation Cross-SP model

Breach Detection: The Cloud federation comprises various workload types that are owned by different autonomous organizations. Breach detection includes a complex data processing pipeline that integrates system signals originated from specific users of a CSP service as well as the potential cloud federation tenants. System signals are comprised of network devices as well as signals from infrastructure services. Only in recent years, after the growing numbers of data breaches and liabilities arising from losses,[12, 13, 14] have organizations started to incorporate business related metrics for breach detection[15]. Both data pipelines need to generate operational security warnings of potential incidents. The output of such warnings usually alerts security operations staff to potential incidents that require the relevant team's triage and response as appropriate.

Such methods are sub-optimal in a Federated Cloud for two main reasons: (1) different data sets are owned by different organization departments that are not integrated physically, schematically or semantically, (2) Lack of unification of both data sets as accomplished by fusion requires a complex transformation of both data sets semantics into a single data set. The above situation exacerbated when migrating the workload to the cloud as it introduces another orthogonal data set that contributes to complexity. The following sections propose a method for breach detection that collapses the three silos into a cohesive semantic-less data set that will enhance the Cloud Federation services detection breaches to an extent limited by available data and their investment in detection .i.e. allowing methods to the tenants to incorporate more data about their workload for more automatic detection.

4 Semantic-less Breach Detection

Malware infected cloud-computing-workloads introduces three core risks for organizations (1) Service unavailability, (2) Data breach , and (3) Data corruption. There is a need for breach detection system that helps to determine whether a workload is infected as well as the type of exploited risk type as enumerated above. Breach detection system effectiveness is influenced by a number of factors. We focused on the human social factor and the emergent public cloud offering. The following paragraph describes the important

factors required for optimized breach detection. This mode of breach detection has to span the heterogeneous schema employed by the various federation members.

4.1 The Human Social Factor

Enterprise IT is typically organized into silos. e.g., IT operations, network operations, database administration, and product engineering. The silos goal is to allow field-based ownership. Usually, silo teams are governed by different management hierarchies, communication styles, and vocabularies i.e. *semantics*. As as far as cyber security goes, semantics manifest by a particular interpretation of intrusion or breach. e.g., malware sending data to C&C might not impact the normal operation of a workload. Thus, product owners are oblivious to that risk while network operations detect unusual egress or ingress traffic usage patterns.

Enterprise IT workloads deployed as SP workload requires adopting unified cyber security best practices that overcome the different management hierarchies, communication styles, system security pans, data scheme, semantics and, vocabularies. The next paragraph shows how Federated Cloud helps enterprise IT improve its cyber security resiliency by offering a prediction tool that allows SP to apply proactive policies to mitigate potential threats.

4.2 The Cloud Federation Factor

Public Cloud services exacerbate the organization's human factor risk by introducing an additional silo that is often separated from the organization it serves. Public cloud operations are agnostic to its tenant's workload semantics by definition. CSPs configure their multi-tenancy to allow business with conflict of interest to run its workload on the same platform. Such practices and policies, augment the lack of cohesive view required for optimized malware detection.

Workloads deployed in public cloud services are not limited to known machine boundaries as traditional on-premise models offer. Although CSPs feature cyber security mechanisms that attempt mimicking the traditional computing workload hosting, workloads artifacts are under the CSP control. As such, the cloud client workloads might be compromised. Thus, there is a need for another cyber security dimension for the SP workload that overcomes the lack of control when running in the cloud.

We proposed a unique self-learning methodology that removes the need for tenant information that streamlines semantic-less information from the various software stacks of the Cloud Federation, including both tenant metrics and control-plane metrics. Also, it streamlines training data of security incidents shared in collaborative platforms outside the Cloud Federation. We also argue that a Cloud Federation optimizes such collaboration and self-learning process. We prototyped a system that implements such self-learning system that resulted in up to 87% True-Positive rate with 93% True-Negative.

Workload data and usage patterns form a critical path for the SP business success. The leakage of some of the workload data and usage patterns impose a threat to the SP business. This challenge represents a new threat of organizational espionage as well as attacks on the SP service that impacts SP business continuity. Therefore, sharing semantics breaks the isolation between the two systems and might hold the hosting system accountable for security attacks in CSP or Cloud Federation platforms. Also, transforming every workload semantics into a coherent model that aggregates numerous SP workloads requires a significant amount of investment. SPs will be reluctant to make such an investment, especially since it doesn't produce income. Therefore, this method has a low likelihood of being implemented. Therefore, enabling a method that eliminates SP investment and business risks is a key for the breach detection system success. Finally, a Cloud-Federation provides a centralized view of cross-CSP operations. Such centralized view allows SP workload deployment to different CSPs to gather a rich data set that will be available for malware identification and later, for predictive analytics. We suggest a method that captures computing resources usage and intra federation traffic and infers potential breach or disruption to proactively alerts CSP security stakeholders about suspicious cyber instances.

4.2.1 From Workload Semantics to Semantic-less

Cloud workloads are broadly composed of two types: online system, and offline system. The former provides low-latency, read/write access to data. For example, a web user requests a web page to load online and serve within a fraction of a second. The latter provides batch-like computing tasks that process the data offline, which is reported later to users by the system servers; for example, the search results based on a pre-calculated index. Offline production workloads are usually comprised of mainly unstructured data sets, such as click stream, web graph, and sensors data[2, 3].

The semantic-less detection will address the polymorphic malware case as its data stream are abstracted from computing activity. More specifically, a tenant's workload in a federated cloud manifested by software containers that are limited to not more than (1) namespace per tenant for isolation and (2) limited to a resources control groups(aka cgroup)[6] Control groups are the mechanism for limiting computing server host CPU, Memory, Disk I/O and Network I/O usage per namespace. That is the foundation of Linux Containers, which alludes to the existing methods of measurements of the metrics set, CPU, memory and I/O usage. We call this set the behavioral attributes set. Access to cgroup and namespace configuration and control is available on the host level i.e. the host OS that runs the multi-tenant workloads i.e. a control-plane component.

4.3 Data Collection

Both Cyber Security leaders and national agencies agree that addressing emerging cyber risks require sharing cyber attacks retrospects and their historical behavior, and discovered vulnerability reports as a foundation for collaboration, predictive time series analysis, risk quantification and risk allocations all leading to safer cyber services [16, 17]. Incidents are often documented in unstructured reports that

[6] https://www.kernel.org/doc/Documentation/cgroup-v1/cgroups.txt

require a manual analysis to identify trends[18].

To assess whether or not a system was breached, it is required to establish malicious system behavior patterns and then decompose those patterns into generic computing system metrics that can later be classified as harmful or safe. The following paragraph includes the source datasets we chose to assess the initial malicious patterns and their detectability by our method. We continued by decomposing the data and removing the tenant semantics. That allows a generic pattern of malicious activity dataset that can be used as a training data for the supervised model.

4.3.1 Source Datasets

We choose the National Vulnerability Database (NVD)[19] and the Vocabulary for Event Recording and Incident Sharing (VERIS) [20]. Both datasets included thousands of reported incidents spanning across various categories. Our model focuses on (1) Unauthorized access attempts, (2) Suspicious Denial of Service, and (3) Data Stealing Malicious Code, including ransomware instances. We filtered the incidents that conform to the categories and performed a qualitative assessment of the identified breach impacts. Lastly, for simplicity, we appllied an additional category that distinguishes the target component reported, service-based or client-based. We included only the service-based incidents. i.e., reported incidents that clearly targeted desktops and workstations were not included in defining tenant semantic structures.

We applied filters for training data accuracy. Filters for VERIS dataset included server workloads as indicated in Section 4.2.1, i.e., Authentication Server, Backup Server, Database Server, DHCP Server, Directory Server(LDAP, AD), Distributed control system, Domain Name Server, File Server, Mail Server, Mainframe Server, Web Application Server, and Virtual Machine Server[20]. Assets operating systems were filtered to Linux and Unix as such operating systems are more prevalent in servers than Windows, MacOSX, and mobile device operating systems.

VERIS dataset includes incident actions. We filtered the action types that fit the paper focus workloads. i.e. Brute Force, Cache Poisoning, Cryptanalysis, Fuzzing, and HTTP Request Smuggling attacks. We excluded Buffer overflow cases as such attacks can be prevented in deterministic methods and common in Windows-based operating systems[21]. The dataset size following the refinement is 5015 incidents. Table 1 summarizes the dataset we used for the training data.

Malware Category	modus operandi	Number of Incidents
Brute Force	Exhaustive effort of data encryption	946
Cache Poisoning	Corrupt data is inserted into the cache database e.g., DNS	894
Cryptanalysis	Exhaustive effort of data encryption	750
Fuzzing	Injects random bad data into an application to break it	946
HTTP Request Smuggling	Exhausting a proxy cache by sending HTTP requests	639
Data stealing malware	Data transmitting across unencrypted network	840

Table 1: Summary of datasets used

4.3.2 Removing the Tenant Semantics

Our approach attempts to detect anomalies in both control-plane and tenant activities that conform to suspicious patterns. In Section 4.3 Data Collection, we defined a categorical dataset that adheres to real incident data. This data applies to potential breaches for server-based workloads. We stipulate, for the purposes of this paper that such server-based workload will obey similar suspicious patterns when deployed in the cloud.

In this paragraph, we transformed the categorical dataset into a multivariate time series data that can be used for supervised anomaly detection. The multivariate set is comprised of general operating system observations that do not include any workload semantics but could be used for contextual anomaly detection. The contextual attributes are used to determine the position of an instance on the entire series. We showed that, based on collected incident data, the conversion of behavior patterns to multivariate time-series satisfies effective breach detection of any malware, conventional or polymorphic.

We gathered the operations reported in the incident reports (Table 1) and inferred about the operating system resources consumed during the malware lifespan. Table 2 depicts the relationship between the malware characteristics and operating system usage. Figure 4 describes a workload sample, video-on-demand. It shows the common pattern of the operating system resources usage that will be used as multivariate time series data sequences. The Evaluation section describes in more details the nature of the data and how it translates into meaningful time series data.

Malware Category	OS Resources Patterns
Brute Force	Extensive CPU, Memory, I/O to disk or network
Cache Poisoning	Extensive I/O to disk or network
Cryptanalysis	Extensive I/O to disk or network
Fuzzing	Network I/O Ingress
HTTP Request Smuggling	Network I/O Egress
Data stealing malware	Network I/O Egress

Table 2: Dataset Classification

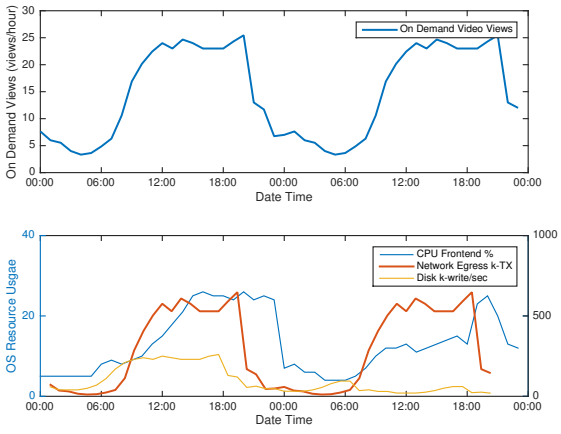


Figure 4: Workload semantics sample transformed into semantic-less training sequences

4.3.3 Prediction Methodology

We used the data gathered in Table 2 for formulating the anomaly detection problem of polymorphic malware[22]. The detection approach includes three distinct methods: (1) Detecting anomalous sequences in OS usages time series events, (2) Detecting anomalous subsequences within OS usages time series, and (3) Detecting anomalous OS usages events based on frequency. Let T denote a set of n training sequences based on OS usage generated by CSPs, SPs, and the Federation control plane. Also, S denote a set of m test sequences generated based on Table 2, we find the anomaly score $A(S_q)$ for each test sequence $S_q \in S$, with respect to T. T mostly includes normal OS usage sequences, while S includes anomalous sequences.

The semantic-less tool output produces a score for a scanned training sequence T using Regression. i.e., forecast the next observation in the time series, using the statistical model and the time series observed so far, and compare the forecasted observation with the actual observation to determine if an anomaly has occurred[22]. For simplicity, our model uses TensorFlow[23] for regression calculation. Our tool is not limited to that tool or the regression type.

5 Evaluation

We prototyped Cloud Federation system that mimics that properties analyzed in section 2, Cloud Federation. The prototyped system includes the component that is depicted in Figure 1. For the scope of the prototype, we enabled semantic-less metrics from both SPs and CSPs to improve correlation efficiency. CSP data sharing limits the effectiveness of any cyber analytical technique and, in practice will represent a compromise between improved cyber security and CSP privacy and confidentiality. With that proviso, in the following section, we evaluate a computer load coordination system component that manages on-demand streaming, generates T, a set of n training sequences based on OS usage generated by CSPs, SPs, and the Federation control plane. We chose on-demand video streaming as Video streaming is expected to constitute up to 85% of Internet consumers traffic within a few years[24]. Also, we showed that video-on-demand streaming follows a pattern of usage that can be monitored for breach detection that can help on-demand SP to seamlessly improve their consumer's privacy and provide their studio's safe e-commerce platform.

5.1 Experiment Planning

Below is a simulation of a cross-regional platform that is comprised of control-plane, workload-plane and coordinating components. This will be embodied in a resource allocation system (Kubernetes). This system provisions resources to be a priority of being near, users. The control-plane enables an effective compute resource provisioning system that spans across different public cloud providers and regions. Also, it collects operating systems usages for both the SP workload and control-planeThe coordinating components will accept user-workload demands as well as green energy availability from various regions and opportunistically seek to process streaming workloads using compute resources provisioned by green energy resources. The workload-plane will be comprised of edge streaming servers that process the end-user on-demand video streaming. It will be built on standard Apache HTTP[7] servers that run on the edge location.

The control-plane software infrastructure is based on Kubernetes[8], it facilitates internal discovery between Apache HTTP server instances so instances can connect across different cloud boundaries and regions. This architecture provides an open architecture that enables continuous monitoring. In a real world federation the data load may require several big data nodes and substantial compute capacity. This paper is a demonstration and proof of concept on a finite scale to permit model and parameter tracking and adjustment.

5.2 Execution

5.2.1 The System Preparation

The prototype experiment included the setup of three virtual datacenters deployed in different regions: (1) Central US, (2) West US and (3) East US. The clusters were sized based on US population distribution[9] by regions i.e. 20% for West US, 40% for East US and 40% Central US. The cluster sizes for West US, Central US, and East US are 3, 7 and 7 machines respectively. Each machine is standard 2-CPU cores with 7.5GB of memory.

The control-plane comprised of Kubernetes API server and controller-manager. The controller coordinator component will need to allocate resources across several geographic regions to different cloud providers. The API server will run a new federation namespace dedicated for the experiment in a manner that such resources are provisioned under a single system. Since the single system may expose external IPs, it needs to be protected by an appropriate level of asynchronous encryption[10].

For simplicity, we use a single cloud provider, Google Container Engine, as it provides a multi-zone production-grade compute orchestration system. The compute instances that process the user workloads are deployed as Docker containers that run Ubuntu 15 loaded with Apache HTTP server. For simplicity, we avoided content distribution by embedding the video content to be streamed in the Docker image. We ran 52 Docker containers that span across the three regions and acted as Content Delivery Network edges.

5.2.2 Baseline and Execution

The baseline execution included data populations for video streaming. The data population was achieved by the Kubernetes Jmeter batch jobs. The loader jobs goal is to generate traffic that obeys the observed empirical patterns depicted in

[7] Apache Web Server reference retrieved from https://httpd.apache.org
[8] Kubernetes reference retrieved from http://kubernetes.io
[9] US Population Distribution retrieved from https://www.census.gov/popclock/data tables.php
[10] Simulation code and data retrieved from https://github.com/yahavb/green-content-delivery-network

Figure 4. The system usage for both control-plane and SP capture, through cAdvisor, a kubernetes resource usage, and performance analysis agent. The agent, from every node in a cluster, populates system usage data to Heapster, a cluster-wide aggregator of monitoring and event data [11].

We labeled the system usage with the semantic-less dimensions, Network egress was measured by thousands of transmitted packets (k-TX), Disk writes per second (k-write/sec) and CPU usage per container (%). The Heapster aggregated the data based on the labels that are later pushed to centralized database, influxDB. We also used the influxDB HTTP API to inject randomized system usage data according to the three labels, CPU, network and disk usage. Those considered as the anomalous sequences $S_q \in S$. We used Figure 4 as a baseline sequence that randomized using NumPy[12]. The randomization followed the malicious usage patterns described in Table 2.

The execution required a TensorFlow session that looped through the dataset multiple times, update the model parameters and obtain the anomaly score $A(S_q)$ for each test sequence $S_q \in S$, on T. The breach and anomaly detection was performed using the following data streams and learning algorithms.

5.2.3 Limitations

We used influxDB because of its seamless integration with Kubernetes Monitoring system. However, our approach is not limited to influxDB or other database systems for that matter. We used TensorFlow for regression and anomaly score calculation. We did not use long training sequences. The maximum duration spanned across 48 hours. Training with longer sequences using long-running jobs and TensorFlow model checkpointing would improve our results. Our test content variety was limited and fixed. That might impact the generated tests sequences stability. Larger content variety would require longer training sequences for optimal detection.

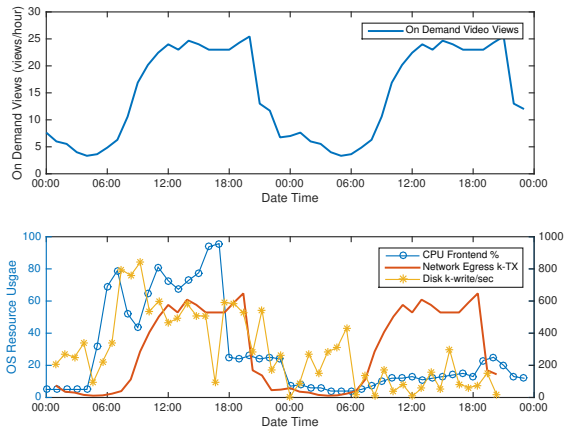


Figure 5: A Ransomware Anomalous Workload semantic-less sample

6 Analysis

The prototype included two core datasets, normal (Figure 4) and malicious (Figure 5). The CPU usage in the normal dataset fit the viewing patterns at the first half of the run. The second half required less CPU due to the caching mechanism applied in the Apache HTTP server that alleviates the need for the CPU when served through a cache. The Disk write pattern manifested similar content caching schema. The network egress ratio was not impacted by the caching schema.

The malicious dataset used the malware classification table (Table 2). Figure 5, shows a semantic-less behavior for ransomware malware that attempts to encrypt data while serving workload. Suspicious signals denoted by a star and o markers for CPU and disk write respectively. Based on the dataset classification, ransomware requires no network egress but CPU for data encryption and writing back to disk the encrypted payload. Our prototype included similar patterns depicted in Table 2 with a similar approach as done for ransomware.

Our model yielded a series of anomaly scores $A(S_q)$ for $S_q \in S$ anomalous patterns. We considered a potential breach of cases of anomaly and when regression produced a sufficient variability factor, i.e., a value that is not close to zero.

7 Conclusions

Security practices traditionally focus on prevention and tightening perimeter boundaries. However, with the advent of disparate, distributed, large scale, multi-tenant environments such as the proposed Cloud Federation, the traditional perimeter boundaries along with traditional security practices are changing. Defining and securing asset boundaries is more challenging and the system perimeter boundaries are more susceptible to breach. In this paper, we proposed a proactive approach for detecting a breach in a cloud workload. Such method requires no upfront investment from the monitored services. Upfront investments are often one of the main barriers to securing cloud service. Our tool eliminates such need and uses general system usage patterns that help to predict potential breach proactively.

Conflict of Interest The authors declare no conflict of interest.

References

[1] Leavitt, N. *Growth* **2009**, *27*, 15–20.

[2] Biran, Y.; Collins, G.; Dubow, J.; Azam, S. Federated Cloud Computing as System of Systems. 2017 Workshop on Computing, Networking and Communications (CNC) (CNC'17). 2017; pp 130–135.

[3] Biran, Y.; Collins, G.; Liberatore, J. Coordinating Green Clouds as Data-Intensive Computing. 2016 IEEE Green Technologies Conference (GreenTech). 2016; pp 130–135.

[4] Nelson, R. *EE-Evaluation Engineering* **2016**, *55*, 18–21.

[11] https://kubernetes.io/docs/user-guide/monitoring/

[12] Package for scientific computing with Python, numpy.org

[5] Kim, D.; Chung, H. R.; Thompson, P. R. Cloud-Based Automation of Resources. 2009; US Patent App. 12/634,050.

[6] Orr, R. J.; Abowd, G. D. The smart floor: a mechanism for natural user identification and tracking. CHI'00 extended abstracts on Human factors in computing systems. 2000; pp 275–276.

[7] Biran, Y.; Collins, G. Open compute-equipment design specification as a standard for cloud computing. Zooming Innovation in Consumer Electronics International Conference (ZINC), 2016. 2016; pp 70–75.

[8] Krutz, R. L.; Vines, R. D. *Cloud security: A comprehensive guide to secure cloud computing*; Wiley Publishing, 2010.

[9] Skorobogatov, S. P. Semi-invasive attacks: a new approach to hardware security analysis. Ph.D. thesis, Citeseer, 2005.

[10] others,, et al. Mayday: Distributed Filtering for Internet Services. USENIX Symposium on Internet Technologies and Systems. 2003; pp 20–30.

[11] Peng, T.; Leckie, C.; Ramamohanarao, K. *ACM Computing Surveys (CSUR)* **2007**, *39*, 3.

[12] Agrawal, T.; Henry, D.; Finkle, J. JPMorgan hack exposed data of 83 million, among biggest breaches in history. 2014.

[13] Krebs, B. *Krebs on Security* **2014**, *6*.

[14] Sidel, R. *Wall Street Journal, Sep* **2014**,

[15] Liu, Y.; Sarabi, A.; Zhang, J.; Naghizadeh, P.; Karir, M.; Bailey, M.; Liu, M. Cloudy with a Chance of Breach: Forecasting Cyber Security Incidents. USENIX Security. 2015; pp 1009–1024.

[16] Jaquith, A. *Security metrics*; Pearson Education, 2007.

[17] Chew, E.; Swanson, M.; Stine, K.; Bartol, N.; Brown, A.; Robinson, W. Performance measurement guide for information security. National Institute of Standards and Technology (NIST) Special Publication 800-55 Revision 1. 2008.

[18] Tufte, S. E. *Department of Management Science and Engineering, Stanford University, School of Information, UC Berkeley* **2015**,

[19] Center, C. C. *CERT Coordination Center* **2004**,

[20] Team, V. R. *Verizon Data Breach Investigations Report (DBIR)* **2015**,

[21] Wartell, R.; Mohan, V.; Hamlen, K. W.; Lin, Z. Binary stirring: Self-randomizing instruction addresses of legacy x86 binary code. Proceedings of the 2012 ACM conference on Computer and communications security. 2012; pp 157–168.

[22] Chandola, V. Anomaly detection for symbolic sequences and time series data. Ph.D. thesis, University of Minnesota, 2009.

[23] others,, et al. *arXiv preprint arXiv:1603.04467* **2016**,

[24] Biran, Y.; Collins, G.; Dubow, J.; Pasricha, S. Coordinating Green Clouds as Data-Intensive Computing. 2016: 3rd Smart Cloud Networks & Systems Conference (SCNS'16). 2016; pp 1–8.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 562-568 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Impact of PSS and SVC on the Power System Transient Stability

Mohammed Omar Benaissa*, Samir Hadjeri, Sid Ahmed Zidi

Laboratory of Intelligent Control and Electrical Power System, University Djillali Liabes, Sidi Bel Abbes 22000, Algeria

ARTICLE INFO

Article history:
Received: 27 March, 2017
Accepted: 01 May, 2017
Online: 01 June, 2017

Keywords:
AVR
Power System
PSS
SVC
Short Circuit
Transient Stability

ABSTRACT

The Static Var Compensator (SVC) is used to improve the stability of the power system because of its role in injecting or absorbing the reactive power in the electrical transmission lines. The Power System Stabilizer (PSS) is also a control device which ensures maximum power transfer and thus the stability of the power system enhancement. The PSS has been widely used to damp electromechanical oscillations occur in power systems. If no adequate damping is available, the oscillations will increase leading to instability. The present work is an original contribution to the problem of transient stability in the electrical power system, the authors have made some efforts to illustrate the flexibility and the importance of inserting the SVC alone or with the PSS the fact that maintain the characteristics of the system within acceptable limits in a very short time. The results show that the system has been developed successfully in terms of transient stability in a bi-machine transmission system only with the presence of PSS when a single-phase fault has been occurred, while the presence of SVC is more than essential when a three-phase fault is occurred.

1. Introduction

Today's world is continuously growing so that generation distribution and transmission of power are also simultaneously required to increase in same manner to fulfill the requirement. Power system stability may be broadly defined as that property of a power system that enables it to remain in a state of operating equilibrium under normal operating condition and to regain an acceptable state of equilibrium after being subjected to a disturbance [1].

Stability of this system needs to be maintained even when subjected to large low-probability disturbances so that the electricity can be supplied to consumers with high reliability. Certain system disturbances may cause loss of synchronism between a generator and the rest of the utility system, or between interconnected power systems of neighboring utilities. Various control methods and controllers have been developed over time that has been used for this purpose.

Recently, there has been a surge of interest in the development and use of FACTS controllers in power transmission systems. These controllers utilize power electronics devices to provide more flexibility to AC power systems. The most popular type of FACTS devices in terms of application is the SVC [2]. This device is well known to improve power system properties such as steady state stability limits, voltage regulation, and damp power system oscillations [3]. The SVC is an electronic generator that dynamically controls the flow of power through a variable reactive admittance to the transmission network, also the SVC regulates voltage at its terminals by controlling the amount of reactive power injected into or absorbed from the power system [4]. When system voltage is low the SVC generates reactive power (SVC capacitive). When system voltage is high, it absorbs reactive power (SVC inductive) [5].

It is known that the power-system stabilizers PSS for generators and the supplementary controllers for flexible ac transmission system (FACT) devices are efficient tools for improving the stability of power systems through damping of low-frequency modes [6], where the frequency of these modes ranges from 0.2 to 2.5 Hz [7]. Power System Stabilizer (PSS) devices are responsible for providing a damping torque component to generators for reducing fluctuations in the system caused by small perturbations [8].

In some cases, when the use of PSS cannot provide sufficient damping for inter-area power swing (0.1-0.7Hz), SVC damping controller is an alternative effective solution. The SVC

* Corresponding Author: Mohammed Omar Benaissa, University Djillali Liabes, Sidi Bel Abbes 22000, Algeria +213558481869,
Email: omarioriquelme10@yahoo.fr

supplemented with damping controller and PSS on synchronous generators are applied to increase the probability of stability [9].

Generally, it is important to recognize that machine parameters change with loading make the machine behavior quite different at different operating conditions. Since these parameters change in a rather complex manner, a set of stabilizer parameters, which stabilizes the system under a certain operating condition, may no longer yield satisfactory results when there is a drastic change in power system, stabilizers PSS should provide some degree of robustness to the variations in system parameters, loading conditions, and configurations [10, 11]. The basic structure of PSS is operating under typical control generator while the basic structure of SVC is operating under typical bus voltage control [13].

This paper is an extension of work originally reported in Proceedings of the International Conference on modeling, identification and control [12]. Firstly we start with an introduction containing an overview of the system under examination, in section two we have given a detailed description of the proposed system, section three is divided into two parties the first one shows the results obtained and discuss the impact of PSS when a single-phase fault is occurred whereas the second one discuss the impact of the SVC when a three-phase fault occurred, finally we have drawn an overall conclusion which reveals the most important point to consider from the current work.

2. System Description

The proposed power system works under abnormal system conditions and has the following characteristics see Figure 1.

- Two synchronous generators
- Two power transformers of 13.8 kV / 500 kV.
- Three bus bars.
- A transmission line of 500 kV and 700 km.
- Two PSS and two AVR (Automatic Voltage Regulator).
- The SVC characteristics: 200 Mvar.
- A purely resistive load of 5000 MW.

Note that the load center is modeled by (5000 MW) resistive load which is fed by the remote of (1000 MVA-plant M1), and a local generation of (5000 MVA-plant M2). A load flow has been performed on this system for G1 and G2 with generating rate of (950 MW and 4046 MW) respectively while the line carries (944 MW) which is closed to its surge impedance loading (SIL = 977 MW). The shunt compensated used in this research is with rate of (200 Mvar) to maintain system stability after faults [14].

The SVC does not have a power oscillation damping (POD) unit. G1 and G2 are equipped with a Hydraulic Turbine and Governor (HTG), excitation system, and Power System Stabilizer (PSS). Two standard types of stabilizer models can be connected to the excitation system: Generic model using the acceleration power (Pa which is the difference between mechanical power (P_m) and electrical output power (P_{eo}) and Multiband model which uses a speed deviation (dw) [14].

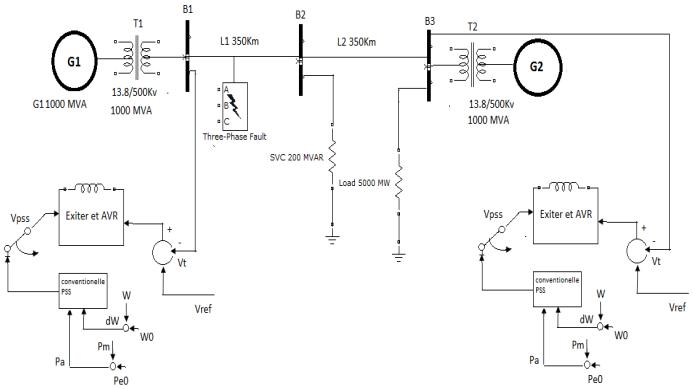


Figure 1. The Simplified Diagram of the Multi-Machine Power System with SVC and PSS

3. Problem Statement

3.1. PSS Modeling and Damping Controller Design

The operating function of a PSS is to produce a proper torque on the rotor of the machine involved in such a way that the phase lag between the exciter input and the machine electrical torque is compensated [15]. The block diagram of the ith PSS with excitation system is shown in Figure 2, where ($\Delta\omega_i$) is the deviation in speed from the synchronous speed which is the input signal. The output signal of the PSS is used as an additional input (ΔU_i) to the Excitation System block [16]. The three basic blocks of a typical PSS model are: The first block is the stabilizer Gain block, which determines the amount of damping. The second is the Washout block, which serves as a high-pass filter. The last one is the phase compensation block, which provides the desired phase-lead characteristic to compensate for the phase lag between the AVR input and the generator electrical (air-gap) torque [17].

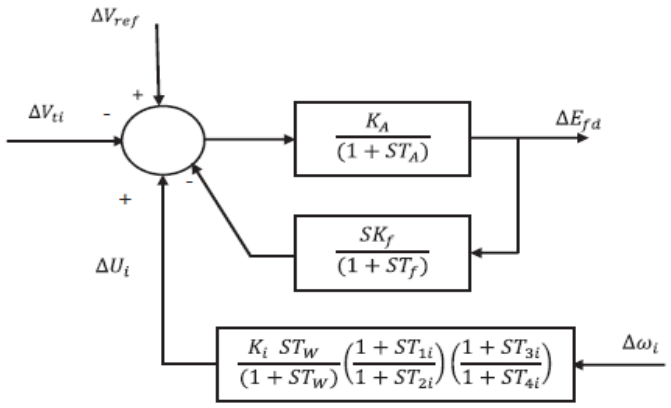


Figure 2. Block diagram of ith PSS with excitation system [16]

3.2. SVC modeling and damping controller design

A Static Var Compensator (SVC) is composed of a fixed capacitor in parallel with a thyristor controlled reactor (FC-TCR), this model consists of a harmonic voltage source in series with a variable source admittance [18]. The SVC adjusts the susceptance in each phase by controlling the conducting angles of the thyristor controlled reactor [19]. Also it can control the unbalanced loads more effectively and can enhance the transient stability of the system by inserting or absorbing instantaneous currents to or from the system [20].

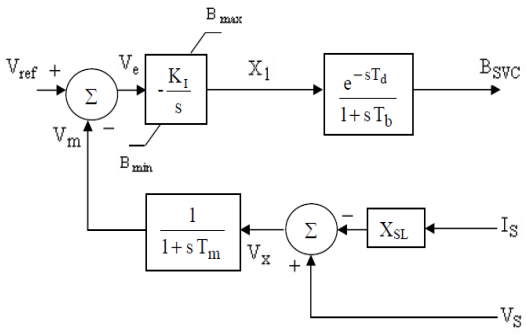


Figure 3. SVC Control Model

The equations describing the SVC controller are:

$$V_m = \frac{1}{T_m}(V_X - V_M) \quad (1)$$

$$X_1 = -V_e K_I \quad (2)$$

$$B_{SVC} = \frac{1}{T_b}(X_1 - B_{SVC}) \quad (3)$$

$$V_e = V_{ref} - V_m \quad (4)$$

$$V_X = V_S - X_{SL}.I_S \quad (5)$$

$$\begin{cases} X_1 = B_{\min}; X_1 < B_{\min} \\ X_1 = B_{\max}; X_1 > B_{\max} \end{cases} \quad (6)$$

Where

B_{min}: is the minimum susceptance

B_{max}: is the maximum susceptance

B_{SVC}: is the Static Var Compensator susceptance

K_I: is the proportional gain

T_m: is the measuring time

V_m: is the measured voltage

V_X: is the voltage at X position

V_{ref}: is the reference voltage.

X_l: is the line reactance

X_{SL}: is the slope.

4. Simulation Results and Discussion

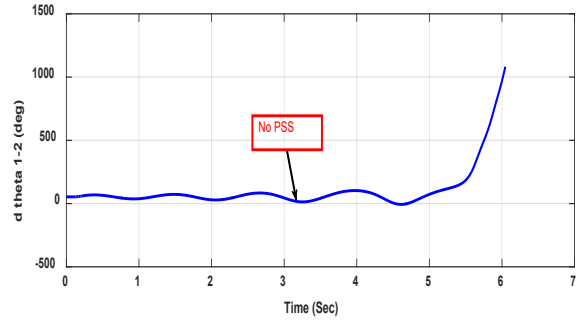
This work illustrates modeling of a simple transmission system containing two hydraulic power plants. A Static Var Compensator (SVC) and Power System Stabilizers (PSS), In order to observe the impact of PSS and SVC on the power system stability a single-phase to ground fault and a three-phase fault have been applied on the first section of the line (L1) see Figure 1.

4.1. Single-Phase Fault - Impact of PSS - No SVC

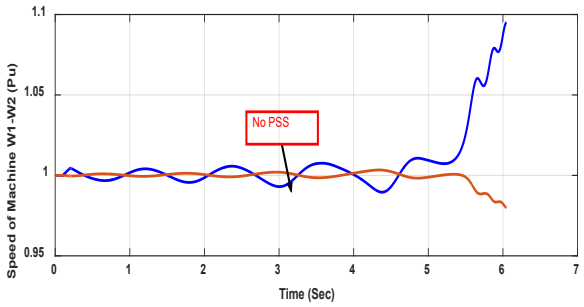
In this part the SVC is set to operate in fixed susceptance mode with (Bref=0) this is equivalent to putting the SVC out of service. Also a single-phase to ground fault have been applied at t=0.1 s and eliminated at t=0.2 s.

4.1.1. PSS out of service

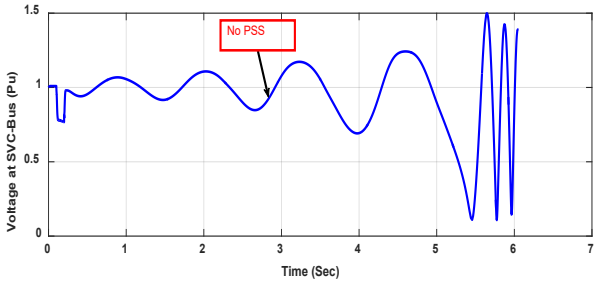
Figure 4 shows the obtained graphs without the impact of the PSS, if a single-phase fault is applied, immediately the system tends to support the oscillations, then these oscillations increased more and more, so that the system loses its steady state.



(a)



(b)

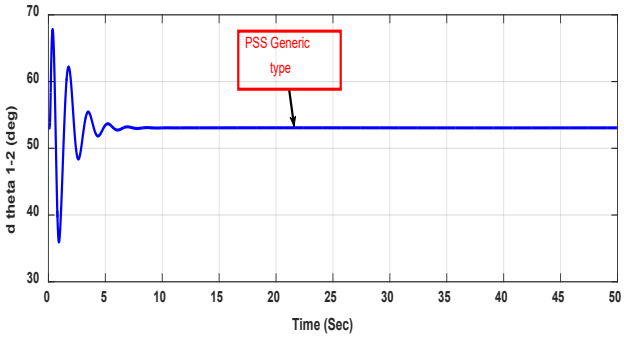


(c)

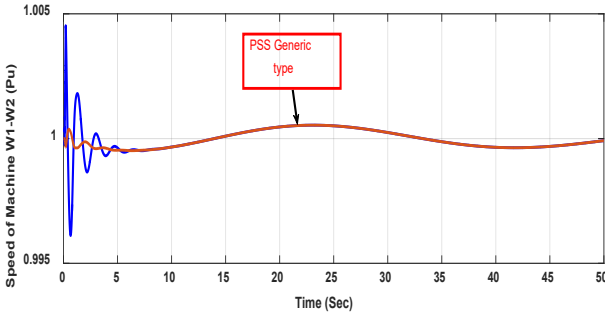
Figure 4. Impact of PSS for 1-phase fault (without PSS): (a) Rotor angle difference d_theta1_2 between the two machines, (b) Machine speeds of G1and G2, (c) Voltage at SVC-Bus.

4.1.2. PSS Generic type in service

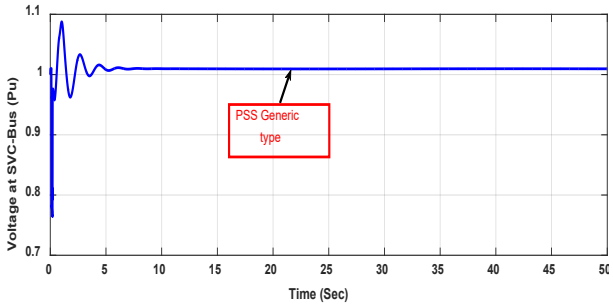
Figure 5 shows the impact of PSS Generic type, if a single-phase fault is applied, Figure 5(a) shows the Rotor angle difference d_theta1_2 between the two machines, it is found that the angle is approximately (50°) at the fault clearing time, Figure 5(b) shows that machine speeds (W1, W2) for G1 and G2 respectively converge to 1 Pu, so the machine speeds oscillate at the same time for low frequencies at 0.025Hz. Figure 5(c) shows that the voltage at SVC-bus is constant.



(a)



(b)

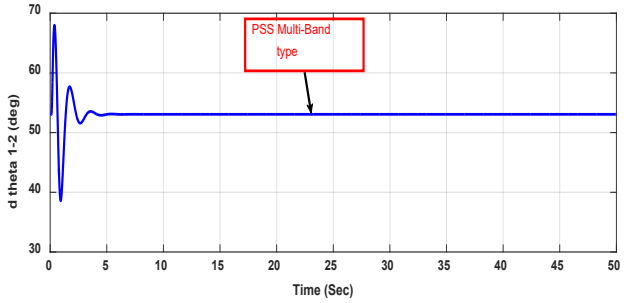


(c)

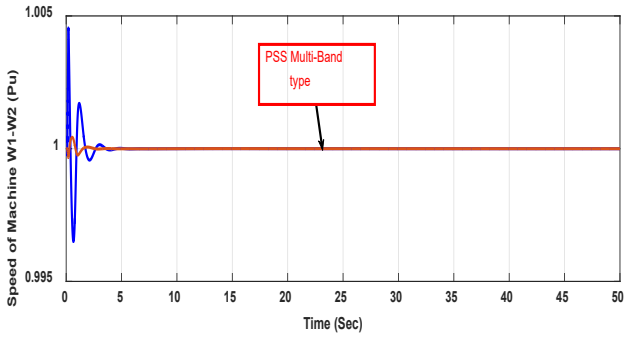
Figure 5. Impact of PSS for 1-phase fault (with PSS Generic type): (a) Rotor angle difference d_theta1_2 between the two machines, (b) Machine speeds of G1and G2, (c) Voltage at SVC-Bus.

4.1.3. PSS Multi-Band type in service

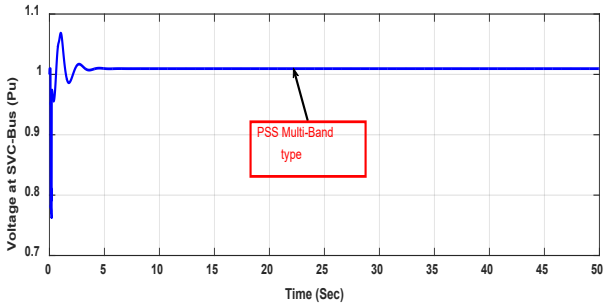
Figure 6 shows the impact of PSS Multi-Band type, if a single-phase fault is applied, the same results shown above with PSS Generic type have been drawn, however the Multi-Band type can attenuate not only the frequencies of 0.8Hz but also 0.025Hz .



(a)



(b)



(c)

Figure 6. Impact of PSS for 1-phase fault (with PSS Multi-Band type): (a) Rotor angle difference d_theta1_2 between the two machines, (b) Machine speeds of G1and G2, (c) Voltage at SVC-Bus.

Now the results have been assembled in the same graph to allow comparison between a various cases (without PSS, with PSS Generic type and with PSS MB type.

The first trace of Figure 7 shows the rotor angle difference d_theta1_2 between the two machines. Power transfer is maximum when this angle reaches 90°, this signal is a good indication of system stability. If d_theta1_2 exceeds 90º for too long a period of time, the machines will lose synchronism and the system goes unstable.

The second trace of Figure 7 shows the speed oscillation of machine G1 notice that machine 1 speed increases during the fault because during that period its electrical power is lower than its mechanical power.

The third trace of Figure 7 shows the positive sequence voltage at the SVC bus, we notice that as soon as the fault has been applied the voltage grows to reach 1.08 then this peak is followed by oscillations.

In the blue waveforms we can clearly see that, if there is no PSS than system can't be able to sustain stability and lose it synchronism (see Figure 7). In the red waveforms notice that when we use PSS Generic type system becomes stable but there is low frequency oscillation present (see Figure 7). In the magenta

waveforms we notice that when we use PSS MB type system may be stable with no low frequency oscillation.

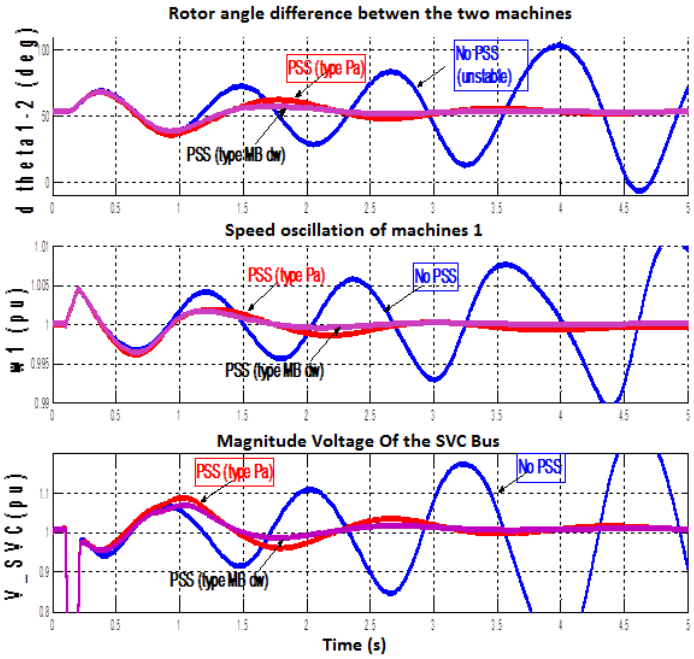


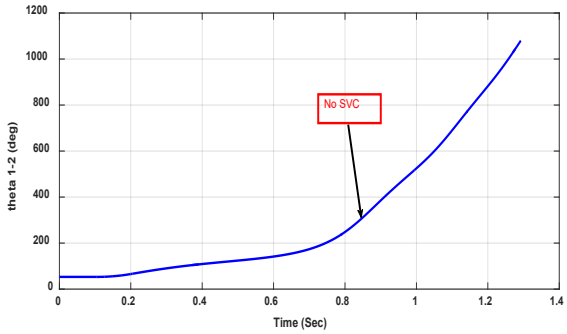
Figure 7. Impact of PSS for 1-phase fault (without SVC)

4.2. Three-Phase Fault - Impact of SVC

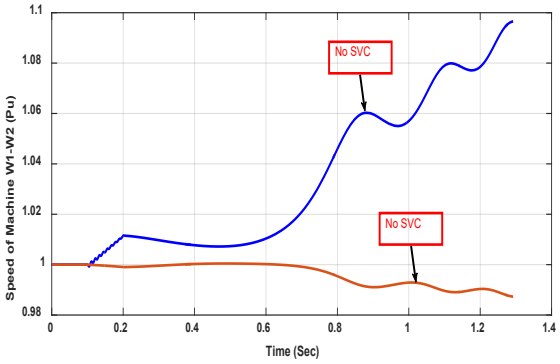
This part contains two tests, in the first one the SVC is set to operate in fixed susceptance mode with (Bref=0) this is equivalent to putting the SVC out of service. The second one, the SVC is set to operate in voltage regulator mode; however, the PSS Generic type has been maintained in service. Also a three-phase to ground fault have been applied at t=0.1 s and eliminated at t=0.2 s.

4.2.1. SVC out of service

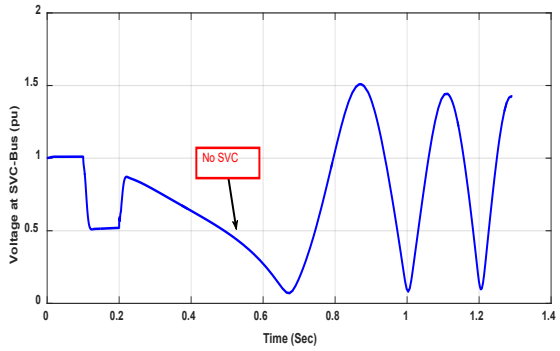
Figure 8 shows the obtained graphs with PSS Generic type in service and without the impact of the SVC (fixed susceptance mode Bref=0), if a three-phase fault is applied, immediately the system tends to support the oscillations, then these oscillations increased more and more, so that the system loses its steady state.



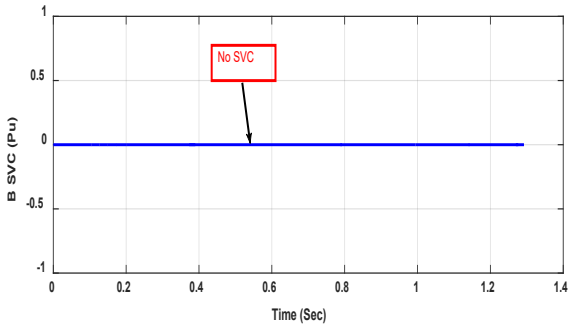
(a)



(b)



(c)

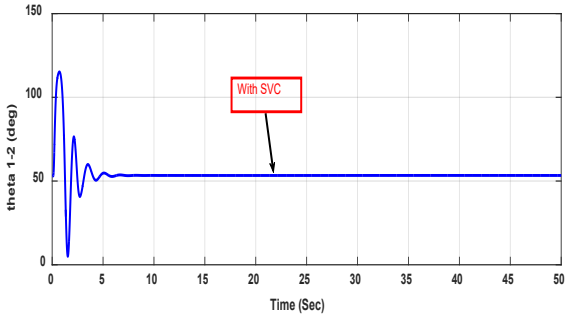


(d)

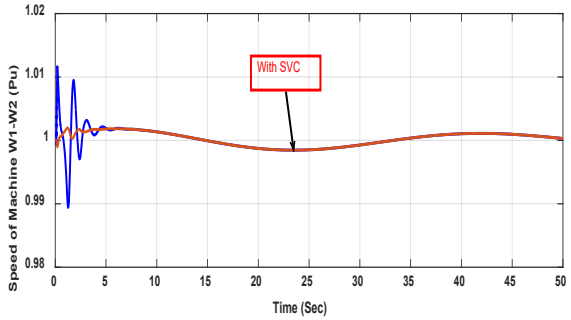
Figure 8. Impact of the SVC for 3-phase fault. (without SVC): (a) Rotor angle difference d_theta1_2 between the two machines, (b) Machine speeds of G1and G2, (c) Voltage at SVC-Bus, (d) Susceptance of SVC

4.2.2. SVC in service

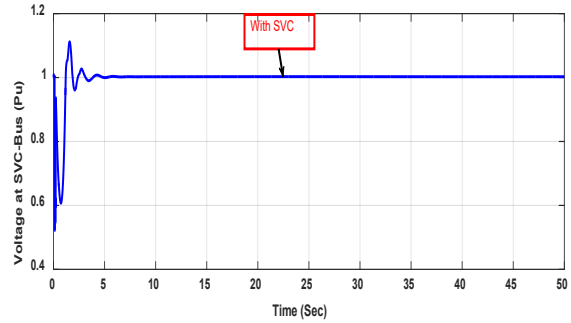
Figure 9 shows the impact of SVC (voltage regulation mode), if a three-phase fault is applied. Figure 9(a) shows the rotor angle difference d_theta1_2 between the two machines, it is found that the angle is approximately (70°) at the fault clearing time. Figure 9(b) shows that machine speeds (W1, W2) for G1 and G2 respectively converge to 1 Pu. Figure 9(c) shows that the voltage at SVC-bus is constant and equals to 1 Pu. Figure 9(d) shows that the SVC tends to support the voltage stability by injecting the reactive power into the transmission line whenever is needed (measured voltage differs from the reference voltage).



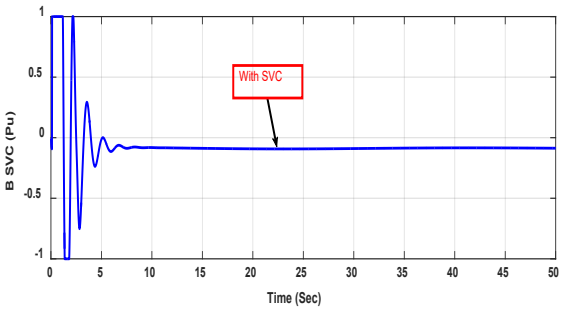
(a)



(b)



(c)



(d)

Figure 9. Impact of the SVC for 3-phase fault. (with SVC): (a) Rotor angle difference d_theta1_2 between the two machines, (b) Machine speeds of G1and G2, (c) Voltage at SVC-Bus, (d) Susceptance of SVC

Now, the results have been assembled in the same graph to allow comparison between the two tests (without SVC and with SVC).

Here we show that if there is three-phase transient fault than without SVC, both PSS are not able to maintain the stability. By looking at the red waveforms we should observe that the two machines quickly fall out of synchronism after fault clearing. In order not to pursue unnecessary simulation, the Simulink 'Stop' block is used to stop the simulation when the angle difference reaches 3*360degrees.

Now the SVC is set to operate in voltage regulation mode. The SVC will now try to support the voltage by injecting reactive power on the line when the voltage is lower than the reference voltage (1.009 pu). In steady state the SVC will therefore be 'floating' and waiting for voltage compensation when voltage departs from its reference set point.

The results of these studies show that the SVC has an excellent capability in damping power system oscillations and enhances greatly the dynamic stability of the power system. We also note that the stabilization time is minimal.

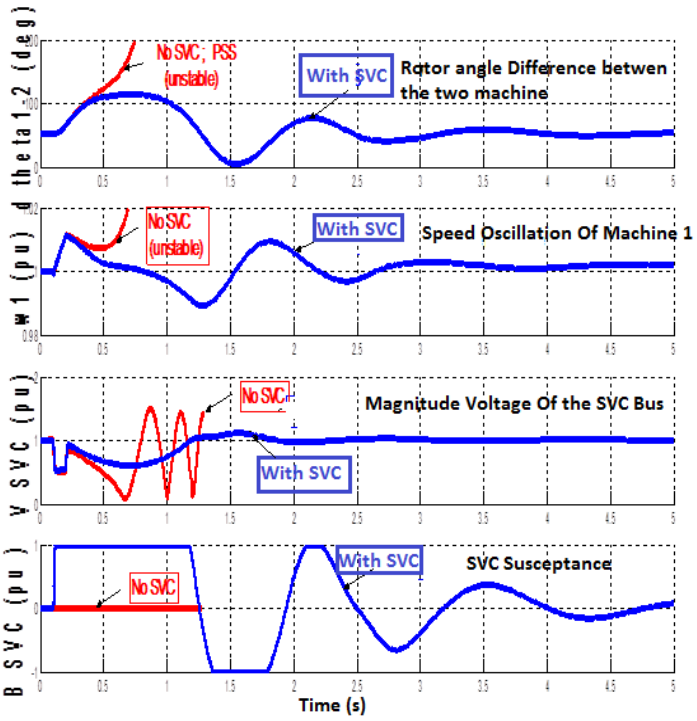


Figure 10. Impact of the SVC for 3-phase fault

5. Conclusion

From simulation results of the proposed model we can conclude:

- The proposed model is oscillatory and instable with absence effects of (PSS) and SVC.
- The selective of (PSS) are capable of proving sufficient damping to the steady state oscillation and transient stability voltages performance over a wide range of operating conditions and various types of disturbances of the system used in proposed model.

- If there is Single line to ground fault than the PSS able to sustain the stability, but using SVC the angle deviation is reduced.
- If there is three phase transient fault then without SVC both PSS are not able to maintain the stability.
- Compare working two types of (PSS), the multiband type oscillation is quickly damped than generic type.

References

[1] N.G. Hingorani, L. Gyugyi, Understanding FACTS: Concepts and Technology of Flexible AC Transmission Systems, IEEE Press, 2000.

[2] T. R. Hussein, "Using Power System Stabilizers (PSS) And Shunt Static Var Compensator (SVC) For Damping Oscillations In Electrical Power System" Journal of faculty MAAMOUN .

[3] S. Panda, N.P. Padhy, "Power system with PSS and FACTS controller modeling, simulation and simultaneous employing genetic algorithm," International Journal of Electrical and Electronics Engineering, 9-18, 2007.

[4] J. Wang , C. Fu, Y. Zhang, "SVC control system based on instantaneous reactive power theory and fuzzy PID" IEEE Trans. Industrial Electronics, **55**(4) , 1658-1665, 2008.

[5] L. Xu, "Coordinated control of SVC and PSS for transient stability enhancement of multi-machine power system" TELKOMNIKA, **11**(2), 1054-1062, 2013.

[6] M. Nadarajah, C. A. Canizares, R. John, G. J. Rogers, "Comparison of PSS, SVC, and STATCOM Controllers for Damping Power System Oscillations" IEEE Trans. Power Syst., **18**(2), 786-792, 2003.

[7] M. Khaleghi, M. M. Hossein Nezamabadi-pour, K. Y. Lee, "Pareto-optimical design of damping controllers using modified artificial immune algorithm" IEEE Trans. Power Syst., **41**(2), 240-250, 2010.

[8] F. Milla, M. A. Duarte-Mermoud, "Predictive optimized adaptive PSS in a single machine infinite bus" ISA Transactions, **63**, 315-327, 2016.

[9] X. Y. Bian, Y. Geng, K. L. Lo, Y. Fu, Q. B. Zhou, "Coordination of PSSs and SVC damping controller to improve probabilistic small-signal stability of power system with wind farm integration" IEEE Trans. Power Syst., **31**(3), 2371-2382, 2016.

[10] M.A. Abido, Y.L. Abdel-Magid, "Coordinated design of a PSS and an SVC-based controller to enhance power system stability" International journal of Electrical Power and Energy Systems, **25**(9), 695-704, 2003.

[11] A. Ali, "Development of system with transitory steadiness of a bi-machine transmission system with power system stabilizers and static Var compensator" International Journal of Engineering Research and Applications, **3**(3), 1121-1125, 2013.

[12] B. M. Omar, H. Samir, Z. S. Ahmed, "Impact of PSS and SVC on The Power System Transient Stability" in 8th International Conference on Modelling, Identification and Control, Medea and Algiers Algeria, 2016.

[13] K. Alok, D. Surya Bhushan, "Enhancement of transient stability in transmission line using SVC facts controller" International Journal of Recent Technology and Engineering, **2**(2), 2013.

[14] S. Gilbert, Transient Stability of a Two-Machine Transmission System with Power System Stabilizers (PSS) and Static Var Compensator (SVC), SimPowerSystems (Guid's), MathWork, MATLAB Simulink version 7.2.0, 2010.

[15] S.M. Abd-Elazim, E.S. Ali, "Coordinated design of PSSs and SVC via bacteria foraging optimization algorithm in a multi-machine power system" Electrical Power and Energy System , **41** (1), 44-53, 2012.

[16] N. Anil kumar, K. Ramesh, "Transient stability improvement using SVC and PSS" International Research Journal of Engineering and Technology, **03**(7), 1305-1311, 2016.

[17] K. Khalid, H. Saleh, E. Ercelebi, "Transient stability improvement in multi-machine system using power system stabilizer (PSS) and static Var compensator (SVC)" International Journal of Electrical, Computer, Energetic, Electronic and Communication Engineering, **09**(12), 1362-1375, 2015.

[18] L. J. Bohmann, R. H. Lasseter, "Equivalent circuit for frequency response of a static Var compensator" IEEE Trans. Power Syst, **1**(4), 1986.

[19] A. Köse, E. Irmak, "Modeling and Simulation of a Static VAR Compensator Based on FC-TCR" in 5th international conference on Renewable Energy Research and Applications, Birmingham UK, 2016.

[20] M. A. Hossain Sadi, M. Hasan Ali, "Combined operation of SVC and optimal reclosing of circuit breakers for power system transient stability enhancement" Electric Power Systems Research, **106**, 241-248, 2014.

Advances in Science, Technology and Engineering Systems Journal Vol.2, No. 3, 569-577 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

The Search for a Convergent Option to Deploy Smart Grids on IoT Scenario

Hamilton da Gama Schroder Filho*, José Pissolato Filho, Vinicius Luciano Moreli

UNICAMP - State University of Campinas, FEEC/DSE, 13083-970, Brazil

ARTICLE INFO

Article history:
Received:29 March, 2017
Accepted:16 May, 2017
Online: 01 June, 2017

Keywords:
Distribution automation
Electric power distribution
LoRaWAN; RF Mesh
Smart city
Smart grid
Smart metering

ABSTRACT

Smart city projects are quickly evolving in several countries as a feasible solution to the urban organization to provide sustainable socioeconomic growth and solve problems that arise as the populations of these cities grow. In this sense, technology application plays an important role in enabling automation of processes, improving the citizen's quality of life and reducing the costs of public services for municipalities and enterprises. However, automation initiatives of services such as electricity, water, and gas which materialize by the so-called smart grids, have emerged earlier than smart city projects, and are consolidating in several countries. Although smart grid initiatives have arisen earlier to projects of smart cities it represents a subset of the great scenario of IoT that is the vision in which the smart city projects are based. The time difference from developments between these two initiatives made the alternatives of communication technologies for infrastructures construction of communication followed different paths. However, in view of the great scenery of IoT is desirable to determine technologies that provide convergence of a single urban communication infrastructure capable of supporting all applications, whether they are typically IoT or traditional smart grid applications. This work is a review which presents and discusses the two main technologies which are currently best positioned to play this role of convergence that is RF Mesh and LoRaWAN. The strengths and weaknesses of each one of them are also presented and propose that in actuality LoRaWAN is a promising option to offer the required conditions to take on this convergent position.

1. Introduction

The first initiatives of smart grids have begun about ten years ago, in the year 2007. At that time the discussions about the intensive use of technology to automate public services in urban centers were very preliminary. In this context, technology options for building communications infrastructure was only scoped to support smart grid projects to the specific demands of utilities in segments such as electricity, water, and gas. In the specific case of the electric sector applications that motivated these initiatives were remote metering of consumption, automation of distribution networks, demand control, and to become viable an automated environment capable of enabling distributed generation development based on renewable sources [2, 3].

More recently, after the first smart grid projects implementation on a large-scale basis, initiatives of intensive use of technology for automation of public services in urban environments has emerged. These initiatives have defined smart city projects that use the concept of Internet of Things (IoT) as a technology to support the communication infrastructure. This new vision has stimulated technologies development aimed to meet applications with a particular profile, such as low data rate and tolerance of high latency. Smart grids typical applications were not primarily in focus [4, 5]. In this way, these two initiatives have created the conditions for industry to develop two different and parallel approaches regarding communication infrastructure. But it didn't take long for the vision about the need for convergence arises.

Considering this new scenario smart grid applications represents a subset in the bigger context of smart cities. As the industry developed initially without this convergent and integrated vision between smart grid applications and smart city the next step was the industry of smart grid communication systems expands the scope to cover other smart city applications. In the other hand, the IoT communication systems industry began to focus their products

*Corresponding Author: Hamilton da Gama Schroder Filho. UNICAMP - State University of Campinas. Cidade Universitária Zeferino Vaz, Barão Geraldo, 13083-970, Campinas, SP, Brazil | E-mail: hamilton.schroder@gmail.com

on meeting the demands typically smart grid with all its peculiarities.

In terms of well-established platform, RF Mesh assumed an outstanding position for smart grid projects. Regarding IoT applications, although there are currently various proposals under development, the technology that has to emerge as the most suitable for IoT and smart grid at the same time is LoRaWAN [6, 7].

This paper presents a review of the main technical requirements for smart grids, the main features of these two prominent technologies (RF Mesh and LoRaWAN) and makes a comparison between them. The goal is to discuss the possibilities of LoRaWAN to take on the role of convergence and interoperable option between different manufacturers required to support demands of smart cities and smart grids.

2. Smart Grid Requirements

The challenges faced by the electricity sector in all world markets have brought about deep changes in all three segments of the industry: generation, transmission, and distribution. From the point of view of the growing demand for electrical energy and urgency in environmental issues comes leveraging the renewable generation technologies development and transforming the traditional centralized array into distributed.

Transmission systems also go through a process of change. Search for technical solutions that reduce system losses is one of the actions. Another change is transmission lines construction to connect efficiently the new array of generation and ensure a suitable protection system [8].

Distribution segment might be the one that has been more challenged to promote changes in their way of operating. The distribution plant that was static becomes endowed with computational intelligence to promote operational gains and improvement of service quality indices, as well as add new services and features to consumers. The concept of smart grids represents this new way to operate and maintain the electrical system. It requires a telecommunications network overlaps grid interconnecting the systems to a centralized computer system. Smart grids add value to the entire industry, but the segments in which its application is most intensive is in distributed generation and mainly in the distribution system [8, 9].

Applications affected by this technological innovation are the consumption measurement and distribution operation. This new concept involves smart metering and Distribution Automation (DA), which evolves into the Advanced Distribution Automation (ADA), which allows the system to reconfigure them in the case of network failure. There are also other applications such as automation of street lights and asset management of the electrical system. Both smart metering and DA have specific technical requirements with regard to telecommunications support systems. As presented in the following items, it demands more stringent parameters from the point of view of data rate available and network latency tolerance.

2.1. Smart Metering

Smart metering means consumption measurements automation eliminating any human intervention in the process. Measuring routines are possible in real-time, every 15 minutes, every hour or in larger intervals. Smart metering processes uses an Advanced Metering Infrastructure (AMI).

Besides the main function, remote metering can provide electricity network, quality parameters and service availability together with the data collected from the SCADA systems of distribution substations.

The parameters contribution of quality and availability extracted from smart meters working as remote sensors enhances the diagnostic map of operation centers due to its capillarity and represent the points of service delivery [9].

2.2. Demand Response (DR)

The main motivation for the control of electrical energy demand is to adjust it to the generation capacity, transmission, and distribution, especially during peak times. In the case of mismatching down between available energy in the system and the total demand, actions are necessary to adjust it and try to balance the electricity system. In the context of smart grids, this interaction between consumers and the distributor of electric energy is possible in real-time.

The main source of consumer information is the smart meter that communicates with the utility systems via a communication network. In this same communication network, the load control takes actions on actuators installed in the facilities of consumers with the goal of providing the necessary adjustments in the consumption level. As the information system for the actions of demand response is the smart metering system, the technical requirement of AMI systems will meet DR's systems as well.

2.3. Distribution Automation (DA)

The basic idea of distribution system automation is to allow management (supervision and control) of network elements remotely from an operations center [9]. This concept is applicable to sub-transmission and distribution substations and distribution network using Automatic Reclosers (ARs).

In the case of DA, actions must be taken by an operator who performs a fault diagnosis and tries to restore as much as possible of network segments [4, 5].

Advanced Distribution Automation (ADA) gives more automatic actions by sending supervisory information to a centralized computer system. This system shall take decisions autonomously to open and close ARs in order to restrict the scope of a failure and restore a maximum number of possible network segments [9].

With regard to the requirements of communications network, three aspects are crucial: maximum data rate available, network availability, and low latency especially in the case of the ADA.

2.4. Other Applications

In addition to typical smart grid applications such as smart metering and distribution automation, there are new demands directly related to electric power sector. Among them are: assets management automation and street light automation. Supervision and control of the street lighting are already a reality in various cities around the world and has been usually supported by RF Mesh networks designed for smart grids. Regarding asset management (transformers, current transformers/potential transformers, capacitors, etc.), increasingly arise projects aimed at

accelerating the information flow to empower management processes. Asset management processes more efficient represent gains in logistics, inventory optimization, and improvement of operational and financial results of companies.

With regard to requirements of telecommunications systems, both streetlight and assets management automation do not require any specific characteristics different of those observed in smart grid applications. Thus, a network Field Area Network (FAN) or Neighborhood Area Network (NAN) designed for smart metering and DA applications will be also capable of supporting them. In this sense, these new applications in the context of a smart grid network reinforce the convergence around a single IoT platform. The question is which technology is best suited to take on this role. Although RF Mesh is in theory capable of supporting all these applications they have a typical IoT profile and suggest the use of a communications infrastructure more adherent to this scenario.

2.5. Telecom infrastructure

Telecommunication infrastructure to support smart grid applications is divided into Outdoor RF Concentrator Devices, and Network Interface Cards (NICs). Concentrators are RF equipment installed on Telecom towers, top of buildings or even poles and are responsible for communicating with endpoints and concentrated those located within its covered area. The NICs are interfaces embedded on endpoints and are responsible for communicating with nearest RF Concentrator and make the interface with endpoint device's application.

Outdoor RF infrastructure is usually divided into private and public networks. Public networks are those provided by third parties, generally, Telcos that share infrastructure with several other users. These are services like mobile network (GPRS, 3G, 4G) and Internet. Private are those in which a network or a network segment dedicated to the exclusive use of a single user. It can be considered private assets those acquired and operated by the company itself or contracted third-party companies, such as telecommunications operators. These are examples of services contracted to third parties: MPLS, dedicated circuits, satellite services, among others [10].

The main options currently available for utility companies to construct private telecommunication networks for smart grid applications are:

i Backbone layer: Optical Systems, and SHF Point-to-Point Radios.

ii Backhaul layer: WiMAX, LTE, and Point-to-Multipoint Radios.

iii Access layer (FAN/NAN): RF Mesh and LoRaWAN, which are the focus of this work.

Considering the criticality of smart grid applications an aspect that needs increasingly be observed in the infrastructure construction of telecommunications and information technology (IT) is the information security.

With regard to communication technology selection to meet the requirements of high availability (on the order of 99.9% or higher), performance, and security are more under control of distribution company if the infrastructure is its own [8]. Table 1 presents the typical requirements of a smart grid network. [8, 9, 11].

Table 1: Typical Smart Grid Requirements – Access Layer

Smart Grid Requirements				
	Smart Metering/DR	DA	ADA	Asset Management/ Street Light
Frequency Range (FAN/NAN)	≤ 2.4 GHz	≤ 2.4 GHz	≤ 2.4 GHz	≤ 2.4 GHz
Transmission Mode	Bidirectional	Bidirectional	Bidirectional	Bidirectional
Maximum Data Rate per Terminal	10 kbps	10 kbps	100 kbps	Best effort
Average Latency End to End	≤ 2 s	< 1 s	< 160 ms	Best effort
Mobility of Endpoints	Not Required	Not Required	Not Required	Required (AM)
Geolocation	Important	Important	Important	Required (AM) Important (SL)

With respect to the defined values for the parameters, it is interesting to define the concept of "best effort". This term refers to the traffic on IP network information according to the features that are available at a given moment, without any compromise with the service quality standards.

Therefore, the information flow on the network will occur in a way that is possible. This operation mode is used to meet applications not sensitive to certain parameters.

In this work, the requirements selection studied and compared to the technologies considers the most critical factors to ensure the minimum acceptable performance by type of application. Each technical aspect listed in Table 1 is presented below with their respective most relevant comments and their impact on smart grid applications.

i Frequency Range (FAN/NAN): The frequency band of operation of wireless systems that support the applications of smart grids is preferably not licensed and must be in such a way that meets the commitment between the longer range with the lowest possible transmission power and less sensitivity to obstructions between transmitter and receiver. Although the unlicensed frequency bands are much more susceptible to disturbance from other systems, they offer the great advantage of not requiring licensing processes before the telecommunications sector regulators, which also impacts in reducing the cost of the project, does not require payment of licenses for the use of the frequency spectrum. The range of the frequency spectrum in most markets that meet these requirements is in the range of 450 MHz to 2.4 GHz. Regarding disturbances caused by other systems operation in unlicensed frequency bands, the technologies used in smart grids applies countermeasures such as signal encoding, frequency-hopping, among others.

ii Transmission Mode: Applications of smart grids imply an interaction between the endpoint and the operation center of electric power distributors for sending a data requested

in the same way as AMI and asset management automation and lighting, or for confirmation of reception and confirmation of a remote command execution just like in the case of the DA/ADA and DR. This makes two-way communication is a mandatory requirement in smart grid support technologies.

iii Maximum Data Rate per Terminal: This parameter indicates the maximum throughput of each endpoint. The data transmission rate in a communication network based on wireless technologies can vary according several factors, of which the main ones are the frequency range, the transmission power, the level of obstruction to the radio frequency signal propagation, the load of traffic to be transmitted over the network in a given time interval, and network latency. Although a typical smart grid application requires a throughput ranging from 1 Kbps to 100 Kbps by endpoints depending on the application, higher values of data transmission rate enable implementation of more elaborate projects and the traffic of more detailed information between the endpoint and the operation center.

iv Average Latency End to End: The delay in data transmission in a communication infrastructure has a critical rate depending on the type of application to be supported. Real-time applications or those which require a quick response time of endpoints, latency must assume values as low as possible. In the case of smart grids, most critical applications are DA and ADA in particular. Another scenario in which communication network latency is a critical factor is when a polling based system requests information from a large number of endpoints. The total time of a polling cycle and information collection from all endpoints need to occur within a certain period of time acceptable and pre-defined in order to make the process feasible. This is the case of the smart metering of consumers and demand control (DR). In some cases, the number of endpoints to be measured and controlled can reach tens of million.

v Mobility of Endpoints: Mobile communications are important to utilities mainly for voice and data communication between the field teams and operation centers. These applications are supported by communication technologies specifically developed for this purpose and are different from those designed to meet the applications that are part of the scope of this work. Although typical applications of intelligent network (AMI, DR, and ADA) do not require communication platform mobility, automating asset management, which is a very important application for the utilities, as commented previously requires this feature to allow the ability to monitor the assets displacement until the positioning at the place of installation and commissioning.

vi Geolocation: Electricity energy distribution networks need georeferenced systems due to its extension and capillarity. This feature is essential for the operation and maintenance of the electricity grids and provides necessary visibility for operation centers to the actions of operation and dispatch processes of service orders to the field teams. The Geographic Information System (GIS) is computational systems responsible for positioning the assets and grid lines itself on the map. In the context of smart grids, a communication platform on which it is possible georeferencing of communications network elements and endpoints represent an important functionality. This feature enables a smart metering platform to work as a sensor network, allowing the mapping of massive or punctual failures. For distribution automation application a georeferenced communications network is an important tool for the network operation according to the DA or ADA vision. For the projects of automation of the asset management geolocation is essential. Associated to the mobile possibility of terminals it is possible to follow and locate a particular asset since the warehouse to its positioning and start-up in the distribution network. Another interesting application of this feature in the plant communication to the public lighting service that allows lamps location on their respective poles.

Based on these main technical features evaluation of communication technology in comparison with the requirements of the smart grid applications, it is possible to determine the adequacy of this particular technology in order to meet the demands. The goal should be map options that provide the wide scope and service convergence in order to optimize the investments required for infrastructure construction, and operation and maintenance costs.

In the following items will be discussed the main features of the two options of communication network technology in the access layer that stand out today: the already established RF Mesh and the emerging LoRaWAN. In the case of LoRaWAN is evaluating its suitability for smart grid applications. Then a comparison is made between these two technologies showing the advantages and disadvantages between them.

3. RF Mesh

This technology for construction of private wireless networks is based on mesh topology in which each network element is a repeater. In this way, each element can be accessed directly from an access point or via another network terminal element through one or several hops. The basic topology of an RF Mesh Network is presented in Fig. 1. It is based on the IEEE 802.15.4g standard in the physical layer, IEEE 802.15.4e standard in the link layer, MAC sublayer, and the IETF 6LoWPAN Protocol on sublayer LLC [10, 11].

For the purposes of the network address, the IPv6 protocol is used which is able to address 3, 4x1038 different IPs. An IP address is associated with a terminal device so that should not exist on the same network one IP address associated with two different terminal devices.

RF Mesh aims to be a technology based on open standards to ensure interoperability between networks and devices from different manufacturers. To operate in this way should be adopted the protocol stack as shown in Fig. 2 [11, 12, 13].

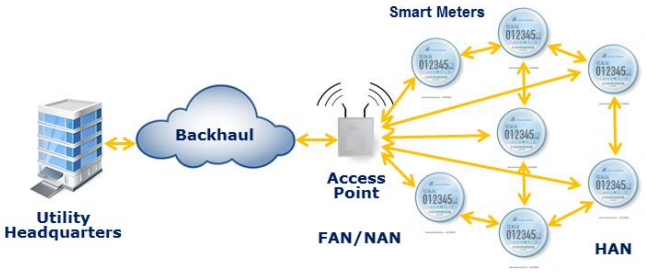


Fig. 1: Basic RF Mesh network topology

The RF Mesh standard was adopted by Wireless Smart Utility Networks Alliance (WI-SUN Alliance). WI-SUN Alliance is a global non-profit organization to promote industry solutions based on open standards and interoperable wireless networks. Officially formed in April 2012 has a mission to test and certify interoperability among different vendors. The focus is smart utility networks and smart cities. It has approximately 95 companies including utilities, government institutions, products manufacturers, and software companies [11]. The current movement of Wi-SUN seeks to expand the scope of its activities to meet IoT services.

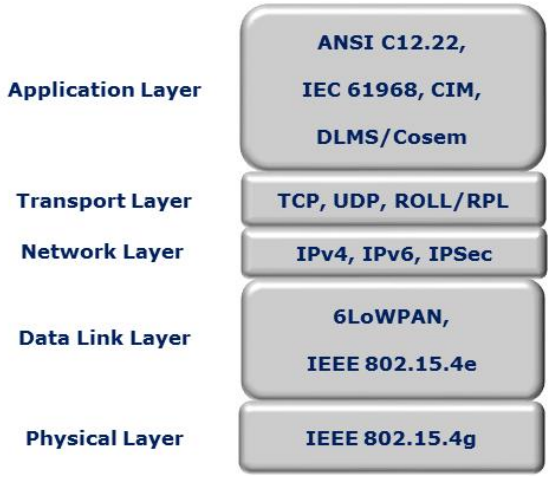


Fig. 2: Wi-SUN Alliance RF Mesh protocol stack

For utility applications, RF Mesh is a fully established technology and of proven effectiveness. The sum of the number of points serviced by the network deployed in several countries reaches over a hundred million. From the point of view of the technical and functional feature, RF Mesh was set to meet the requirements of utility companies' essential applications, such as smart metering and distribution automation, in the case of the electric sector companies [14].

The main characteristics are presented in Table 2 [11, 13, 14]. Data were based on theoretical information that was confirmed through field measurements on an RF Mesh network deployed in seven electric power distribution companies of CPFL Energia Group in State of São Paulo, Brazil deployed for smart metering of all its C&I customers and distribution automation.

Table 2: Main RF Mesh features

RF Mesh Features	
Application	Smart Grid (Smart Metering, DA) and Public Lighting. IoT in development.
Topology	Mesh
Frequency Range	863-876 MHz/915-921 MHz (Europe) 865-867 MHz (India) 902-928 MHz (North America and Brazil) 470-510 MHz (China) 920-928 MHz (Japan) 917-923.5 MHz (Korea)
Maximum Data Rate per Terminal	10-100 kbps
Average Latency	700 ms per hop (recommended up to 3 hops)
Maximum Aggregation per Concentrator	10,000 terminals
Urban Range (without repetition)	3-5 km
Rural Range (without repetition)	10-15 km
Technology Maturity Level	Smart Grid: Established IoT: In development
Mobility of Endpoints	Possible with restrictions
Geolocation	Not possible

Some highlights should be made regarding the technical RF Mesh specifications. First is related to the frequency band. The 900 MHz band is non-licensed in most markets. It dispenses with the obtaining of an operation license from the regulator agency of the telecommunications industry. The counterpart is the interference possibility, but that is outlined using advanced modulation techniques. Another aspect is the coverage area. It is a very interesting frequency range to bypass obstacles and provides a very convenient compromise between transmission power and available bandwidth.

The second point is about the network latency that is related to the topology. Mesh topology offers the advantage of range extension and creation of alternative routes automatically, on the other hand, can cause increased latency in the network. For smart metering applications, this feature does not represent a problem but produce an undesirable impact for DA applications. In order to overcome this restriction for DA applications, RF Mesh designs usually build specific paths in the network or use additional specific network elements, which lead to the rise of network costs. Endpoints can connect to the central Server via any network concentrator. This provides mobility in all coverage areas, but with restrictions in the sense that there is a trend of increase routing information traffic due to neighbor endpoints interaction.

4. LoRaWAN

This newly created technology won momentum with the LoRa Alliance foundation in March 2015. This non-profit organization gathers currently about 330 companies from various countries of the world between telecom operators, equipment manufacturers, semiconductor manufacturers, software companies, computer companies and consulting firms [15].

LoRaWAN is a fully convergent technology based on open standards, low-cost, and was designed from the start to build urban platforms for it. Thus the scope of application is very comprehensive. Among the possible applications are smart grids for utilities, including companies of the electric power sector. The offer of embedded devices grows constantly and includes portable device monitoring, public and patrimonial security solutions, management of urban infrastructure, healthcare, and smart meters [16].

Unlike RF Mesh topology LoRaWAN has a star topology which simplifies the operation, and significantly reduces traffic on network destined to routing information. On the other hand, it does not count with the coverage possibility of extension through the relay on the neighboring terminal device. This loss of functionality as well as being convenient for the context of it does not represent a problem due to the fact a LoRaWAN concentrator device has an average cost of approximately 5 times less than a Mesh RF concentrator.

As the frequency range of operation is also located on non-licensed 900 MHz spectrum, covered area tends to be very similar to RF Mesh. Advanced modulation techniques and access (CDMA) are also used to make the network virtually immune to interference and increased sensitivity of LoRAWAN embedded interfaces on endpoint devices [6, 17].

If a coverage expansion is needed, it can be done by using additional concentrators in RF shadow regions without causing a significant rise in project costs, considering the margin of investments reduction if compared to RF Mesh. Fig. 3 shows the reference topology of LoRaWAN and Fig. 4 the protocol stack [18].

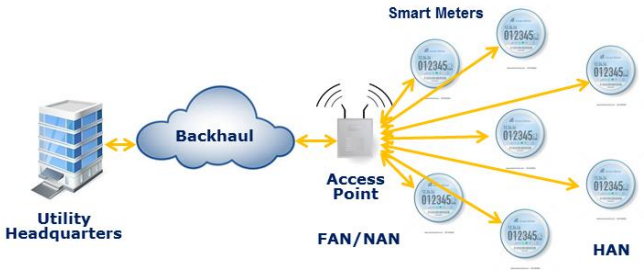


Fig. 3: Basic topology of LoRaWAN network

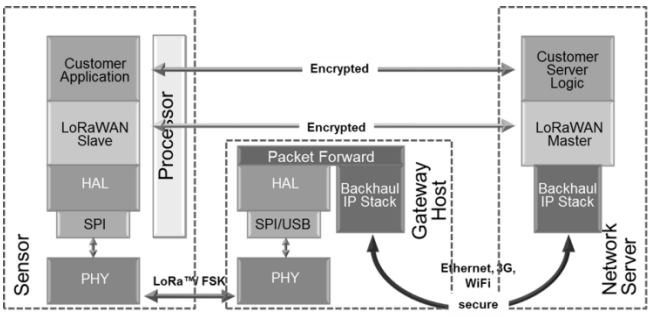


Fig. 4: LoRaWAN protocol stack

A prominent feature of this technology is the extremely low power consumption. The NICs in LoRaWAN endpoint devices are capable of detecting RF signals with power up to 20 dB below noise level consuming a minimum of electric energy. These LoRaWAN NICs can operate autonomously with the same internal battery for 10 to 20 years. This feature is extremely interesting to equip the utility's smart meters.

Regarding the performance characteristics given in Table 3 [6, 17, 19] it is observed significant advantages such as data rate available per terminal, uniform latency due to multiple hops absence to connect endpoints to a concentrator and aggregation capacity of terminals per concentrator. In the current version of LoRaWAN, this feature is 50% higher if compared to RF Mesh current version.

Table 3: Main LoRaWAN features

LoRaWAN Features	
Application	IoT, Smart Cities, Smart Grid (Smart Metering. DA needs to be studied)
Topology	Star
Frequency Range	867-869 MHz (Europe and India) 902-928 MHz (North America and Brazil) 470-510 MHz (China) 920-925 MHz (Japan and Korea)
Maximum Data Rate per Terminal	50 kbps
Average Latency	1 s
Maximum Aggregation per Concentrator	15,000 terminals
Urban Range (without repetition)	2-5 km
Rural Range (without repetition)	10-15 km
Technology Maturity Level	In positioning
Mobility of Endpoints	Possible
Geolocation	Possible

A highlight needs to be made regarding DA applications. Although the theoretical data suggest that this application can also be supported further studies must be done considering all requirements. It is necessary to consider the various situations of communication networks and the topology of distribution grid to determine a recommendation more assertive. The LoRAWAN suitability to any situation of DA projects still needs to be proven. A LoRAWAN differential is the three different classes of service available to allocate the applications according to its technical requirements [20]. Follows below a summary description of these three classes of services:

i. Bidirectional end-devices (Class A): The terminal's uplink transmissions are based on an ALOHA-type protocol. The downlink from the server can only be made in two short receive windows that open after the uplink transmission. This class is that it provides the lowest consumption of energy in the terminal. It can be used for street light, smart meters of the residential segment and Asset Management.

ii. Bidirectional end-devices with scheduled receive slots (Class B): A pre-programmed transmission window is opened, which is managed from a timing signal (Beacon) that indicates when the receiver is ready to receive. It can be more convenient for smart metering of a C&I customer, smart metering substations located at the border of distribution and sub-transmission networks, or to load firmware updates of smart meters.

iii. Bidirectional end-devices with maximal receive slots (Class C): A transmission window continuously opens for transmission. This class is best suited for the possible allocation of the applications (Reclosers) and sub-transmission management systems based on SCADA.

Mobility is an inherent feature of technologies designed for IoT. In this sense, LoRaWAN is totally suitable for applications that require this functionality, such as asset management.

In the following item, a comparison is made between these two technologies for construction of smart grids FAN/NAN. The objective is to determine if LoRAWAN represents an appropriate option considering its technical characteristics of data rate available, low complexity, the guarantee of interoperability among different vendors, and mainly by its convergence around the IoT applications for smart cities.

5. Technologies Comparison

As demonstrated on items III and IV both technologies RF Mesh and LoRaWAN meet the requirements for the most part of smart grid applications. There is no restriction to use LoRaWAN for smart metering. An exception must be made for distribution automation (DA/ADA) in the case of LoRaWAN that still must be better studied. Table 4 presents a comparison between RF Mesh and LoRaWAN considering the parameters studied on this work.

Especially with regard to network latency, it is needed to evaluate boundary conditions about the possibility to use this technology, due to the fact that its performance on this item operates on the borderline of the requirements. For ADA the use of LoRaWAN in the current version of technology presents an even greater constraint also with regard to maximum permissible latency. For these two applications, a solution would be the use of technologies in Backhaul layer able to perform this function. In this scenario, the endpoint of distribution automation would be accessed and serviced directly by this layer of the network, without using the access layer LoRaWAN. Possible telecommunications technology for DA/ADA would be LTE, WiMAX, Point-to-Multipoint radio systems, or even an RF Mesh System specifically designed for this purpose or serving as a Backhaul solution for a LoRaWAN urban network. Table 4 presents a comparison of technical features of these two Network Access Layer technologies. It is proven the requirements compliance for smart metering and other applications such as street light and assets management automation.

Even if a restriction is observed for distribution automation applications, there is a boundary condition to be used which would be to build a network specifically designed to support this application while maintaining other smart grid services via LoRaWAN.

Table 4: RF Mesh X LoRaWAN Comparison

RF Mesh x LoRaWAN		
	RF Mesh	LoRaWAN
Topology	Mesh	Star
Maximum Data Rate per Terminal	10-100 kbps	50 kbps
Average Latency	700 ms per hop (recommended up to 3 hops)	1 s
Maximum Aggregation per Concentrator	10,000 terminals	15,000 terminals
Outdoor RF Concentrator Average Cost per Terminal	US$ 0,50	US$ 0,07
Technology Maturity Level	S.G: Established IoT: In development	In positioning
Mobility of Endpoints	Possible with restrictions	Possible
Geolocation	Not possible	Possible

One relevant advantage of LoRAWAN technology is the fact that is a fully convergent technology, open standards-based, low-cost, and ready to meet all IoT applications in the context of a smart city project. It is desirable that investments in a smart city platform meet all applications and services with an IoT vision, in order to optimize resources and simplification of telecommunications infrastructure. The LoRAWAN technology simplicity is another important aspect in comparison with RF Mesh. The fact that each terminal accesses directly a concentrator excludes the need for complex routing protocols present on RF Mesh. In the context of IoT, this functionality is not even desirable. There are various situations in which it is not even appropriate a terminal access a concentrator through another terminal.

Mesh topology provides the advantage compared to star topology extending the coverage area through NICs that act as a repeater. The endpoints of a neighborhood area network (NAN) can be aggregated in one NIC closer to the concentrator. If a failure event or unavailability of a NIC acting as an aggregator, another one can assume that role. Star topology does not have this functionality, but as discussed earlier, it does not represent a problem for the IoT scenario and offers the advantage of becoming latency uniform of the network. This occurs because star topology does not have hops among endpoints to access a concentrator as happens with RF Mesh.

The endpoints mobility is possible in both technologies. Endpoints of LoRaWAN and RF Mesh are not necessarily bound to a specific concentrator, but in RF Mesh there is a trend of increase traffic of routing information. In this sense, mobility is possible with restrictions. RF Mesh was initially designed to support applications such as smart metering, DA and street light, in which mobility is not an essential issue. Thus, this feature was not initially a priority. However, new smart grid applications with an IoT profile like asset management require a communication network capable of supporting mobility.

The geolocation possibility is an important aspect to be considered in the communication technology selection designed to support applications on smart grids. Utilities use georeferenced systems in their operational processes in order to plot on maps, geographical areas of distribution plant and locate through geographical coordinates their electrical grid assets. Geolocation becomes an important functionality for a communication network that aims to automate these assets, whether for operational processes or management. In this sense, LoRaWAN presents a significant advantage over the RF Mesh. Even though the studies point to restrictions on the use of LoRaWAN for DA, the ability to georeference smart meters in measuring plant extends its functionality to a sensor network which is able to geographically map failures occurrence in delivering of electric power. In addition to this application, the geolocation associated with communications mobility plays a key role in the projects of automation of management of electrical assets.

With regard to the frequency of operation and coverage, both urban and rural environments are well served by both technologies. They have similar characteristics and meet the requirements of the smart grids. However, considering a wider vision in the context of possible partnerships between utilities and municipalities to attend public services, a real scenario points out to the communication infrastructure construction in which coexist technologies such as LoRaWAN in the Access Layer, RF Mesh with mainly functions of Backhaul Layer besides LTE, WiMAX, PMP radios and Optical Fiber systems. Eventually, it will be necessary to complement the coverage with contract services from Telcos based on the public platform of NarrowBand IoT (NB-IoT) for RF shadow regions of the proprietary networks in which the investment on the part of the utilities and municipalities to cover those regions are not financially viable.

Data rate available per terminal in RF Mesh is typically 10 kbps, but some versions go up to 100 kbps. For IoT applications including smart grid, the data rate required by applications rarely exceeds this value. LoRaWAN technology that offers a maximum data rate per terminal of 50 kbps also meets this requirement. Certainly, the data rate effectively available at each point of the coverage area of a concentrator will depend on the level of obstruction to RF propagation, and the distance between the endpoint and the concentrator. In this sense optimization of effective data rate available on each endpoint will depend on the RF design.

Aggregation capacity is a parameter that defines how many endpoints can be connected using a single concentrator device. This implies directly in RF design which will affect the total cost of outdoor RF infrastructure and distribution of this cost per endpoint.

Outdoor RF Concentrator Average Cost per Terminal considers one RF concentrator distribution average cost among endpoints and the maximum aggregate capacity per concentrator. The cost of embedded NICs in the endpoint is not included. A comparison shows that LoRaWAN presents a lower cost in part due to the fact that the average price of the LoRaWAN concentrator device is about 5 times lower than in RF Mesh. Another fact is that aggregation capacity is higher reducing in about 7 times the cost of RF concentrator devices per endpoint.

The maturity level of technologies is different, mainly due to the time of startup of each one: Wi-SUN Alliance in April 2012 and LoRa Alliance in March 2015. This suggests that although the LoRaWAN is advancing fast in various markets, there is still much to be developed. The great interest of the industry demonstrated in just 1 year of organization, especially due to the adequacy of LoRaWAN on IoT projects makes the future of this very promising technology. On the other hand, RF Mesh is well established for the smart metering and automation. Although LoRaWAN is suitable for smart metering, there is an uncertainty regarding its support for distribution automation due to its high latency.

The next steps which certainly will follow from now on point out to an increasing engagement of industry; services companies, including utility sector; government entities; research agencies and academia. Studies of new applications that can be supported by LoRaWAN are advancing in various countries. In the context of smart grids, main applications that require focus to adjust this technology are related to electric grid automation.

6. Conclusion

Smart city projects in the urban environment introduce innovation in the infrastructure services provisioning, making applications accessible directly and fast anywhere. In this context, the concept of Internet of Things (IoT) plays a key role in the telecommunications platform construction that do these projects implementation possible on a large-scale basis. Smart grid initiatives for utilities represent part of this big scenario of smart cities.

Construction of a platform for IoT becomes reality by private and public networks combination, including the Internet as the main platform. Formation of private networks for IoT has two technologies that stand out currently: RF Mesh and LoRaWAN. In this sense, it is desirable to determine an option that provides smart grid and other smart city applications convergence around the same technology. In order to support smart grid applications, utility companies have widely adopted RF Mesh that aims to be an open and interoperable option to build private networks. Regarding interoperability that is essential for IoT projects, RF Mesh has yet work to be done.

For IoT services including smart grids the emerging technology LoRaWAN represents a really convergent alternative which also adopts open and interoperable standards, but at a lower cost if compared to RF Mesh. LoRaWAN technology provides the features necessary to take on the role of an adequate alternative infrastructure to support all smart cities applications, including smart grids. An exception must be made for DA and ADA, which needs to be better studied and tested under a LoRaWAN platform.

A boundary condition would be a separate network deployment specifically designed for this purpose. Another possibility is to build a LoRaWAN Backhaul network also capable of supporting DA and ADA. The currently available technology options that could play this role are WiMAX, LTE, Point-to-Multipoint Radio Systems, or even an RF Mesh System.

Finally, taking into account what was studied on this work it is recommended that utility companies that have not invested yet in a smart metering network based on RF Mesh should consider LoRaWAN inclusion as a possible option in their projects. This can help to find a viable way to build a smart grid infrastructure, taking into account the new context of smart cities, in order to provide a greater gain in scale, preparing the path for possible partnerships between companies providing public services and municipalities. The ultimate goal as well as to make feasible business plans that enable the projects, is to provide citizens quality public services based on technology-intensive application, at the lowest cost as possible.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This work was supported by the Department of Electrical Systems (DSE), Institute of Electrical Engineering and Computing (FEEC), State University of Campinas – UNICAMP, SP, Brazil.

References

[1] H. G. S. Filho, J. P. Filho and V. L. Moreli, "The adequacy of LoRaWAN on smart grids: A comparison with RF mesh technology," *2016 IEEE International Smart Cities Conference (ISC2)*, Trento, 2016, pp. 1-6.doi: 10.1109/ISC2.2016.7580783.

[2] TITLE XIII- SMART GRID SEC. 1301- 1308 STATEMENT OF POLICY ON MODERNIZATION OF ELECTRICITY GRID. Energy Independence & Security Act '07.

[3] Y. Yang, F. Lambert and D. Divan, "A survey on technologies for implementing sensor networks for power delivery systems", IEEE Power Eng. Soc. Gen. Meeting, pp. 18, Jun. 2007.

[4] J. Jin, J. Gubbi, S. Marusic and M. Palaniswami, "An information framework for creating a Smart City through Internet of Things", *IEEE Internet Things J.*, vol. 1, no. 2, pp. 112-121, 2014.

[5] A. Zanella, N. Bui, A. Castellani, L. Vangelista, M. Zorzi, "Internet of Things for Smart Cities", IEEE Internet Things J., vol. 1, no. 1, pp. 22-32, Feb. 2014.

[6] M. Centenaro, L. Vangelista, A. Zanella and M. Zorzi, "Long-Range Communications in Unlicensed Bands: the Rising Stars in the IoT and Smart City Scenarios", IEEE Wireless Communications, vol. 23, October 2016.

[7] P. Diduch and L. Chang, "Smart Grid: Challenges, research directions and possible solutions", Power Electronics for Distributed Generation Systems (PEDG), *2010 2nd IEEE International Symposium*, pp.670,673, 16-18 June 2010.

[8] F. B. Beidou, W. G. 6 M. Centenaro, L. Vangelista, A. Zanella and M. Zorzi, "Long-Range Communications in Unlicensed Bands: the Rising Stars in the IoT and Smart City Scenarios", IEEE Wireless Communications, vol. 23, October 2016.

[9] V. C. Gungor, D. Sahin, T. Kocak, S. Ergut, C. Buccella, C. Cecati, G. P. Haneke, "A Survey on Smart Grid Potential Applications and Communication Requirements", *IEEE Trans. Ind. Inf.*, vol. 9, no. 1, pp. 28-42, 2013.

[10] H. Li, "Efficient and secure wireless communications for advanced metering infrastructure in Smart Grids", *IEEE Trans. Smart Grid, vol. 3 no. 3 pp. 1540-1551 Sept 2012.*

[11] *IEEE Guide for Smart Grid Interoperability of Energy Technology and Information Technology Operation With the Electric Power System (EPS), End-Use Applications*, and Loads, IEEE Std 2030-2011, 2011, pp. 1–126.

[12] H. G. Schroder Filho, J. Pissolato Filho and A. J. G. Pinto, "New Methodology for Smart Grids in Brazil". *IEEE Chilecon 2015.* Santiago, Chile, 2015, p. 573-578.

[13] Phil Beecher. (2016, Mar.). *Wi-SUN Alliance - Interoperable Communications Solutions*. Available: https://www.wi-sun.org/images/assets/docs/20160315-Wi-SUN-interoperability-standards-r5.pdf. Accessed in: 05/16/2017.

[14] K.-H. Chang, B. Mason, "The IEEE 802.15.4g standard for smart metering utility networks," Proc. SmartGridComm 2012, pp.476–480, Nov. 2012.

[15] B. Lichtensteiger, B. Bjelajac, C. Mller and C. Wietfeld, "RF Mesh Systems for Smart Metering: System Architecture and Performance". In: *Proc. 1st IEEE Int. Conf. SmartGridComm*, pp. 379-384, 2010.

[16] LoRa Alliance. LoRaWAN™ Certified Products. Availible: https://www.lora-alliance.org/Products/Certified-Products. Accessed in: 05/19/2017.

[17] F. Adelantado. Xavier Vilajosana, Pere Tuset-Peiro, Borja Martinez, Joan Melià-Seguí and Thomas Watteyne, "Understanding the Limits of LoRaWAN". IEEE Communications Magazine. 2017, In press.

[18] LoRa Alliance. *LoRa Alliance – Wide Area Network for IoT*. Available: https://www.lora-alliance.org/. Accessed in: 05/17/2017.

[19] N. Sornin, M. Luis, T. Eirich, T. Kramp and O.Hersent. (2015, Jan). *LoRaWAN Specification v1.0*. Available: https://www.lora-alliance.org/portals/0/specs/LoRaWAN%20Specification%201R0.pdf Accessed in 05/16/2017.

[20] LoRa Alliance. *LoRa Technology*. Available: https://www.lora-alliance.org/What-Is-LoRa/Technology. Acessed in: 05/17/2017.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 578-586 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Optimizing the Decoding Complexity of PEG-Based Methods with an Improved Hybrid Iterative/Gaussian Elimination Decoding Algorithm

Reem Alkanhel[1*], Marcel Adrian Ambroze[2]

[1]*Department of Information Technology, Princess Nourah Bint Abdulrahman University, Riyadh, Saudi Arabia*

[2]*Department of Computing, Electronics and Mathematics, Plymouth University, United Kingdom*

ARTICLE INFO

Article history:
Received: 04 April, 2017
Accepted: 15 May, 2017
Online: 04 June, 2017

Keywords:
Hybrid decoding
Early stopping criterion
Group integrity
Missing tag recovery
Progressive edge-growth (PEG) methods

ABSTRACT

This paper focuses on optimizing the decoding complexity of the progressive-edge-growth-based (PEG-based) method for the extended grouping of radio frequency identification (RFID) tags using a hybrid iterative/Gaussian elimination decoding algorithm. To further reduce the decoding time, the hybrid decoding is improved by including an early stopping criterion to avoid unnecessary iterations of iterative decoding for undecodable blocks. Various simulations have been carried out to analyse and assess the performance achieved with the PEG-based method under the improved hybrid decoding, both in terms of missing recovery capabilities and decoding complexities. Simulation results are presented, demonstrating that the improved hybrid decoding achieves the optimal missing recovery capabilities of full Gaussian elimination decoding at a lower complexity, as some of the missing tag identifiers are recovered iteratively.

1. Introduction

Checking the integrity of groups containing radio frequency identification (RFID) tagged objects or recovering the tag identifiers of missing objects is important in many activities to find missing tagged objects. Autonomous verification of a group of RFID tags is an essential feature in RFID systems, because it ensures standalone operations as well as scalability, privacy and security. This can be achieved by treating a group of tag identifiers as packet symbols which are encoded and decoded as in binary erasure channels (BECs). This paper is an extension of work originally presented in [1], by optimizing the decoding complexity of the progressive-edge-growth-based (PEG-based) method for the extended grouping of RFID tags while achieving significant missing recovery enhancements compared to other extended grouping methods presented in [2–4].

Various decoding algorithms are devised to recover erasures over BECs, each with different performance and complexities. The most prominent decoding algorithms are maximum likelihood (ML) and iterative (IT) decoding algorithms. ML decoding for a BEC can be implemented as Gaussian elimination (GE), which provides the optimum decoding in terms of erasure recovery capabilities at the price of high decoding complexity [5]. On the other hand, IT decoding based on belief propagation features a linear decoding complexity, but it remains suboptimal in recovering erasures, as it is affected by the existence of short cycles [6, 7]. This suggests a hybrid (HB) decoding strategy which combines the above-mentioned IT with GE decoding algorithms in order to compromise between performance and complexity. References [8–12] proposed and implemented HB decoding algorithms in BECs with the basic idea of employing IT decoding first. If the IT decoding fails to recover all the erased symbols, the GE decoding is activated to complete the decoding process and resolve the simplified system. The GE decoding algorithm can fully recover the erasures in the simplified system if the columns of the decoding matrix are linearly independent. This paper focuses on optimizing the decoding complexity of the PEG-based method using an HB decoding algorithm. To further reduce the decoding time, the HB decoding is improved by including an early stopping criterion to avoid unnecessary iterations of iterative decoding for undecodable blocks. Simulation results are presented, demonstrating that the improved HB decoding achieves the optimal missing recovery capabilities of full GE decoding at a lower complexity, as some of the missing tag identifiers are recovered iteratively.

*Reem Alkanhel, Department of Information Technology, Princess Nourah Bint Abdulrahman University, Riyadh, Saudi Arabia, rialkanhal @pnu.edu.sa.

The remainder of this article is organized as follows. A brief overview of ML and IT decoding algorithms is provided in Section 2. Section 3 introduces the early termination of IT decoding, while Section 4 presents the improved HB decoding. Section 5 evaluates the performance achieved with the improved HB decoding using simulations. Conclusions and future work follow in Section 6.

2. Maximum Likelihood and Iterative Decoding Algorithms

2.1. Maximum Likelihood Decoding Algorithm

The ML decoding algorithm on the BEC involves solving a system of linear equations and finding the unknowns (erasures) in

$$TID_{k`}\ A_{k`} = TID_k\ A_k \quad (1)$$

where $TID_{k`}$ and $TID_k$ are the arrays of lost and known symbols (tag identifiers), respectively. Similarly, $A_{k`}$ is the matrix which contains the columns of A related to $TID_{k`}$, while $A_k$ is the matrix which contains the columns of $A$ related to $TID_k$. A system of linear equations has a unique solution, if and only if the columns of $A_{k`}$ have full rank (i.e. they are linearly independent). This ML decoding can be implemented as GE by reducing $A_{k`}$ to row echelon form. The process of the algorithm consists of two steps. The forward elimination step first reduces the system to an upper triangular form, which can be done using elementary row operations. The algorithm continues with a backward substitution step. It recursively resolves erased symbols until the solution is found. The last symbol of an upper triangular matrix is given the value of the symbol on the right-hand side. Then, this value is replaced in all the equations in which it is included. GE provides the optimum decoding in terms of erasure recovery capabilities at the price of the high decoding complexity of $O(n^3)$, where n is the number of variables [13]. GE is used to represent ML throughout this article.

2.2. Iterative Decoding Algorithm

IT decoding algorithms or message-passing algorithms are commonly used in BEC decoding [14]. IT decoding based on belief propagation features a linear decoding complexity of $O(n)$, but it remains suboptimal in recovering erasures over the BEC, as it is affected by the existence of short cycles [6, 7]. IT decoding algorithms consist of solving (1) by recursively exchanging messages along the edges of a Tanner graph between variable nodes and check nodes. All parity check equations having only one erased symbol are found, and the erased symbols are recovered. Then, new equations which have only one erased symbol may be created and used to recover further erased symbols. The decoding process is as follows:

1) Each variable node v sends the same message $M_v \in \{0, 1, e\}$, where e denotes the erased symbol, to each of its connected check nodes.
2) If a check node receives only one e, it replaces the erased symbol with the bitwise exclusive or XOR of the known symbols in its check equation.
3) Each check node c sends different messages $E_{c,v} \in \{0, 1, e\}$ to each of its connected variable nodes v. These messages represent the results of the previous step.
4) If the variable node of an erased symbol receives $E_{c,v} \in \{0,1\}$, the variable node updates its value to the value of $E_{c,v}$.

The IT decoding iterates the above process until all the erasure symbols are recovered successfully or until a predetermined maximum number of iterations (I_{max}) has passed without successful decoding [15].

3. Early Termination of Iterative Decoding

Early termination of IT decoding has been extensively researched within correcting errors introduced by channels. At each iteration, the IT decoding detects decodable blocks by checking parity check constraints. If all parity check constraints are satisfied, the decoding terminates. Otherwise, the decoding always completes $I_{\max}$ iterations for undecodable blocks. The decoding time and complexity increases linearly with the number of decoding iterations. To avoid unnecessary decoding iterations when processing undecodable blocks, a proper stopping criterion is required to terminate the decoding process early and therefore to save decoding time and power consumption [16]. Different early stopping criteria have been proposed for IT decoding, aiming to detect undecodable blocks and to stop the decoding process in its early stage. These criteria can be divided into two types. One type utilizes log-likelihood ratios (LLRs), for example the criterion in [17]. It identifies undecodable blocks using variable node reliability, which is the addition of the absolute values of all variable node LLRs. The decoding process is stopped if the variable node reliability remains unchanged or is decreased within two consecutive iterations. This is a consequence of the observation that a consistent increase of the variable node reliability is expected from a decodable block. The criterion is disabled if the variable node reliability is larger than a channel-dependent threshold. Similar criteria have been proposed in [18–20] which monitor the convergence of the mean magnitude of the LLRs at the end of each iteration to identify undecodable blocks. The other type of criteria are those which observe the numbers of satisfied or unsatisfied parity check constraints; for example the criteria in [21] and the improved criterion in [22] compare the numbers of satisfied parity check constraints in two consecutive iterations. If they are equal, the decoding process is halted. On the other hand, the criteria in [23, 24] use the number of unsatisfied parity check constraints to detect undecodable blocks.

4. Improved Hybrid Decoding Algorithm

In the HB decoding process, the decoding time of IT decoding for undecodable blocks becomes longer as the decoding iterates for $I_{\max}$. Figure 1 shows the iteration histogram as a function of the missing amounts experienced for two variants of group sizes, including 49 and 169 tags. I_{max} is set to 50 and 200, respectively. It is explicit that as the number of missing tags increases, iteration for $I_{\max}$ starts to occur due to the existence of undecodable blocks. This can clearly be seen in the region where IT decoding has failed, for example when recovering more than 15 and 70 missing tags from a group of 49 and 169, respectively. Therefore, the taking into account of an early stopping criterion, which adaptively adjusts the number of decoding iterations, arises as an attractive solution to avoid unnecessary iterations and thus to save decoding time. Since the main advantage of HB decoding is to reduce GE complexity, the frequency of GE decoding usage should be as small as possible,

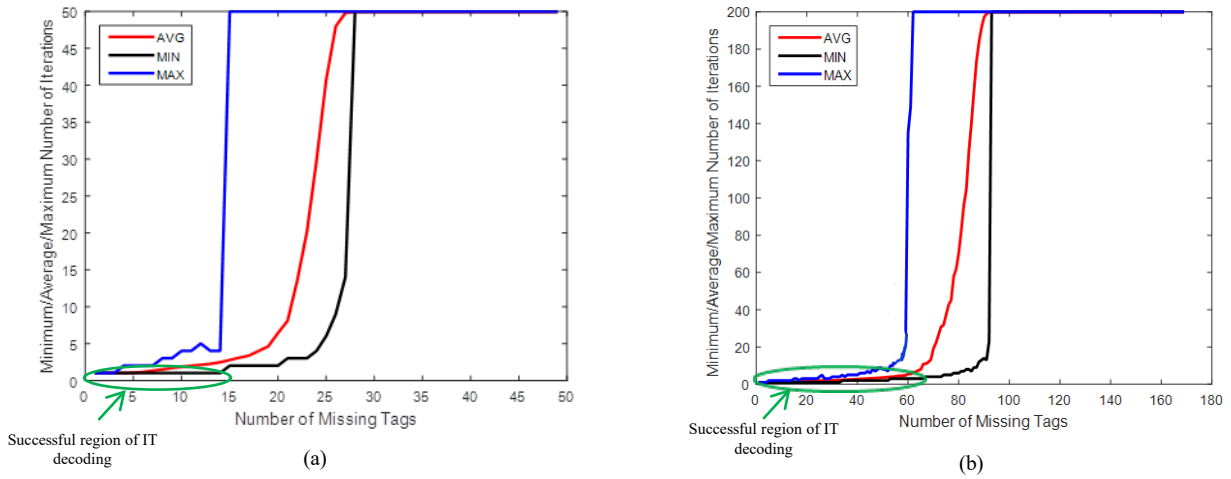


Figure 1. The iteration numbers required for decoding different group sizes: (a) $n = 49$; (b) $n = 169$ where $j = 3$ for the PEG-based method

even for a large number of missing tags. When GE decoding is utilized, the size of the simplified system should also be as small as possible. This suggests a stopping criterion for the PEG-based method which permits partial decoding of undecodable blocks up to a point where IT decoding cannot recover any missing tag identifiers. The proposed criterion can distinguish between decodable and undecodable blocks. It does not cause any computation overheads, as it utilizes the information available during the decoding process. Algorithm 1 outlines the improved HB decoding algorithm. The input is the received codeword $y = [y_1, y_2, \dots, y_n]$ of the transmitted codeword $x = [x_1, x_2, \dots, x_n]$ where $x_i \in \{0,1\}$ and $y_i \in \{0, 1, e\}$ for $i = 1 \dots n$. The output is $M = [M_1, \dots, M_n]$ where $M_i \in \{0, 1, e\}$. Let S_j denote the set of symbols in the j-th check equation of the code, n_e is the number of erasures at the input to the decoder, and $E_{j,i}$ is the outgoing message from the check node j to the variable node i. The improved algorithm is based on the observation of the checksum of erasure symbols n_e^I at the end of each iteration I. It halts the decoding process if the checksum remains unchanged or does not decrease within two consecutive decoding iterations, as in Step 24. This results from the observation that decodable blocks exhibit a consistent decrease in checksum values.

Algorithm 1 Improved Hybrid Decoding

1: $I = 0$
2: $n_e^{I-1} = n_e$
3: **for** $i = 1$ to n **do**
4: $M_i = y_i$
5: **end for**
6: **repeat**
7: **for** $j = 1$ **to** m **do**
8: **for** $i \in S_j$ **do**
9: **if** a check node j receives only one 'e' **then**
10: $E_{j,i} = \sum_{i' \in B_{j,i' \neq i}} (M_{i'} \mod 2)$
11: **else**
12: $E_{j,i} =$ 'e'
13: **end if**
14: **end for**
15: **end for**
16: **for** $i = 1$ to n **do**
17: **If** $M_i =$ 'e' **then**
18: **If** there is at least one $E_{j,i} = 0$ or 1 **then**
19: $M_i = E_{j,i}$
20: **end if**
21: **end if**
22: **end for**
23: $n_e^I =$ count 'e' in $M_i \ \forall\ i = 1 \dots n$
24: **if** ($M_i \neq$ 'e' $\forall\ i = 1 \dots n$) or ($I = I_{max}$) or ($n_e^{I-1} = n_e^I \dots$) **then**
25: Terminate IT decoding and start GE decoding
26: **else**
27: Increment I
28: $n_e^{I-1} = n_e^I$
29: **end if**
30: **until** terminated

5. Simulation Results

This section gives an account of the various simulations which have been carried out to analyse and assess the performance of the PEG-based method under the improved HB decoding, both in terms of missing recovery capabilities and decoding complexities.

5.1. Missing Recovery Capability Analysis

To illustrate the performance of the improved HB decoding in recovering missing tags, Figure 2 shows the average error rate as a function of the number of missing tags for different values of group sizes n where the column weight j varies between 3 and 4. It can be clearly shown that increasing j greatly increases the performance of the improved HB decoding in recovering missing tags. The main reason for this behaviour is that a large value of j produces a large number of rows and thus increases the rank of the matrix. For example, if five tags are missing from a group of n = 25, all their identifiers can be recovered with at least j = 3 under the improved HB decoding. However, if ten identifiers are missing, the column weight needs to be adjusted to 4 in order to achieve 100% reliable recovery (0% error rate). Table 1 compares the missing recovery capabilities of the simulated PEG-based method under the improved HB decoding algorithm with other extended grouping methods presented in [2–4]. It is clear that the improved algorithm achieves significant missing recovery enhancements compared to other extended grouping methods. It can be concluded from the results that the improved HB decoding achieves the optimal missing recovery capabilities of full GE decoding, since the remaining systems of IT decoding containing short cycles are solved with GE decoding. The results also reveal that the recovery performance of the improved HB decoding increases by increasing the group size. This results from the fact that the PEG-based method accounts for short cycles in Tanner graphs by maximizing the local girth, i.e. the length of the shortest cycle, at variable nodes, which is more achievable when the group size is large [1]. Applying the improved HB decoding to groups of 25 and 49 tags, for example, results in the recovery of missing tags of up to 48% and 65%, respectively, of the group size, with a 0% error rate. On the other hand, for groups of 121 and 169 tags, the recovery capabilities increase to 72% and 74.5%, respectively.

5.2. Decoding Complexity Analysis

The relative complexities of the improved HB decoding are analysed in terms of decoding time. The decoding time is measured on a 2 GHz Intel i7 processor. The entire decoding complexity is equal to the complexity of IT decoding plus that of GE decoding if short cycles exist. This is mainly determined by the efficiency of the first stage of IT decoding; the better the performance of the IT decoding, the less the complexity and thus the time of the improved HB decoding. During the decoding process, IT decoding is first applied on the parity check matrix. If the IT decoding fails, GE decoding is activated to solve the remaining systems with a complexity less than the complexity of decoding the original matrix.

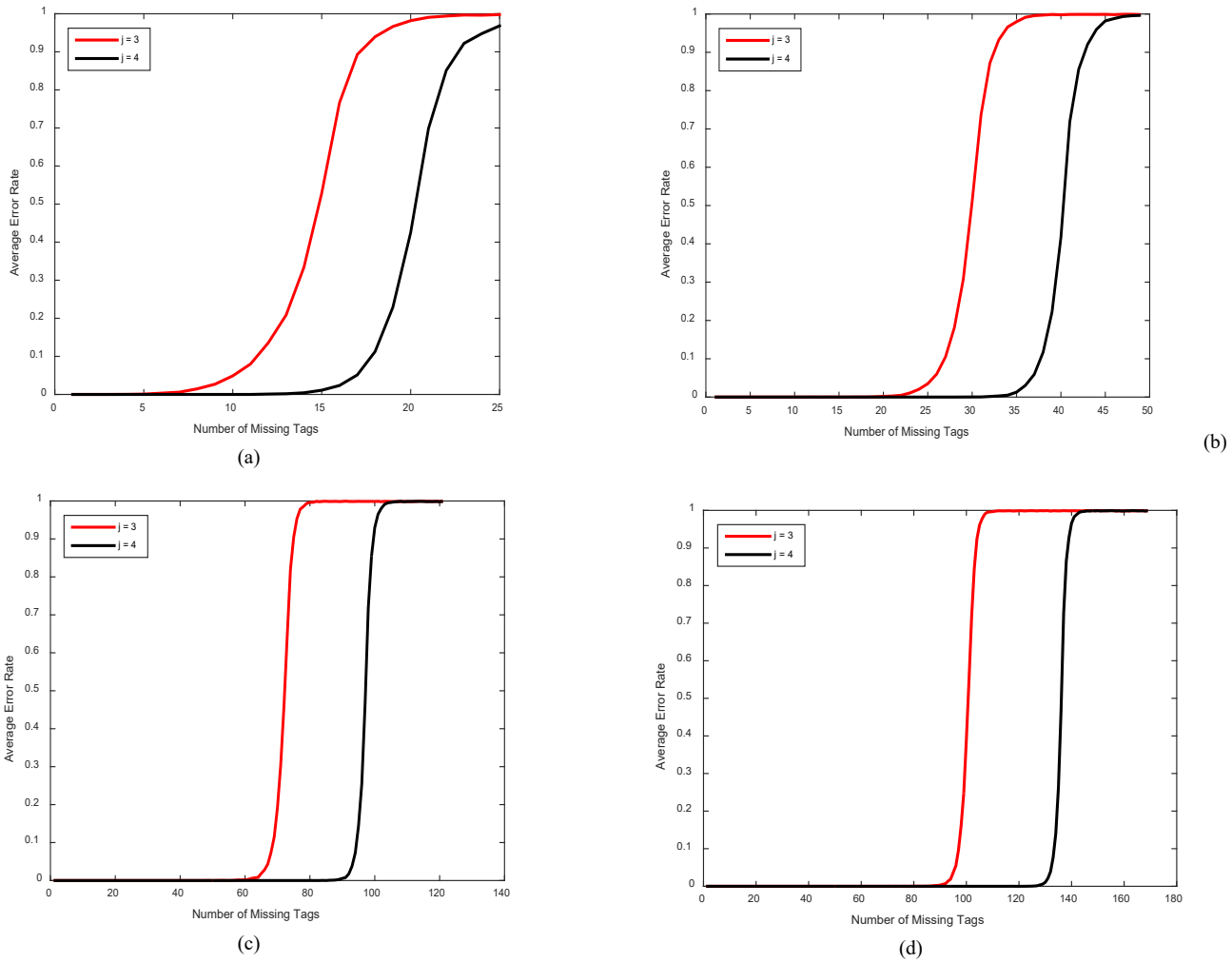


Figure 2. The average error rate for different group sizes: (a) n = 25; (b) n = 49; (c) n = 121; (d) n = 169 using the PEG-based method under the improved HB decoding

Table 1. Comparisons of missing recovery capabilities of the PEG-based method under the improved HB decoding algorithms with other extended grouping methods

Parameter	Methods	Largest number of missing identifiers which can be recovered within 0% error rate	Parameter	Methods	Largest number of missing identifiers which can be recovered within 0% error rate
(25, 3, 5)	[2]	3	(121, 3, 5)	[2]	1
	[3]	5		[3]	5
	[4]	5		[4]	5
	PEG-based method under GE decoding	6		PEG-based method under GE decoding	64
	PEG-based method under HB decoding	6		PEG-based method under HB decoding	64
(25, 4, 5)	[2]	5	(121, 4, 5)	[2]	6
	[3]	7		[3]	13
	[4]	9		[4]	15
	PEG-based method under GE decoding	12		PEG-based method under GE decoding	87
	PEG-based method under HB decoding	12		PEG-based method under HB decoding	87
(49, 3, 5)	[2]	1	(169, 3, 5)	[2]	1
	[3]	5		[3]	5
	[4]	5		[4]	5
	PEG-based method under GE decoding	23		PEG-based method under GE decoding	93
	PEG-based method under HB decoding	23		PEG-based method under HB decoding	93
(49, 4, 5)	[2]	3	(169, 4, 5)	[2]	1
	[3]	7		[3]	9
	[4]	8		[4]	16
	PEG-based method under GE decoding	32		PEG-based method under GE decoding	126
	PEG-based method under HB decoding	32		PEG-based method under HB decoding	126

Two experiments are conducted:

1) Decoding time measurements when IT decoding fails and thus GE is required. This is the worst case from a decoding point of view.

2) Decoding time measurements as a function of the number of missing tags to compare two cases, one where IT decoding is successful and the other where IT decoding fails and GE is needed.

5.2.1. Decoding Time when Gaussian Elimination is Required (Worst Case)

Figure 3 illustrates the decoding time during the GE process of the improved HB decoding for a group of 49 and 169 tags where $j = 3$. By changing the number of missing tags, the resulting average simplified system size changes, which affects the decoding time. The figure illustrates this time as a function of the simplified system size and the related histogram. The region where IT decoding recovers some missing tag identifiers is only considered in this analysis. However, in the second region, where IT decoding cannot recover any missing tag identifiers, GE decoding time is equal to that of full systems. The improved algorithm efficiently decreases the average decoding time of GE decoding, because the complex GE decoding is only utilized when required over a simplified system resulting from IT decoding. For a group of 49 tags, the histogram demonstrates that the maximum decoding time remains below 2 ms related to an average decoding time of 1.8 ms. When the group size increases to 169 tags, the GE decoding time also increases, following a $O(n^3)$ law. It can clearly be seen that the maximum decoding time remains below 13 ms related to an average decoding time around 12 ms. In contrast, GE decoding on full systems, as illustrated in Figures 5 and 6 (presented in the next section), is slower, especially for a large group of tags. For a group of size 49 and 169 where $j = 3$, it is evident that the decoding time lasts on average 2 and 15 ms, respectively.

To illustrate the impact of the proposed early stopping criterion on decoding time, Figure 4 shows iteration numbers which are taken for various missing amounts to be compared to those of the conventional IT decoding depicted in Figure 1. One can observe that the proposed stopping criterion greatly reduces the average number of iterations in the region where IT decoding is not successful. Figure 4 shows that undecodable blocks need at most 17 and 33 iterations, respectively, when recovering more than 15 and 70 missing tags from a group of 49 and 169.

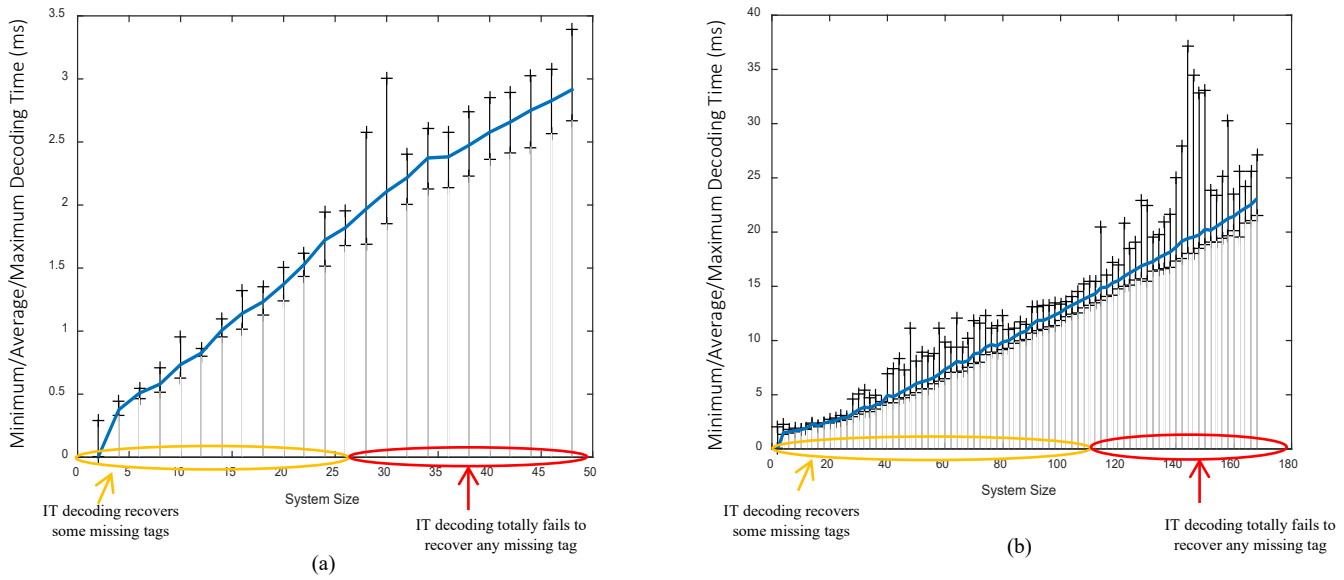


Figure 3. Histogram of GE decoding time with respect to the simplified system size for different group sizes: (a) $n = 49$; (b) $n = 169$ where $j = 3$ for the PEG-based method

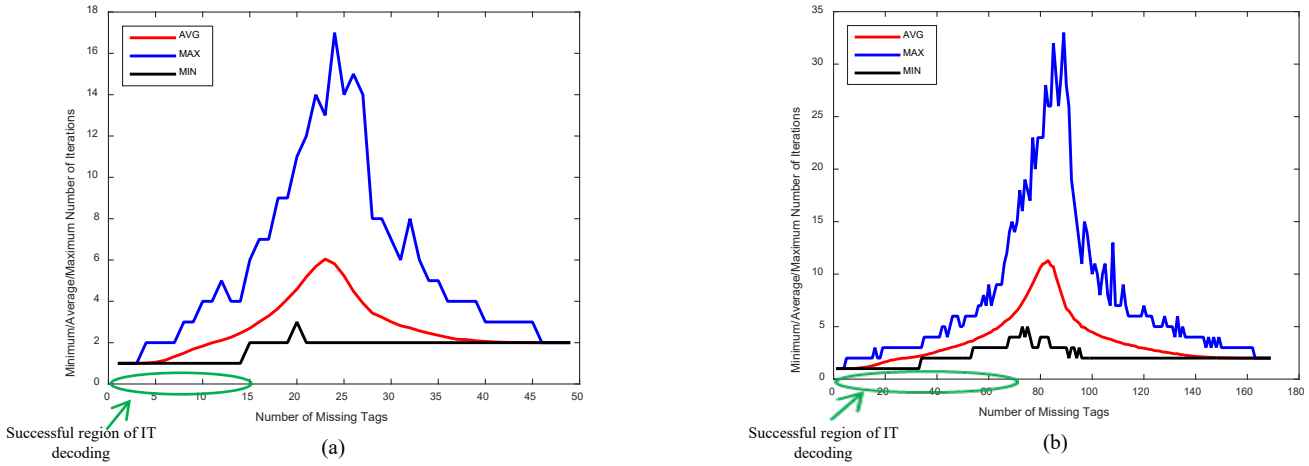
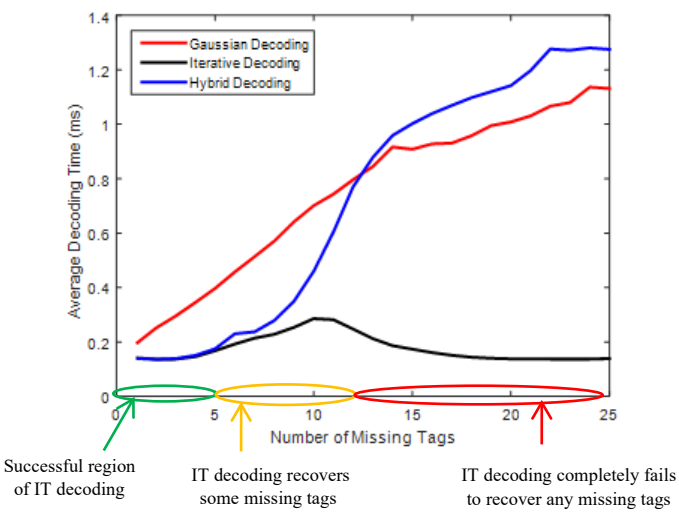


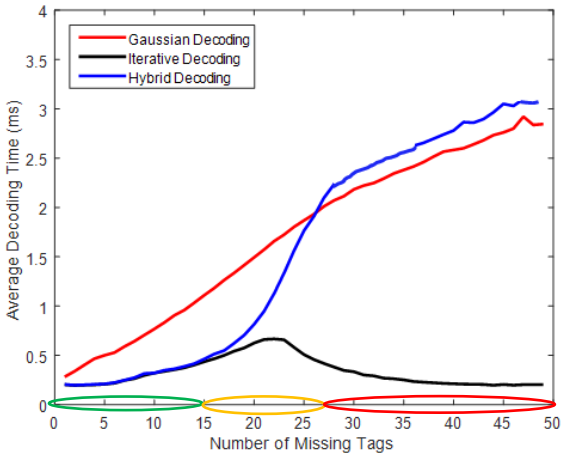
Figure 4. The enhanced iteration numbers required for decoding different group sizes: (a) $n = 49$; (b) $n = 169$ where $j = 3$ for the PEG-based method

5.2.2. *Decoding Time of the Improved Hybrid Decoding*

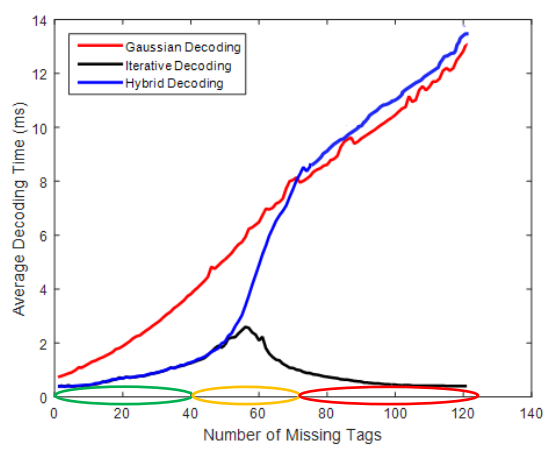
Figures 5 and 6 show the average decoding time as a function of the number of missing tags for different values of n where j is equal to 3 and 4 under the IT, GE and improved HB decoding algorithms. It is obvious that the decoding time depends on the missing amounts experienced. With small numbers of missing tags, the decoding time of the improved HB decoding is fast and is similar to that of IT decoding. This is clear in a region where IT decoding is successful. As the number of missing tags increases, the decoding time of the improved HB decoding also progressively increases, because GE decoding is more and more often required. This can be seen in the curves: after the successful region of IT decoding, the average decoding time of HB decoding increases progressively, since IT decoding turns out to be ineffective, and the size of the simplified system increases accordingly. Then, above a certain threshold which depends on the missing recovery capabilities of IT decoding, the decoding time of the improved HB decoding gradually approaches the full GE decoding. Since the entire decoding time of the improved HB decoding is equal to the decoding time of IT decoding plus that of GE decoding, the improved HB decoding algorithm can easily be tailored to satisfy operational requirements. More precisely, the IT decoding phase can be avoided when the aforementioned condition is met, to save HB decoding time. It is clear that the n and j parameters also have major impacts on the decoding time. As n or j increase, the improved HB decoding time increases too. This is due to specific phenomena. The complexity of IT decoding depends on the number of XOR operations, which is determined by the number of variables in each equation, whereas the complexity of the GE decoding depends on the number of linear equations and the number of variables. Therefore, by increasing n or j, the number of operations required by the IT and GE decoding algorithms is increased accordingly. If decoding time is an issue with a large n value, choosing a smaller j value is possible. If the HB decoding time is compared with the GE decoding time for



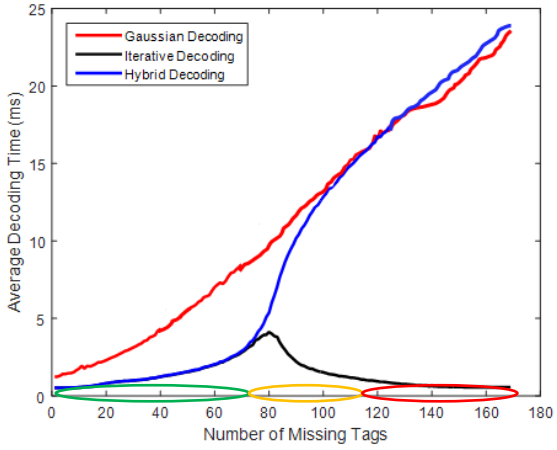
(a)



(b)

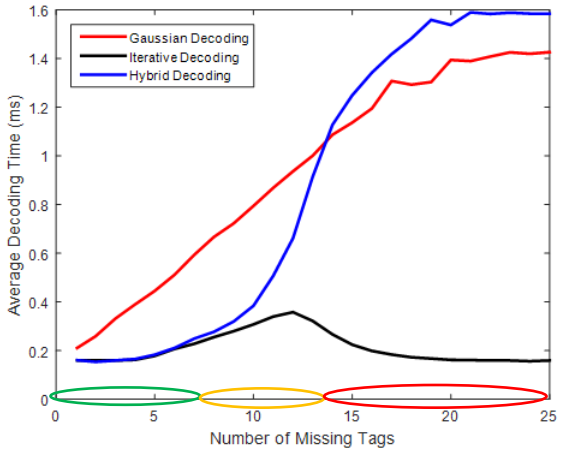


(c)

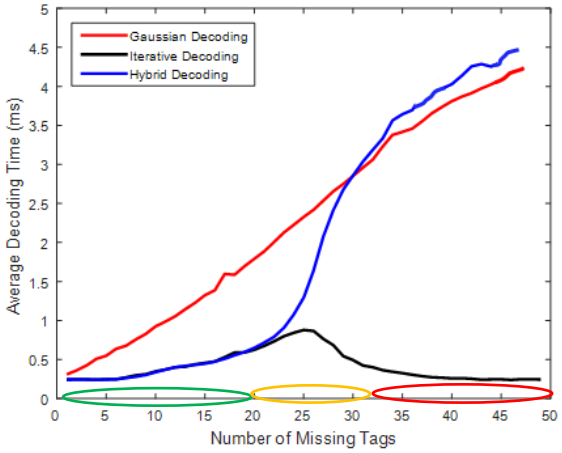


(d)

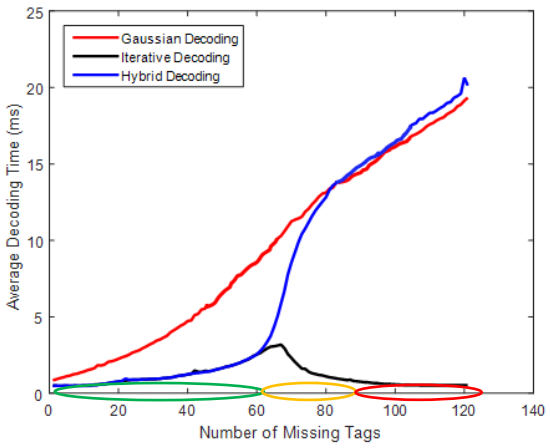
Figure 5. The average decoding time for recovering missing tag identifiers from different group sizes: (a) $n = 25$; (b) $n = 49$; (c) $n = 121$; (d) $n = 169$ where $j = 3$ for the PEG-based method under the IT, GE and improved HB decoding algorithms



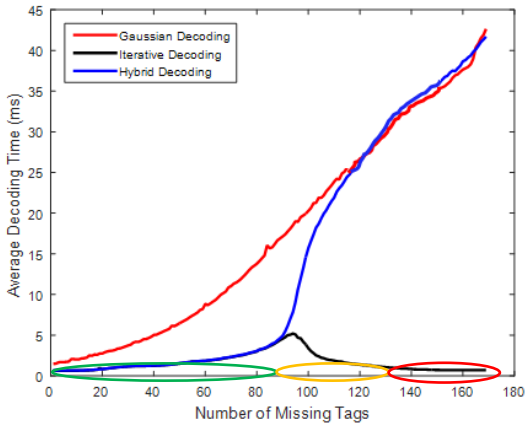
(a)



(b)



(c)



(d)

Figure 6. The average decoding time for recovering missing tag identifiers from different group sizes: (a) $n = 25$; (b) $n = 49$; (c) $n = 121$; (d) $n = 169$ where $j = 4$ for the PEG-based method under the IT, GE and improved HB decoding algorithms

the full system, it can be concluded that the HB decoding is faster for small to medium numbers of missing tags even when GE decoding is required.

6. Conclusions and Future Work

This paper has focused on optimizing the decoding complexity of the PEG-based method using a hybrid iterative/Gaussian elimination decoding algorithm. The hybrid decoding is improved by including an early stopping criterion to avoid unnecessary iterations of iterative decoding for undecodable blocks and thus saving decoding time. Simulation results have been presented showing that the improved hybrid decoding algorithm achieves the optimal missing recovery capabilities of full Gaussian elimination decoding, since the remaining systems of iterative decoding containing short cycles are solved with Gaussian elimination decoding. This recovery performance significantly increases with an increase in the group size of tags. It has been shown that the proposed stopping criterion greatly reduces the average number of decoding iterations for undecodable blocks and therefore reduces decoding time when compared to the conventional IT decoding algorithm. The improved algorithm is very efficient in decreasing the average decoding time of Gaussian elimination decoding for small to medium amounts of missing tags up to a certain threshold which depends on the missing recovery capabilities of IT decoding. This makes the improved hybrid decoding a promising candidate for decoding the PEG-based method for extended the grouping of RFID tags.

For future work, more investigation is needed into the joint use of the two decoding algorithms according to missing amounts and group sizes.

References

[1] R. Alkanhel, M. Ambroze, "Extended Grouping of RFID Tags Based on Progressive Edge-Growth Methods" in 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA), Ras Al Khaimah, UAE, 2016.

[2] Y. Sato, Y. Igarashi, J. Mitsugi, O. Nakamura, J. Murai, "Identification of Missing Objects with Group Coding of RF Tags" in 2012 IEEE International Conference on RFID, Orlando, USA, 95–101, 2012.

[3] Y. Su, J. Lin, and O. K. Tonguz, "Grouping of RFID tags via strongly selective families" IEEE Commun. Lett., **17**(2), 1–4, 2013.

[4] Y. Su, O. K. Tonguz, "Using the Chinese remainder theorem for the grouping of RFID tags," IEEE Trans. Commun., **61**(11), 4741–4753, 2013.

[5] G. Liva, B. Matuz, E. Paolni, M. Chiani, "Achieving a near-optimum erasure correction performance with low-complexity LDPC Codes" Int. J. Satell. Commun. N., **28**(5–6), 291–315, 2010.

[6] R. Gallager, "Low density parity check codes," IEEE Trans. Inf. Theory, **8**(1), 21–28, 1962.

[7] T. Richardson, R. Urbanke, "Efficient encoding of low-density parity-check codes" IEEE Trans. Commun., **47**(2), 638–656, 2001.

[8] M. Cunche, V. Roca, "Improving the decoding of LDPC codes for the packet erasure channel with a hybrid Zyablov iterative decoding/Gaussian elimination scheme" INRIA Research Report, 1–19, 2008.

[9] M. Cunche, V. Roca, "Optimizing the Error Recovery Capabilities of LDPC-Staircase Codes Featuring a Gaussian Elimination Decoding Scheme" in 10th IEEE International Workshop on Signal Processing for Space Communications (SPSC), Rhodes Island, Greece, 2008.

[10] E. Paolini, G. Liva, B. Matuz, M. Chiani, "Generalized IRA erasure correcting codes for hybrid iterative/maximum likelihood decoding" IEEE Commun. Lett., 12(6), 450–452, 2008.

[11] L. Yang, M. Ambroze, M. Tomlinson, "Decoding turbo Gallager codes for the erasure channel" IEEE Commun. Lett., 1(2), 1–3, 2008.

[12] W. Huang, S. Member, H. Li, J. Dill, "Fountain Codes with Message Passing and Maximum Likelihood Decoding over Erasure Channels" in IEEE Wireless Telecommunication Symposium, New York City, USA, 1–5, 2011.

[13] M. Cunche, V. Saviny, V. Roca, "Analysis of Quasi-Cyclic LDPC codes under ML Decoding over the Erasure Channel" in IEEE International Symposium on Information Theory and its Applications (ISITA), Taichung, Taiwan, 2010.

[14] W. Ryan, S. Lin, Channel Codes Classical and Modern, Cambridge University Press, 2009.

[15] M. Luby, M. Mitzenmacher, M. Shokrollahi, D. Spielman, "Efficient erasure correcting codes" IEEE Trans. Inf. Theory, 47(2), 569–584, 2001.

[16] T. Xia, H. Wu, H. Jiang, "New stopping criterion for fast low-density parity-check decoders" IEEE Commun. Lett., 18(10), 1679–1682, 2014.

[17] F. Kienle, N. Wehn, "Low Complexity Stopping Criterion for LDPC Code Decoders" in 61st IEEE Vehicular Technology Conference, Stockholm, Sweden, 25–28, 2005.

[18] J. Li, X. You, J. Li, "Early stopping for LDPC decoding: convergence of mean magnitude (CMM)" IEEE Commun. Lett., 10(9), 667–669, 2006.

[19] Z. Cai, J. Hao, L. Wang, "An Efficient Early Stopping Scheme for LDPC Decoding" in 13th NASA Symposium on VLSI Design, Post Falls, Idaho, 1325–1329, 2008.

[20] T. Chen, "An Early Stopping Criterion for LDPC Decoding Based on Average Weighted Reliability Measure" in Cross Strait Quad-Regional Radio Science and Wireless Technology Conference, Harbin, China, 123–126, 2012.

[21] D. Shin, K. Heo, S. Oh, J. Ha, "A Stopping Criterion for Low-Density Parity-Check Codes" in 65th IEEE Vehicular Technology Conference, Dublin, Ireland, 1529–1533, 2007.

[22] D. Alleyne, J. Sodha, "On Stopping Criteria for Low-Density Parity-Check Codes" in 6th International Symposium on Communication Systems, Networks and Digital Signal Processing (CSNDSP), Graz, Austria, 633–637, 2008.

[23] G. Han, X. Liu, "A unified early stopping criterion for binary and nonbinary LDPC codes based on check-sum variation patterns" IEEE Commun. Lett., **14**(11), 1053–1055, 2010.

[24] T. Mohsenin, H. Shirani-Mehr, B. Baas, "Low Power LDPC Decoder with Efficient Stopping Scheme for Undecodable Blocks" in IEEE International Symposium on Circuits and Systems, Rio de Janeiro, Brazil, 1780–1783, 2011.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 587-597 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

FAG - a management support tool with BI techniques to assist teachers in the virtual learning environment Moodle

Luciana Silva Zapparolli[1*], Itana Stiubiener[2]

[1]*Faculdade de Tecnologia, Informática, Fatec Mauá e Zona Sul, Brasil*

[2]*Universidade Federal do ABC, Centro de Matemática, Computação e Cognição, UFABC, Brasil*

ARTICLE INFO

Article history:
Received : 05 April, 2017
Accepted : 15 May, 2017
Online: 04 June, 2017

Keywords:
Business Intelligence
Learning Analytics
Monitoring in virtual environments

ABSTRACT

One of the great challenges in distance-learning is to follow the actions of the teachers/tutors and also the actions of students during the process of teaching and learning. This article presents the FAG Tool integrated with the LMS Moodle was developed to help managers of the Distance Education environment to monitor the actions of teachers/tutors and also teachers/tutors in the follow-up of student actions. Through the techniques of Business Intelligence (BI) and Learning Analytics (LA), the tool generates analytical reports and dashboards, presenting a holistic and transversal view, being this vision the differential of this tool. The use of FAG allows teachers/tutors to monitor the participation of all their students in all virtual rooms under their responsibility and thus take corrective measures in the teaching and learning process, such as reducing the risk of avoidance. For the managers, it can be considered as a support tool for decision making regarding the faculty, maintaining or not the teacher/tutor in the process of teaching and learning or even be a base to enlarge or reduce their classes depending on their performance in the virtual environment. Through the use of the FAG, this decision-making can happen during the teaching and learning process and not only after the end, as is usual, because the reports are easy to understand and present accurate information in time to ensure the success of the teaching and learning process.

1. Introduction

This publication is an extension of the article entitled "Management support tool in virtual leaming environments using MoodIe as a case study" that is published in 2016 XI Latin American Conference on Learning Objects and Technology (LACLO)[1].

Distance education is a widely used teaching modality in the teaching and learning process [2]. In this modality of teaching, there are several actors, and the teacher is responsible for motivating and stimulating the learning process in order to transform the information into knowledge by the students, who are considered active subjects in this process [3]. The tutor is responsible for guiding the student, explaining and clarifying questions related to the subject and participating in the evaluation activities [4] and the distance-learning environment and process manager is responsible for planning, organizing, directing and controlling [5]. In this work the role of a teacher and a tutor will be considered as a single role, called, teacher/tutor, since not necessarily a course needs a teacher and a tutor. This decision is related to three situations: digital fluency of a teacher, an excessive number of students enrolled and teachers availability of time to carry out and follow the course.

Nothing prevents a teacher from acting as a tutor in online courses [6]. In order to guarantee better results in the process of distance teaching and learning, one of the needs is the monitoring process of teachers/tutors' actions, often done by a manager of many or all courses of an institution. Usually, the management reports provided by an LMS are confusing, offering scattered data that do not allow a global view of the observed processes, thus hindering efficient management and decision-making [7]. Among the several existing LMSs Moodle will be used, once it is an LMS widely used in thousands of institutions in the world and consequently a

*Corresponding Author: Luciana Silva Zapparolli, Fatec de Mauá e Zona Sul, Informática, Brasil, +55 11 45433221; Email: luciana.zapparolli@fatec.sp.gov.br

very safe and very tested system. Moodle offers Learning Analytics (LA) tools that can be defined as a set of tools geared to the measurement, collection, analysis and reporting of student data in order to improve the teaching and learning process through analysis of generated data. These tools generate analytical reports and/or graphs of specific resources as classrooms or participants. In this way, the importance of this work is the identification of an auxiliary tool that, integrated to Moodle LMS, offers a better follow-up of the students' participation, besides an efficient monitoring of the actions of the teachers/tutors in relation to all the activities and contexts in which they work.

Section 2 presents concepts related to the knowledge area of work, such as Business Intelligence (BI), Learning Analytics (LA), some Moodle tools associated with LA, as well as the Pentaho BI and Qlik Sense Cloud tools, which will be used in the case study. Section 3 will present the methodology and case study. Section 4 presents the tool developed to solve the problem detected in the case study and partial result. In section 5 the conclusion will be presented and in section 6 future works will be proposed. Finally, bibliographical references are presented.

2. Concepts referring to the knowledge area of work

2.1. Business Intelligence (BI)

BI is a tool that assists in the selection, presentation and analysis of data, is a set of processes, technologies and tools necessary to transform data into information, information into knowledge [8].

Throughout their existence, organizations generate and store a large amount of data related to the activities performed, thus making it necessary to apply BI techniques in the strategic decision-making process, since mistaken decisions can jeopardize the future of an organization. In this scenario, there are also educational institutions, which need to be able to subsidize the optimization of their decision-making processes [9].

2.2. Business Intelligence Architecture Components

They are components that aid in the visualization and manipulation of large amounts of data:

- Data warehouse (DW): is defined as a database that serves as a repository of an organization built for the purpose of generating reports and business analysis [10].

- ETL (Extract Transform Load): The ETL process consists of reading the data from one or more databases, the process of which is called the extraction, after the extraction has the conversion of the extracted data to the form that is required, in order to be loaded on a date warehouse or other database, called transformation and finally the loading of the data in the data warehouse [11]. The ETL feeds DW and BI, and if this process does not work properly, the BI will not have the necessary information [12].

- OLAP (On-Line Analytical Processing): OLAP tools enable online generation of information in the form of reports, rankings, ad-hoc queries (uniquely mounted by the user), etc. [13].

The OLAP technology allows making an instance of the data in a multidimensional view, thus allowing the presentation of the information in different perspectives [19]. OLAP information is stored in subsets called information cubes, updated periodically from the DW, according to Figure 1.

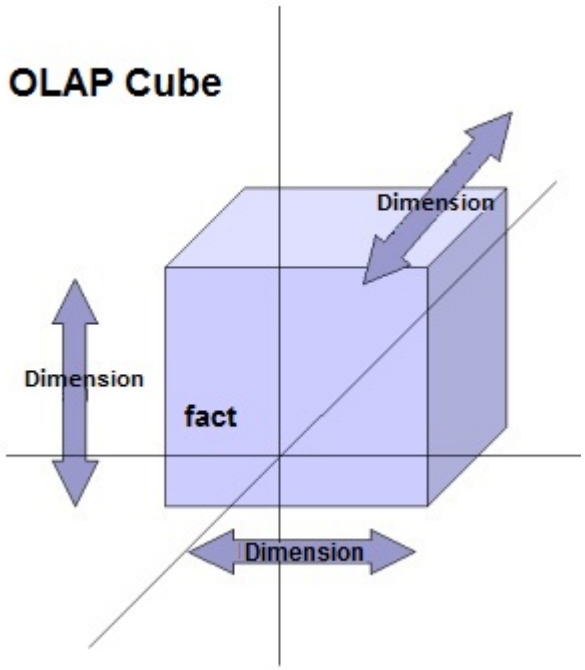


Figure 1. OLAP cubes

2.3. Business Intelligence in Learning Management System (LMS)

The concept of Business has changed. With the adoption of LMS in the learning support, the volume of information to be managed by teachers/tutors and also by the managers of LMSs systems increases, which makes it difficult to manage in these environments [14]. A derivation of the principles of BI used in information systems applied to the educational field, called Learning Analytics [15].

2.4. Learning Analytics (LA)

The term Learning Analytics appeared for the first time in 2009 in The Horizon Report. It is a field of research that is growing and presents several tools implemented in educational environments aim to offer educators resources so that they recognize and evaluate their educational process [17-18].

The main objective of the learning analysis is to generate information for those who create learning situations, such as teachers/tutors, managers and educational institutions [8]. It is a powerful way to inform students, teachers/tutors and managers about student performance and the progress of the learning process [19].

2.5. Learning Analytics (LA) in Moodle

Among the tools with LA resources offered by the Moodle can be cited:

- GISMO: consists in a visualization tool used to analyze the learning process of students comprehensively, which gives a view of the whole class and not just a specific student or a specific resource. It is useful to teachers/tutors.

- MOCLog: it is a set of tools used to analyze and present Moodle data. The development of this tool was based on GISMO,

through this tool users have access to summary reports of interactions related to actions in resources and tools, such as questionnaires [16].

- Learning Analytics Enhanced Rubric (LAe-R): consists a plugin created for Moodle LMS being an advanced classification method used for the careful evaluation. Contains some related criteria and classification levels associated with data from the analysis of student interaction and learning behavior in a course such as access time to learning materials, message numbers, and others [17].

In addition to the tools presented there are other ways to see the trends, analysis, and data on the platform from Moodle 2.8, among them:

- Event monitoring: Issue notifications to administrators and teachers/tutors when certain events occur in Moodle;

- General statistics: this is a plugin that generates site graphics and course reports, including user login, countries, preferred languages, number of courses per category, number of courses per size and registered users;

- Comprehensive report (site-wide Report): they are accessed by administrators are reports of the entire site;

- Compromise Analysis: The Engagement Analytics module provides information about student progress in the course, choosing activities that will be monitored by Moodle;

- Activities (Logs): These are reports that can be viewed by administrators;

- Forums Reporting Charts (Report Graph Forum): Displays the interactions in a single Forum activity and creates a targeted graph;

Through the tools with LA resources offered by Moodle it is not possible to extract analytical and consolidated reports containing information about all the rooms under the responsibility of a certain teacher / tutor, which makes it extremely difficult for the EaD manager to group information from room to room, considering That, most of the time, a teacher/tutor participates in several rooms simultaneously.

2.6. Pentaho BI

The Pentaho Open BI platform is a suite of open source applications for creating BI solutions, distributed through the open source Pentaho Public License (PPL). The Pentaho suite is one of the most used and reputed in the market [18].

The most important Pentaho modules that will be used in this work are:

- Pentaho Data Integration (PDI): one of the most powerful components responsible for the ETL process, is a collection of tools designed to fill the Data Warehouse with data [22].

- Platform Pentaho User Console (BiServer): an analysis platform that will run under a Tomcat server in a MySQL database, Linux environment.

- Pentaho Report Designer: tool that allows you to generate reports in a consolidated and analytical way.

2.7. Qlik Sense Cloud

The Qlik Sense Cloud platform focuses on self-service, lets you quickly create views.

The tool does not work with inflexible cubes, it creates a large file with all associations, so any table can be "fact" or "dimension" and all relevant data is available in RAM, a technology known as AQL (Associative Query Logic), Different from OLAP, where it is necessary to define the necessary dimensions and metrics [20].

Qlik Sense Cloud lets you share Qlik Sense applications in the cloud for free. It is a Software as a Service (SaaS) service or program as a service[20].

The great differential of this tool, besides the ease of generating graphical reports is that it adapts to tablet and smartphone[21].

3. Methodology and case study

For the development of the work the following steps were performed:

- Extraction of real data of Moodle, referring to the disciplines of four undergraduate courses that occurred in the second half of 2015, being, Human Resources Management, Management Process Management, Administration and Accounting Sciences. Statistics of the interactions of teachers/tutors with students in the tools were presented: forum and questionnaire, as well as student participation, through a case study.

- Analysis of the results obtained in order to verify the problems pointed out by the bibliographic research.

- Implementation of the tool using the platforms Pentaho BI and Qlik Sense Cloud.

The name of the teacher/tutor will be omitted, and in the case study, it will be referenced as tutor "A".

The data were obtained from the three disciplines, under the responsibility of tutor "A", named: MKP - Personal Marketing, MTA - Methods and Techniques of Applied Research and PO - Organizational Psychology. These disciplines were chosen by totaling a larger number of registered students.

3.1. Profile of tutor "A" in the courses and disciplines analyzed

The responsibilities of teachers/tutors in this environment:

- Track students' daily access,

- Attend the Doubts at Forum within 24 hours of posting,

- Post informational and alert messages in the News Forum,

- Prepare and distribute the programmatic contents of the course according to the specific program specified in the course plan, respecting the academic calendar of the institution,

- Prepare and publish activities that can be:
  - o Questionnaires,
  - o Evaluation forums,
  - o Online types or file uploads,
- Give feedback on activities,
- Prepare and correct the evaluations.

In addition, teachers/tutors are encouraged to motivate the participation of students by sending weekly messages, alerting them to the timing of the proposed activities and orienting them in the studies. The more classes the teacher/tutor has under their responsibility, the more time they will have to spend to fulfill those responsibilities.

3.2. Follow-up of student participation and tutor actions "A" in the subjects MKP, MTA and PO

Table 1 presents an overview of student participation (access) in the three virtual rooms, under the responsibility of tutor "A". Students who, at least once, accessed the virtual room in the last month of the course in December 2015 were considered active ".

Table 1. Participation of students (accesses) in virtual rooms

Discipline	Students				
	Total amount	Active amount	(%)	Never attended the course amount	(%)
MKT	34	27	79,4%	0	0,0%
MTA	188	172	91,5%	11	5,8%
PO	95	88	92,6%	3	3,1%

Figure 2 presents the analytical report offered by Moodle, referring to student accesses, in a specific room where the teacher/ tutor acts.

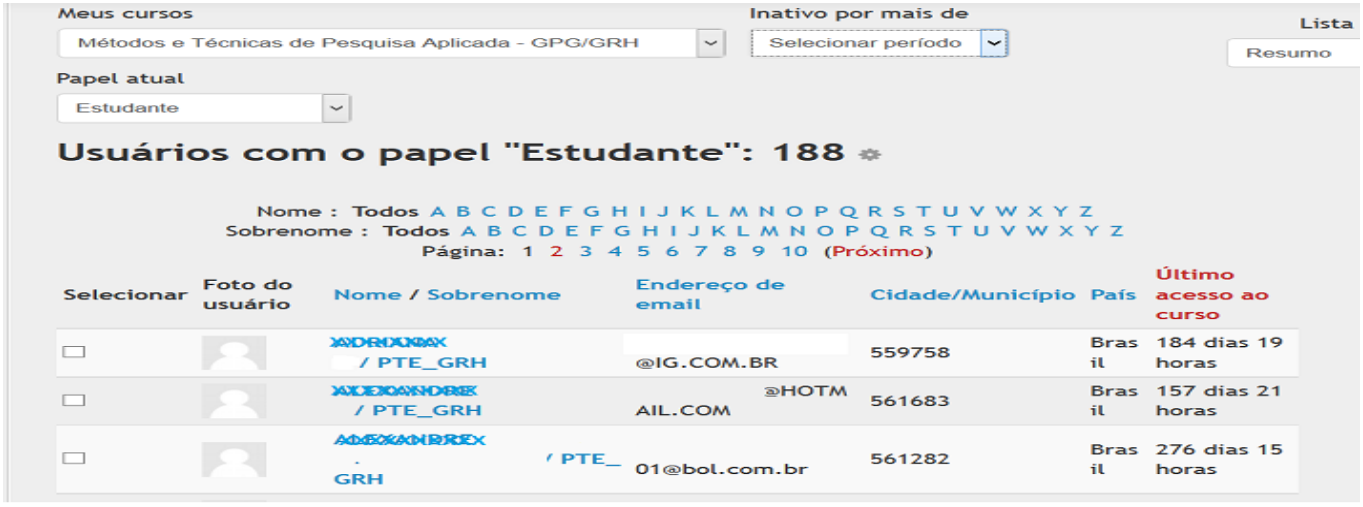


Figure 2. Students access

This report is offered by Moodle LMS, however, the teacher/ tutor must access each virtual room under his responsibility, issue the report and check the last access of the students registered in it. There is no cross-sectional view.

Table 2 presents an overview of the participation of students in the Forum of News, in the three virtual rooms, the responsibility of tutor "A".

Table 2. Participation of students in the News Forum

Discipline	Students		
	Total	Access	(%)
MKT	34	17	50,0%
MTA	188	38	20,2%
PO	95	21	22,1%

Table 3 presents an overview of students' participation in the Doubt Forum, in the three virtual rooms, under the responsibility of tutor "A".

Table 3. Student participation in the Doubts Forum

Discipline	Student		
	Total	Access	(%)
MKT	34	23	64,6%
MTA	188	62	32,9%
PO	95	33	34,7%

Figure 3 presents the analytical report regarding the accesses in the News Forum by students in a specific room that the teacher /tutor acts.

The report extracted from Moodle regarding student access in the Doubt Forum is identical to the one shown in Figure 3.

It is important to note that all participants of the course receive the messages that are posted in the forums also by e-mail and, since it is not possible to count the accesses to the e-mail messages, only the accesses in the virtual classrooms were considered.

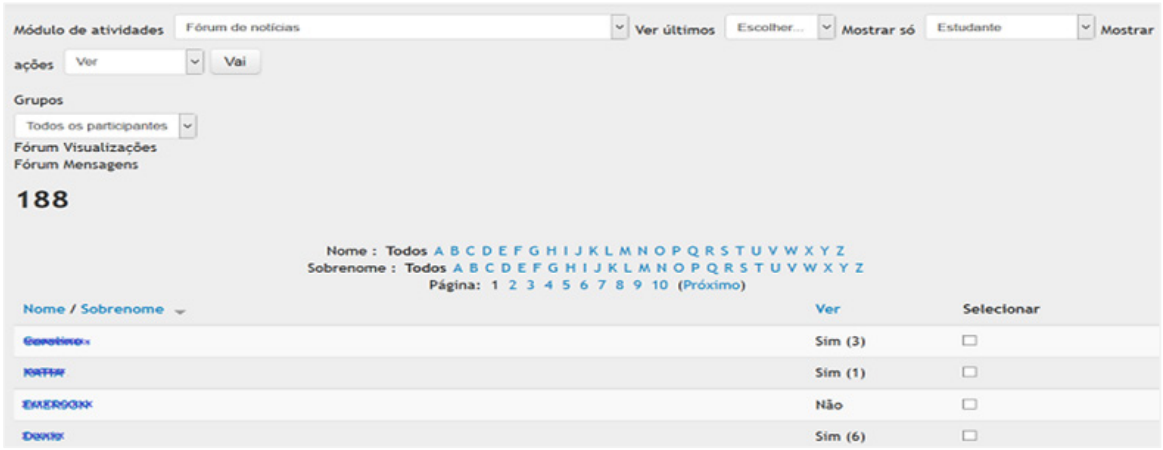


Figure 3. Student Accesses in the News Forum

Figure 4 presents the report made available by Moodle, regarding the participation of students in the activities of the questionnaire type, of a specific room in which the teacher/tutor acts.

Table 4 presents an overview of students' participation in Questionnaire-type activities in the three virtual classrooms, under the responsibility of tutor "A".

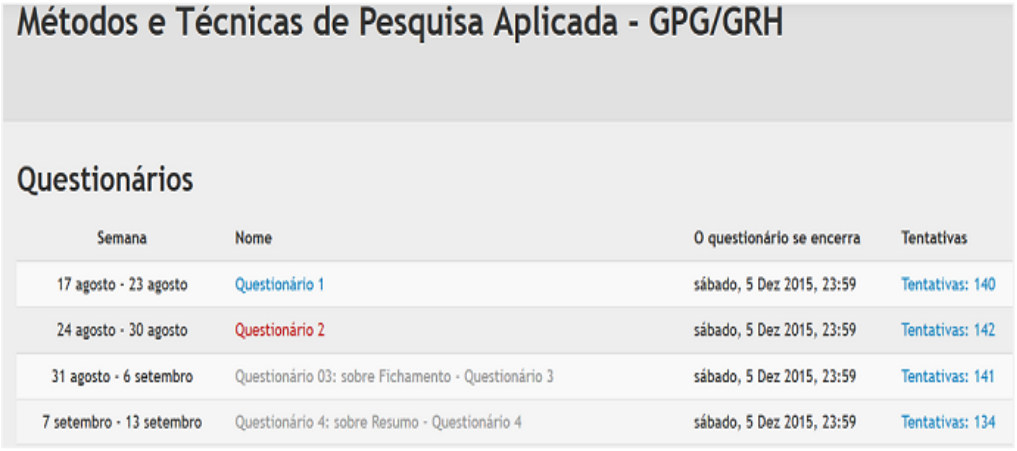


Figure 4. Participation of students in the activities of the Questionnaire type

Table 4. Participation of students in the activities of the Questionnaire type

Activity	MKT		MTA		PO	
	Total students: 34		Total students: 188		Total students: 95	
	attempts	(%)participation	attempts	(%)participation	attempts	(%)participation
Quest1	28	82,3%	140	74,4%	68	71,5%
Quest2	28	82,3%	142	75,5%	72	75,7%
Quest3	27	79,4%	141	75,0%	77	81,0%
Quest4	23	67,6%	134	71,3%	70	73,7%
Quest5	25	73,5%	121	64,3%	65	68,4%
Quest6	25	73,5%	128	68,0%	67	60,0%
Quest7	21	61,7%	132	70,2%	65	63,1%
Quest8	20	58,8%	128	68,0%	57	60,0%
Quest9	23	67,6%	128	68,0%	60	63,1%
Quest10	21	61,7%	126	67,0%	57	60,0%
Quest11	18	52,9%	121	64,3%	51	53,7%
Quest12	18	52,9%	128	68,0%	51	53,7%

Table 5 presents an overview of the participation of the tutor "A" in the Doubt Forum, in the three virtual classrooms.

Table 5. Participation of tutor "A" in the Doubts Forum

Discipline	Number of messages posted by students	Number of messages read by tutor	Number of messages commented by the tutor	(%) Participation rate
MKT	15	15	13	86,6%
MTA	48	48	44	91,6%
PO	26	26	25	96,0%

Table 6 presents an overview of the participation of tutor "A" in the Forum of News, in the three virtual classrooms.

Table 6. Participation of tutor "A" in the Forum of News

	MKT	MTA	PO
Messages posted to the News Forum	4	4	4

3.3. Analysis of extracted data

It is observed in the study, through Figures 2, 3 and 4, that Moodle allows the extraction of reports of inactive students for more than one period, specified in the system, student participation in forums, activities, however, are not easy interpretation.

The teacher/tutor used the News Forum tool to keep the students informed about the events in the course, such as term of activities, calendar of tests, alerts and also this Forum was used for messages of welcome to the students, however, In the case of tutor "A" only four messages were posted in the News Forum of each virtual room, which is considered little due to the importance of this tool in transmitting information to students. Access to this News Forum is important because it is published relevant topics such as: notices, alerts, important dates, deadlines, etc.

One of the main roles of tutors is to be responsible for explaining and clarifying issues related to the discipline [4]. The tool used for this purpose in the examples presented was the Doubts Forum. In the presented results it was observed that the participation of the teacher/tutor was representative, through the messages posted in the Doubts Forum, in the sense of commenting them, the participation of the students was not significant, according to the number of messages posted by them.

In this study, the teacher/tutor offered twelve evaluation questionnaires during the course and, through the BOX Activities, was able to extract information on how many attempts were made in each questionnaire, but difficult to interpret.

The monitoring of the actions in Moodle becomes very laborious for both the teacher/tutor and the manager since it is necessary that these monitoring must be done in each virtual classroom in which the tutor acts, in isolation, for later the information be grouped and provided an overview of the actions.

The same deficiency of the environment is verified for the environment manager and learning processes, which does not have a way to obtain a visualization of the actions of the teachers / tutors in a global way, grouping all the virtual rooms under their responsibility, being necessary that they access each Virtual room in isolation and extract the information of your interest.

A competent tutorial action could minimize many flaws in this process of teaching and learning, and allow the manager to make decisions regarding the faculty during the course, so it is interesting that:

- The teacher/tutor is monitored weekly through access statistics in virtual rooms;
- Messages are posted in the News and Doubts Forums;
- Evaluation activities are posted weekly, such as: questionnaires, online tasks, file submission task and discussion forums.

4. Monitoring tool for students and teachers/tutors

A Fag tool was developed integrating 2 software: BI Pentaho and Qlik Sense Cloud. Integrated into the Moodle LMS and this tool is responsible for generating analytic and consolidated reports that show a transversal view of all system's actors (students and teachers/tutors) and the interaction between these actors with the system. It is important to notice that this is the main contribution of the tool: provide a wide and transversional vision system's vision. Figure 5 shows the architecture of the FAG Tool.

From these reports, it is expected that:

- The teacher/tutor can act quickly and efficiently in his / her role of motivating and stimulating student participation;
- distance-learning environment and process managers improve the time in decision making regarding the faculty and can achieve better results in distance-learning.

4.1. Fag Tool Architecture

Figure 5 shows the architecture of the developed FAG tool. The data source used is MySQL, extracted from Moodle LMS, which represents the data layer. In this layer, it was necessary to move the useful data for the generation of reports and dashboards to an intermediate data repository called the data warehouse (DW), this process is called ETL (Extraction, transformation and load). Performed through the Pentaho BI tool and scheduled to run in the D-1 period, updating the DW always with the data from the previous day.

From DW, the software Pentaho BI and Qlik Sense Cloud were used to generate the analytical reports and dashboards presenting a holistic and transversal view of the participation of students and teachers/tutors LMS Moodle.

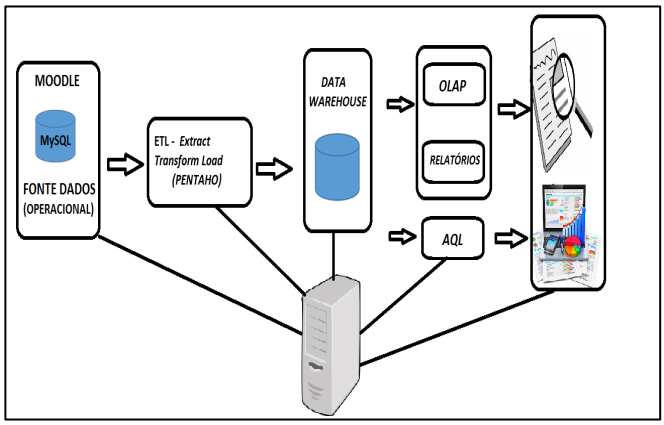


Figure 5. Fag tool architecture

4.2. Partial result

In this section, it will be presented some analytical reports generated by the tool, in addition to the dashboards, all with a holistic and transversal view. Figure 6 presents the percentage participation report of students linked to the courses of the selected teacher/tutor.

The system manager selects the desired teacher/tutor, the related courses are listed with the number of students enrolled, the number of active students and consider as active the students who accessed the course in the last 30 days and calculated and displayed the % of student participation in each course.

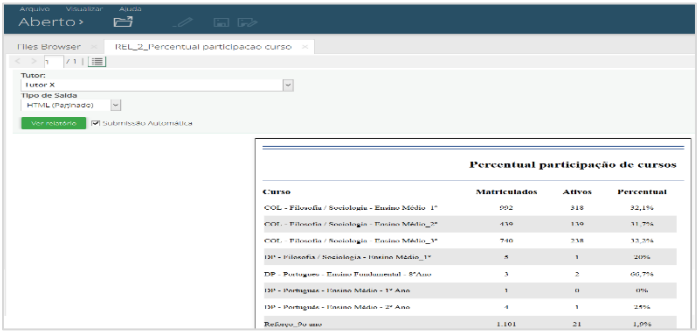


Figure 6. Percentage of student participation report

Figure 7 presents the report with the participation percentage of students in the weekly activities proposed by the teacher/tutor.

The teacher/tutor selects his / her name, the respective courses are listed, the number of students enrolled and the number of students who did not perform any of the weekly activities proposed and is calculated and displayed the percentage of effectiveness, that is, percentage of students who Carried out all the proposed weekly activities.

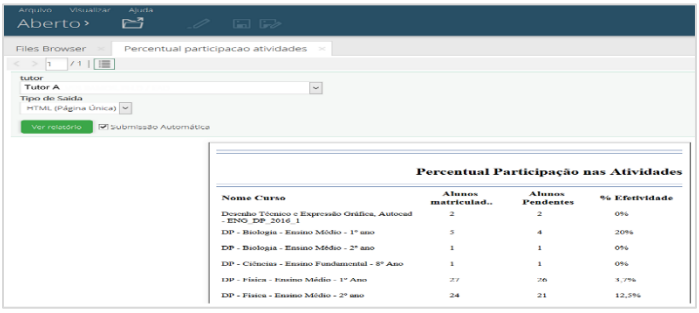


Figure 7.Percentage of participation in weekly activities

Figure 8 presents the report with the percentage of students who have, never attended the course.

The teacher/tutor selects his / her name, lists the respective courses linked to him/her with the number of students enrolled, the number of students who have never attended the course, it is calculated and displayed the percentage of students who never attended the course and in sequence Are listed name and email of students who have never attended the course, if the number of students they have never accessed is greater than zero.

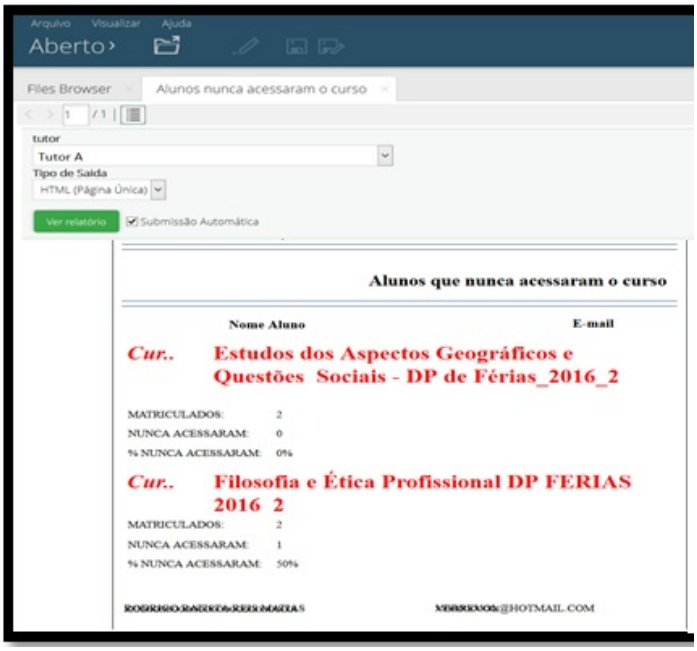


Figure 8.Percentage of students who have never attended the course

Figure 9 shows the report with the total time of access of the teacher/tutor.

The system manager selects the desired teacher / tutor and the month, the corresponding teacher / tutor courses are listed with the days of the week and the time in hours, minutes and seconds, which the teacher / tutor has connected to each course linked to his or her profile in that selected month.

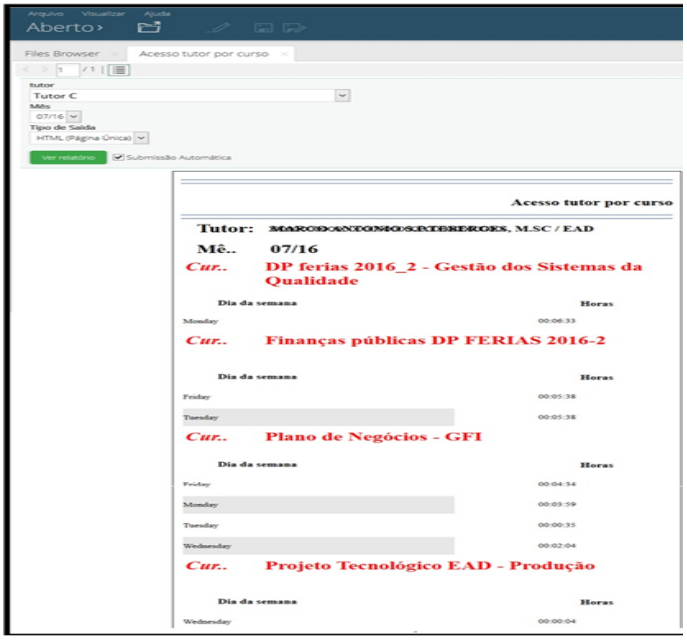


Figure 9.Total time of teacher/tutor access

Figure 10 presents to the system manager a dashboard with the total number of students enrolled in Moodle LMS (Qtd. Students), users in the role of teachers / tutors and the registered courses, as well as a graph that shows the students who have accessed the courses for teachers / tutors in the last 30 days.

Through queries that use AQL technology, a specific teacher/tutor can be selected and information automatically updated according to the courses linked to that teacher/tutor, according to Figure 11.

Through the Tutor dimension, the courses are related to the selected teacher/tutor (Tutor A), the number of students enrolled in the courses linked to it (144), and the graph with the number of students who have accessed the courses in the last 30 days and the list of names, e-mail and courses of the respective students.

In addition to the selection of the teacher/tutor, one can select a specific course linked to it, consequently, the information will be updated, as shown in Figure 12.

According to Figure 12, the vision for the COURSE dimension is presented and it can be observed that in the Course "COL - Filosofia / Sociologia – Ensino Médio_1_2016 " selected, of the teacher/tutor "Tutor B", 290 students are enrolled and of these students 246 attended the course in the last 30 days, therefore, they are considered active.

In addition to the selection of teacher/tutor and courses, the tool also allows verifying the accesses in the course linked to the teacher/tutor, through the selection of the access date (s), according to Figure 13.

This shows the number of students who have accessed the selected course linked to the teacher/tutor on that specific date (s), in addition to the students' names and emails. The result is also displayed in graphic form.

The Forums view illustrated in Figure 14 shows the percentage of students who accessed the Doubts and News Forums through the FORUMS dimension.

The view with the dimension QUESTIONS selected shows the percentage of students who accessed the Doubts Forums.

The vision with the NEWS and COURSE dimension selected presents the percentage of students who accessed the News Forum in the selected course.

The view News Forum with the dimension STUDENT selected presents the courses in which "Student A" is linked and the percentage of accesses in the News Forums in each course.

The view with the dimension QUESTIONS and COURSE selected presents the percentage of students who accessed the Doubts Forum in the selected course.

The Forums view with NEWS dimension selected shows the percentage of students who accessed the News Forums.

The view Forum of Doubts with the dimension STUDENTS selected presents the courses in which "Student A" is linked and the percentage of accesses in the Doubt Forums in each course.

The Tutor Management view presents the average time of access of the teachers/tutors to the courses in Moodle LMS.

The tool also presents the general average time of access of the teachers/tutors to the courses per day of the week or the general average time of access of the teachers/tutors to the courses when selecting the name of the tutor in the dimension TUTOR.

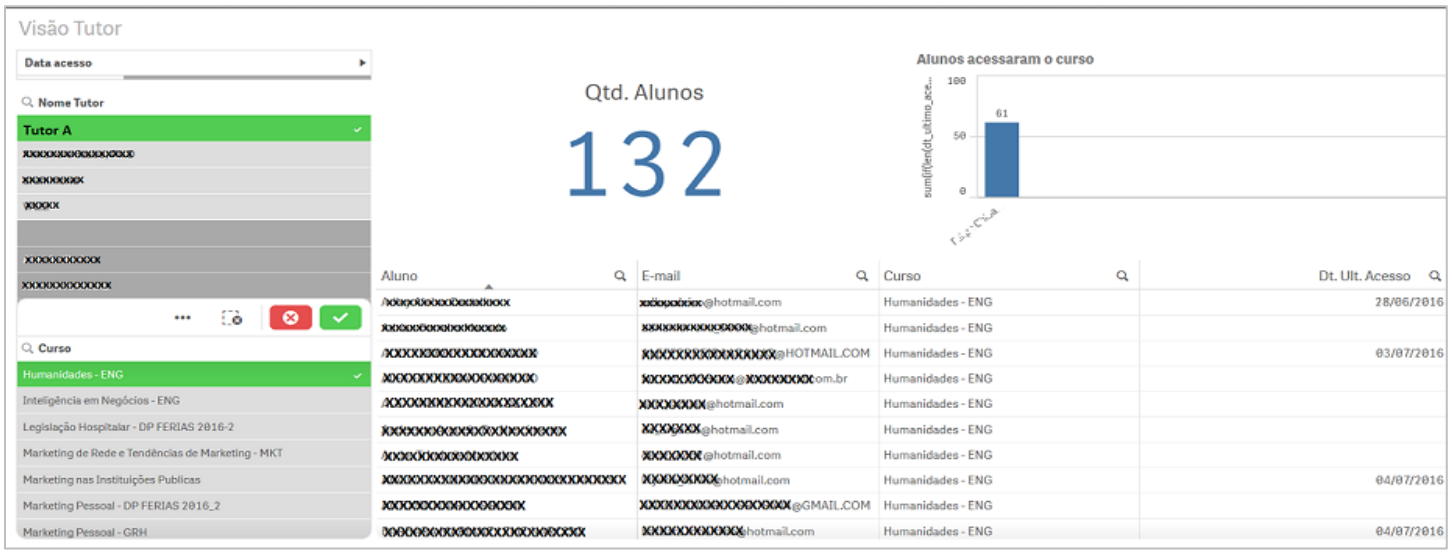


Figure 10. Vision of the participation of the students linked to the courses in Moodle LMS

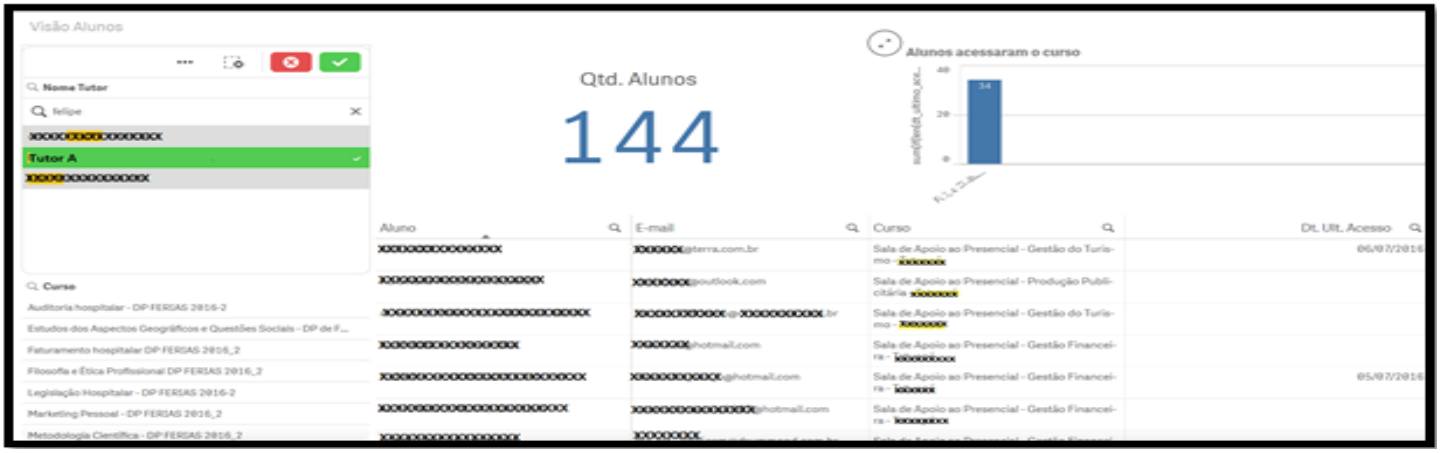


Figure 11.Participation of students through selection of a teacher/tutor

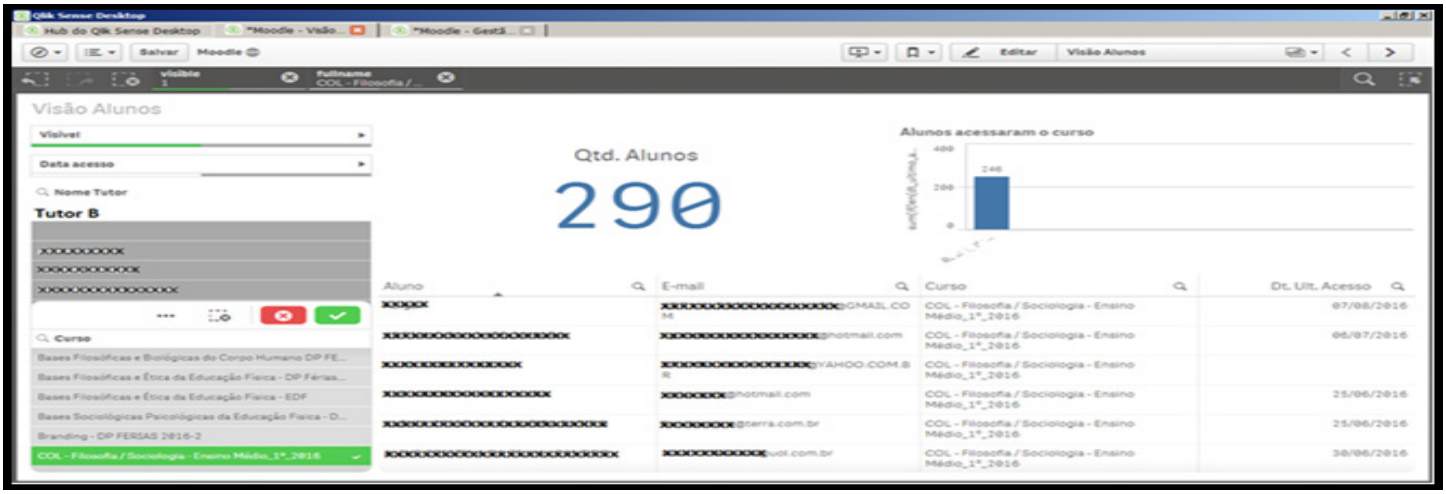


Figure 12.Participation of students in selection of a teacher/tutor and course linked to him/her

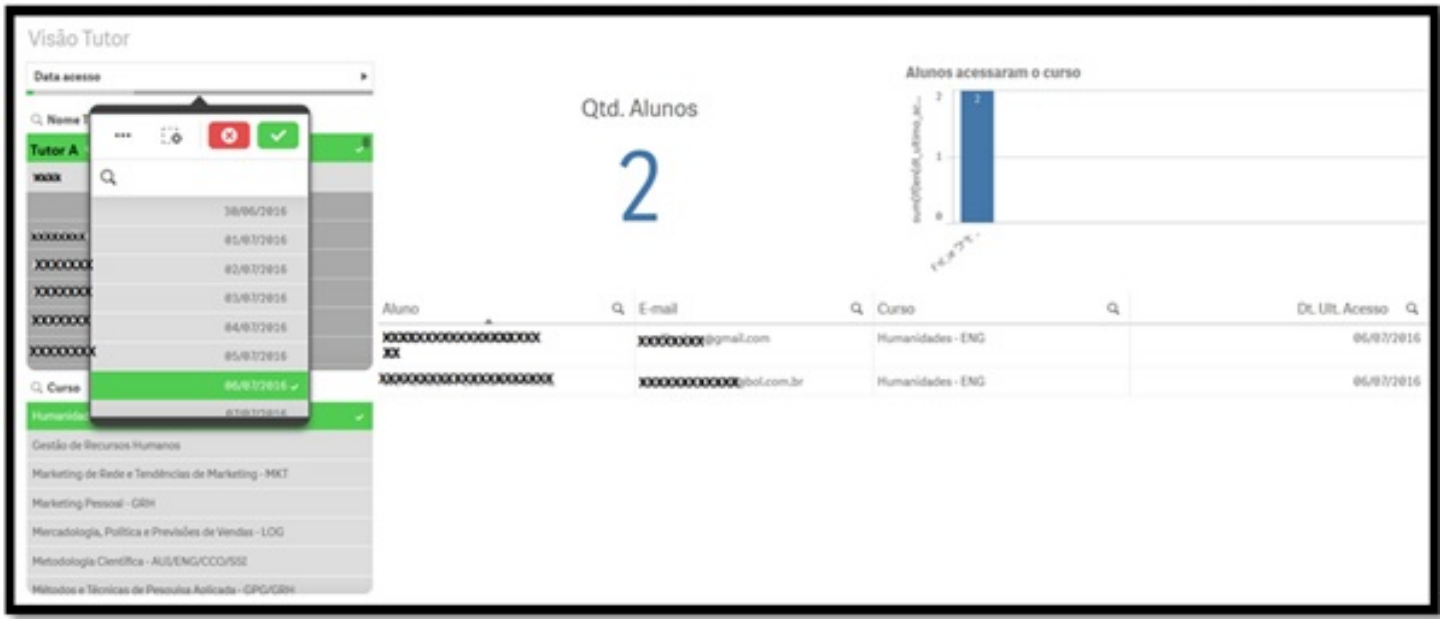


Figure 13.Participation of students in selection of a teacher/tutor, course linked to him and date (s) of access

Figure 14.View Forum Leaders

5. Conclusion

In this study was present some Moodle LMS's tools showing that they do not provide resources to the teacher/tutor to efficiently monitor the participation of the students and nor does it provide the tools necessary for educational process managers to monitor the actions of teachers/tutors in all activities that are involved.

The work also presents a tool - FAG that provides reports, through the use of BI techniques, to automatically generate analytical and consolidated views and necessary dashboards for both the teacher/ tutor in relation to student participation as well as the system manager in relation to the actions of the teachers/tutors, simultaneous in all the courses or rooms linked to them in Moodle LMS.

The FAG tool, integrated to Moodle LMS, provides information in a very simple and fast way, minimizing the time needed to obtain meaningful information and thus facilitating the interpretation work of the teachers/tutors, through the holistic and transversal view.

Using the FAG tool, teachers/tutors are able to verify the participation of the students in activities of the Questionnaire type, in Forums of doubts and news, as well as access to the courses. By using this information, it is possible to stimulate students participation, motivating them and reducing the level of avoidance and dropout.

The managers, through the reports offered by the FAG tool, are able to monitor the actions of teachers/tutors, identifying their participation in many activities like Forums of doubts, publication of activities, time of access to the courses and interaction with students.

The FAG Tool assists teachers/tutors in the follow-up of students' participation in the courses related to them and also supports the LMS's managers in the decision making in a timely and assertive way, improving the management process in LMSs systems.

6. Future Work

As future work, it is suggested to create plug-ins in Moodle LMS to be added to the LA resources offered by the platform. In addition, it is suggested the creation of a parametric measurable tool that will allow the generation of the visualizations in any LMS, through the creation of scripts that use parameters, thus to be able to feed the DW with the necessary information.

References

[1] L.S. Zapparolli, I. Stiubiener. "Management support tool in virtual learning environments using moodle as a case study." Learning Objects and Technology (LACLO), Latin American Conference on. IEEE, 2016.

[2] V. Bremgartner, J.F. De Magalhães Netto, C. S. De Menezes. Adaptation resources in virtual learning environments under constructivist approach: A systematic review. In: Frontiers in Education Conference (FIE), 2015. 32614 2015. IEEE. IEEE, 2015. p. 1-8.

[3] S. B. Affonso; E. Quinelato. Educação a Distância: Algumas Considerações a Respeito do Autogerenciamento da Aprendizagem pelos Estudantes. EAD em FOCO, v. 4, n. 2, 2014.

[4] M. L. Belloni. Educação a Distância – 7ª edição. São Paulo: Autores Associados, 2015.

[5] G. Rumble. A gestão dos sistemas de ensino a distância. Ed. da UnB, 2003.

[6] A. Clementino. "Gestão pedagógica de cursos em EAD online." Acessível em: http://www. abed. org. br/congresso2005/por/pdf/041tcc5. pdf (2013).

[7] A.H.G.Rêgo, A.V.Gomes, M.Lucena and A.C.Brito. Um módulo gerador de relatórios no auxílio às atividades docentes no Moodle. Anais do XXII SBIE-XVII WIE, 2011.

[8] A. M. R. Almeida, S.S. Camargo. Aplicando Técnicas de Business Intelligence sobre dados de desempenho Acadêmico: Um estudo de caso. XI Escola Regional de Banco de Dados. Caxias do Sul-RS, 2015.

[9] R.R. Gonçalves. "Integração de Dados na Prática." São Paulo: Érica (2012).

[10] E. Turban, R.Sharda, J.E.Aronson. Business Intelligence: um enfoque gerencial para a inteligência do negócio. Bookman Editora, 2009.

[11] V.V. dos Santos. Data Warehouse: análise da performance de ferramentas de ETL.2013. 52 f. Monografia (Especialização) - Curso de Especialização em Engenharia de Projetos de Software, Universidade do Sul de Santa Catarina, Florianópolis, 2013. Cap. 7. Disponível em: <http://www.uniedu.sed.sc.gov.br/wp-content/uploads/2013/10/Valdinei-Valmir-dos-Santos.pdf>. Acesso em: 10 fev. 2016.

[12] A. Bentes. TI Update: A Tecnologia da Informação nas grandes empresas. Rio de Janeiro: Brasport, v. 2, 2008.

[13] L. M.M. Giraffa. Recursos De Learning Analytics Para Compor Indicadores Auxiliares Na Avaliação Dos Estudantes. Revista EmRede, v. 2, n. 2, p. 32-43, 2016.

[14] M.Á. Conde and Á. H. García. A promised land for educational decision-making?: present and future of learning analytics."Proceedings of the First International Conference on Technological Ecosystem for Enhancing Multiculturality. ACM, 2013.

[15] L. Johnson, R.Smith, A.Levine and K.Haywood. The 2010 Horizon Report. New Media Consortium. 6101 West Courtyard Drive Building One Suite 100, Austin, TX 78730, 2010.

[16] O. Petropoulou, K Kasimatis, I. Dimopoulos and S. Retalis. LAe-R: A new learning analytics tool in Moodle for assessing students' performance. Bulletin of the IEEE Technical Committee on Learning Technology, v. 16, n. 1, p. 1, 2014

[17] A. Marinheiro, J. Bernardino. Experimental Evaluation of Open Source Business Intelligence Suites using OpenBRR. Latin America Transactions, IEEE (Revista IEEE America Latina), v. 13, n. 3, p. 810-817, 2015.

[18] J. Nascimento, G.B.Paulus, D.L.V.G.Rubert and R.L.Antoniazzi. "Comparação da eficiência de bancos de dados relacionais como datawarehouse em um contexto de business intelligence". Revista interdisciplinar de ensino, pesquisa e extensão, v. 3, n. 1, 2016.

[19] E. Reis, F. Teixeira, M. A. Araújo. Implementando uma solução de Business Intelligence com o Microsoft SQL Server 2005–Parte 1. SQL Magazine, Rio de Janeiro, n. 59, p. 52-66, 2009.

[20] QLIKVIEW: conheça a ferramenta de B.I. Disponível em: <http://www.devmedia.com.br/qlikview-conheca-a-ferramenta-de-b-i/30123>. Acesso em: 15 ago. 2016.

[21] IN. Inteligência em Negócios. Disponível em: <http://www.in1.com.br/qlik-sense> . Acesso em: 15 ago. 2016.

[22] Ping-Tsai Chung and Sarah H. Chung. "On data integration and data mining for developing business intelligence." Systems, Applications and Technology Conference (LISAT), 2013 IEEE Long Island. IEEE, 2013.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 598-604 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

A comparative between CRISP-DM and SEMMA through the construction of a MODIS repository for studies of land use and cover change

Herman Jair Gómez Palacios*, Robinson Andrés Jiménez Toledo, Giovanni Albeiro Hernández Pantoja, Álvaro Alexander Martínez Navarro

Department of System Engineering, Mariana University, 520002, Colombia

ARTICLE INFO

Article history:
Received: 04 April, 2017
Accepted: 12 May, 2017
Online: 04 June, 2017

Keywords:
MODIS
CRISP-DM
SEMMA

ABSTRACT

Among the most popular methodologies for development of data mining projects are CRISP-DM and SEMMA, This research paper explains the reason why it was decided to compare them from a specific case study. Therefore, this document describes in detail each phase, task and activity proposed by each methodology, applying it in the construction of a MODIS repository for studies of land use and cover change. In addition to the obvious differences between the methodologies, there were found other differences in the activities proposed by each model that are crucial in non-typical studies of data mining. At the same time, this research determines safely the advantages and disadvantages of each model for this type of case studies. When the MODIS product repository construction process was completed, it was found that the additional time used by CRISP-DM in the first phase was composed in the following phases, since the planning, definition of mining goals, and generation of contingency plans, allow developing the proposed phases without inconvenience. It was also demonstrated that CRISP-DM is presented as a true methodology in comparison with SEMMA, because it describes in detail each phase and task through its official documentation and concrete examples of its application.

1. Introduction

This paper is an extension of work originally presented in 11CCC IEEE 2016 [1].

Within the set of business intelligence technologies is data mining, which in itself is a phase of the KDD process, and is responsible for the data preparation and the interpretation of the results. Keeping in mind that KDD is a process that has a high level of complexity, it was necessary the use of methodologies that allow a systematic approach in the construction of data mining projects, today there are two methodologies for the information treatment that are avant-garde in their models. The SAS company suggests the use of SEMMA methodology, likewise in 1999, a large consortium of European companies, NCR (Denmark), AG (Germany), SPSS (England) and OHRA (Netherlands) joined forces to develop a free distribution methodology called CRISP-DM (Cross Industry Standard Process for Data Mining) [2]. Those methodologies structure the data mining project in phases that are interrelated, converting the process into iterative and interactive.

The CRISP-DM methodology structures the life cycle of a data mining project in six phases [3], which interacts with each other in an iterative way during development of the project as shown in Figure 1. The methodology in a general description is organized in phases, where each phase is structured in several generic tasks of second level, and these make a projection towards specific tasks that described the actions that should be developed [3]. E.g., if the second level has the generic task "data cleaning", the third level indicates the tasks that have to be developed for a specific case, such as "numerical data cleaning" or "categorical data cleaning". Finally, the fourth level includes a set of actions, decisions and results of the data mining project in question.

On the other hand, the SEMMA methodology takes its name from the different stages that lead the process of information exploitation, sample, explore, modify, model, and assess. The methodology is itself a cycle whose internal steps can be performed iteratively according to the needs. SEMMA provides an easy-to-understand process that allows the development and maintenance of information exploitation projects [4].

Both methodologies structurally have similarities, e.g., the specific task of data cleaning, however they differ in several aspects, like tasks, activities and phases, e.g., in CRISP-DM there

*Corresponding Author: Herman Jair Gómez Palacios, Department of System Engineering, Mariana University, 520002, Colombia
Email: hjgomez@umariana.edu.co

is a phase dedicated to problem understanding in terms of business, defining objectives, resources, roles, etc., for setting data mining goals, while SEMMA proposes to directly access the data to be analyzed, defining a sample and applying techniques of data mining without considering the business objectives. Therefore, this causes certain differences in the data mining process, and this document aims to identify them.

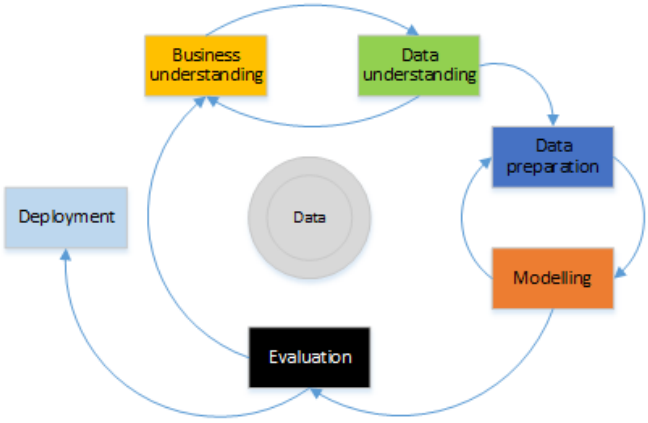


Figure 1: CRISP-DM methodology phases. Adaptation of: Chapman, P. [3].

The case study proposed for the comparison focuses on the construction of a MODIS repository for studies of land use and cover change. Considering that vegetation is a primary indicator of the state of an ecosystem because of its fundamental role in the water and carbon cycles [5], in addition, photosynthetic activity is a measure of the transformation of radiant energy into chemical energy, and any observable change in it may indicate alterations in the environment [5]. An agile and low-cost way to monitor large areas and analyze vegetation behavior is the use of vegetation indices obtained from remote sensors, e.g., from the MODIS sensor.

MODIS (MODerate resolution Imaging Spectroradiometer), which is one of the most important for the monitoring of the processes of change in the earth for its spatial and spectral characteristics. The MODIS sensor was designed by an interdisciplinary team of scientists with extensive experience in remote sensing. This team worked for about 10 years to finally define requirements for data collection, calibration and processing [6]. This sensor produces a total of 648 tiles, which about 290 are classified as soil and therefore produce vegetation indices [7]. However, a vegetation index can be defined as a parameter calculated from reflectance values at different wavelengths, and is particularly sensitive to vegetation cover [8]. According to the previous definition, the variation of the vegetation index over time allows to determine if a field in terms of coverage or biomass is better or worse according to its historical values.

This study was used the MOD13Q1 products generated by the sensor on board the TERRA satellite, and as an alternative was used the MYD13Q1 products from AQUA satellite. The NDVI (Normalized Difference Vegetation Index) was used as an index to estimate the quantity, quality and development of vegetation based on measurement, by means of remote radiation sensors of certain bands of the electromagnetic spectrum that vegetation emits or reflects [7]. The EVI (Enhanced Vegetation Index) also was used, which is considered to be the most robust vegetation index, because it is more robust compared to the contribution of soil and atmospheric influences [7].

From the files downloaded for the study area, the first three products were extracted (NDVI, EVI and VI Quality); the quality of the data present in the spectral bands will determine the quality of the studies of land use and cover change. The reason why it is necessary to carry out a cleaning process following the steps proposed by a recognized methodology of data mining, where the quality of the data before applying mining techniques is essential. Therefore, it was decided to make a comparison between the CRISP-DM and SEMMA methodologies to determine the most appropriate for non-typical data mining studies such as the one referred to in this research. Thus the process was developed with a high level of detail phases, tasks and activities of each methodology, and it was finally possible to obtain the best quality of the MODIS products that would be used in studies of land use and cover change.

2. Research background

Nowadays, some worldwide researches are based on the MODIS sensor, e.g., to temporal and spatial comparison of the frequency of heat sources as a fire indicator in Colombia in recent years [9], as well as in techniques of data mining, physical and statistical models, based on satellite remote sensing methods [10], also in remote sensing systems that allow exploring diverse fields such as agricultural, forestry, or to evaluate land use and cover change [11]. However, in the found referents, it was evident the insufficiency of researches in which to guarantee the highest quality of the data to be used, following a data mining model, and even less if it is included in studies of land use and cover change using MODIS products. On the other hand, although there are several comparative studies between different models of data mining, including CRISP-DM and SEMMA [12, 13, 14], it is not clear which to use for researches that obtain data from MODIS products.

3. Methodology

The area of analysis of this research includes the territory of the department of Nariño; therefore the MODIS products downloaded correspond to tile h10v08. The tiles are distributed within a Cartesian system, starting at the position (0, 0) corresponding to the upper left corner and ending at the position (35, 17), corresponding to lower right corner [6]. The coverage area of each tile is 10x10 deg. The rows are defined by the letter "v" and the columns by the letter "h", in this way is possible to make a suitable selection of the area of interest.

The MODIS products that work with vegetation indices are those corresponding to the MOD13 and MYD13 series, from which were selected those that provide a spatial resolution of 250 meters and a temporal frequency of 16 days, to finally obtain 355 MOD13Q1 products from the years of 2000 to 2015. For the identification of MODIS products with excellent quality for studies of land use and cover change, CRISP-DM and SEMMA data mining methodologies were applied faithfully following each phase, task and activity suggested.

With respect to the download, reprojection, transformation, cleaning and storage of MODIS products, some “scripts” were used in all cases to optimize the process, among them ModisDownload [15]. The scripts were created for the statistical program R, which is free (GNU) and through the RGDAL and Raster libraries can directly process the HDF files. The scripts

make use of the MRT reprojection tool, which were granted free of charge by NASA, this tool allows the download, reprojection, conversion and integration of MODIS products [16], in addition to extracting the layers from each of them.

4. Result

The detailed process of the application of the mining methodologies presented in the previous sections is described below, and the similarities and differences between both models, are presented in a parallel, spotlighting the key elements of phases and previous tasks.

4.1. Phase 1

The development of this first phase in CRISP-DM is known as "Business Understanding" [3], which allows obtaining a greater level of understanding of the case study, this being a key component to construct the MODIS repository for studies of land use and cover change, because it is possible to plan a project and define clear and concrete data mining goals.

Task 1: denominated by CRISP-DM as "Determining business objectives" [3], in this general task the methodology suggests knowing what the client wants to obtain, thus giving the analyst a business perspective. In this first task the objectives were defined, based on the context, area of the problem and current solution, therefore the objective was: to classify regions in the department of Nariño in agricultural, forest and livestock, as defined by the IGAC [17]. Task 2: the methodology defines it as "Assess situation" [3], which is done by taking into account the inventory of resources (hardware, knowledge and data), identification of requirements, assumptions, constraints, risks and contingency plans, as well as including the costs and benefits of the study. For this research, it was identified among other elements: the different alternatives for the acquisition of MODIS products, among which standout 1) Reverb, online tool for the discovery of scientific data of the earth; 2) USGS Global Visualization Viewer which is a tool for searching and downloading satellite data 3) LP DAAC Data Pool, which allows direct access to MODIS products. In addition to these free tools provided by NASA, there are payment tools and scripts developed by experts with free access, the latter being the best alternative for the project. Task 3: CRISP-DM defines it as "Determine data mining goals" [3], in the specific tasks the data mining objectives and the evaluation criteria of the model are determined, which for the case study were respectively: identification of types of regions in Nariño, according to kind of vegetation found and the evaluation of the accuracy of the model compared with real data provided by IGAC - Nariño. Task 4: is known as "Produce project plan" [3], and is the last general task proposed by CRISP-DM for the first phase, in the specific tasks, the project plan was constructed through a table in which it stands out for each phase: time, resources, risks and contingency plans that are required for the successful project implementation, and to create the table, all the information obtained in previous tasks was required. Therefore the planning of the phases necessary to construct the repository of MODIS products can be visualized in the Table 1.

On the other hand, SEMMA in its first phase "Sample" [4], proposes to extract a sample of the total data, in this way allowed to make a previous analysis of the problem of mining. For the case study it was obtained that:

For data access, it was defined for this study that it would work with the MOD13 algorithm, specifically with the MOD13Q1 products, because in addition to be suitable for studies of land use and cover change, they do not produce pixels on oceans or soil under water. The study population was 355 HDF files and 1065 products of interest; B) to establish a small sample, 10 percent of the total population was chosen, therefore the sample size was 107 products; C) to form the sample was chosen a simple random method, obtaining 28 NDVI, 53 EVI and 26 VI Quality layers. For the extraction of the first three layers (NDVI, EVI and VI Quality) of interest in the case study, the MRT tool was used, as well an R script that optimizes the process.

Table 1: Project planning

Phase	Resources	Risks	Contingency plans
Business understanding	Documentation about MOD13Q1 product. Documentation about vegetation indices. Documents of IGAC from Nariño. Visualization tool: Quantum GIS	Limited information in Spanish about these topics.	English and Portuguese documentation
Data understanding	Website with products for soil studies as LP DAAC Scripts to download, reprojection and conversion of MODIS products. Tool to reproject and convert MODIS products (MRT). Environment and R programming language to execute the scripts.	Bugs or limitations in the scripts.	Manual FTP downloading.
Data preparation	NDVI, EVI and VI Quality products. Environment and R programming language to execute the scripts. PostgreSQL - PostGIS to store the layers. Programming language: Python	MODIS products that cannot be used.	Downloading MODIS products from AQUA satellite.

At the end of the phase, it is concluded that CRISP-DM benefits the achievement of results because it allows for fully understanding of the problem in terms of the business and its equivalent in terms of data mining, even proposes to make a detailed plan that guarantees the success in the process.

4.2. Phase 2

In CRISP-DM the second phase is known as "Data Understanding" [3]. In this phase the methodology proposes data selection criteria, obtains and explores them, in a way that identifies the elements that determine their level of quality.

Task 1: denominated by CRISP-DM as "Collect initial data" [3], where the following criteria were defined: 1) Products of the MOD13 series with a resolution of 250 meters. 2) The study region should contain the area of Nariño (available in the h10v08 tile). 3) Download available products between February 2000 and June

2015 (for a total of 373 files in HDF format). For the download of MODIS products it is necessary to keep in mind parameters of the script to select a specific date, or a range of time; The coordinates to specify the tile to be downloaded, or several tiles when joining several coordinates; The script named "ModisDownload" of free access, originally created by Naimi [15], later adapted and improved by the community, uses the GDAL library, which allows reading and writing raster data compatible with spatial references, also the reading of vector data and HDF extension files belonging to the MODIS products. The MRT reprojection tool is free and developed by NASA [16]; this tool allows the download, reprojection, conversion and integration of MODIS products, in addition to the extraction of the layers of each product. The HDF files were obtained from the website: land data products and services LP DAAC [16]. The execution of the script was done per year and the time required to download the products of a particular year was 161 minutes. The total time for the download of all MODIS products was 2310 minutes. Adding the download time of AQUA sensor products would add 126 minutes. Task 2: known in the methodology as "Describe data" [3], for the case study was determined the description of the type and range of values of the attributes and the volume of the data (1119 files in TIF format and reprojected to the coordinate system accepted in Colombia. The files correspond to first three layers of the MOD13Q1 products). Task 3: named by CRISP-DM as "Explore data" [3], focused on the VI Quality product, which allows the determination of the quality of the data processed per pixel in the layers, using parameters such as the usability of the pixel, amount of aerosol, detection of adjacent clouds, atmospheric correction, among others. Task 4: defined as "Verify data quality" [3], it was decided to work with the first VI Quality layer parameter, identifying special values and cataloging its meaning (Table 2).

Table 2: Description of the VI Quality parameter used

Bits	Parameter Name	Value	Description
0-1	VI Quality	00	VI produced with good quality.
		01	VI produced, but checks other QA.
		10	Pixel produced, but most probably cloudy.
		11	Pixel not produced due to other reasons than clouds.

The "Explore" phase [4] of SEMMA was developed with an exploration of the sample data, allowing complement the initial understanding of the problem, later the process was restarted and applying for all the data, thus sufficient inputs were available for the next phase. For the case study, the MODIS products of interest were the first three: NDVI, EVI and VI Quality. For the NDVI and EVI layers the valid range of values is -2000 to 10000 and its scale factor is 0.0001, which means that the values of 10,000 in a raster should be multiplied by 0.0001 in order to achieve the current value. As for the data quality criteria, the first four attributes of the first parameter of the Quality VI layer were chosen as indicated in Table 2.

The first phase of CRISP-DM is very important because it offers clarity on the problem to be solved and how to do it, whereas SEMMA does it, but in a very superficial way depending on several later phases of this, nevertheless SEMMA has as advantage that work with a smaller amount of data (however this first phase is optional). Regarding the data selection and evaluation criteria, the CRISP-DM methodology proposes to define them in the third task of the first phase, however, each of them had to be reviewed and improved in several of the following tasks in the same phase. Through the criteria it was possible to optimize the process of data selection and the cleaning of data. In SEMMA the definition of selection criteria was quite questionable at the beginning, and each criterion had to be improved and corrected in the course of the tasks and even in the following phases.

4.3. Phase 3

This phase is known as "Data Preparation" [3] in CRISP-DM, it is composed of four general tasks and the same number of specific tasks.

Task 1: named by CRISP-DM as "Select data" [3], allows determining which data will be included for the study and the criteria to exclude the data with low quality. For SEMMA, this task is called "Transformation and selection of data" [4] and consists, as in CRISP-DM, of identifying the data to be used in the next task, however SEMMA did not consider the inclusion of MYD13Q1 products, thus obtaining 1065 files, unlike the 1119 obtained by CRISP-DM, despite this, the development of the task was similar in the two methodologies and described below.

For each downloaded MOD13Q1 product the first three layers (NDVI, EVI and VI Quality) were extracted, by means of the adaptation of the script developed by Golicher [18], to transform the image to the TIF format and at the same time reproject it to the coordinate system accepted in Colombia, obtaining a total of 1119 new files. The VI Quality product allows determining the quality of the data processed per pixel, identifying which data has an excellent quality, an intermediate quality, a low quality and which cannot definitively be used [7]. For the case study it defined the first 4 bits of the VI Quality layer described above in Table 2, determining that the total of non-usable pixels will be the result of the sum of the probably clouded pixels and the pixels not produced, Therefore, if this value exceeds 30% of the pixels in the image, it should be discarded [19].

In order to determine the quality of each image for the NDVI and EVI products, the calculation was performed by adapting the script developed by Lydholm [20] for the statistical program R. The process was performed by each VI Quality layer, determining the number of non-usable pixels in the study region and keeping in mind an analysis is not possible if the amount of unusable data exceeds 30% [19].

In the script, the necessary libraries were loaded to use methods to create raster images and functions to handle HDF files. The next step was to set the directory in which the TIFF files corresponding to the Quality VI layer are located. After extracting the addresses of each of the files in the list, it was proceeding to work each of them in an iterative structure to determine if the quality of the associated layers has a value greater than 70%. Finally, the directory was set to store the images under the accepted quality value.

In the script was also defined an iterative structure which used the list of addresses of the files with the VI Quality layer obtained in the previous steps, for each VI layer, a raster file was generated and stored in memory, then, a function was defined that It will take the first bits inside the raster, transform them to integers and reverse them. Then, the previous function was used in each raster, thus obtaining the total amount of pixels with a high, medium and low quality. These are pixels with such a high amount of clouds that they cannot be used, and pixels not produced or not usable for other reasons. On the other hand this also determines the amount

of pixels corresponding to land, ocean, land under shallow water (rivers) and land under moderately deep water (lakes).

Once the parameters for each raster previously defined in Table 2 were obtained, it was extracted, transformed to a percentage and adds the values corresponding to cloudy and non-produced pixels. Then a logical expression was used to determine if the previous sum is greater than 30, if it is true, a text file was generated with the date of the VI Quality layer analyzed as its name, and inside it stored the quantity of pixels of high quality, average quality, pixels probably cloudy and unprocessed. Likewise in the following line its equivalent in percentage, the sum of the last two values and finally the exact name of the file containing the VI Quality layer. This text file allows the investigator to know in detail why that particular image was discarded by the script. After this, a PNG image (Figure 2) was generated with the analyzed raster and stored in the specified destination at the beginning of the code.

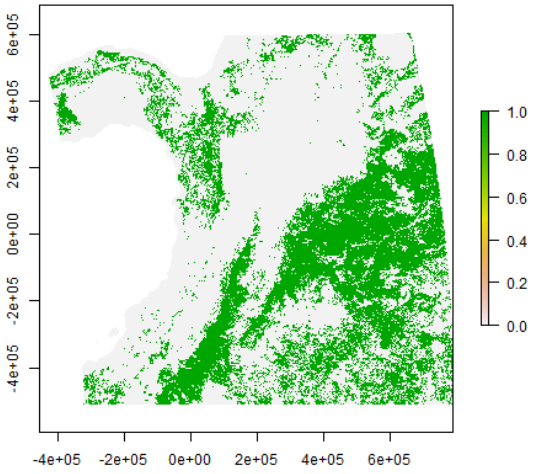


Figure 2: Image generated by the R script of the VI Quality layer on June 10, 2015.

However, due to the size of the MODIS tile and considering that the study region was smaller than the size of the image, it was decided to re-evaluate the images discarded by the script, therefore it was constructed in such a way that it generates a report with the images that did not exceed the minimum of non-usable pixels, besides an image in PNG format with the date of the VI Quality layer analyzed, in which represented in a green color are the clouds present in the region, and a text file with the quality values. Delegating responsibility to the researcher to decide which were still valid and which should definitely be withdrawn from the study.

Sometimes the script discarded the two images obtained in the same month, leaving several months in a year without data to analyze, for this reason the MODIS products generated by the sensor in AQUA satellite were downloaded and used for those special months. This situation had already been foreseen in the first phase of the CRISP-DM methodology, so that a contingency plan was included in the overall planning of the data mining project and it was carried out in this phase.

B) Task 2: named by CRISP-DM and by SEMMA as "Clean data" [3, 4], this task classified the MOD13Q1 products with the sufficient quality to perform studies of land use and cover change, and the products to be discarded. In contrast to SEMMA, in the CRISP-DM methodology, the MYD13Q1 products were established as a contingency plan in view of the risk of running out of products in a particular month. In the following section, the task is presented considering the contingency plan.

Keeping in mind the selection criteria, (percentage of usable pixels, spatial and temporal resolution, and MODIS product type) it was obtained: 746 images, which were valid images only 534 therefore the quantity of discarded images was 212. The time required for this process was 2804 minutes, being the process that required most time during the entire investigation. By year the time used was about 179 minutes; in addition the time for the 18 images of the satellite AQUA was about 147 minutes.

Table 3 and Table 4 present the data cleaning report for two years in particular, highlighting in gray the images that despite reporting a higher percentage of unusable pixels to be accepted can be used in the study region. Also highlighted in orange are the months in which all images were excluded, thus the MYD13Q1 products were downloaded and analyze the result of the process can be seen in Table 5 and in Table 6.

Table 3: Data cleaning process MOD13Q1 between 2014 and 2015 (part 1)

Item	Name	Date
1	2014-02-18.250m_16_days_VI_Quality.tif	2014-02-18
2	2014-03-06.250m_16_days_VI_Quality.tif	2014-03-06
3	2014-04-07.250m_16_days_VI_Quality.tif	2014-04-07
4	2014-05-25.250m_16_days_VI_Quality.tif	2014-05-25
5	2014-06-10.250m_16_days_VI_Quality.tif	2014-06-10
6	2014-08-13.250m_16_days_VI_Quality.tif	2014-08-13
7	2014-09-30.250m_16_days_VI_Quality.tif	2014-09-30
8	2014-10-16.250m_16_days_VI_Quality.tif	2014-10-16
9	2014-11-01.250m_16_days_VI_Quality.tif	2014-11-01
10	2015-03-06.250m_16_days_VI_Quality.tif	2015-03-06
11	2015-03-22.250m_16_days_VI_Quality.tif	2015-03-22
12	2015-04-07.250m_16_days_VI_Quality.tif	2015-04-07
13	2015-04-23.250m_16_days_VI_Quality.tif	2015-04-23
14	2015-05-09.250m_16_days_VI_Quality.tif	2015-05-09
15	2015-05-25.250m_16_days_VI_Quality.tif	2015-05-25
16	2015-06-10.250m_16_days_VI_Quality.tif	2015-06-10

Task 3: named by CRISP-DM as "Construct data" [3], after completing this task, the structure of the tables that store the data resulting from the previous task were obtained. For SEMMA, this task is called "Construction of the repository" [4] and actually covers in its development in the fourth task of CRISP-DM, because it not only identifies the structure of the repository, it also builds it. In CRISP-DM that process is divided into two tasks, therefore the description of the third task of SEMMA corresponds to two sections dedicated to third and fourth task of CRISP-DM.

PostGIS was used for the construction of the repository, because is a spatial extension of PostgreSQL, that in addition to be a spatial database compatible with "OpenGIS Simple Features for SQL" [21], it allows us to manage spatial objects (roads, rivers, forests, etc.) as objects of the database, allowing to calculate relationships between objects that are very difficult to model without using spatial objects [21], such as proximity, adjacency, containment, among others. The database was created in PostgreSQL, however a PostGIS extension must be added to the DBMS so that it recognizes, e.g., raster type attributes.

Table 4: Data cleaning process MOD13Q1 between 2014 and 2015 (Part 2)

Item	Pixel produced, but most probably cloudy		Pixel not produced due to other reasons than clouds		No usable pixels (%)	Exclude
	Quantity	(%)	Quantity	(%)		
1	8135534	36.6	1319373	5.9	42.5	Si
2	8618282	38.7	1318783	5.9	44.6	Si
3	6370910	28.6	1320358	5.9	34.5	Si
4	6102452	27.4	1320584	5.9	33.3	Si
5	6627708	29.8	1317959	5.9	35.7	No
6	7669296	34.5	1319543	5.9	40.4	Si
7	5441419	24.5	1318093	5.9	30.4	Si
8	5706722	25.7	1318820	5.9	31.6	No
9	5596434	25.2	1320674	5.9	31.1	Si
10	6657345	29.9	1305071	5.9	35.8	Si
11	6245216	28.1	1304754	5.9	34	Si
12	9507546	42.7	1302703	5.9	48.6	Si
13	5593483	25.1	1304982	5.9	31	No
14	6682094	30	1304743	5.9	35.9	Si
15	9714449	43.7	1305965	5.9	49.6	Si
16	5624849	25.3	1305321	5.9	31.2	No

Table 5: Data cleaning process MYD13Q1 (part 1)

Item	Name	Date
1	2015-03-14.250m_16_days_VI_Quality.tif	2015-03-14
2	2015-03-30.250m_16_days_VI_Quality.tif	2015-03-30
3	2015-05-01.250m_16_days_VI_Quality.tif	2015-05-01
4	2015-05-17.250m_16_days_VI_Quality.tif	2015-05-17

Table 6: Data cleaning process MYD13Q1 (part 2)

Item	Pixel produced, but most probably cloudy		Pixel not produced due to other reasons than clouds		No usable pixels (%)	Exclude
	Quantity	(%)	Quantity	(%)		
1	7161513	32.2	1321807	5.9	38.1	Si
2	7846294	35.3	1320976	5.9	41.2	Si
3	5662943	25.5	1317536	5.9	31.4	No
4	7794798	35	1324353	6	41	Si

Considering that database will be filled with identical information, it was inappropriate to create a single table that contains them; the best way to handle large amounts of data is by partitioning tables. In order to perform the partitioning of tables, a parent table must be created, which will later inherit the characteristics of the child tables [22]. For the case study corresponds to one table per year and vegetation index.

Task 4: named by CRISP-DM as "Integrate data" [3], this task allows the combination of data resulting from previous tasks. For the case study it was necessary to use the structures defined in the repository construction as described below.

To create the database, a script was developed based on Golicher's approach [18], the Python-built script for this work allowed defining one table per year, inheriting the characteristics of the main table, and allowing the entry of a layer in TIF format for each table depending on the year in which it was obtained. At the end of the process, 15 tables were obtained from the NDVI vegetation index and another 15 tables from the EVI vegetation index. Finally, a total of 32 tables were obtained, each with an average number of records of 1517.

In the script used, the corresponding libraries to handle SQL statements and file handling were loaded. Later, it selected the folder in which the files for the NDVI or EVI vegetation index were stored, then a list was extracted with all the files and was classified according to the year of each layer, finally, an iterative expression was built which runs the data by each year, beginning in the year 2000 and ending in the year 2015. In the iterative structure a new table was created with the name of the vegetation index and the respective year, a restriction of dates in such a way as prevent entering different values than the selected year. The other attributes are inherited from the main table, finally, it was proceeded to define another cycle that will be responsible for filling each record of the table with the corresponding indices (usually two for each month). In the next iterative expression, each product NDVI or EVI was evaluated, creating a temporary table with the raster of each one, for this the command program "raster2pgsql" offered by PostgreSQL was used, including the parameters and access data for the required database. Then, a new attribute of date type was added to the temporary table to insert the exact date of the MODIS product. Data of the temporary table was transferred to the table defined in the previous cycle. Finally, if it leaves the iterative structure, it means that for that particular year, all the corresponding layers have already been registered; therefore an index can be defined per year.

5. Conclusion

The data cleaning process for MODIS products in studies of land use and cover change using the CRISP-DM methodology, was easier than SEMMA, mainly because CRISP-DM is presented as a true methodology of data mining with detailed phases, tasks and activities. Also, it has documentation produced by typical case studies which the methodology was applied; guaranteeing that CRISP-DM methodology for the cleaning of the MODIS products would work without any problems. Another important point in the success of the application of CRISP-DM was the preliminary tasks that clearly defined the process, including predictions about possible problems in the data cleaning stage.

It should also be noted that the use of R scripts helped and optimized the process of cleaning, downloading and reprojection of MODIS products, because the time invested in each phase of the methodology was smaller than the one invested on the traditional way. It can also be concluded that managing a high level for understanding the problem was a key to build the MODIS repository for studies of land use and cover change, because it defined a clear and concrete mining goals, where the CRISP-DM methodology dedicates an entire phase to transform the problem in terms of data mining.

On the other hand, SEMMA is designed to work with the SAS® Enterprise Miner™ tool, and all available documentation focuses on this tool, therefore, when working with non-typical mining cases as the case study presented in this article, the amount of SEMMA disadvantages increase, demonstrating that SEMMA is not suitable for applying in a similar studies as presented in this manuscript. In contrast, CRISP-DM methodology is recommended when attempting to perform a data mining project, because it has all the available documentation, detailed phases, tasks and activities, and the development of the first phase, that facilitate the problem understanding and its transforming to a data mining problem. It is a better approach if the knowledge of the

problem in terms of business is insufficient and if the problem does not fit within the typical cases solved with mining techniques.

For future work, it is recommended to use all available MODIS products obtained from the sensor in the TERRA satellite and from the sensor in AQUA satellite. Both satellites combined generate four images every month, however, according to the results it was common that in the months with worse weather in Colombia, sometimes all the images were discarded by excessive cloudiness for those months.

It is also worth noting that the time for the data cleaning process was quite long, mainly due to the physical characteristics of the computer used, then, the best way to optimize the process would be through a computational cluster of high performance, improving computing capacity and it would significantly reduce the total time invested at the end of the data cleaning process, and in general of the whole study. In the same way, it is suggested to optimize the scripts used in the research, making them more efficient, allowing to analyze all the required years in a single execution, and to facilitate the inclusion of different phases in sequence and in parallel. Finally, a suitable graphical interface for scripts would be very helpful for researchers who need them, facilitating their use and improving their "usability".

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The authors wish to express their gratitude to department of systems engineering and to department of environmental engineering of Mariana University, also to all the individuals who have been sources of help throughout the research period.

References

[1] H.J. Gómez, et al "Comparativa entre CRISP-DM y SEMMA para la limpieza de datos en productos MODIS en un estudio de cambio de cobertura y uso del suelo" in Computing Conference (CCC), 2016 IEEE 11th Colombian. IEEE, 2016. p. 1-9.

[2] Gondar Nores, Jose Emilio. Redes Neuronales Artificiales, Data Mining Institute, 2001

[3] P. Chapman, "CRISP-DM Step by Step Data Mining," 76 p, 1999.

[4] Data Mining and the Case for Sampling. Solving Business Problems Using SAS® Enterprise Miner™ Software. SAS Institute Inc. 36 p, 1998.

[5] P. Vázquez, E. Adema, y B. Fernández, "Dinámica de la fenología de la vegetación a partir de series temporales de NDVI de largo plazo en la provincia de La Pampa," pp 1-2, Agosto 2013.

[6] J. Francois M., "Aplicaciones del sensor MODIS para el monitoreo del territorio," pp 16-98, 2011.

[7] R. Solano, K. Didan, M. Barreto y A. Huete, "MODIS Vegetation Index User's Guide (MOD13 Series)," The University of Arizona, version 3.0 Junio 2015.

[8] V. Torres, P. Paredes, y P. Rial, "Análisis de la anomalía del índice de vegetación mejorado (EVI), del sensor MODIS, para la provincia de Santa Cruz," pp 1-2, 2010.

[9] D. Armenteras, F. González, y K. Franco A., "Distribución geográfica y temporal de incendios en Colombia utilizando datos de anomalías térmicas," Bogotá, 16 p. 2009.

[10] J. García G., "Remote Mining: Aplicando minería de datos a teledetección sobre LIDAR," España, 86 p. 2010.

[11] A. Sánchez, A. Etter, y M. CLARK, "Land Cover Change in Colombia: Surprising Forest Recovery Trends between 2001 and 2010," 16 p. Agosto 2012.

[12] H. Camargo, y M. Silva, "Dos caminos en la búsqueda de patrones por medio de minería de datos: SEMMA y CRISP," Rev. Tecnol. – Journal of Technology • Volumen 9 No. 1. Bogotá, 8 p. 2011

[13] J. Moine, S. Gordillo, y A. Haedo, "Análisis comparativo de metodologías para la gestión de proyectos de minería de datos," CACIC Argentina, 8 p. 2011.

[14] A. Azevedo, y M. Santos, "KDD, SEMMA AND CRISP-DM a parallel overview," AIDIS Amsterdam, 5 p. 2008.

[15] B. Naimi, "ModisDownload: an R function to download, mosaic, and reproject the MODIS images. R-GIS," 2015.

[16] USGS Earth Resources Observation and Science (EROS) Center, Land processes DAAC, "MODIS reprojection tool user's manual," pp 42-54 Abril 2011.

[17] Instituto Geográfico Agustín Codazzi, "Estudio general de suelos y zonificación de tierras Departamento de Nariño. Capítulo 8: Zonificación del medio biofísico," pp 251-253, 2015.

[18] D. Golicher, "Loading Modis NDVI time series into PostGIS raster," 2012.

[19] C. Mulleady, y D. Barrera, "Estimación de la tasa de evapotranspiración a partir de datos satelitales," 2013. Meteorologica vol.38 no.1 Ciudad Autónoma de Buenos Aires ene./jun. 2013.

[20] M. Lydholm R., "Geographic Information Systems. StackExchange: Parse modis MOD13Q1 quality layers in R," 2015.

[21] PostGIS 2.0 Manual, SVN Revision (14671), pp 7, 27-29. 2015.

[22] The PostgreSQL global development group. PostgreSQL 9.0.22 Documentation. Cap. 5: Partitioning, 2015.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 605-611 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Power requirements and battery life measurement for wireless transmission between two nodes in different mediums

Radouane Karli[1], Abdellatif Bouchalkha[2,*], Khalid Alhammadi[1]

[1]*Electrical Engineering Department, Khalifa University of Science and Technology, P.O. Box 127788, Abu Dhabi, United Arab Emirates*

[2]*Physics Department, Khalifa University of Science and Technology, P.O. Box 127788, Abu Dhabi, United Arab Emirates*

ARTICLE INFO

Article history:
Received : 05 April, 2017
Accepted : 12 May, 2017
Online: 04 June, 2017

Keywords:
Wireless sensor
Battery life
Power consumption
ZigBee terminal

ABSTRACT

One of the most important roles of the wireless sensor networks (WSN) is to avoid wiring costs, be self-sustainable and be able to function for several years. However, due to the slow progress in battery technology, power continues to be a limited resource in wireless sensor communication and electric energy storage remains to be an important issue. On the other hand, if batteries must be replaced often, many remote sensing applications may become impractical. Therefore, batteries with long life on the order of several years are needed. This paper is an extension of work originally presented in The 5th International Conference on Electronic Devices, Systems and Applications to investigate further the power requirements for wireless data transfer between two nodes using batteries with different capacities (55 mAh, 550 mAh and 5500 mAh). In particular, the effect of a propagation medium such as air, distilled water and engine oil on the wireless communication inside a one meter long metallic pipe was investigated. Our first result shows a successful transmission of wireless signal through air, distilled water and oil medium with very low transmission losses. The second result shows that an increase in the battery capacity will increase the two-node wireless sensor operation time even in different propagation medium. This result can be used to determine the required battery capacity for extending the WSN operation time.

1. Introduction

Pipelines in oil and gas industries constitute generally the most economical way to transport large quantities of oil, refined oil products or natural gas. The management of oil and gas transport in a pipeline is a challenging task with the rapid growth of the required length of pipelines. Challenges such as oil and gas leaks, pipeline corrosion, wax precipitation and environmental pollution are treated with highest priority in oil industries. Hence, the most common and important applications of WSN in oil industry are related to real-time monitoring of the production performance through process control, safety, and rigorous maintenance operations [1, 2].

Because of its crucial importance, the continuous operation of a WSN is essential and must be maintained at all time. The lifetime of a WSN can be defined as the life of the shortest living node in the network. However, depending on the application, the density of the network, and the possibilities of the network reconfiguration, it can also be defined as the life of some other (main or critical) nodes [2]. In order to prolong a WSN life, it is required to reduce the energy consumption of the nodes as much as possible [3].

However, for many important WSN applications, such energy option is not feasible, and specific power-management strategies are necessary for WSN nodes that are powered by non-rechargeable batteries. Energy consumption still remains one of the main obstacles to the development of WSN technology, especially in cases where long and reliable network operation is required [3].

In most cases once the WSN has been deployed, it is impossible to replace each and every battery present in the network after discharge. One important challenge in the design of a large WSN is energy efficiency where the design of a WSN should extend its life without sacrificing its reliability [4].

*Corresponding author: Abdellatif Bouchalkha, Physics Department, Khalifa University of Science and Technology, P.O. Box 127788, Abu Dhabi, United Arab Emirates | E-mail: abouchalkha@pi.ac.ae

To reduce the overall energy consumption of the network, one should use some power management (PM) techniques at the level of the sensing, communication, or processing unit. Another option is to consider using some energy harvesting techniques to extend the WSN operation time [5].

The main goal of this paper is to study the power consumption of a two-node wireless sensor communication for batteries with different capacities. Two experiments were considered for this study. The first one was the free space wireless communication in open air without metallic pipe and the second one was the wireless communication inside a metallic oil pipe acting as a waveguide. The measurement results showed that the data can be transmitted successfully through air inside a pipe with very low transmission loss. It was also shown that an increase in the battery capacity increases the two-node wireless sensor operation time. This result can be used to determine the required battery capacity for extending the wireless sensor system life time.

2. Linear Network

The architecture and proposed work for a general WSN may not be suitable for linear WSN applications such as monitoring pipelines. This requires a good review and a thorough study of the applications needing linear WSN to determine the specific design requirements for it.

There are several advantages for using linear alignment of nodes to provide protection and monitoring of linear topology such as oil, water, and gas pipelines. Some of these advantages include faster and less costly network deployment in addition to the ability to introduce flexible multi-hop routing which can overcome intermediate node failures.

In a linear distributed detection system, an initial sensor starts the detection process and make decisions then transmit both the data and the decisions to its neighbor as shown in Figure 1. Based on the received decisions from their predecessors, the succeeding sensors again form decisions and transmit them together with the data until the final sensor or gateway node is reached [6].

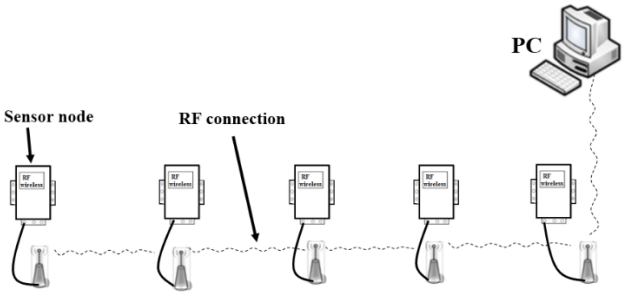


Figure 1.Linear structure of wireless sensor network.

Since the pipeline network extends generally along a line, it is important for the network to be continuously connected to collect and transfer information from the sensor nodes distributed over the pipeline to the control station. It will also be required to transfer control commands to the actor and sensor nodes, which are often located inside the pipeline. The sensor network architectures generally used for reliable communication in pipeline systems are based on wired networks, wireless networks, or a combination of both. Typical applications [7, 8], include the case of sensors embedded in the outer surface of a pipeline. This linear sensor arrangement is the result of their linear topology.

Linear WSN is a special category of WSN where the nodes are placed in a linear form. Kevin et al., [9] gave two different definitions for the network topology model of a linear WSN. The first definition describes it as a connected non-cyclic diagram where each node has exactly one or two distinct neighbors. The two nodes with only one neighbor are referred to as the endpoints of the network. While the second definition presents it as a linear network topology containing N nodes where each node can communicate with other nodes up to a number M of hops [10, 11].

Lin et al., [12] have proposed an application of linear network in power transmission lines. In this case, they used a clustering algorithm to balance energy consumption and leverage hybrid media access control for monitoring the power transmission lines. This proposed technique makes the system complex to implement and less responsive. Akbar et al. [13], in their contribution to linear WSNs, proposed another technique. This consists of mobile sinks, and courier nodes that work underwater in a 3-D linear formation which minimizes energy consumption as well as accurately computing required routing information. Such 3-D linear system is, though accurate in routing, yet also too complex for real-time sensing and monitoring. This is especially the case for the implementation in mission critical applications such as pipeline monitoring.

Linear WSNs may be simplistic in topology but the decision of routing and deployment of nodes to conserve resources is not as simple as it may seem to be. A detailed hierarchical and topological classification and implementation of linear sensor networks is provided by Jawhar et al., and Liu et al., [14, 15]. In both cases it is difficult to implement in pipeline monitoring due to the surrounding harsh environment. Therefore, certain design requirements and research issues need to be addressed and investigated, given that few research papers have discussed the special requirements for linear WSN. These requirements have strong impact on the network performance.

On the other hand, extending the network lifetime is still a critical issue in many WSN cases. Different research groups have discussed this issue [16, 17], from the point of view of network power management software and protocol but not in terms of extending battery lifetime.

In this paper, we investigate the wireless transmission between two consecutive nodes to address the issues related to wireless communication in a linear WSN for long pipelines. The results of this work are very useful and can be extended to a multi-node linear system in the future.

3. Battery life

The two most important properties of a battery are its voltage level and its capacity (expressed in Ampere-hour, Ah) which represents the amount of energy stored in a battery. For an ideal battery, the voltage stays constant over time until the moment when it is completely discharged. Ideally, the battery's capacity remains the same for any load. However, in a real situation, the voltage of a battery drops during the discharging phase and the

effective capacity of a real battery is lower under a higher load. In this case, the high current draw decreases the battery capacity in a nonlinear behavior. In the case of an ideal battery, it would be easier to calculate the battery life L in the case of a constant load as follows [18, 19].

$$L = \frac{C}{I} \quad (1)$$

Where C is the battery capacity and I is the load current.

The above simple relation does not hold for predicting the life of a real battery due to the various nonlinear effects neglected. A better and still simple approximation for the battery life calculation under constant load can be obtained from Peukert's law. The later gives an expression for the capacity of a battery in terms of the rate of discharging. As a result, when the rate of discharging increases, the battery's available capacity decreases. The formula for Peukert's law can be expressed in different forms with the most common one used is shown below [19, 20]:

$$L = H\left(\frac{C}{IH}\right)^{k} \quad (2)$$

Where;

L – The time, in hours, that the battery will last for a given rate of discharge.

H – The discharge time in hours.

C – The battery capacity in Amp-hours based on a specified discharge time.

I – Current in Amp.

k – Peukert exponent parameter which is a unique number for each battery.

4. Experimental setup

4.1. Wireless communication experiment

The wireless transmission experimental setup is shown in Figure 2. This setup consists of two wireless communication nodes using a fixed digital data set provided by an analog input voltage. The two nodes are standard ZigBee terminals powered by a lithium battery and communicating in the frequency range 2.4-2.5 GHz. The digital data is transmitted by the transmitter terminal and received by the receiver terminal which is interfaced to a PC through a National Instrument NI-ELVIS DAQ. An IP Modem Configure-F8114 and the SScom32E software were used for collecting data on the computer. The principle of operation is that the PC sends a command to the wireless ZigBee terminal F8114 connected through RS232 to a PC to communicate with wireless terminal F8914 and read its digital output data. The communication was established via simple dipole antennas to ensure unidirectional transmission between the two terminals. These directional antennas help reduce the effects of interference as well as extend the communication range between the two terminals. The dipole antennas have an impedance of 50 Ω, a gain of 3 dB and a length of 125 mm. The data is then processed by the PC which is connected as shown in Figure 2.

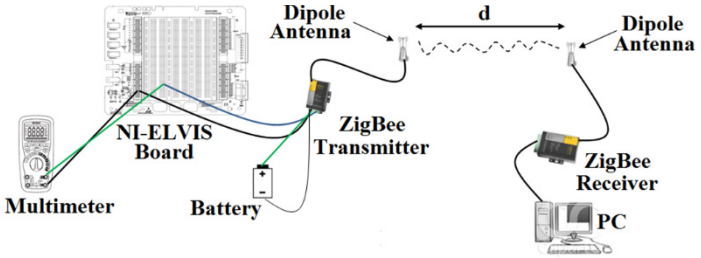
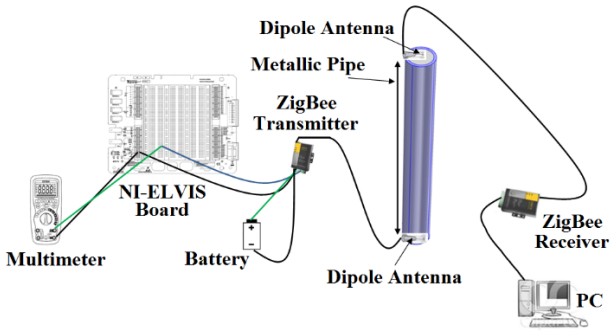


Figure 2. Wireless communication setup

The transmitter/receiver terminal uses the standard ZigBee communication protocol which is based on offset quadrature phase shift keying to transmit digital data [21]. The acquired data was in hexadecimal (HEX) format which was then converted to decimal (DEC) format. The corresponding value was then calculated, as described in the manual of the F8914 ZigBee Terminal.

For the study of the effect of a different medium on the wireless communication, we used a one meter long hollow metallic pipe made of iron as shown in Figures (3a) and (3b). We have also used the pipe as a waveguide for transmitting and receiving data wirelessly between the two dipole antennas driven by the two ZigBee terminals. The diameter and the thickness of the pipe were respectively 0.1 m and 0.003 m. The metallic pipe was placed between the two antennas and filled, alternatively, with distilled water or engine oil. The initially analog voltage data was transformed to ZigBee format and sent through these different mediums. The received values for air, distilled water, and oil mediums were recorded and compared to the transmitted values as will be shown later in Figure 5. The temperature was also kept constant at approximately 21±0.5 °C throughout the experiment.



(a)

(b)

Figure 3. Experimental setup for wireless communication in a mettalic pipe as a diagram (a) and the actual laboratory setup (b).

4.2. Wireless node power requirement experiment

A common way of specifying the energy capacity of the wireless sensor is to provide the battery capacity as a function of elapsed time until it is fully discharged. The experimental setup used in this part is shown in Figure 4.

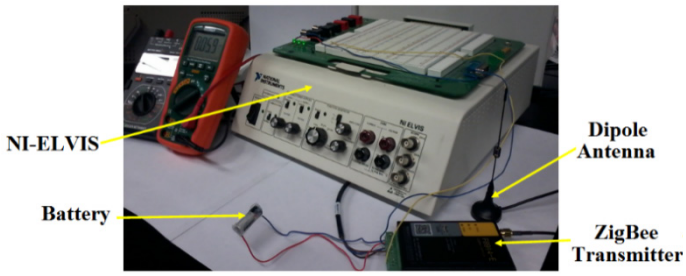


Figure 4. The battery voltage measurement setup.

In this case, we maintained a good communication between the transmitter and receiver terminals but measured the load voltage of the battery connected to the transmitter as a function of the elapsed time. Three different batteries with different nominal capacities (*55 mAh, 550 mAh and 5500 mAh*) were used in our experiment. The power level of the transceivers was kept constant at 2.82 mW and the distance between the transmitter and receiver was fixed to either 1 m or 9 m.

The analog voltage value was converted to digital value first then transmitted wirelessly. The transmission was done continuously and after each transmission the battery voltage was recorded. The battery voltage level measurements were conducted at regular time intervals at the transmitter side. The communication between the emitter and the receiver was maintained and checked before each measurement to make sure it is successful.

5. Results and discussions

5.1. Two-nodes communication in different mediums

The results from the wireless communication experiment described earlier are shown in Figure 5. This experiment was conducted to establish wireless communication and compare the transmitted and received data between the two ZigBee terminals. From the figure it can be clearly seen that the voltage data was successfully transmitted through the three different mediums namely air, distilled water and engine oil with very low relative transmission losses. It can also be seen from Figure 5 that the wireless transmission is independent of the medium inside the pipe and the digital data was well received as expected and confirmed. This is a good confirmation that digital data transmission has experienced practically no losses even through a wide range of different propagation mediums such as air, distilled water and engine oil.

In oil wells, we are always monitoring the temperature and the pressure downhole as the most critical parameters. In general, the temperature is high and can easily reach 150 °C. In order to validate the above results, we have adjusted the previous experimental setup to transmit and receive real measured temperature data. The experimental setup used is shown in Figure 6 with wireless transmission through the 1 meter metallic pipe. In this case we have replaced the constant voltage value by a temperature sensor immerged in a variable temperature water bath. The temperature sensor used is a type K thermocouple (Chromel/Alumel) with a temperature range of -30 °C to 300 °C and an output voltage range of -6.4 mV to 54.9 mV.

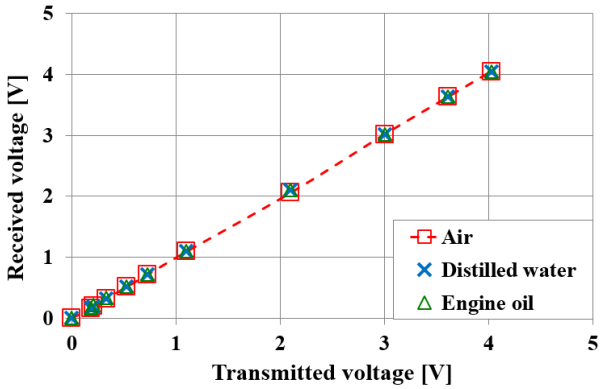


Figure 5. Comparison between the wirelessly transmitted and received voltage in a 1 meter long mettallic pipe for the different cases indicated.

The multimeter used is capable of measuring very low voltages across the two wires of the thermocouple as shown in Figure 6. The measured voltage was multiplied by the Seebeck constant to get the corresponding temperature [22]. This temperature was recorded as the input temperature. Since the thermocouple voltage was very low, it was fed to an AD620 instrumentation amplifier before the wireless transmittion and further processing. The received wireless signal is transformed back to temperature data and the results are shown in Figure 7.

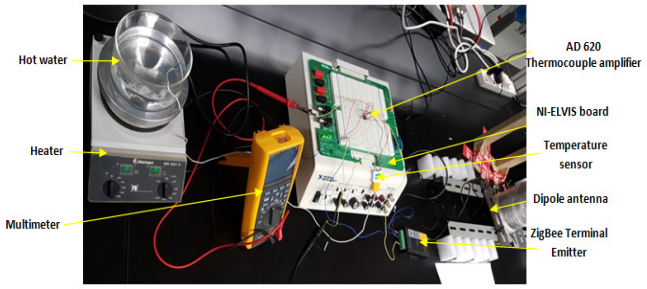


Figure 6. Wireless communication setup with Temperature sensor (Thermocouple K)

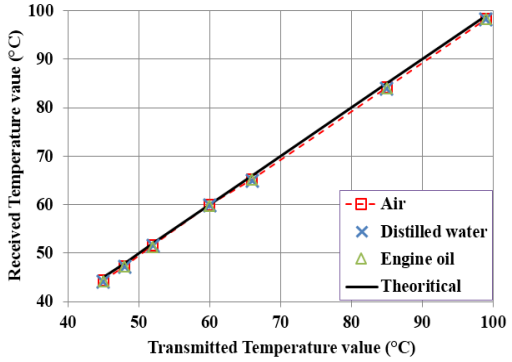


Figure 7. Comparison between the wirelessly transmitted and received temperature data in a 1 meter long mettallic pipe for the different cases indicated.

From Figure 7, it can be seen that the temperature data is successfully transmitted through air, distilled water and oil mediums with very low transmission loss. The maximum

temperature limit of 100 ºC is due to water evaporation temperature. Higher temperatures could be reached with other fluids such as oil if needed. From the graph we can estimate the error on the transmitted temperature to be about 0.7 % at 100 ºC. This is very close to the expected K-type thermocouple uncertainty of 0.75 % at the same temperature. In comparison with the previous experiment, using constant voltage, where the error was negligible, we can conclude that the observed error in transmitting temperature data was mainly introduced by the thermocouple.

5.2. Battery life measurements

For the power requirement and battery life study, we consider the first results using a single battery B1 with a nominal capacity of 55 mAh. Figure 8 shows a comparison between the battery voltage level as a function of continuous wireless transmission time and distance between the nodes for the battery B1. The wireless transmission was conducted in air with and without a metallic pipe. Using the Terminal ZigBee powered by B1 with a distance of 1 m between the transmitter and receiver, we observed that the battery B1 had lasted for about 79 hours of continuous transmission. This is in good agreement with the value of 81 hours listed by the manufacturer. This is due to the special cutoff voltage of 8.8 V below which this battery is considered to be non-reliable.

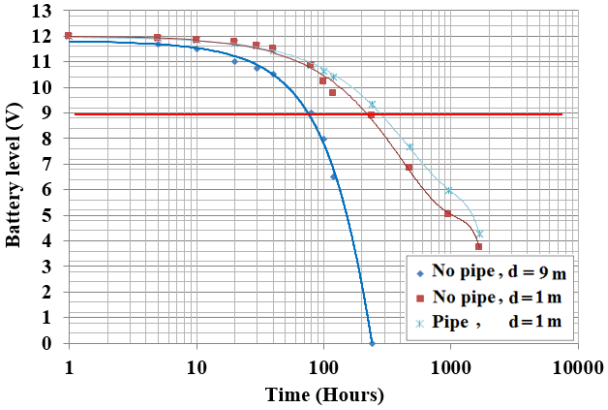


Figure 8. Comparison between the battery voltage level as a function of continuous wireless transmission time and distance between the nodes for the battery B1 with 55 mAh nominal capacity . The transmission was conducted in air with and without a metallic pipe.

It is also observed from Figure 8 that the distance between the nodes has an important effect on the data transmission. From the figure we can see that when the distance was changed from 9 m to 1 m without using the metallic pipe, the battery lasted for about 250 hours as opposed to 79 hours with the 9 m distance. This increases the battery life by a factor of 3. On the other hand, we noticed that using the metallic pipe as a cylindrical waveguide has also enhanced the battery life. This is due to the reduced attenuation effect from the open environment and the enhancement due to the waveguide effect making the transmission more directional. In this case according to Figure 8, the battery life measured for the wireless transmission in the metallic pipe under the same conditions was about 300 hours. Comparing the two results we can see that we have achieved a battery life enhancement of at least 20 % for this particular case. This is quite encouraging and will be investigated and optimized further in our future work.

In order to investigate further the battery life as a function of its capacity, we have chosen to conduct additional experiments with two other batteries B2 and B3 with higher nominal capacities of 550 mAh and 5500 mAh respectively. These experiments were conducted in open air for batteries B1, B2, and B3 as indicated in Table 1. The temperature of the environment was kept the same at around 21±0.5 °C and the distance between the transmitter and receiver was fixed to 9 m.

Table 1. Battery life measured at the cutoff voltage of 8.8V with B4 and B5 as indicated.

Battery	Nominal capacity (Ah)	Battery life (years)
B1	0.055	0.009
B2	0.55	0.02
B3	5.5	0.16
B4 [23]	34	1
B5 [24]	700	20

The battery voltage level measurements at the wireless transmission node were recorded as a function of the continuous transmission time for the three different batteries B1, B2, and B3. The results are shown in Figure 9 with a cutoff voltage of 8.8 V as discussed earlier. From the graph we immediately notice that the battery life was longer for the higher capacity batteries as expected and confirmed by our measurements. We also notice that the measured battery voltage decreases more slowly for the batteries with larger capacities (B2 and B3) hence providing longer battery life according to the indicated cutoff voltage of 8.8 V.

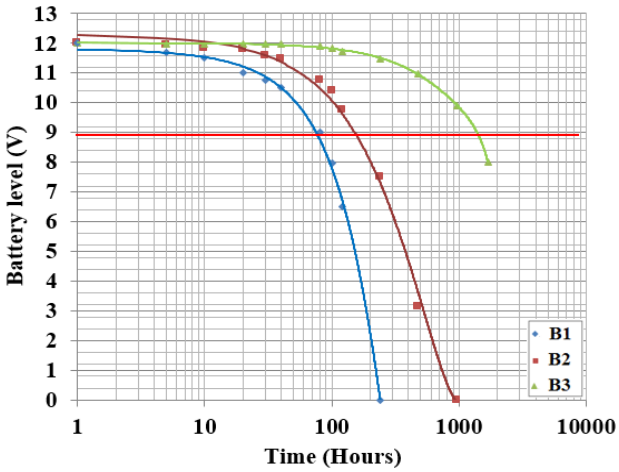


Figure 9. Comparison between the battery voltage level as a function of continuous wireless transmission time for the batteries B1, B2, and B3 with nominal capacities 55, 550, and 5500 mAh respectively, in open air for a fixed distance of 9 m between the two nodes.

For the battery B2 with nominal capacity 550 mAh, it's observed that the life of transmission is about twice greater than that of the battery B1 with nominal capacity 55 mAh. In addition, we found that the battery B2 lasted for about 170 hours of continuously transmitting data before reaching the minimum required cutoff voltage of 8.8V. For the third battery B3 with the nominal capacity of 5500 mAh, it was observed that after 1400 hours, the transmitter has consumed only 17 % of the battery's capacity. In Table 1 we report the measured battery lives of the three different batteries obtained from Figure 9 at the cutoff voltage of 8.8 V.

The results in Table 1 were used to obtain the very useful graph shown in Figure 10. Note that the data for batteries B4 and B5 were obtained from the literature [23, 24]. The curve in Figure

10 is obtained from Equation (2) with I = 2.3 mA and k = 1.25 as the average value for most lead acid batteries [25]. Notice the good agreement between the measured and calculated battery life.

The graph in Figure 10 can be used to predict and estimate the remaining battery life at any moment in time during battery operation. This is quite useful for WSNs that require long-term operation with single batteries. As shown in Figure 10, the requirement for a specific battery life can also be determined. For example, a 15 years operation time can be reached with a 500 Ah capacity battery while a 10 years operation time requires a 320 Ah capacity battery as shown in the figure below. From the literature review and the available products in the market, these battery lifetimes are now accessible and currently used in certain applications. This, certainly, makes our predictions very useful to the users of these batteries. With our plot in Figure 10, one can estimate the remaining time for these long-life batteries and plan the required battery replacements.

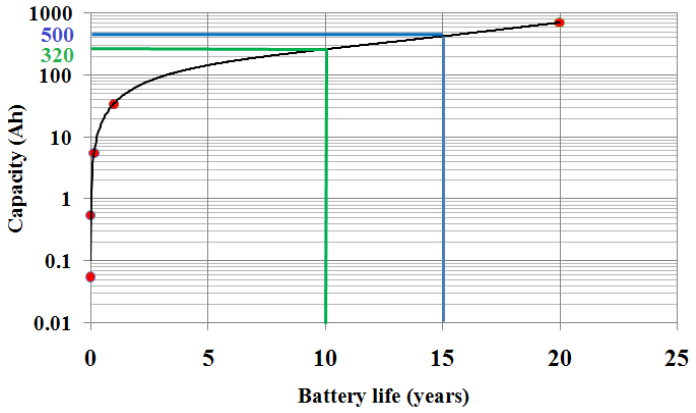


Figure 10. A plot of the different battery capacities as a function of their corresponding life. The line is a theretical curve fit as discussed in the text.

In summary, we conclude that the transmitting WSN nodes will perform well with long life batteries and small discharge time such as the case of the high capacity batteries. Also, from the measurement results, the digital data is successfully transmitted through air, distilled water, and engine oil mediums with very low transmission loss relative to air. We can conclude and confirm that the wireless transmission is well established and was not affected much by the two different mediums considered in our study which are distilled water and engine oil compared to air. We have also shown that the wireless transmission in a metallic pipe contributed at least 20 % more to the battery life of the wireless transmitter according to our current experimental setup. On the other hand, our results are more general and could be adapted to different types of wireless sensor networks to predict and monitor the required battery power.

6. Conclusion

In this paper, we have presented the study of a two-node wireless communication considering both the medium of propagation and the power requirements. The two terminals used were ZigBee terminals with dipole antennas placed at the ends of a metallic pipe filled with air, distilled water or engine oil. Our experimental results showed that efficient communication between the two wireless nodes was well established in air, distilled water and engine oil at the standard communication band of 2.4-2.5 GHz.

We have also presented the wireless sensor power requirement study and the required battery life for sustained wireless communication of a continuously transmitting node. The study of the battery life allowed us to estimate the required battery capacity for a specific long-term wireless sensor operation. For example, for a 15 years operation time one needs a battery of at least 500 Ah capacity. This estimation can be conducted for any specific wireless node power requirement. This study enables predictions concerning the service life for long-term use of batteries. The experimental results showed that the rate of energy discharge varies from one battery to another and can be quite long for large capacity batteries. When the battery capacity is large, it ensures good life of the wireless sensor transmission and the WSN.

This work is the first phase of a big project that will be continued and implemented. We would expect to extend our work in the future to include other components such as temperature sensors, pressure sensors, gateways and other models for wireless sensor networks. This will be mainly targeting the oil pipeline monitoring for addressing specific flow assurance challenges. We will also investigate further the 20 % battery life enhancement observed in the case of the guided wave conditions that we setup in the metallic pipe.

Acknowledgment

This work was funded through the Project No. *RDProj.041-FA* by Abu Dhabi National Oil Company (ADNOC), R&D, Abu Dhabi, United Arab Emirates (UAE).

References

[1] R. Karli, A. Bouchalkha and K. Alhammadi, “Power consumption and battery life study of a two-node wireless sensor system”, The 5th International Conference on Electronic Devices, Systems and Applications, December 6-8, Ras Al Khaimah, UAE, 2016.

[2] M. A. Akkaş, R. Sokullu and A. Balcı,”Wireless Sensor Networks in Oil Pipeline Systems Using Electromagnetic Waves”, The 9th International Conference on Electrical and Electronics Engineering (ELECO), 2015.

[3] A. Moschitta and I. Neri, Power consumption Assessment in Wireless Sensor Networks, ICT - Energy - Concepts Towards Zero - Power Information and Communication Technology: Dr. Giorgos Fagas (Ed.), InTech, DOI: 10.5772/57201, 2014.

[4] Y. Chen and Q. Zhao, “On the Life of Wireless Sensor Networks”, IEEE Communications Letters, Vol. 9, Issue 11, pp. 976–978, 2005.

[5] I. Dietrich and F. Dressler, “On the life of wireless sensor networks”, ACM Transactions on Sensor Networks, Vol. 5, issue 1, pp. 1-38, 2009.

[6] Gernot Fabeck, Rudolf Mathar, “Optimization of Linear Wireless Sensor Networks for Serial Distributed Detection Applications”. Vehicular Technology Conference (VTC 2010-Spring), IEEE 71st, 2010.

[7] S. Yoon, W. Ye, J. Heidemann, B. Littlefield, and C. Shahabi, “SWATS: Wireless sensor networks for steamflood and waterflood pipeline monitoring,” Netw. IEEE, Vol. 25, issue 1, pp. 50–56, 2011.

[8] T. T.-Te Lai, W.-J. Chen, K.-H. Li, P. Huang, and H.-H. Chu, “Triopusnet: Automating wireless sensor network deployment and replacement in pipeline monitoring,” ACM/IEEE 11th International Conference on Information Processing in Sensor Networks (IPSN), pp. 61–71, 2012.

[9] K.J. Henry, D.R. Stinson, “Resilient aggregation in simple linear sensor networks”, IACR, Cryptology ePrint Archive, 2014.

[10] S. Ali, S. Bin Qaisar and E. A. Felemban, “Dimensioning of linear and hierarchical wireless sensor networks for infrastructure monitoring with enhanced reliability”, KSII Transactions on internet and information systems, Vol. 8, No. 9, Sep. 2014.

[11] Changqing Xia, Wei Liu, and Qingxu Deng, “Cost Minimization of Wireless Sensor Networks with Unlimited-lifetime Energy for Monitoring Oil Pipelines”, IEEE/CAA Journal of Automatia Sinica, Vol. 2, No. 3, 2015.

[12] J. Lin, B. Zhu, P. Zeng, W. Liang, H. Yu, and Y. Xiao, "Monitoring power transmission lines using a wireless sensor network," Wireless Commun. Mobile Comput., Vol. 15, No. 14, pp. 1799—1821, 2015.

[13] M. Akbar, N. Javaid 1, A. Hussain Khan, M. Imran, M. Shoaib, and A. Vasilakoset al., "Efficient data gathering in 3D linear underwater wireless sensor networks using sink mobility," Sensors, Vol. 16, No. 3, pp. 404–431, 2016.

[14] I. Jawhar and N. Mohamed, "A hierarchical and topological classification of linear sensor networks," in Proc. IEEE Symp. Wireless Telecomm., pp. 1–8, 2009.

[15] Y. Liu, J. Feng, O. Simeone, J. Tang, A. M. Haimovich and M. C. Zhou, "Topology discovery for linear wireless networks with application to train backbone inauguration," IEEE Transactions on Intelli. Transp. Syst., Vol. 17, No. 8, pp. 2159–2170, 2016.

[16] Xia C Q, Guan N, Deng Q X, Yi W. Maximizing lifetime of threedimensional corona-based wireless sensor networks. International Journal of Distributed Sensor Networks, Article ID 149416, 2014.

[17] Keskin M E, Altinel K, Aras N, Ersoy C. Wireless sensor network lifetime maximization by optimal sensor deployment, activity scheduling, data routing and sink mobility. Ad Hoc Networks, 17, pp. 18−36, 2014.

[18] M. R. Jongerden and B. R. Haverkort, "Which battery model to use?" IET, Vol. 3, No. 6, pp. 445-457, 2009.

[19] D. Linden and T. B. Reddy, Handbook of Batteries, 3rd Edition, McGraw-Hill, ISBN 0-07-135978-8, 3-12, 2002.

[20] G. Wu, R. Lu, C. Zhu and C. C. Chan, "Apply a Piece-wise Peukert's Equation with temperature correction factor to NIMH battery state of charge estimation" Journal of Asian Electric Vehicles, Vol. 8, No. 2, pp. 1419–1423, 2010.

[21] Ha H. Nguyen and Ed Shwedyk, "A First Course in Digital Communications", Cambridge university Press, ISBN 0521876133, 2009.

[22] Arman Molkim, "Simple Demonstration of the Seebeck Effect", Science Education Review, Vol. 9, No. 3, pp. 103–107, 2010.

[23] LTC batteries, SL-2790/S Sandard, catalogue No. 11 2 17901 00, Tadiran Batteries GmbH.

[24] Thomas Dittrich, Chen Menachem, Herzel Yamin and Lou Adams, "Lithium Batteries for Wireless Sensor Networks", Tadiran Batteries GmbH.

[25] D. Doerffel and S. Abu Sharkh, "A critical review of using the Peukert equation for determining the remaining capacity of lead-acid and lithium-ion batteries", Journal of Power Sources, Vol. 155, No. 2, pp. 395–400, 2006.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 612-616 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

A Novel Dual-Band Rectenna for Ambient RF Energy Harvesting at GSM 900 MHz and 1800 MHz

Dinh Khanh Ho[1,2], Van-Duc Ngo[1] Ines Kharrat[2], Tan Phu Vuong[2], Quoc Cuong Nguyen[1], Minh Thuy Le[1*]

[1] *Department of Instrument and Industrial Informatics, School of Electrical Engineering, Hanoi University of Science & Technology (HUST), Hanoi, 100000, Vietnam*

[2] *IMEP-LAHC laboratory, UMR 5130 INPG-UJF-CNRS, Grenoble INP-MINATEC, 38016 Grenoble Cedex 1, 38100, France*

ARTICLE INFO

Article history:
Received : 05 April, 2017
Accepted : 05 May, 2017
Online: 04 June, 2017

Keywords:
Rectenna
RF energy harvesting
Dual-band antenna
Monopole antenna
RF-DC conversion

ABSTRACT

This paper presents a novel dual-band rectenna for RF energy harvesting system. This rectenna is created from a dual-band antenna and a dual-band rectifier which operates at GSM bands (900 MHz and 1800 MHz). The printed monopole antenna is miniaturized by two meander-lines. The received signal from the receiving antenna is rectified by a voltage double using Schottky diode SMS-7630. The rectifier is optimized for low input power level of -20dBm using harmonic balance. Prototype is designed and fabricated. The simulation is validated by measurement with power conversion efficiency of 20% and 40.8% (in measurement) at the input power level of -20dBm. The proposed rectenna has output voltage from 183-415 mV. From the measured results, this rectenna provides the possibility to harvest the ambient electromagnetic energy for powering low-power electronic devices.

1. Introduction

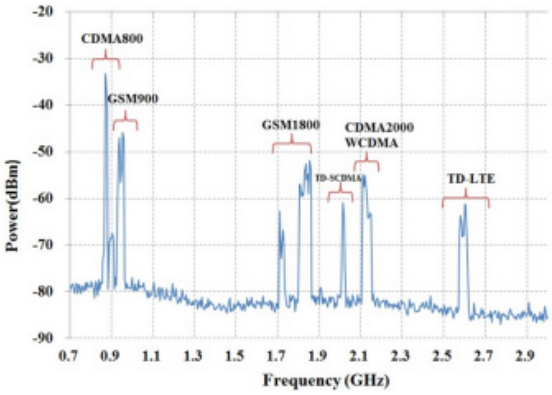
Nowadays, with the development of semiconductor, microelectronic and wireless technology, the wireless electronic devices are getting more compact, wireless, low-power and highly autonomous (i.e.: sensor nodes in Wireless Sensor Networks). In general, almost all electronic devices are powered by batteries, resulting in some disadvantages such as larger size and weight of device, regular maintenance due to limited battery lifetime, which can be more challenging and expensive if the device operates in severe or even difficult to access environment. Therefore, the ambient energy harvesting techniques are proposed as a solution.

The energy existing in the natural environment is of many different kinds, each has a different power density in ambient environment and requires a suitable harvesting technique, such as solar energy (100 mW/cm^2), thermal energy (60 µW/cm^2), ambient RF energy (0.0002 – 1 µW/cm^2), energy from vibration (200 µW/cm^3), etc. Among these energy harvesters, the ambient RF energy harvester has drawn considerable interests in recent years because of some advantages such as continuous RF energy source, low profile, low cost, simple fabrication and compatibility with MMIC (Monolithic Microwave Integrated Circuit).

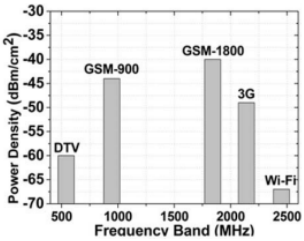
In recent years, various rectenna designs have been published for RF energy harvesting applications. They can operate at single-band [2,3] , dual-band [4-7], multi-band [8], or broadband [9-11]. A narrow band rectenna can give a higher power conversion efficiency (PCE), but the amount of harvested power is low, whereas a multi-band rectenna can has the lower PCE, but more harvested power. The promising solution is to design a dual-band or multi-band rectenna to maximize PCE at the specific frequencies where maximum ambient signal level is available [12]. However, most of designs use an input power level higher than the real power level of RF energy in ambient environment. The spectrum surveys in [13-15] indicate that the RF energy in the downlink channels of GSM-900, GSM-1800, UMTS-2100 and TD-LTE bands is dominant in ambient environment with an expected input power level of -20dBm, that is presented in Fig. 1.

In this work, a dual-band rectenna operating at GSM 900 MHz and 1800 MHz is proposed. The rectenna consists of a dual-band monopole antenna miniaturized by two meander-lines, and a voltage doubling circuit using SMS-7630 Schottky diode to rectify the received signal from the antenna. The RF – to – DC conversion efficiency is optimized for low input power level of -20dBm using Harmonic Balance (HB) simulation in ADS software [17]. Each part of the rectenna is designed and tested separately, and then integrated in a circuit to evaluate the overall performance.

*Corresponding Author: Minh Thuy Le, Dept. of 3I, School of Electrical Engineering, HUST, Email: thuy.leminh@hust.edu.vn



(a)



(b)

Fig. 1. RF power receive: (a) Outside shopping mall [13]; (b) Distributed RF power density in ambient [16]

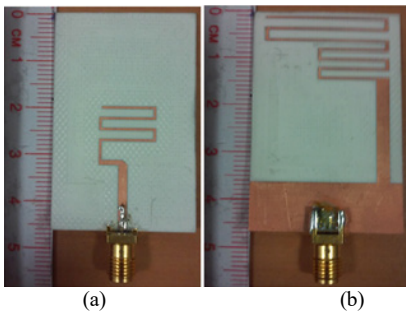
2. Receiving Antenna Design

The fabricated antenna is shown in Fig. 2. The proposed antenna has dimension of 46 mm of length, 30 mm of width. The used material for substrate is Rogers RO4003 with dielectric constant of 3.55 and thickness of 0.8 mm.

This structure is inspired from the antenna design in [17,18] . Paper [18] presents a low-profile dual-band printed Inverted-L (IL) shaped monopole antenna with Parasitic Inverted-F (PIF) shaped element in the ground plane for 2.4/5.2/5.8 GHz WLAN applications, the overall size of this antenna is 40 mm x 30 mm. We can apply directly the technique in [18] to obtain an antenna operating at GSM 900 MHz and 1800 MHz, but the size of antenna will be increased considerably. Therefore, the miniaturized technique in [19] is used to reduce the size of our structure. In [19], a miniaturized antenna using meander-line strips with the overall size of 32 mm x 32 mm is presented for operating at GSM-850, GSM- 900, UMTS-2100, WiMAX-2.3 and WiMAX-3.3 bands. However, the gain of this antenna is quite low almost for all frequency bands, even negative in GSM bands. Our antenna has dual-band with gain of 1.46 dB, 1.91 dB at GSM 900 and GSM 1800.

The return loss and far-field characteristics of the developed antenna are shown in Fig. 3 and Fig. 4. As Fig. 3, the return loss of antenna between simulation and measurement are not similar.

The measurement result is not well matched at GSM 900MHz and shifted a little toward the right at GSM 1800 MHz. This can be explained that as the coupling between two meander-lines in practice weaken or lost, the meander-line for GSM 900MHz was not appropriately excited. Therefore, antenna efficiency is not as high as we expected. This aspect must be improved in the next work.



(a) (b)

Fig. 2. Antenna prototype
(a) To view, (b) Back view

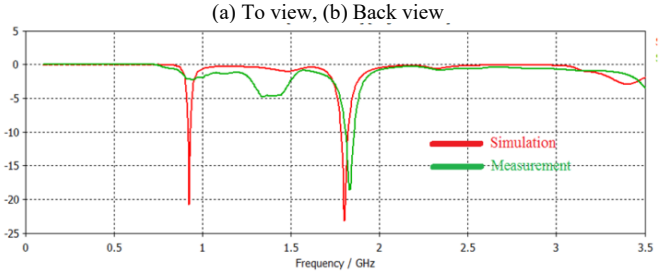
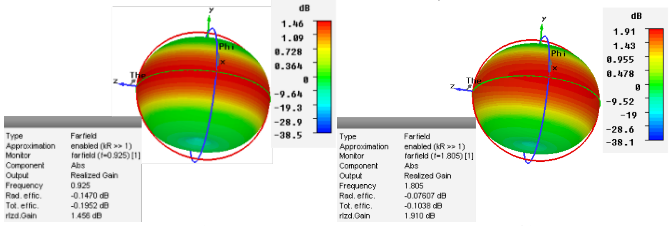


Fig. 3. Return loss of proposed antenna (red-line:simulation result; green-line: measurement result)



(a) (b)

Fig. 4. Radiation patten of proposed antenna
(a) GSM 900; (b) GSM1800

3. Rectifying Circuit Design

The structure of a classical rectenna is shown in Fig. 5. A rectifier consists of a harmonic filter, an impedance matching network to maximize the power transfer from the receiving antenna to AC – to – DC conversion circuit, a conversion circuit converting the received signal by antenna to DC signal, a DC filter to smooth DC signal and to bring the harmonic signal back to rectify and a resistive load. However, this proposed rectifier does not use the Harmonic filter.

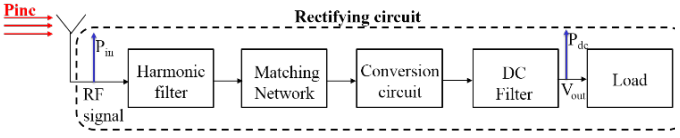


Fig. 5. Structure of rectenna

The RF signal is rectified by a voltage doubling circuit using Schottky diode. The topology of voltage double and the model of diode are shown in Fig. 6. The voltage double consists of two stages, the first one includes capacitor C_1 and diode D_1 clamping the negative part of input signal and the second one includes capacitor C_2 and diode D_2 for peak rectification. C_1 is used in order to block the harmonics generated by the diode (due to its nonlinear behavior) and return to the receiving antenna to avoid the unwanted radiation. C_2 plays a role as a DC filter, and also storages energy, which would be then replaced by a dual radial stub DC filter. The output voltage is defined as:

$$V_{out} = 2*(V_{in} - V_d) \quad (1)$$

Where V_{out} is output dc voltage, V_{in} is input voltage, V_d is forward voltage of diode.

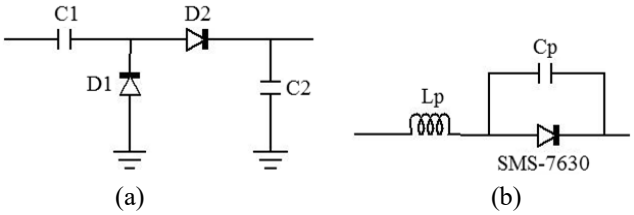


Fig. 6. (a) Voltage double circuit; (b) Diode SMS-7630 model

The diode is the most important component in the design of a rectifying circuit. As shown in paper [20], diode SMS-7630 is chosen. SPICE model parameters of diode SMS-7630 are taken from [21]. The parasitic phenomenon due to package of diode is considered by a parasitic inductor $L_p = 0.7$ nH and a parasitic $C_p = 0.045$ pF. First, the impedance matching network can be (made from?) implemented by lumped elements or distributed elements. In our project, we use a dual-stub matching network to match the receiving antenna and the rectifier at two frequencies. Fig. 7 illustrates the impedance matching network after the total size reduction by meander-line. Fig. 8 shows simulated result of matching network. The good impedance matching is achieved at GSM 900 MHz and GSM 1800 MHz, proving the capability of the dual-frequency matching circuit.

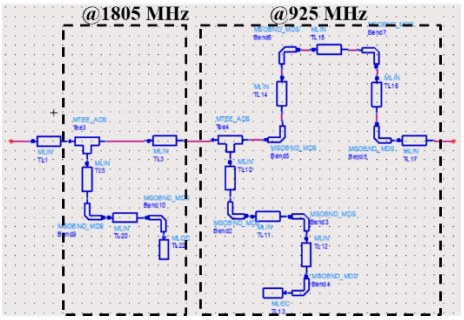


Fig. 7. Impedance matching network

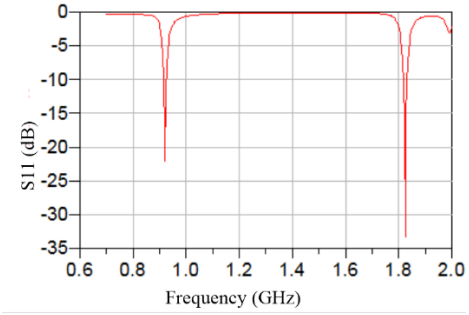


Fig. 8. Simulated result of matching network

Second, a DC filter helps smoothing the DC signal as well as bringing the harmonics back to rectify and its filtering characteristics. The DC filter with dual radial stub has the function of attenuating the fundamental frequency of GSM 900 MHz and 1800 MHz. Besides, the second and the third harmonics are well filtered with less than -3dB transmission coefficient S21 (cf. Fig. 9).

The rectifying circuit is simulated by Harmonic Balance (HB) and Large Signal S-Parameter (LSSP) simulation of ADS in . It is optimized for low input power level of -20 dBm at 925 MHz and 1805 MHz (two frequencies begin the downlink channels of GSM 900 MHz and 1800 MHz, respectively). An optimized load of 8.2kΩ is selected for maximum conversion efficiency. The Fig. 10 shows the photograph of the fabricated rectifier with the total sizes of 10 cm x 6.5 cm. The substrate Roger 4003C is used.

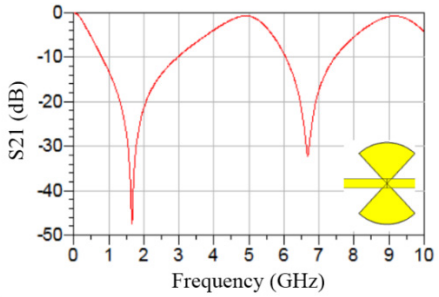


Fig. 9. DC filter and its characteristic

The measured and simulated results indicate that at the low input power level of -20 dBm, the rectifier operates at two frequencies of 925 MHz and 1805 MHz with conversion efficiency of 30.4% and 20%, and output voltage of 158 mV and 128 mV, respectively.

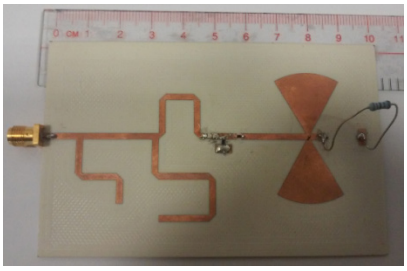


Fig. 10. Photograph of the fabricated rectifier

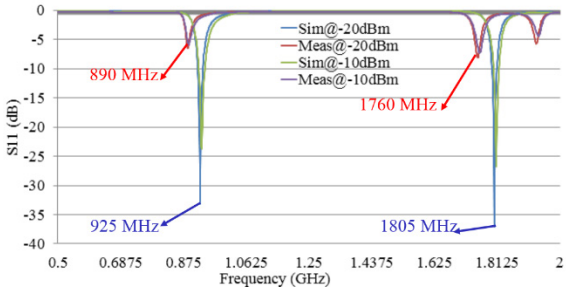


Fig. 11. Return loss of rectifier with difirent incident power level

Fig. 11 plots the measured and simulated return losses (S11) of rectifier at different input power levels. The measurement results are shifted a little toward the left and not well matched. It can be explained by two reasons. Firstly, the dielectric constant of Rogers RO4003 can vary from 3.38 to 3.55. When the dielectric constant increases, the resonant peak shifts to lower frequency. Secondly, it can be due to diode model, especially the junction capacitance *Cj* [20]. By increasing *Cj*, we find that the resonant frequency shifts to lower frequency and the power conversion efficiency degrades drastically.

Fig. 12 presents the results of efficiency and output voltage of the dual-band rectifier when varying incident power level from – 20 dBm to 5 dBm. The maximum output DC voltage is 3.3 V with an input power incident of 0 dBm. However, the incident power of GSM900 and GSM1800 can never be 0 dBm. Table I compares the simulated and measured results at the input power level of -20 dBm.

As the results are shown in Table I, the measured efficiency is haft of simulation result. The inaccurate diode model can be the reason as described above. However, this circuit gives the acceptable output voltage with the low input power level. The exact diode is expected to give higher power conversion efficiency.

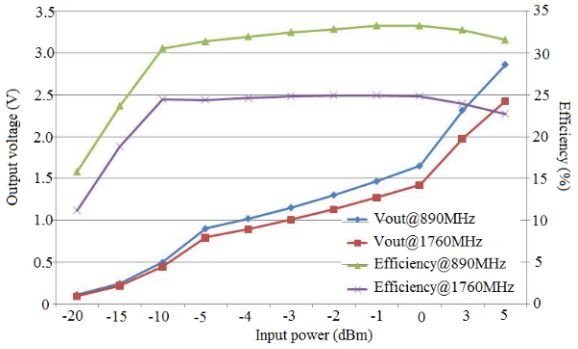


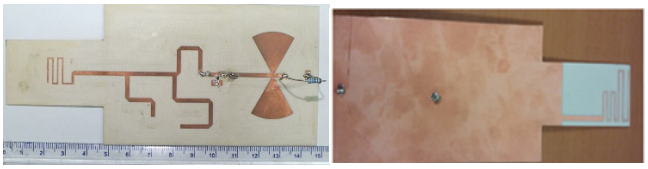
Fig. 12. Efficiency and ouput voltage of rectifier

TABLE I. COMPARISON OF SIMULATION AND MEASUREMENT OF RECTIFIER WITH Pin = -20dBm

	Frequency(GHz)	Efficiency (%)	Out voltage (mV)
Sim	0.925 1.805	30.4 20	158 128
Meas	0.890 1.760	15.8 11.2	114 96

4. Rectenna performance

The receiving antenna and rectifier are designed and evaluated separately. Then they are integrated into the rectenna prototype that is fabricated and evaluated, as presented in Fig. 13:



(a) (b)

Fig. 13. Photograph of the fabricated rectenna
(a) Top view; (b) back view

The rectenna measurement is performed in an anechoic chamber as in Fig. 14. The receiving antenna is fixed in front of the transmitting antenna, the received RF power of antenna, P_{RF}, will be measured by spectrum analyzer. After that, the receiving antenna will be replaced by rectenna at the same position, the output voltage of rectenna (V_{out}) will be measured by a voltmeter. The conversion efficiency of rectenna $\eta_{RF\text{-}DC}$ is defined as the ratio of the DC power P_{DC} and the RF power P_{RF}:

$$\eta_{RF-DC} = \frac{P_{DC}}{P_{RF}} \times 100\% = \frac{V_{out}^2}{R} \times \frac{1}{P_{RF}} \times 100\% \qquad (2)$$

The results of rectenna are presented in Fig. 15 and Fig. 16. Our rectenna recovers the voltage from 183 mV to 415 mV when it is placing near the mobile and in ambient environment (cf. Fig. 16). The efficiency of rectenna is 40.8% and 20% at 1834 MHz and 890 MHz, respectively, with incident power of – 20 dBm.

Table II shows the comparison of various rectifiers published in several literatures. This work provides a dual-band design with higher conversion efficiency. As observed from the table, this work provides a dual-band design with higher conversion efficiency compared to others at the input power level of -20 dBm. However, the total size is bigger than [16,22-24] and smaller than [7] because of the large dimension of DC filter and dual-band coplanar receiving antenna.

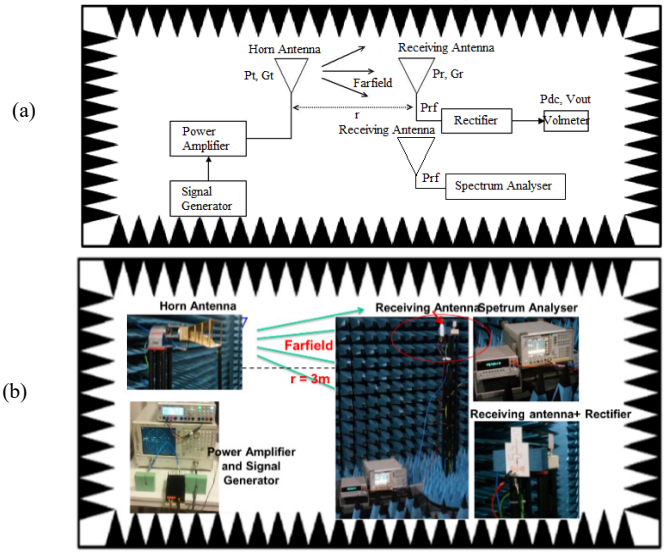


Fig. 14. Rectenna measurement setup in the anechoic chamber at IMEP-LAHC laboratory

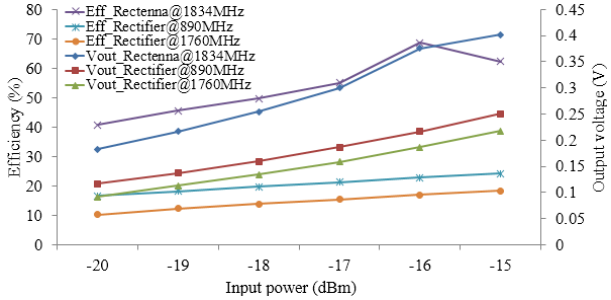


Fig. 15. Measured output voltage versus input power of only-rectifier and rectenna

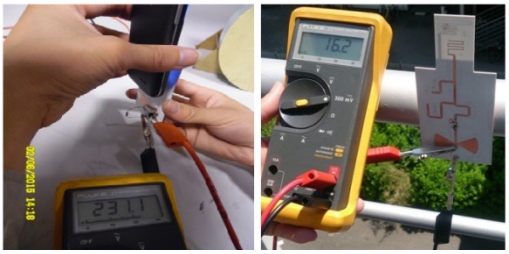


Fig. 16. Rectenna experimen in ambient environement

Table II. COMPARISON OF THE PROPOSED RECTENNA AND THE RELATED WORKS

Ref	Freq (GHz)	Diode	Input power (dBm)	η (%)	Total sizes*
[7]	1.8 2.5	HSMS285C	-20	*33.5*	0.72 x 0.57
[22]	0.915	HSMS2860	-17	*40*	0.14 x 0.13
[16]	0.915 1.8	HSMS 2850	-15 to 20	>30	0.17 x 0.1
[23]	2.45	SMS 7630	-20	27.7	0.37 x 0.33
[24]	1.8 2.45	SMS 7630	-20	13	0.5 x 0.5
This work	**0.925 1.805**	**SMS 7630**	**-20**	**20 to 40.8**	**0.67 x 0.27**

N/a *is not available*
η is conversion efficiency of rectenna
** is the distance calculated by the wavelength*

The rectenna efficiency is quite high compared to rectifier efficiency at section 3. This can be explained by the effect of ground layer of rectifier to monopole antenna performance when the antenna is integrated into rectifier. This aspect is being discussed and reviewed in our current research. Nonetheless, these results demonstrate the potential application of this rectenna in autonomous sensor or self-powered electronic device where additional storage of energy circuit is required.

5. Conclusion

A dual-band rectenna was designed and tested in ambient environment. Simulation results show that the presented rectenna can achieve good voltage. Measurement results on a prototype rectenna show that the operating frequency was well-matched; the rectenna harvests the maximum of 415 mV in ambient environment, which implies that the proposed rectenna is suitable for autonomous sensor application. However, as we mentioned in Section 2, the receiving antenna of this rectenna can be optimized in the future researches to give higher efficiency. A super capacitor can be inserted into the circuit to store energy that is harvested by our rectenna.

Acknowledgment

This research is funded by Ministry of Education and Training (MOET) under grant number B2015-01-93.

References

[1] S. Kim *et al.*, 'Ambient RF Energy-Harvesting Technologies for Self-Sustainable Standalone Wireless Sensor Platforms', *Proc. IEEE*, vol. 102, no. 11, pp. 1649–1666, Nov. 2014.

[2] G. Andia Vera, A. Georgiadis, A. Collado, and S. Via, 'Design of a 2.45 GHz rectenna for electromagnetic (EM) energy scavenging', 2010, pp. 61–64.

[3] R. A. Rahim, S. I. S. Hassan, F. Malek, M. N. Junita, and M. F. Jamlos, 'An investigation of ambient radio frequency as a candidate for energy harvesting source', 2012, pp. 95–99.

[4] R. Scheeler, S. Korhummel, and Z. Popovic, 'A Dual-Frequency Ultralow-Power Efficient 0.5-g Rectenna', *IEEE Microw. Mag.*, vol. 15, no. 1, pp. 109–114, Jan. 2014.

[5] K. Niotaki, S. Kim, S. Jeong, A. Collado, A. Georgiadis, and M. M. Tentzeris, 'A Compact Dual-Band Rectenna Using Slot-Loaded Dual Band Folded Dipole Antenna', *IEEE Antennas Wirel. Propag. Lett.*, vol. 12, pp. 1634–1637, 2013.

[6] H. Sun, Y. Guo, M. He, and Z. Zhong, 'A Dual-Band Rectenna Using Broadband Yagi Antenna Array for Ambient RF Power Harvesting', *IEEE Antennas Wirel. Propag. Lett.*, vol. 12, pp. 918–921, 2013.

[7] S. S. Sarma and M. J. Akhtar, 'A dual band meandered printed dipole antenna for RF energy harvesting applications', 2016, pp. 93–94.

[8] S.Keyrouz, H.J.Visser, and A.G. Tijhuris, 'Multi-band simultaneous radio frequency energy harvesting', presented at the 7th European Conference on Antennas and Propagation, Gothenburg, Sweden, 2013.

[9] M. Arrawatia, M. S. Baghini, and G. Kumar, 'Broadband rectenna array for RF energy harvesting', 2016, pp. 1869–1870.

[10] M. Arrawatia, M. Baghini, and G. Kumar, 'Broadband Bent Triangular Omnidirectional Antenna for RF Energy Harvesting', *IEEE Antennas Wirel. Propag. Lett.*, pp. 1–1, 2015.

[11] Hongxian Zhang and Xinen Zhu, 'A broadband high efficiency rectifier for ambient RF energy harvesting', 2014, pp. 1–3.

[12] N. Shariati, W. S. T. Rowe, J. R. Scott, and K. Ghorbani, 'Multi-Service Highly Sensitive Rectifier for Enhanced RF Energy Scavenging', *Sci. Rep.*, vol. 5, p. 9655, May 2015.

[13] A. S. Andrenko, Xianyang Lin, and Miaowang Zeng, 'Outdoor RF spectral survey: A roadmap for ambient RF energy harvesting', 2015, pp. 1–4.

[14] J. Tavares *et al.*, 'Spectrum opportunities for electromagnetic energy harvesting from 350 mhz to 3 ghz', 2013, pp. 126–130.

[15] M. Pinuela, P. D. Mitcheson, and S. Lucyszyn, 'Ambient RF Energy Harvesting in Urban and Semi-Urban Environments', *IEEE Trans. Microw. Theory Tech.*, vol. 61, no. 7, pp. 2715–2726, Jul. 2013.

[16] Z. Liu, Z. Zhong, and Y.-X. Guo, 'Enhanced Dual-Band Ambient RF Energy Harvesting With Ultra-Wide Power Range', *IEEE Microw. Wirel. Compon. Lett.*, vol. 25, no. 9, pp. 630–632, Sep. 2015.

[17] 'Advanced Design System (ADS) | Keysight (formerly Agilent's Electronic Measurement)'. [Online]. Available: http://www.keysight.com/en/pc-1297113/advanced-design-system-ads?cc=VN&lc=eng. [Accessed: 02-Mar-2017].

[18] C. G. Permana and A. Munir, 'Printed multiband antenna for mobile and wireless communications', 2011, pp. 236–240.

[19] S. Verma and P. Kumar, 'Printed inverted-L shaped monopole antenna with Parasitic Inverted-F element for dual band applications', *Microw. Opt. Technol. Lett.*, vol. 56, no. 5, pp. 1163–1167, May 2014.

[20] 'Harvesting Wireless Power: Survey of Energy-Harvester Conversion Efficiency in Far-Field, Wireless Power Transfer Systems', *IEEE Microw. Mag.*, vol. 15, no. 4, pp. 108–120, Jun. 2014.

[21] 'SKYWORKS : Products : SMS7630 Series'. [Online]. Available: http://www.skyworksinc.com/Product/511/SMS7630_Series. [Accessed: 27-Oct-2016].

[22] A. M. Almohaimeed, M. C. E. Yagoub, and R. E. Amaya, 'Efficient rectenna with wide dynamic input power range for 900 MHz wireless power transfer applications', 2016, pp. 1–4.

[23] T. J. Lee, P. Patil, C. Y. Hu, M. Rajabi, S. Farsi, and D. M. M.-P. Schreurs, 'Design of efficient rectifier for low-power wireless energy harvesting at 2.45 GHz', 2015, pp. 47–49.

[24] C. Song, Y. Huang, J. Zhou, J. Zhang, S. Yuan, and P. Carter, 'A High-Efficiency Broadband Rectenna for Ambient Wireless Energy Harvesting', *IEEE Trans. Antennas Propag.*, vol. 63, no. 8, pp. 3486–3495, Aug. 2015.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 617-622 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Optimisation of Software-Defined Networks Performance Using a Hybrid Intelligent System

Ann Sabih*, Yousif Al-Dunainawi, H. S. Al-Raweshidy, Maysam F. Abbod

Electronic and Computer Engineering Department, College of Engineering, Design and Physical Sciences Brunel University London, United Kingdom

ARTICLE INFO

Article history:
Received: 06 April, 2017
Accepted: 11 May, 2017
Online: 04 June, 2017

Keywords:
ANN
Evolutionary Optimisation
SDN

ABSTRACT

This paper proposes a novel intelligent technique that has been designed to optimise the performance of Software Defined Networks (SDN). The proposed hybrid intelligent system has employed integration of intelligence-based optimisation approaches with the artificial neural network. These heuristic optimisation methods include Genetic Algorithms (GA) and Particle Swarm Optimisation (PSO). These methods were utilised separately in order to select the best inputs to maximise SDN performance. In order to identify SDN behaviour, the neural network model is trained and applied. The maximal optimisation approach has been identified using an analytical approach that considered SDN performance and the computational time as objective functions. Initially, the general model of the neural network was tested with unseen data before implementing the model using GA and PSO to determine the optimal performance of SDN. The results showed that the SDN represented by Artificial Neural Network ANN, and optmised by PSO, generated a better configuration with regards to computational efficiency and performance index.

1. Introduction

Over the past few years, upon entering the era of 'big data, there has been a change in the traditional internet traffic to a more complex traffic form. This complexity has required the increased flexibility and scalability of the modern data centre. In addition to the utilisation of numerous device types within the same area and an increase of advanced network applications, it is now possible for a multitude of end-point devices to share and exchange varying network traffic patterns. As a result, there is a need for a change in the infrastructure of the current network in order to modernise it in line with these differing network traffic patterns. As such, a new approach has been proposed [1].

In the recent years, hybridisation or combination of different machine learning and adaptation techniques has been employed for a large number of new intelligent system designs. The main aim of integrating these techniques is to overcome individual limitations and to achieve synergetic effects [2]. These techniques including, Neural Networks (NNs), the Adaptive Network Fuzzy Inferences System (ANFIS) are used for mapping and modelling purposes. Whilst evolutionary based optimisation approaches, such as

the Genetic Algorithm (GA) and Particle Swarm Optimisation (PSO) have been applied widely to produce powerful and optimised intelligent systems [3].

In many modern systems, artificial intelligence methods, such as intelligent transportation, have taken on notable roles. This adoption of new technology has encouraged the consideration of improvements in the conventional computer networks. The SDN paradigm's abstraction concept and AI techniques have a complex relationship that can be utilised to develop network elements with adaptive behaviours that can also introduce contemporary mechanisms that can overcome the common network issues as well as new issues related to SDN [1].

*Corresponding Author: Ann Sabih, Electronic and Computer Engineering Department, College of Engineering, Design and Physical Sciences Brunel University London, United Kingdom, Email: Ann.Sabeeh@brunel.ac.uk

Recently, with the fast development of the Internet, network topology, and number of applications have been changed gradually, which has led to be more complicated in structures and functions. A network based on the traditional TCP/IP architecture faces many challenges, especially the router, as the network core, takes on too many efforts to deal with. As a result, the validity and efficiency of data forwarding is being threatened. Hence, we need to find a new kind of network architecture to solve the existing problems. To this end, the study of future networks is proceeding all over the world, and the Software Defined Network (SDN) is one of them [4]. SDN is an architecture enabling rapid innovation, while hiding much of the complexity of the networking design. As such, it is a promising solution for network virtualisation that decouples control from the forwarding or data plane [5]. In doing so, it can provide the capability of remote and centralised control of the network forwarding-plane through the network control-plane.

SDN, as a network platform, has been studied widely in the literature and many researchers have proposed soft computing methods to model and optimise the network. Yilan and et al. [6] provided a genetic algorithm for solving the bandwidth-constrained multi-path optimisation problem in the SDN. Gao and et al. [7] contributed a Particle Swarm Optimisation algorithm to solve the control placement problem that takes into consideration both the latency between controllers and their capacities. Zhang and Fumin [8] explored applications of the SDN technique in the direction of the automation of network management, the unified control of optical transmission and IP bearing, smooth switching in a wireless network as well as network virtualisation and QoS assurance. Risdianto and et al. [5] evaluated the performance of an SDN-based virtual network on different virtualisation environments, including operating system based virtualisation, hardware-assisted virtualisation, and par virtualisation. Sgambelluri and et al. [9] proposed novel mechanisms that have been specifically introduced to maintain working and backup flows at different priorities and to guarantee effective network resource utilisation when a failed link is recovered. Ionita and Victor-Valeriu identified a method of avoiding DDoS attacks in the SDN environment. These scholars utilised a cyber defence system to assess risk that was structured around the neural network and the biological danger theory. This demo platform is able to achieve full packet capture in the SDN in addition to mitigating any attacks, providing it is considered necessary by the central command component. The benefit of these packet captures is that they can be utilised for forensic analysis to identify the attacker [10].

A novel open flow controller, which was structured around the intracerebral neural network was suggested by Wu and Huang in order to generate a independent media handover with wireless operation. As such, binary decisions are made by the controller that are determined by the link control parameters. These are instantly generated and gathered from the trained neural network, which is reliant on the interactions between mobile speed and the wireless link performance parameters. Using a mutated PSO, these authors managed to train the fundamental controller equation for the media independent handover that is interspersed with the intracerebral neural network's sigmoid and radial activation functions [11].

A hybrid intelligent system is proposed in this paper to optimise the performance (i.e. throughput, delay) of the SDN. The proposed system includes ANN to model inputs (flows of the flow tables)-outputs (throughput and delay) of the network. After, the evolutionary algorithms have been employed to find the optimal set of inputs that maximise this performance. This method grants to lessen the burden of SDN switch by making the network controller adaptively providing the rules/flows to the switches instead of making the switch suspends for the controller events every period of time.

2. Software-Defined Networks

It was previously noted that a software-defined network (SDN), could be centrally organised around the principle of the compartmentalisation of the control plane and data plane. The SDN is a budding architecture that is promoted by the ONF due to its ability in forwarding functions and network control decoupling. This decoupling allows the direct programming of the network control and the abstraction of the underlying infrastructure for applications and network services. The networking devices have their intelligence removed in this architecture and positioned within a centralised controller, which manages the functionality of the entire network. Software based SDN controllers also have this centralised network intelligence allowing them to maintain a global view of the network. This allows the infrastructure devices to take on a forwarding role that is able to process incoming packets. This role is determined by a set of rules generated instantaneously by the controller within the control layer that is based on some predefined program logic. A remote commodity server is a usual manner in which these controllers are run. A secure connection allows communication with the forwarding elements utilising a standardised command set. A high-level architecture for SDN is presented by Open Networking Foundation ONF [12], that has three main layers, application, control and infrastructure layers that are vertically split as detailed in Figure 1

As demonstrated by [9], the SDN controller serves to maintain the stream tables so as to engage in the management of every stream that is conveyed through the system. Application layer service requests are mapped by the SDN control layer into defined commands and directives to data plane switches. The SDN control layer then supplies applications with data plane topology and activity information. As such, the control layer acts as either a server or a collaborating set of servers, which are recognised as SDN controllers.

The OpenFlow procedure can be regarded as an open SDN specification that is constituted of a pair of foundational elements: first, the OpenFlow switch and, second, the controller. The former facilitates the execution of the data plane while the latter realises

the control plane. It should be noted that the OpenFlow procedure as a totality is employed by the interchange that takes place with regard to the controller and the switches via a safe channel. The OpenFlow switch and controller communicate via a safe channel that is realised in the context of a distinctive practice and, according to [10], this is also referred to as OpenFlow. The controller transmits a Flow Table towards the OpenFlow switches and each stream passage in the former element is constituted of a regulation crafted from an organisation of fields, thereby coordinating the approaching bundles. It is notable that this is an activity that characterises the issue of how the coordinating parcels can be prepared, for example, by transmitting on a specific yield port.

In addition, a number of counters are employed in order to collect information relating to the stream. In a similar way, each passage can be linked to alternative information, one case of this being the requirement level and the hard and idle timeout of the pair of timing devices.

The bundle is captured and conveyed to the controller that employs the safe channel in those instances where a switch receives a parcel that is not engaging in the coordination of the passages that have been introduced. Having captured the bundle, the activity of the OpenFlow switch is regulated by the controller as a result of the overhaul of its Flow Tables, thereby transmitting the coordinating parcels to the end point [8]. One implication of this is that the relative casings are not required to partake in the controller another time as a result of the potential initiation of the Flow Tables' reformulation [3].

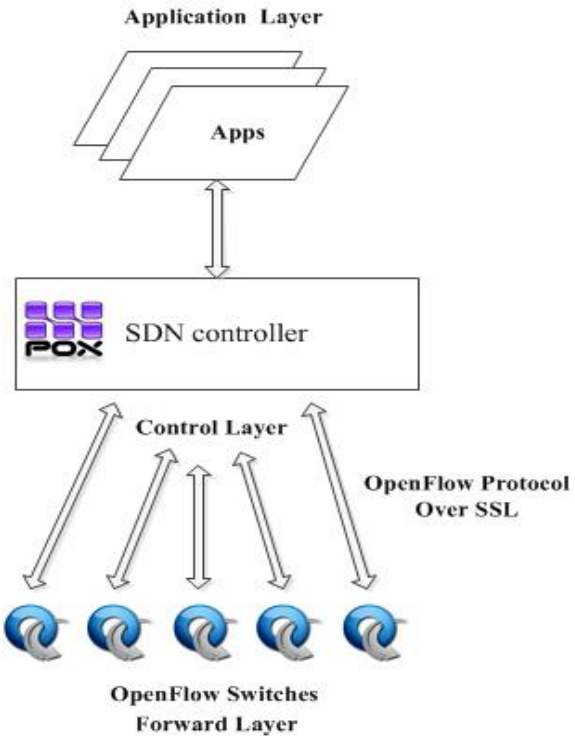


Figure 1. The software-defined network's architecture

3. Artificial Intelligence

3.1. Neural Networks

A branch of artificial intelligence, disseminated at a considerable pace in recent times owing to its suitability with regard to the modelling and forecasting ability it has in relation to dynamic systems, is the artificial neural network (ANN). In light of their promising potential, ANNs have emerged as a central branch of research into artificial intelligence. The registration of the input-output relationships of nonlinear and synthesis systems have been identified as one of the key advantages offered by ANNs and, notably, this relationship can be straightforwardly, rapidly, and cost- and time-effectively discerned by lowering the error with regard to the network output(s) and the actual output(s). A defining feature is that, following the appropriate preparation of the network, outputs can be estimated within a very short space of time (namely, in a matter of seconds). Frameworks that centre on artificial neural networks are currently seeing effective application in a range of areas – a few examples being including adaptive control, laser applications, nonlinear system identification, robotics, image and signal processing, medical areas, pattern recognition, error detection, process logging, and renewable and sustainable energy areas – in order to surmount obstacles faced by engineers [11]. The only appropriate connections in a feedforward neural network structure are between the outputs and inputs of each layer. As such, there are no connections between the outputs of one layer and the inputs of the same or previous layers [13].

3.2. Evolutionary Algorithms

Taking inspiration and motivation from natural processes and, in addition, basing the developments on factors relating to iterative and probabilistic processes, a range of evolutionary algorithms – including genetic algorithms (GAs), particle swarm optimisation (PSO), and simulated annealing – have been formulated in recent years. The primary purpose of such developments is for application in optimisation issues. Two multi-purpose and frequently employed algorithms include GA and PSO, and these are utilised in every domain [12].

- **Genetic algorithm**: Formulated by Holland (1975), GAs are self-modifying global optimisation probability search algorithms, the fundamental concept of which was inspired by the genetic mechanisms that form the basis of the theory of Darwinian natural selection and biological evolution. GAs operate by simulating the biological processes that are observed in the natural world as driving the phenomena of genetic and evolutionary development; according to the concept of natural selection, GAs provide solutions to deep problems by employing code technique and reproduction processes [13]. GAs has been extensively employed in a variety of domains with considerable efficacy in recent years, and this is primarily attributed to their almost universal relevance and promising results. The difference between the more traditional search algorithms and GAs[4] is that the latter have numerous candidate solutions rather than just one partial or candidate solution. With GAs, each problem's candidate solution is portrayed by a data

structure that is termed an 'individual'. There are two parts to each individual; these are the chromosome and the fitness. An individual's chromosome is constructed from genes with each value assigned to the gene being referred to as alleles. These individuals combine to form a population, the size of which remains constant for the duration of the search for most GAs. Out of the current population, a number of 'parent' individuals are selected, based on their fitness, which are allowed to create offspring. Individuals with above average fitness have a higher than average chance of selection for parenting. Following selection, the parents are subjected to a number of reproductive operators, such as crossover and mutation. Those subjected to crossover have a copy of their genes taken to create an offspring's chromosome. This is comparable to the creation of living organisms that are created from a genetic mixture of both parents from sexual reproduction. However, only one parent is required for mutation. In this manner, the offspring is often an almost exact replica of the parent but with a few altered genes. Following the generation of the offspring, their represented candidate solutions can be evaluated, and the offspring's fitness is determined. As the population size remains largely static, before the offspring can be incorporated into the population, it is necessary for some individuals in the current population to be removed, or die. Removal of individuals is often decided from their fitness with those individuals with a below average fitness being more likely to be selected for removal than those with an average or above average fitness. Again, this is reminiscent of the evolution of living organisms and is termed natural selection. As such, those individuals that display better fitness are allowed to procreate and live longer. Interestingly, this process of fitness selection means that the original population does not need to be very good. Indeed, it is often the case that each individual in the initial population represents a randomly generated candidate solution; however, the repetitive application of natural selection and reproduction allows GAs to generate rapid and efficient solutions [13].

- **Particle Swarm Optimisation**: First developed by Kennedy and Eberhart (1995) [14] and built on by the researchers several years later (Ibid, 2001) [15], PSO algorithms have been applied with enormous success in optimising a broad range of applications [16]. PSO operates by locating all individuals and particles (usually in the range of 10-100) in randomised positions and, following this, intending that each particle engages in random motion in a determined direction in the search space. Following this, the direction of each particle is incrementally modified in order to proceed according to the optimal previous positions, thereby identifying more favourable positions on the basis of specifications or an objective function (i.e. fitness). The original particle speed and location are chosen randomly and, in turn, the velocity formula presented below is used to provide updates:

$$V_{c_i+1} = wV_i + C_1R_1 \times (Pb_i - x_i) + C_2R_2 \times (Gb - x_i) \quad (1)$$

Contrastingly, the new particle is computed by summing the new velocity to that which precedes it, as presented below:

$$x_{i+1} = x_i + V_{c_i+1} \quad (2)$$

Where: Vc denotes the particle's velocity; X denotes the particle's position; R1 and R2 are independent random variables uniformly distributed in [0, 1]; C1 and C2 are the acceleration coefficients, and w represents the inertia weight. The particle's new velocity can be calculated by employing Eq. 13, and the information required includes the previous velocity, the distance of the particle's present position from its optimal position (Pb), and the global best position (GB). Following this, on the basis of Eq. 14, the particle is conveyed to a new location in the search space and, notably, the way in which every particle performs is evaluated in relation to a predetermined objective function known as the performance index.

4. Equation Simulation and Results

4.1. SDN Simulation

SDN simulation has been performed using the Mininet platform which is the common emulation for SDN which is used by researchers to collect all datasets of inputs and outputs. Mininet has the ability to imitate various types of system components, for example, have, layer-2 switches, layer-3 switches, and interfaces. In addition, simulation experiments were carried out using POX controller with OpenFlow 1.0 and monitoring of the flows was required for all events. Regarding which, in order to build the learning system, the SDN controller gives the orders to SDN switch to monitor the flows, and the switch keeps on monitoring them to detect the events. When the flows are monitored, the entire events whether they occur only frequently or periodically are stored in the database to be used in the ANN learning system. The flows/inputs include the rules coming from the controller. In turn, the output will be represented by the system throughput and the network delay. Consequently, the data which are collected from the ANN learning system is considered as being an efficient input to the optimisation algorithms. This paper presents a new method to minimize the load of the SDN switch by making it changing adaptively by the controller, instead of wait for the event all the time.

4.2. SDN Identification

When the ANN training started, the dataset had been preprocessed by normalising them into the range between -1 and 1. The dataset of the inputs and outputs was divided randomly into three subsets: a training set, valediction set and testing set. The first subset was for establishing the gradient as well as updating the network weights and biases. The error regarding the second subset was observed during the training development. The validation error is usually reduced in the initial training phase, as is the training set error. Nevertheless, when the network overfits the data, the error in the validation set invariably starts to rise. In the current case, the network parameters were saved at the minimum of the validation set error.

Input values need to be normalised in the range [-1, 1], which corresponds to the minimum and maximum actual values. Subsequently, testing the ANN requires a new independent set (test sets) in order to validate the generalisation capacity of the prediction model. A multilayer feedforward network was implemented to estimate the performance of the SDN. In order to obtain a maximum accuracy of prediction, the network was trained in different topologies. For each network architecture, the training was run ten times for various random initial weights and biases using the Levenberg Marquardt algorithm (LMA). After investigating the performance of different architectures using the exhaustive search method, the best trained ANN with one hidden layer was found to consist of 17 neurons in this hidden layer, which gives the comparably better performance of MSE, with 2.488 ×10-8, see Figure.. 2. While, Figure 3 shows the performance of the network as a mean square error (MSE) versus the network architecture for the single hidden layer. Also, Figures 4 and 5 show the simulated and predicted SDN performance for both the training and test sets. It is noticed that the ANN model is efficiently accurate and the network is accepted as a general model to be integrated, as the next step, with GA or PSO so as to produce the proposed intelligent hybrid system.

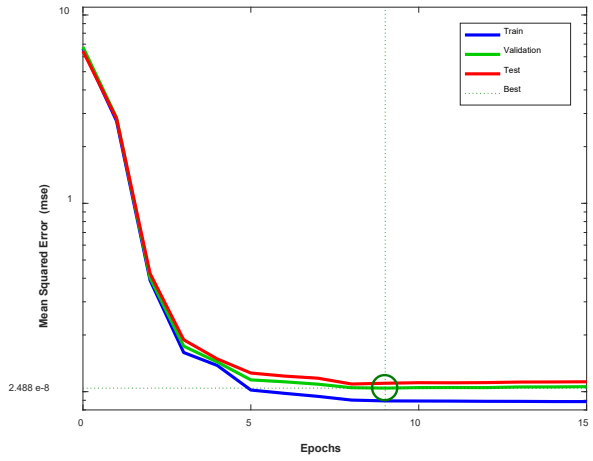


Figure 2. the progress of the ANN performance.

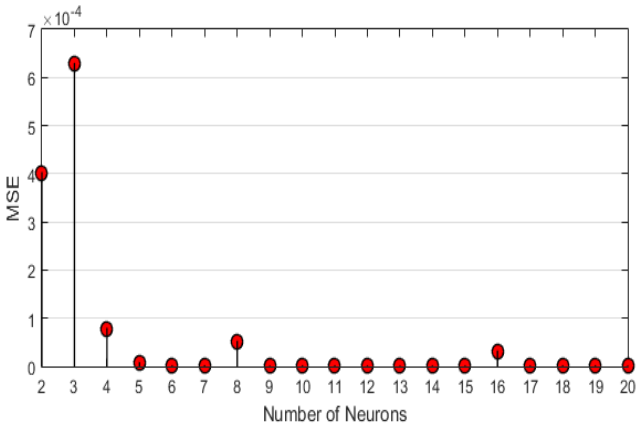


Figure 3 Performance of a single hidden layer ANN

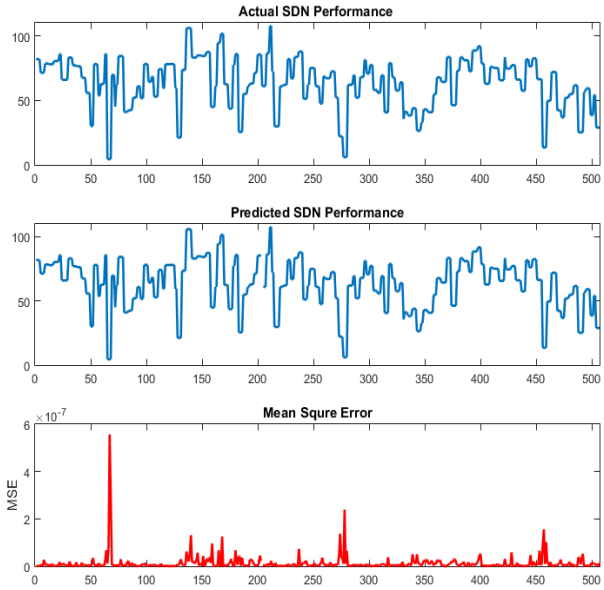


Figure 4 Comparing predicted with actual SDN performance for the training sets

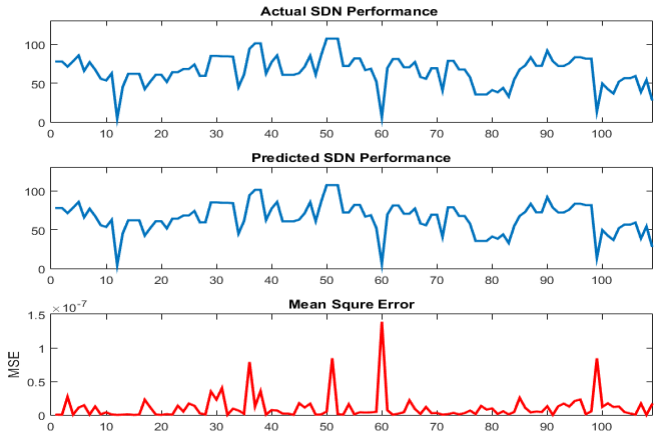


Figure 5 Comparing predicted with actual SDN performance for the testing sets

4.3. SDN Optimisation

GA and PSO were integrated separately with the trained ANN model to select the optimal set of inputs that make the network work as efficient as possible. Figure 6 shows the architecture of the system including the trained general ANN model as well as PSO and GA as an optimizer.

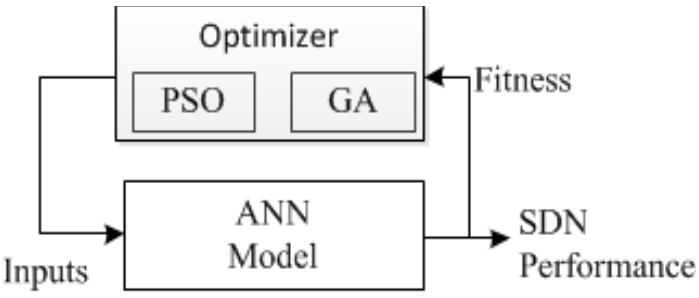


Figure 6 The architecture of the proposed method

TABLE I. PERFORMANCE AND COMPUTATION TIME COMPARISONS FOR GA AND PSO

Optimisation method	SDN Performance %	Computation Time (min)
PSO	96.321	3.23
GA	94.846	6.75

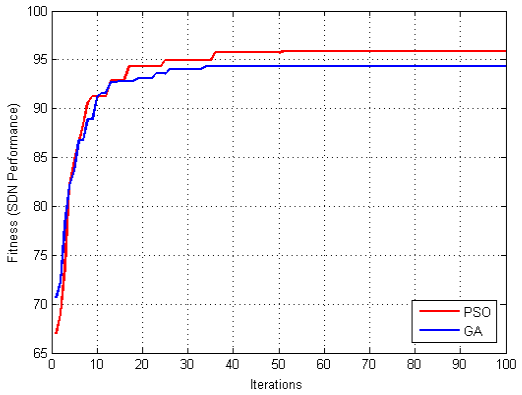


Figure 7 Convergence comparison of GA and PSO

The simulation experiments were carried out by MATLAB platform. PSO was employed to find the optimal structure of the network, and the best operational parameters of this and the GA algorithm were chosen after extensive simulations, which were set as follows:

- Size of the population or swarm: 50
- Maximum iterations or generations (max) :100
- Cognitive acceleration (C1): 1.2
- Social acceleration (C2): 0.12
- Momentum or inertia (w): 0.9

A comparison of the results of the SDN performance is provided in Table 1 and Fig. 7. PSO has outperformed GA regarding the performance, and computational time, the convergence is faster with fewer iterations, and the obtained fitness is higher. However, the results provided were obtained after running PSO/GA 20 times.

5. Conclusion

This paper has presented a novel hybrid intelligent system for the modelling and perfecting of the Software Defined Network (SDN). This involved the training of an artificial neural network (ANN), which had a single layer in the hidden zone, to map the inputs and network performance. An acceptable MSE was shown by the network that was demonstrated as being below 2.466×10^{-8}. The application of unseen data as a test set to the trained ANN model proved its generality; this ANN model was then coordinated with evolutionary algorithms (EAs) to create the presented intelligent hybrid system. In order to optimise the EA, PSO and GA were utilised in order to identify the optimal input set for the SDN. This optimisation was based on the fitness function of the individuals. The comparison results showed that PSO was a more effective option in terms of computational time, convergence and performance.

References

[1] M. Latah and L. Toker, "Application of Artificial Intelligence to Software Defined Networking: A Survey," Indian J. Sci. Technol., vol. 9, no. 44, Nov. 2016.

[2] Y. Al-Dunainawi and M. F. Abbod, "Hybrid Intelligent Approach for Predicting Product Composition of Distillation column," Int. J. Adv. Res. Artif. Intell., vol. 5, no. 4, pp. 28–34, 2016.

[3] D. Ruan and P. Wang, "Intelligent hybrid systems: fuzzy logic, neural networks, and genetic algorithms". Springer, 1997.

[4] W. Xia, Y. Wen, C. H. Foh, D. Niyato, and H. Xie, "A Survey on Software-Defined Networking," IEEE Commun. Surv. Tutorials, vol. 17, no. 1, pp. 27–51, 2015.

[5] A. C. Risdianto and E. Mulyana, "Implementation and analysis of control and forwarding plane for SDN," in 2012 7th International Conference on Telecommunication Systems, Services, and Applications (TSSA), 2012, pp. 227–231.

[6] Y. Yilan Liu, Y. Yun Pan, M. Muxi Yang, W. Wenqing Wang, C. Chi Fang, and R. Ruijuan Jiang, "The multi-path routing problem in the Software Defined Network," in 2015 11th International Conference on Natural Computation (ICNC), 2015, pp. 250–254.

[7] C. Gao, H. Wang, F. Zhu, L. Zhai, and S. Yi, "A Particle Swarm Optimization Algorithm for Controller Placement Problem in Software Defined Network," Springer International Publishing, 2015, pp. 44–54.

[8] Z. F. Zhang Shunmiao, "Survey on software defined network research," Appl. Res. Comput., vol. 30, pp. 2246–2251, 2013.

[9] A. Sgambelluri, A. Giorgetti, F. Cugini, F. Paolucci, and P. Castoldi, "OpenFlow-Based Segment Protection in Ethernet Networks," J. Opt. Commun. Netw., vol. 5, no. 9, p. 1066, Sep. 2013.

[10] I. Mihai-Gabriel and P. Victor-Valeriu, "Achieving DDoS resiliency in a software defined network by intelligent risk assessment based on neural networks and danger theory," in 2014 IEEE 15th International Symposium on Computational Intelligence and Informatics (CINTI), 2014, pp. 319–324.

[11] Q. Wu, J. Huang, and O. Yang, "An Open Flow Controller Based on the Intercerebral Neural Network for Media Independent Handover," Int J Swarm Intel Evol Comput, 2014.

[12] D. Kreutz, F. M. V. Ramos, P. Esteves Verissimo, C. Esteve Rothenberg, S. Azodolmolky, and S. Uhlig, "Software-Defined Networking: A Comprehensive Survey," Proc. IEEE, vol. 103, no. 1, pp. 14–76, Jan. 2015.

[13] M. Jamshidi and A. Zilouchian, "Intelligent control systems using soft computing methodologies". CRC Press, 2001.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 623-633 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Data Error Detection and Recovery in Embedded Systems: a Literature Review

Venu Babu Thati*[1], Jens Vankeirsbilck[1], Jeroen Boydens[1], Davy Pissoort[2]

[1]*Reliability in Mechatronics and ICT, Department of Computer Science, KU Leuven, 8200 Brugge, Belgium*

[2]*Reliability in Mechatronics and ICT, Department of Electrical Engineering, KU Leuven, 8200 Brugge, Belgium*

ARTICLE INFO

Article history:
Received :05 April, 2017
Accepted :11 May, 2017
Online: 04 June, 2017

Keywords:
Data flow errors
Error detection
Error recovery
Reliability
Embedded systems

ABSTRACT

This paper presents a literature review on data flow error detection and recovery techniques in embedded systems. In recent years, embedded systems are being used more and more in an enormous number of applications from small mobile device to big medical devices. At the same time, it is becoming important for embedded developers to make embedded systems fault-tolerant. To make embedded systems fault-tolerant, error detection and recovery mechanisms are effective techniques to take into consideration. Fault tolerance can be achieved by using both hardware and software techniques. This literature review focuses on software-based techniques since hardware-based techniques need extra hardware and are an extra investment in cost per product. Whereas, software-based techniques needed no or limited hardware. A review on various existing data flow error detection and error recovery techniques is given along with their strengths and weaknesses. Such an information is useful to identify the better techniques among the others.

1. Introduction

In general, a critical aspect of any computer system is its reliability. Computers are expected to perform tasks not only quickly, but also correctly [1]. Recent trends in embedded systems attract industries to use them more and more in day-to-day life for an increasing number of applications. Application areas include, but are not limited to, mechatronic industries, medical equipment, smart energy consumers, mobility. Reduced size and reduced supply voltage make systems more susceptible to disturbances. Since there are more systems in use and the environment becomes more harsh, a system failure or a system crash is more likely to occur. A system failure could lead to serious consequences such as human injury, environment pollution and a huge amount of money loss for industries [2].

The rise in usage of electronics under harsh conditions significantly increases the probability of all kinds of disturbances from the environment. Such disturbances are glitches, electromagnetic interference, temperature variations, etc. [3–6]. It is proven that decreasing the size and supply voltage of the components in circuits and increasing their complexity leads to less reliable systems [7]. The corresponding systems are susceptible to soft errors (bit flips) which are typically transient. Transient faults do not cause any permanent physical damage and can be restored by overwriting the introduced bit flip or by a system restart. Still these faults are categorized as systematic faults since given the exact same circumstances these faults will re-appear in exactly the same way. Because the environment changes, transient faults don't occur continuously, unlike design and manufacturing faults [1,8,9].

Errors in embedded systems can cause unusual behavior and degrade systems integrity and reliability [7]. A number of hardware and software techniques have been developed to make embedded systems fault-tolerant against transient faults [10,11]. Fault tolerance is a two step process. The first step is fault detection, indicating that somewhere in the system fault has occurred. The second step in the process is fault recovery, restoring the system from fault state to the normal state [12].

Today, fault tolerance is mainly achieved via hardware solutions. Such hardware-based solutions are hardware redundancy approaches to meet the requirements of the reliability. Such hardware redundancy techniques are expensive since they have to be implemented on every product produced. A commonly used hardware-based technique for error detection in embedded systems is N-modular redundancy. This technique uses N (N>2) parallel modules for comparing the original and redundant process results. This hardware redundancy technique introduces a 100* (N- 1)% performance and memory overhead but does achieve a fault coverage of 100% [12,13]. To reduce the overheads in hardware-based fault tolerance techniques different software-based redundant techniques have been proposed and implemented [12,13]. Such software solutions would lead to a more cost-efficient solution in many situations. Due to its

*Venu Babu Thati, KU Leuven; Spoorwegstraat 25; 8200 Brugge, Belgium
Contact No: +32 (50)664805 & venubabu.thati@kuleuven.be

flexibility and cost, software-based solutions are used in a number of applications. Software redundancy increases the system's reliability but requires extra memory space and processing time to execute redundant instructions [15]. A number of software-based data error detection and recovery techniques have been proposed and implemented in literature for fault-tolerant embedded systems [1,7,8,12–22].

According to recent studies, soft errors are one of the primary sources of failure in embedded systems [7,16,23–25]. These soft errors (bit flips) may further affect the system during program execution, leading to a faulty system. Such bit flips have an effect on data flow or control flow of the program. Generally, data flow errors lead to corruption of variables in the program causing a wrong intermediate or output result. In contrast, control flow errors lead to a jump in the program execution order [26–29]. This review paper focuses on various data flow error detection and recovery techniques existing in literature to make embedded systems fault-tolerant against bit flips. Since a number of data error detection and recovery techniques exist in literature, it is important to review and identify the strengths and weaknesses of each of these techniques for a fault-tolerant embedded system. Figure 1 gives an overview of the software-based data protection techniques that will be discussed in this paper.

The remainder of this paper is organized as follows: Section 2 describes and reviews the different data flow error detection techniques. Section 3 describes and reviews the different error recovery techniques. Section 4 provides future work plans and Section 5 concludes this paper.

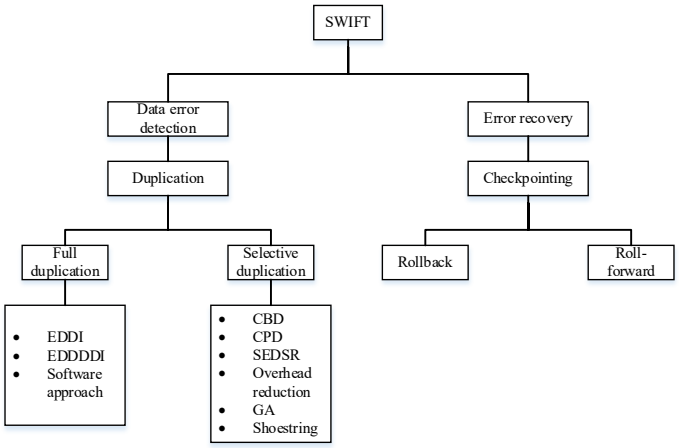


Figure 1. Overview of software-based data protection techniques.

2. Related Work

Soft errors usually occur due to heavy radiation, power supply distortion, environmental temperature fluctuations, and other factors. The introduced soft errors can corrupt the data of the program in execution. To counter this data corruption, a number of data flow error detection and recovery techniques have been proposed in literature [1,7,8,12–17,19–21]. In previous work [4], we listed several data flow error detection techniques and discussed their reported results. The contribution of this review paper is 1) that we list not only detection techniques but also recovery techniques, 2) that we discuss the considered techniques more in depth and 3) that we give several strengths and weaknesses per technique. The provided strengths and weaknesses have been determined based on the technique's inner working and reported results. By looking at the strengths and weaknesses of each of the technique presented under error detection and error recovery, one can easily identify the better technique immediately with clear reasoning.

3. Data Flow Error Detection

This section presents and reviews various existing data flow error detection techniques such as EDDI (Error Detection by Duplicated Instructions), ED^{4}I (Error Detection by Diverse Data and Duplicated Instructions), Software approach, CBD (Critical Block Duplication), CPD (Critical Path Duplication), SEDSR (Soft Error Detection using Software Redundancy), Checking rules, GA (Genetic Algorithm) and Shoestring approach. Strengths and weaknesses of each of these techniques will be discussed. All of the presented data flow error detection techniques are software-based.

Duplication is the basic mechanism involved in data error detection techniques [4]. A number of data flow error detection techniques have been developed based on a unique duplication mechanism for better fault coverage or lower overhead in terms of memory consumption. The duplication can be applied at various levels such as a full duplication and selective duplication [1,4,7,8,12–15,17,18]. Full duplication techniques and selective code duplication techniques are discussed in Sections 2.1 and 2.2 respectively.

In order to evaluate the data error detection techniques, authors of the corresponding techniques [1,7,12–17] have chosen various case studies for the experiments. Bubble sort, quicksort, insertion sort, and matrix multiplication are the most used case studies in previous research in this field [13,16,17,26,30]. Of course, some of the techniques uses other case studies such as FFT, differential equation solver, mean, vortex, etc. Further, a fault injection mechanism has been used to inject the faults in hardened case studies for validation. All of the provided information in Table 1 and Table 3 such as error detection techniques, case studies, injected faults, detected faults, fault coverage, performance overhead, and memory overhead are considered from literature [1,7,12 – 17]. Fault coverage, performance overhead, and memory overhead are defined in (1), (2), and (3).

$$Fault\ coverage = \frac{Number\ of\ detected\ faults}{Number\ of\ injected\ faults} \quad (1)$$

$$Memory\ overhead = \frac{Final\ memory - Initial\ memory}{Initial\ memory} \quad (2)$$

$$Performance\ overhead = \frac{Final\ time\ - Initial\ time}{Initial\ time} \quad (3)$$

3.1. Full Duplication

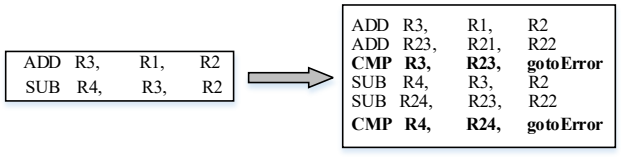
This section presents various existing full duplication techniques for data flow error detection. The basic mechanism involved in all of the full duplication techniques is duplicating the entire code and comparing the original and duplicated output to detect errors. Full code duplication has been performed in different ways for different techniques as in [7,13,15,31].

Table 1. Representation of error detection technique with case studies, case studies length, injected faults and detected faults from literature [1,7,12–17].

Error detection technique	Case studies used	Case studies length (in lines)	Injected faults	Detected faults
EDDI	Insertion sort Quicksort Matrix mul.	30 36 47	500 500 500	495 491 496
ED^{4}I	Quick sort Matrix mul. Insertion sort	47 36 30	500 500 500	479 481 482
Software approach	Constant mod-ulus algorithm	20	19520	19520
CBD	Bubble sort Matrix mul. Quick sort	25 36 47	784 784 784	615 561 553
CPD	Differential equation solver	30	784	575
SEDSR	Bubble sort Matrix mul. Quick sort	25 36 47	1000 1000 1000	950 940 956
Checking rules	Bubble sort Matrix mul. Dijkstra's	25 36 19	10000 10000 10000	9560 9600 9350
GA	Bubble sort Regression Quicksort	25 23 47	40000 40000 40000	32800 32000 33600
Shoestring	Vortex Crafty Gap	SPEC2000 benchmark suit	12000 12000 12000	11050 6000 9000

Error detection by duplicated instructions

EDDI is one of the most often used error detection techniques in research [12]. The EDDI technique states three different instructions for program execution: a master instruction (MI), a shadow instruction (SI), and a comparison instruction (CI) [13], as shown in Figure 2b. Figure 2a represents two different master instructions. The master instruction is the original instruction of the source code, while the shadow instruction is the duplicated instruction added to the source code. Validation of correct operation is accomplished by comparing registers and memory values of master instructions and those of shadow instructions. In Figure 2b, first three instructions refer to MI, SI, and CI. If there is any mismatch between the master and shadow output, the comparison instruction reports an error. To achieve the highest fault detection ratio, EDDI is applied at the assembly level [8,13]. In order to evaluate the effectiveness of the proposed technique, quicksort, matrix multiplication, insertion sort, and FFT were used as case studies.



a) Original instructions b) Transformed instructions

Figure 2. Master, shadow, and comparison instructions representation.

Strengths

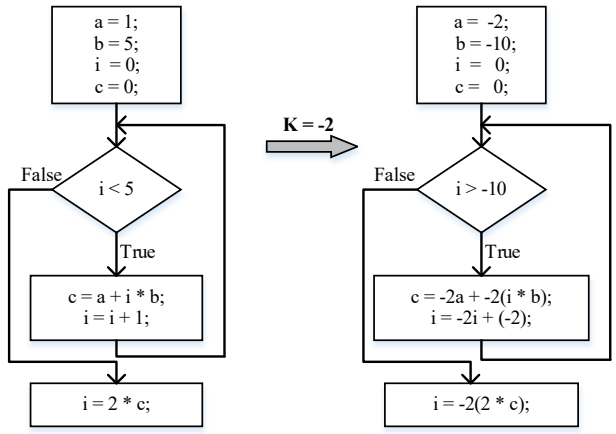
With the final computation results from MI and SI in the program, the error can be detected by placing a comparison instruction. The EDDI technique achieves nearly 98.8% [13,16] of fault coverage by placing a redundant comparison instruction after each MI and SI in the program. The remaining percentage of undetected errors comes from the faults that create an infinite loop in the program. EDDI is one of the techniques that has highest fault coverage in this field.

Weaknesses

Since it is a full duplication technique, all of the instructions presented in the program need to be duplicated. Next to the original (MI) and duplicated (SI) instructions, a comparison instruction has to be placed to report errors. Since each of the original instructions is converted to three instructions, so performance and memory overhead of EDDI are 104.7% and 200% [13,16], as shown in Table 3.

Error detection by diverse data and duplicated instructions

ED^{4}I detects errors by executing two different programs, the original and a transformed (duplicated) program, and comparing their results. The comparison gives an error if the original and duplicated programs do not lead to the same result. The transformation of ED^{4}I technique representation is shown in Figure 3, where $x' = k.x$ for integer numbers is used. Where k is the fault detection probability of the program, $\boldsymbol{x}$ is the original program and $\boldsymbol{x'}$ is the transformed program [7]. In the presented ED^{4}I technique, the optimum value for k that maximizes the fault coverage probability is calculated. After performing the validations with case studies, the authors identified that $k = -2$ is the optimum value to have a maximum fault detection. EDDI and ED^{4}I techniques are comparable because of their common case stuides.



a) Original program b) Transformed program

Figure 3. Original and transformed program with k = -2.

Strengths

The ED^{4}I technique presents a transformation algorithm for the program that transforms an original program (integers or floating point numbers) x to a new program $\boldsymbol{x'}$ with diverse data. This technique achieves a 96.1% [7,16] fault coverage when using the optimum value for k. This result is approximately equal to that of the EDDI technique, as shown in Table 3.

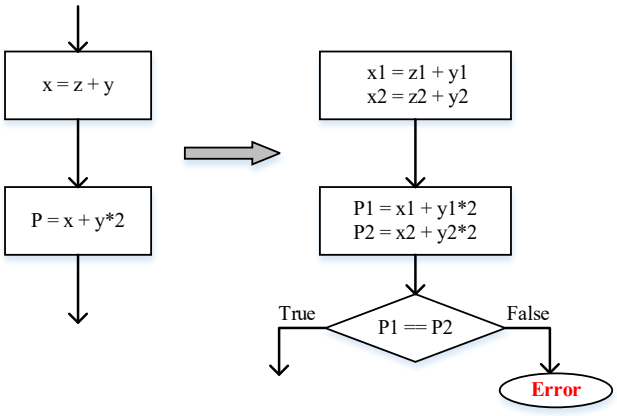
Weaknesses

As in EDDI, ED^{4}I also needs to duplicate the entire original program with diverse data. This technique requires a number of redundant instructions for duplication and comparison, which causes an increment in overhead. The performance and memory overhead imposed by this technique is nearly 126.6%

and 160% [16], as given in Table 3. ED^{4}I transformation is only good for either integers or floating point numbers, but not for both. For example, if a program has mixed data types such as floating point numbers and integers, in that case, we need multiple transformations with different *k* values for each type. The drawback of such a multiple transformations is that it will introduce a more performance and memory overhead.

Detecting soft errors by a pure software approach

The error detection mechanism in the proposed technique is based on a set of transformation rules. These transformation rules are classified into three basic groups: 1) error affecting data, 2) error affecting basic instructions, and 3) error affecting control instructions to detect the errors [31]. Error affecting data rules are used to detect data flow errors, whereas error affecting basic and control instructions rules are used for control flow error detection. In error affecting data, the motive is to identify and define a dependency relationship between variables of the program. Furtherly, classifying them into intermediary variables and final variables based on their role in the program [31]. From Figure 4a, variables *x*, *y*, and *z* are classified as intermediary variables, which are used for other variables operation. Whereas variable *P* is classified as a final variable, which does not participate in any operations. After each write operation to the final variables, both original and duplicated values are compared for a consistency check, if any inconsistency is identified an error detection procedure is activated. By applying the different transformation rules to each of the original variables, this technique is able to detect errors that occur in data, basic instructions, and control instructions.



a) Original program b) Transformed program

Figure 4. Transformations rules targeting error affecting data in a sample program [31].

Strengths

The presented technique is mainly based on a set of transformation rules. Error affecting data rules are used to detect data flow errors with full duplication scheme. Fault coverage achieved with this technique is 100% [31], because of duplicating the entire program and comparison after each write operation to the final variables. Software approach is one of the techniques that has highest fault coverage in this field.

Weaknesses

Usage of more redundant instruction for duplication and the comparison lead to increase in overhead. The appeared performance and memory overhead in this technique are 244% and 362% [31], as shown in Table 3.

3.2. Selective Duplication

This section presents a number of existing selective code duplication techniques for data flow error detection. The main difference with full duplication techniques is that selective duplication techniques first analyze the program to detect the most important parts and only duplicate those important parts. Defining and identifying the important part of a program can be done in different ways, leading to different techniques [1,12,14–17].

Error detection by critical block duplication

The presented selective code duplication technique is named CBD. The CBD technique follows three different steps to detect data flow errors. The first step is, to identify the critical blocks in the control flow graph. These critical blocks are the most vulnerable in the program because its output has an influence on the other blocks. The second step is to duplicate the identified critical blocks. The final step is to compare the original and duplicated critical blocks to detect errors. The authors of this proposed technique introduced a simple way for critical block detection from the example of control flow graph, as shown in Figure 5. A block that has the most number of outgoing edges to the other blocks in the control flow graph is considered, as a critical block [12]. In Figure 5, block 1, has three outgoing edges to the other blocks, whereas others have less than three. In this case, the highlighted block 1 is identified as a critical block. Furthermore, the critical block, block 1, is duplicated and compared to the original block. If any mismatch is identified between the original and duplicated instructions, it will report an error.

Strengths

In Section 2.1, we have reviewed various full duplication techniques and their advances in fault coverage. But, because of increased performance and memory overhead, it seems that full code duplication is not a good option. Limiting the code duplication scope is useful in real-time and general purpose applications where cost is the primary factor. In CBD, performance and memory overhead are decreased because they use less redundant instructions for duplication and comparison. The appeared performance and memory overhead in this technique are 50% and 101.6% [12], as shown in Table 3.

Weaknesses

With regards to CBD, redundant instructions are inserted only in the critical block, so there is a possibility of missing undetected errors in other blocks leads to a reduction in fault coverage. The achieved fault coverage with CBD technique is only 73.5% [12]. Another major drawback of this technique is that it is a compiler and/or case study dependent and could just act as a full duplication technique. For example, let's consider a control flow graph with 8 basic blocks, 5 of them have two outgoing branches and other 3 have only one. Since 5 of them are critical blocks, they are expected to have at least 80% of the code. Such a control flow graph with CBD approach is duplicating 80% of the original program.

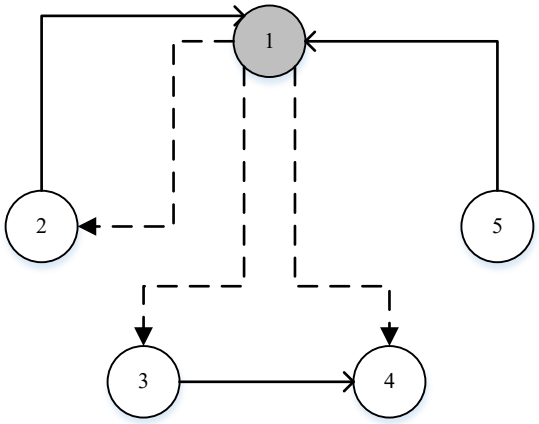


Figure 5. Example of control flow graph with CBD [12].

Error detection by critical path duplication

In this CPD technique, the data flow graph is used instead of using the control flow graph. The data flow graph is used to derive the interconnection of variables and their dependencies and effects on each other. In a data flow graph example, nodes represent the operands and vertices represent the variables of the program, as shown in Figure 6 [14]. The basic idea behind this technique is to identify and duplicate the critical path in the data flow graph. The first step is, to identify the critical path in the data flow graph. The authors of [14,17,32] who proposed CPD technique introduced a simple principle for critical path detection. The longest path in the data flow graph is considered as a critical path because of the great possibility of error occurrence on that long path. According to the principle proposed by the authors, the longest path in the data flow graph is identified and kept in the box, as shown in Figure 6. Next, the identified critical path will be duplicated and the comparison instructions will be placed after each write operation in the final variables. If the final variables in the program are not equal to each other, it reports an error.

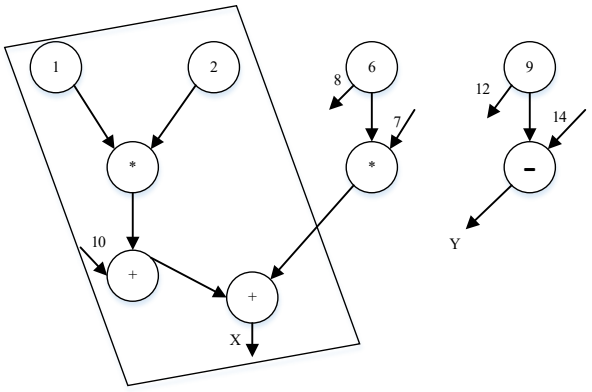


Figure 6. Example of data flow graph with CPD [14].

Strengths

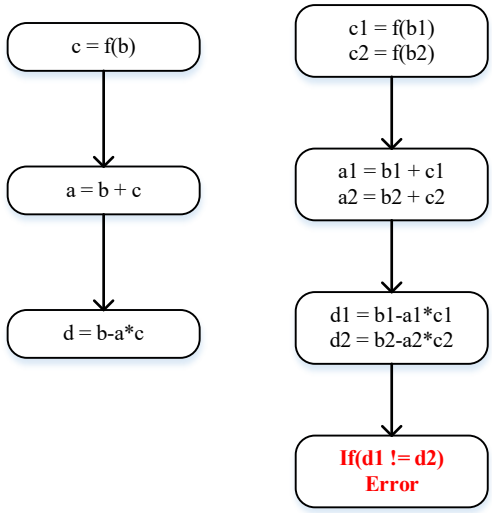
With regard to CPD, only instructions presented in the critical path are duplicated to detect the data flow errors with minimum overhead. In CPD, performance and memory overhead are decreased because they use less redundant instructions for duplication and comparison. The appeared performance and memory overhead in CPD are 60% and 103% [14], as shown in Table 3.

Weaknesses

As far as CPD is concerned, redundant instructions are inserted only in the critical path, so there is a possibility of missing out on the undetected errors in the other small paths in the data flow graph. This lead to reduce in a fault coverage. Fault coverage achieved with CPD is only 73.3% [14]. In CPD, creating a data flow graph is much harder in assembly than C and C++. It is also difficult to perform the duplication and maintain the control flow graph in order when only data flow graph is given.

Soft error detection using software redundancy

This technique is named SEDSR. In this technique, the critical block is duplicated as in CBD. As in [12,17], the critical block is the block with the most number of outgoing edges to the other blocks in the control flow graph, as shown in Figure 5. In this technique, critical block variables are further divided into two categories: (1) middle variables: important in computing the other variables and (2) final variables: they don't perform any computations [17]. In the critical block, a redundant comparison instruction is placed after the final variables to compare these parameters in original and duplicated blocks. Figure 7a represents the sample (original) program of the critical block and variables *a, b,* and *c* are considered as the middle variables and *d* is considered as the final variable. Figure 7b is the duplicated version of a sample program with comparison instruction for the critical block. If any mismatch between original and duplicated variables is identified during the comparison, an error is reported and the program execution is halted. SEDSR and CBD techniques are comparable because of their common case studies.



a) Sample program b) Duplicated program

Figure 7. Sample program of critical block with duplication and comparison [17].

Strengths

SEDSR is one of the critical block duplication techniques. In comparison with CBD, in this technique, the critical block variables are not directly duplicated but furtherly classified into two types such as middle variables and final variables. In SEDSR, by placing a comparison instruction after writing to the classified final variables detect a lot of errors. Fault coverage achieved with this technique is 94.85% [17], which is increased in comparison to the CBD, as shown in Table 3.

Weaknesses

As in CBD, this technique also considers only the critical block with further improvements in the process over CBD, as mentioned in strengths. At the same time, performance and memory overhead are increased because of placing extra comparison instructions. The imposed performance and memory overhead in SEDSR are 112.3% and 134.6% [17], they are increased in comparison to the CBD. Since SEDSR has a similar kind of duplication mechanism as in CBD, this technique is also a compiler and/or case study dependent and could just act as a full duplication.

Overhead reduction in data- flow software-based fault tolerance techniques

The aim of this technique is to provide low overhead with the same level of reliability as in EDDI, ED^{4}I , and Software approach [7,13,31]. This technique provides an alternative implementation of software-based techniques. The alternative overcomes the drawback of the massive overhead introduced by other techniques for soft error detection. In the presented technique, a set of rules for the data protection are explained as shown in Table 2 such as, 1) global rules: each register used in the program should have its replica, 2) duplication rules: (a) duplicating all instructions except branches, (b) duplicating all instructions except branches and stores, and 3) checking rules: to compare the values of a register with its replica at different positions [15]. Such rules are applied to various methods with the choice to detect the errors. Table 2 explains the purpose of each rule. Checking rules (overhead reduction), EDDI, and ED^{4}I techniques are comparable because of their common case studies.

Table 2. Checking rules description [15].

Global rules	(Valid for all techniques)
G1	Each register used in the program has a spare register as a replica
Duplicated rules	Performing the same operation on the registers replica
D1	All instructions except branches
D2	All instructions except branches and stores
Checking rules	Compare the value of a register with its replica
C1	Before each read on the register
C2	After each write on the register
C3	Before loads, the register that contains the address
C4	Before stores, the register that contains the data
C5	Before stores, the register that contains the address
C6	Before branches

Strengths

Different methods are implemented in this technique by considering the choice of duplication and checking rules, as shown in Table 2. A couple of methods have equal fault coverage with changes in overhead. Selecting the right checking rules is important because they have an influence on fault coverage and overhead. The method with highest fault coverage and lower overheads is considered as the best method. In this technique, the best method has a fault coverage of 95% with performance and memory overhead of 72.3% and 82% [15]. The appeared performance and memory overhead in this technique are decreased in comparison to the EDDI, and ED^{4}I techniques, as shown in Table 3.

Weaknesses

Compared to full duplication error detection techniques such as EDDI, and ED^{4}I, this technique has slightly reduced fault coverage. Fault coverage achieved with this technique is 95% because of using less number of redundant instructions with the choice of checking rules in comparison to the EDDI and ED^{4}I techniques.

Method for hardening a program against soft error using genetic algorithm

In this technique, GA has been used to identify the most vulnerable blocks of the program through input data [16]. The identified vulnerable blocks have to be strengthened against errors through duplication and comparison. The proposed technique follows three different steps to detect errors as shown in Figure 8. Those three steps are, 1) preprocessing of the input program: with regard to the results obtained from the related researchers such as [33,34], a considerable number of program instructions does not have any effect on the program output results. This step includes a method called program slicing [35], which eliminates some of those instructions that do not have an impact on the program output results. The first step improves the speed of proposed GA in the second step, 2) identifying the most vulnerable blocks: GA has been proposed to identify vulnerable blocks. GA takes the source code of the program as an input to find out the smallest subset of the basic blocks which are more vulnerable. The most vulnerable blocks are identified based on initial population, selection, crossover, mutation, evaluation, and replacement processes introduced in GA, as clearly explained in [16], and 3) strengthening the identified vulnerable blocks: based on the required level of reliability, most vulnerable blocks in the program are strengthened against errors [16], is shown in Figure 8.

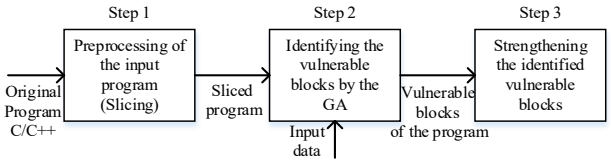


Figure 8. Representation of proposed method [16].

Strengths

Due to initial preprocessing and then selective vulnerable block duplication and comparison, the presented technique uses a less number of redundant instructions. Usage of a less number of redundant instructions decreases its performance and memory overhead. Performance and memory overhead presented in this technique are 24.3% and 60.3% [16].

Weaknesses

As other selective duplication techniques presented in this section, this technique considers only the most vulnerable blocks in the program for duplication. By duplicating only the vulnerable blocks in the program, most of the faults can be detected but not all. There is a possibility of undetected errors in the other normal blocks which lead to a reduction in fault coverage. Fault coverage achieved with this is technique is 82% [16].

Shoestring: Probabilistic soft error reliability

In the program, any instruction that can potentially influence global memory is considered as a high-value instruction [1]. In fact, if it consumes a wrong input, they are likely to produce outputs that result in user-visible corrupted results. In this

technique, high-value instructions are defined as the most vulnerable instructions and will have a huge impact on program output. Shoestring technique contributes in different issues for detecting the errors. Such issues are: 1) a transparent software solution for addressing soft errors, 2) a new reliability-aware compiler analysis, and 3) a selective instruction duplication that leverages compiler to identify and duplicate a small subset of vulnerable instructions [1]. Code duplication begins by selecting a single high-value instruction, from the set of all high-value instructions in the program. The selected single high-value instruction then proceeds to duplicate and then compare with comparison instruction. Duplication process is terminated when no more producers exist for duplication or the producer is already duplicated. Then the inserted comparison instructions are used for checking the errors. In Figure 9, the shaded parts represent the code duplication chains and the dashed circles indicate high-value instructions.

Strengths

Shoestring is a minimally invasive software solution, which results in very low overhead. Since duplicating and comparing only high-valued instructions, less number of redundant instructions are used which leads to decrease in overhead. Performance and memory overhead introduced in this technique are 20.16% and 40% [1], these are better than any other techniques in this field.

Weaknesses

Shoestring approach initially identifies the most unsafe instructions, as high-value instructions. Only duplicating and comparing the high-valued instructions, produce better overhead. At the same time, fault coverage is reduced because of possible undetected errors in the unduplicated instructions. Fault coverage achieved with this technique is 80.6 [1]%.

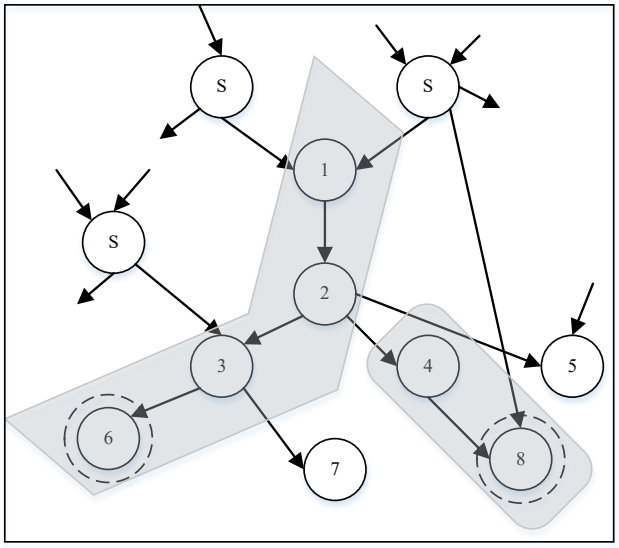


Figure 9. Example of data flow graph illustrating shoestrings code duplication chains [1].

4. Error Recovery

Error recovery techniques have been implemented to recover from the identified errors to keep systems in an error-free state with minimum overhead. Error recovery is generally based on the checkpointing concept [19–21,36]. Checkpoints are saved at regular intervals in the program based on the program execution behavior.

This section presents and reviews general checkpointing techniques for rollback error recovery and roll-forward error recovery. Strengths and weaknesses of rollback and roll-forward error recovery policies with checkpointing techniques are discussed.

4.1. Rollback Error Recovery

Rollback error recovery is one of the most used error recovery policies to recover the errors by using the checkpointing techniques in embedded systems. Bashiri, et al. propose a checkpointing technique for rollback error recovery. In rollback error recovery, in the case of an error, the processor state is restored to the error-free state with lower overhead [21]. In general, cost, performance, and memory overhead are primary factors for any error recovery technique.

The primary step for developing an error recovery technique is defining the correct error model. In [21], the proposed technique is based on control flow error model. During compilation time, the program is partitioned into basic blocks. The basic block is a set of instructions in a program without a jump instruction. Thereafter, an error detection mechanism needs to be added to the basic blocks presented in the program. Figure 10 shows the example of placing checkpoints in the control flow graph of the program. Usually, the checkpoint is stored in memory for rolling back the system with an immediate effect whenever an error is detected. Such a memory must be a fault-tolerant memory. The checkpoint contains the content of the registers, stack pointer and memory locations like stack region, constants, and variables [21]. For the considered benchmark programs such as bubble sort, matrix multiplication, and linked list copy, a checkpoint capturing is inserted to each of the basic blocks individually.

For example, a control flow graph is constructed with six blocks based on the program. Then checkpoints are added to the blocks based on the program execution order as shown in Figure 10. Since there will be a possibility of error occurrence before the first checkpoint location, it is mandatory to put a checkpoint at the beginning of the program. Remaining checkpoints are placed at the locations based on the program's vulnerability. In Figure 10, locations of the second and third checkpoints contain the vulnerable information. During the program execution time, whenever an error is detected, a detection mechanism informs the recovery routine and recovery routine recovers the error from the previously restored checkpoint. To evaluate the presented checkpointing technique, a pre-processor has been implemented that selects and adds the checkpoints to blocks [21].

Strengths

Bashiri, et.al, proposes a general checkpointing technique for rollback error recovery to recover from detected errors. The advantage of rollback error recovery is that if the error is detected, the processor state is restored into error free state without using the spare processor. The number of redundant instructions needed for rollback error recovery is very low. In the presented checkpointing technique for rollback error recovery, appeared memory overhead is low and also a cost efficient.

Weaknesses

In the presented checkpointing technique for rollback error recovery, whenever an error is detected, immediately the system

must rollback to its previous checkpoint of the corresponding program. Due to the fact that considerable time overhead is the main drawback of the rollback error recovery, as shown in Table 4. This technique is not fit for the typical time critical applications.

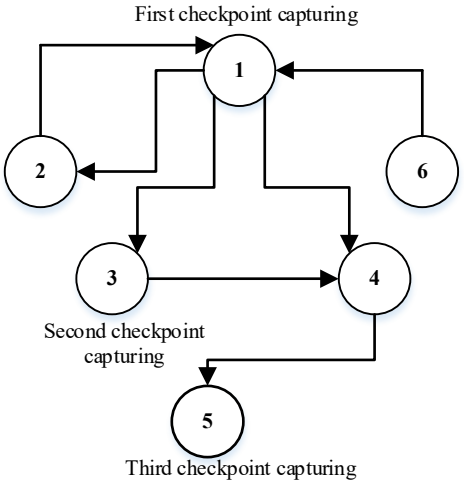


Figure 10. Periodic checkpointing representation in control flow graph.

4.2. Roll-Forward Error Recovery

Roll-forward error recovery is another error recovery policy to recover the errors. Roll-forward schemes are developed to increase the possibility that a given process completes within a given time. In that case, a couple of roll-forward schemes uses a spare processor for removing the rollback to save the time. At the same time, in time critical applications, redundancy is an important factor to consider because of cost, power, memory, and other factors. However, both schemes for roll-forward recovery with and without spare processor are discussed in this section.

Roll-forward recovery with dynamic replication checks (with spare processor)

The presented roll-forward recovery scheme uses dynamic replication checks to detect errors and is named RFR-RC (Roll-Forward Recovery with dynamic Replication Checks). This scheme is organized based on the isolated checkpoint intervals. For any isolated checkpoint intervals, a task is executed on two independent processors such as processor P1 and processor P2 as shown in Figure 11 [20]. In the presented scheme, at every checkpoint, the duplicated task records its state in the storage and the recorded state is forwarded to the checkpoint processor. Thereafter, at the end of the checkpoint interval, the checkpoint processor compares the two states from the processors. If the compared checkpoint states match with each other, the checkpoint will be committed and both the processors P1 and P2 continue their executions into the next checkpoint interval [20,21]. During the comparison, if any mismatch is detected, a validation step starts immediately. During the validation process, processors P1 and P2 continue their execution. At the same time, a spare processor is occupied to retry the last checkpoint interval using the previously committed checkpoint. Once the spare processor is ready with its process, the state is compared with the previous states of processors P1 and P2. The faulty processor among two processors such as P1 and P2 will be identified after this comparison. Then the identified faulty processor state will be made identical to that of the other processor. Now, both processors duplicating the task need to be in the correct state. As from the assumption [20] of single independent faults, a further validation process is not required.

Strengths

In the presented RFR-RC scheme, a spare processor is used to save time. With an extra processor, there is no need of rolling back to restore the system from error state. In RFR-RC, during the validation, the spare processor is used to identify the faulty processor and recovery action will be taken immediately. Appeared time overhead in RFR-RC scheme is decreased in comparison to the rollback scheme and RFR-BC scheme, as shown in Table 4.

Weaknesses

Because of using a spare processor to avoid the rollback, the cost is getting high. Memory overhead appeared in this technique is increased in comparison to the rollback recovery scheme.

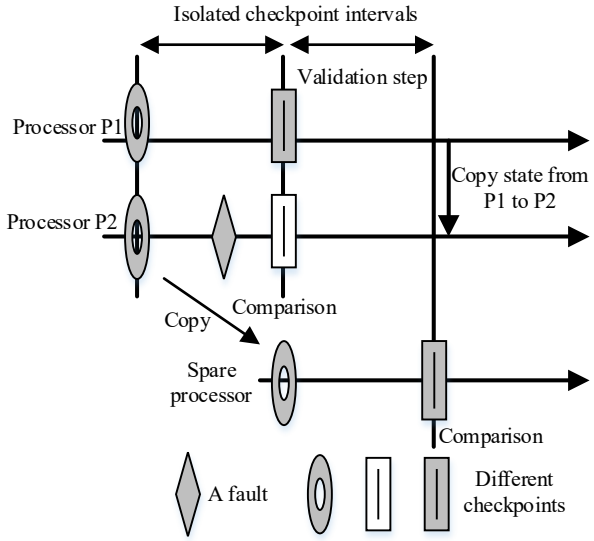


Figure 11. Roll-forward recovery in RFR-RC scheme [19,20].

Roll-forward recovery with behaviour-based checks (without spare processor)

In order to avoid the rollback, self-checks have been inserted to identify the faulty processors [20]. Such a self-detection methods are behaviour-based checks, such as control flow monitoring, detecting an illegal instruction, and memory protection. In [20,37], a new scheme has been proposed and implemented for roll-forward recovery named RFR-BC (Roll-Forward Recovery with Behavior based Checks). The proposed scheme uses a process pair approach to avoid the rollback to reduce the time. The intuitive idea presented in this scheme is, whenever an active task fails, the spare task becomes active to provide the necessary services [20]. The information sending through the active and spare task do not differ. However, information passing through the spare task need to be verified by the acceptance test before sending. Thereafter, the states of two processors (processor1 and processor 2) are verified at the end of a checkpoint interval to declare passing the test is committed, as shown in Figure 12.

Basically, acceptance test for sending the information validates a couple factors such as timing, coding, reasonableness, structural and diagnostic checks [20,38]. In the presented scheme, checkpointing is used for fault identification and roll-forward error

recovery. Whenever a faulty processor is located, then its state is made identical to the checkpoint state of the error-free processor. Because of this, both processors (processor1 and processor 2) will be in the correct state at the beginning of the next checkpoint interval. Figure 12 demonstrates the REF-BC scheme.

Strengths

RFR-BC scheme in roll-forward recovery does not need a spare processor as in RFR-RC to avoid rollback. In the RFR (Roll-Forward Recovery) schemes, the continuity of the executing program will be maintained so that the recovery delay will be removed [20,21]. The advantage of RFR-BC scheme is that time overhead is decreased in comparison to the rollback scheme, as shown in Table 4. It is also a more cost efficient solution compared to RFR-RC because of no spare processor.

Table 3. Results of the presented data error detection techniques from literature [1,7,12–17].

Error detection technique	Case studies used	Fault coverage (%)	Performance overhead (%)	Memory overhead (%)
EDDI	Insertion sort	99	113.90	200
	Quicksort	98.2	89.3	200
	Matrix mul.	99.2	111.1	200
	(Average)	**(98.8)**	**(104.7)**	**(200)**
ED^{4}I	Quicksort	96.4	110	164
	Matrix mul.	95.9	133	170
	Insertion sort	96.2	137	146
	(Average)	**(96.1)**	**(126.6)**	**(160)**
Software approach	Constant mod-ulus algorithm	100	244	362
CBD	Bubble sort	78.54	51	94
	Matrix mul.	71.53	57	109
	Quicksort	70.57	42	102
	(Average)	**(73.5)**	**(50)**	**(101.6)**
CPD	Differential equation solver	73.3	60	103
SEDSR	Bubble sort	95.01	112	127
	Matrix mul.	93.95	121	146
	Quicksort	95.59	104	131
	(Average)	**(94.85)**	**(112.3)**	**(134.6)**
Checking rules	Bubble sort	95.6	70	82
	Matrix mul.	96	82	89
	Dijkstra's	93.5	65	75
	(Average)	**(95)**	**(72.3)**	**(82)**
GA	Bubble sort	82	23	64
	Regression	80	24	57
	Quicksort	84	26	60
	(Average)	**(82)**	**(24.3)**	**(60.3)**
Shoestring	Vortex	92	12	31
	Crafty	75	24	48
	Gap	75	24.5	41
	(Average)	**(80.6)**	(20.16)	**(40)**

Table 4. Results of the presented error recovery techniques from literature [19– 21,36].

Method	Time overhead
Rollback recovery	31.1%
Roll-forward recovery (RFR-RC)	1.23%
Roll-forward recovery (RFR-BC)	2.08%

Weaknesses

In the RFR-BC, appointed self-check detection has an inaccurate error coverage and can't detect certain types of faults such as faults causing infinite looping. Time overhead with RFR- BC scheme is increased in comparison with the RFR-RC scheme for roll-forward recovery.

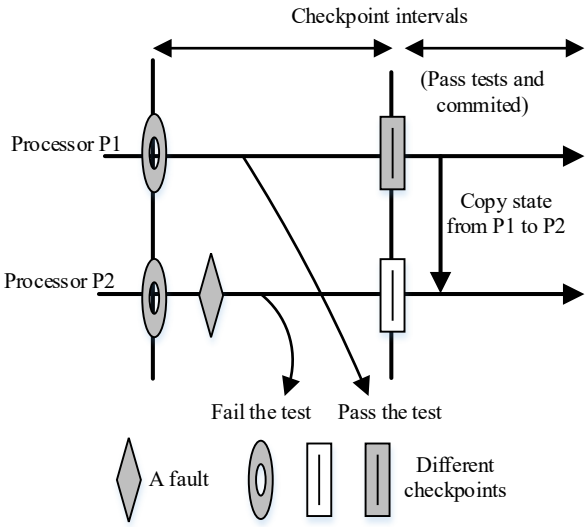


Figure 12. Roll-forward recovery in RFR-BC scheme [19,20].

Table 5. Strengths and weaknesses comparison of error detection methods.

Error detection technique	Strengths	Weaknesses
EDDI	The EDDI technique achieves high fault coverage by placing a comparison instruction after each MI and SI in the program.	Each of the original instruction is converted into three instructions, so performance and memory overhead is increased.
ED^{4}I	This technique achieves a high fault coverage when using the optimum value for *k*.	Good for either integers or floating point numbers, but not for both.
Software approach	Fault coverage achieved with this technique is very high because of duplicating the entire program.	More number of required instructions for duplication and comparison lead to increase in overhead.
CBD	Performance and memory overhead are decreased because they use less redundant instructions for duplication	This technique is a compiler and/or case study dependent and could just act as a full duplication technique.
CPD	Due to less number of instruc-tions required for duplication, performance and memory over-head is decreased.	In CPD, creating a data flow graph is much harder in assembly than C and C++.
SEDSR	In SEDSR, by placing a comparison instruction after writing to the classified final variables detect a lot of errors.	This technique is also a compiler and/or case study dependent and could just act as a full duplication.
Checking rules	The appeared performance and memory overhead in this technique are decreased in comparison to the EDDI, and ED^{4}I techniques.	Fault coverage achieved with this technique is reduced because of using less number of redundant instructions with the choice of checking rules.
GA	Because of selective vulnerable block duplication, the presented technique uses a less number of redundant instructions.	There is a possibility of undetected errors in the other normal blocks which lead to a reduction in fault coverage.
Shoestring	By duplicating and comparing only high-valued instructions, leads to decrease in overhead.	Fault coverage is reduced because of the possible undetected errors in the unduplicated instructions.

Table 6. Strengths and weaknesses comparison of error recovery methods.

Error recovery	Strengths	Weaknesses
Rollback error recovery	If the error is detected, the processor state is restored into error free state with out using spare processor.	Time overhead is the main drawback of the rollback error recovery, because of its rollback.
RFR-RC	Time overhead in RFR-RC scheme is decreased.	Usage of a spare processor to avoid the rollback, the cost is getting high.
RFR-BC	RFR-BC scheme in roll-forward recovery does not need a spare processor as in RFR-RC to avoid rollback.	Self-check detection has an inaccurate error coverage and can't detect certain types of faults.

5. Future Work

The strengths and weaknesses given in this paper for each technique were determined theoretically, by analyzing the technique and determining what data flow errors they detect, which they neglect and which overhead the techniques introduce.

To guide researchers and embedded systems engineers, we will perform an experimental comparison of the mentioned techniques. This experimental comparison will allow to evaluate the techniques on the same base: same hardware, same case studies, and same fault injection process. We'll perform this comparison both for data flow detection techniques and data flow recovery techniques. The outcome of the experimental comparison will allow other applicants of the techniques to quickly determine which existing technique is the best, in general or for their application.

Finally, we'll use the gathered data from the experimental comparison to develop a technique that can detect and recover from data flow errors, without introducing abnormal overhead.

6. Conclusions

This review paper lists and reviews various data flow error detection and recovery techniques existing in literature in the field of embedded systems. Each of the considered data error detection and correction technique has been discussed in terms of strengths and weaknesses. The discussion is summarized in Tables 5 and 6.

After thoroughly reviewing the strengths and weaknesses of error detection techniques, we have found that some of the techniques such as *ED^4I, EDDI and Software approach* are good for fault coverage but come with high overhead both in memory and performances. On the other hand, some other techniques such as *CBD, CPD, GA and Shoestring* are good for overhead but come with a reduction in fault coverage. At the same time, there exist a couple of techniques such as *SEDSR and Checking rules* which are good for fault coverage with satisfactory overhead.

As error recovery is concerned, based on the identified strengths and weaknesses from the methodology, we have found that *checkpointing technique for rollback error recovery* is better if memory overhead is the main concern. *Checkpointing technique for roll-forward error recovery schemes* are better if time overhead is the main concern.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This work was funded by a technology transfer (TETRA) grant from the Flemish government in Belgium (VLAIO), under grant number IWT-150128 RELIM (Real-life immunity for embedded systems).

References

[1] Feng S, Gupta S, Ansari A, Mahlke S. Shoestring: probabilistic soft error reliability on the cheap: In: ACM SIGARCH Computer Architecture News.
[2] IEC 61508; Available from: http://www.iec.ch/functionalsafety/explained/.
[3] Li A, Hong B. Software implemented transient fault detection in space computer. Aerospace science and technology 2007;11(2):245–52.
[4] Thati VB, Vankeirsbilck J, Boydens J. Comparative study on data error detection techniques in embedded systems: In: 2016 XXV International Scientific Conference Electronics (ET), Sozopol, Bulgaria.
[5] S. Jagannathan, Z. Diggins, N. Mahatme, T. D. Loveless, B. L. Bhuva, S-J. Wen, R. Wong, and L. W. Massengill (ed.). Temperature dependence of soft error rate in flip-flop designs: IEEE; 2012.
[6] R. Baumann. Soft errors in advanced computer systems 2005;22(3):258–66.
[7] Oh N, Mitra S, McCluskey EJ. ED 4 I: error detection by diverse data and duplicated instructions. Computers, IEEE Transactions on 2002;51(2):180–99.
[8] Reis GA, Chang J, Vachharajani N, Rangan R, August DI. SWIFT: Software implemented fault tolerance: In: Proceedings of the international symposium on Code generation and optimization.
[9] Seyyed Amir Asghari, Okyaya Kaynak, and Hassan Taheri. An investigation into Soft Error Detection Efficiency at Operating System Level 2014.
[10] Dhiraj K. Pradhan, Nitin H. Vaidya: 24th International symposium on fault-tolerant computing Papers: FTCS-24; 1994.
[11] Chris Inacio. Software Fault Tolerance; Available from: https://users.ece.cmu.edu/~koopman/des_s99/sw_fault_tolerance/.
[12] Abdi A, Asghari SA, Pourmozaffari S, Taheri H, Pedram H. An Optimum Instruction Level Method for Soft Error Detection. International Review on Computers and Software 2012;7(2).
[13] Oh N, Shirvani PP, McCluskey EJ. Error detection by duplicated instructions in super-scalar processors. Reliability, IEEE Transactions on 2002;51(1):63–75.
[14] Abdi A, Asghari SA, Pourmozaffari S, Taheri H, Pedram H. An Effective Software Implemented Data Error Detection Method in Real Time Systems: In: Advances in Computer Science, Engineering & Applications: Springer; 2012, p. 919–26.
[15] Chielle E, Kastensmidt FL, Cuenca-Asensi S. Overhead Reduction in Data-Flow Software-Based Fault Tolerance Techniques: In: FPGAs and Parallel Architectures for Aerospace Applications: Springer; 2016, p. 279–91.
[16] Arasteh B, Bouyer A, Pirahesh S. An efficient vulnerability-driven method for hardening a program against soft-error using genetic algorithm. Computers & Electrical Engineering 2015;48:25–43.
[17] Asghari SA, Abdi A, Taheri H, Pedram H, Pourmozaffari S, others. SEDSR: soft error detection using software redundancy. Journal of Software Engineering and Applications 2012;5(09):664.
[18] Oh N, McCluskey EJ. Error detection by selective procedure call duplication for low energy consumption. Reliability, IEEE Transactions on 2002;51(4):392–402.
[19] Pradhan DK, Vaidya NH. Roll-Forward Checkpointing Scheme: A Novel Fault-Tolerant Architecture. IEEE Trans. Comput. 1994;43(10):1163–74.
[20] Jie Xu B. Roll-forward error recovery in embedded real-time systems: 1996 international conference on parallel and distributed systems; 1996.
[21] Mohsen Bashiri, Seyed Ghassem Miremadi and Mahdi Fazeli (ed.). A Checkpointing Technique for Rollback Error Recovery in Embedded Systems: IEEE Xplore; 2006.
[22] Nicolescu B, Savaria Y, Velazco R. Software detection mechanisms providing full coverage against single bit-flip faults. Nuclear Science, IEEE Transactions on 2004;51(6):3510–8.
[23] Karnik T, Hazucha P. Characterization of soft errors caused by single event upsets in CMOS processes. IEEE Trans.Dependable and Secure Comput. 2004;1(2):128–43.
[24] Arasteh B, Miremadi SG, Rahmani AM. Developing Inherently Resilient Software Against Soft-Errors Based on Algorithm Level Inherent Features. J Electron Test 2014;30(2):193–212.
[25] M. Li, P. Ramachandran, S. Sahoo, S. Adve, V. Adve, and Y. Zhou (ed.). SWAT: An Error Resilient System; 2008.
[26] Vankeirsbilck J, Thati VB, Hallez H, Boydens J. Inter-block jump detection techniques: A study: In: 2016 XXV International Scientific Conference Electronics (ET), Sozopol, Bulgaria.
[27] Oh N, Shirvani PP, McCluskey EJ. Control-flow checking by software signatures. Reliability, IEEE Transactions on 2002;51(1):111–22.

[28] Z. Alkhalifa, V.S.S. Nair, N. Krishnamurthy and J.A. Abraham. Design and evaluation of system-level checks for on-line control flow error detection 1999;10(6):627–41.
[29] O. Goloubeva, M. Rebaudengo, M. Sonza Reorda and M. Violante (ed.). Soft-error detection using control flow assertions: IEEE; 2003.
[30] Mei-Chen Hsueh, Tsai TK, Iyer RK. Fault injection techniques and tools. Computer 1997;30(4):75–82.
[31] Nicolescu B, Velazco R. Detecting soft errors by a purely software approach: method, tools and experimental results: In: Embedded Software for SoC: Springer; 2003, p. 39–51.
[32] Rebaudengo M, Reorda MS, Torchiano M, Violante M. Soft-error detection through software fault-tolerance techniques: In: Defect and Fault Tolerance in VLSI Systems, 1999. DFT'99. International Symposium on.
[33] Cook JJ, Zilles C. A characterization of instruction-level error derating and its implications for error detection: In: 2008 IEEE International Conference on Dependable Systems and Networks With FTCS and DCC (DSN), Anchorage, AK.
[34] Wang N, Fertig M, Sanjay Patel. Y-branches: when you come to a fork in the road, take it: In: 12th International Conference on Parallel Architectures and Compilation Techniques. PACT 2003, New Orleans, LA, USA, 27 Sept.-1 Oct. 2003.
[35] Weiser M. Program slicing: In: Proceedings of the 5th international conference on Software engineering.
[36] Dhiraj K. Pradhan, Nitin H. Vaidya. Roll-forward and rollback recovery: performance-reliability trade-off 1994;43(10).
[37] Silva JG, Silva LM, Madeira H, Bernardino J. A fault-tolerant mechanism for simple controllers: In: Goos G, Hartmanis J, Leeuwen J, Echtle K, Hammer D, Powell D, editors. Dependable Computing — EDCC-1. Volume 852. Berlin, Heidelberg: Springer Berlin Heidelberg; 1994, p. 39–55.
[38] Fault Tolerance: Principles and Practice: Springer-Verlag New York Inc; 2013.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 634-641 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Augmented Reality Prototype HUD for Passenger Infotainment in a Vehicular Environment

Shu Wang[1], Vassilis Charissis[1,*], David K. Harisson[2]

[1] Glasgow Caledonian University, Department of Computer Communications and Interactive Systems, G40BA, UK

[2]Glasgow Caledonian University, Department of Engineering, G40BA, UK

ARTICLE INFO

Article history:
Received : 06 April, 2017
Accepted : 15 May, 2017
Online: 04 June, 2017

Keywords:
Head Up Display
Augmented Reality
In-Vehicle Infotainment

ABSTRACT

The paper presents a prototype Head Up Display interface which acts as an interactive infotainment system for rear seat younger passengers, aiming to minimize driver distraction. The interface employs an Augmented Reality medium that utilizes the external scenery as a background for two platform games explicitly designed for this system. Additionally, the system provides AR embedded information on major en route landmarks, navigational data, and local news amongst other infotainment options. The proposed design is applied in the peripheral windscreens with the use of a novel Head-Up Display system. The system evaluation by twenty users offered promising results discussed in the paper.

1. Introduction

Head-Up Display (HUD) interfaces are currently emerging as an increasingly viable alternative to traditional Head-Down Displays (HDD). HUDs present fresh opportunities for the presentation of information using symbolic and/or alphanumeric representation. By occupying larger estate directly on the driver's field of view, they can provide crucial information swiftly and without distracting the driver. As such they can enhance significantly a driver's information-retrieval capacity and response times in near collision situations [1,2,3].

Prototype HUD design interfaces and devices have significantly mitigated this issue of the driver's attention being diverted from his/her field of view, as shown in previous studies [3,4 & 5]. Yet, the requirements of the passengers' user group have not been adequately explored. In particular, rear seat passengers, specifically children, can seek to attract their parents' attention during long distance trips or whilst commuting. Such actions have a detrimental impact on the driver's attention, and could potentially lead to an accident. A UK survey that collected 2000 British parents' opinion about their children's behaviour whilst they were driving reported findings of around 62% of the parents feeling more at ease without their children in the car, 43% feeling tense and irritated with their children, and about 55% admitting that they were prone to losing their temper while driving long distance [6]. Furthermore, some parents used mirrors, not to check external road conditions, but to glance at and check their children's behaviour in the back seat.

This action could lead to hazardous driving and could potentially cause traffic collisions [7,8]. In light of the aforementioned facts and observations this project aims to collate and explore the current state of technology in infotainment car devices, as a base for launching the design and evaluation of the Human-Computer Interaction (HCI) for applications to rear-side HUD displays. The latter would enhance passenger's entertainment in the vehicular environment and provide visual and auditory information regarding the external environment. The utilization of HUD draws from our previous experiments with collision avoidance interfaces that achieved significant results towards the reduction of the collision probability in adverse weather conditions [4,5,9].

Overall, the paper is structured as follows: The paper introduces the target group and current driver distraction issues. The above form the framework requirements used for the development of the proposed HUD interface. In turn the paper presents a description of the proposed system and related design considerations. The evaluation process and results of twenty users are discussed. Finally, conclusions are summarized and a future plan of work is presented.

*Corresponding Author: Vassilis Charissis, Glasgow Caledonian University, Department of Computer Communications and Interactive Systems, G40BA, UK
E-mail: vassilis.charissis@gcu.ac.uk

2. Current In-Vehicle Distraction Issues and Solutions

Previous studies have observed that long-distance journeys in a car, constrain family members in the confined vehicle space for extensive periods of time. Typically, the adult members of the family are trying to keep the younger passengers occupied through various activities such as chatting, singing songs or playing on mobile phones [7]. Adults and children, however, frequently have different expectations during travel time, with children being inclined to expect the time to be enjoyable and playful, whilst adults prefer it to be relaxing and quiet [7]. Children usually start becoming bored and losing interest after about 30 minutes of long distance driving, resulting in negative feelings of parents as noted above, with 60% of parents dealing with these situations by lying about the journey time and 70% choosing to buy food and drinks in an attempt to divert attention and thus resolve the situation [6].

Despite of these, recent studies also demonstrate that an increasing number of families with dependent children choose to take their children to school by car or to travel long distances with the family [7,8].

Evidently the in-vehicle interactions between the passengers, particularly the younger ones, can distract a driver's attention and increase dramatically the collision probability. As such, the following section will present a potential solution to this issue that could be mutually beneficial for both the driver and passengers.

A number of electronics and automotive manufacturers have made tentative plans and experiments in order to maintain the rear passengers occupied during long distance traveling or commuting. Early examples include "Backseat Playground" designed in 2009 by Mobility Studio in Sweden. The game aims to improve the travel experience for the passengers. The most significant feature of the game was to combine the game's sounds with the actual travel location sounds, to enhance the experience of the game, and many participants pointed out that this feature was so vivid that it made them confuse the virtual game with the actual geographical location. Hardware including a pocket PC, microphone, magnetometer, gyroscopes, headphones, laptop, GPS and wireless LAN were used to achieve this feature [10].

The game utilizes landmarks and transforms then into an imaginary land occupied by virtual creatures and treasures. The system utilizes a gaming device which operates by pointing it towards road objects as they pass by. The players can play individually or collaborate with other players in meeting road traffic. Although it has many mini stories, the whole game is more suitable for travelling for a short time, with an increase in distance travelled leading to the story becoming more unclear and potentially confusing [10].

In 2011 "nICE" was designed by a German University aiming to increase the communication experience during travel time. This application was specially designed for BMW; in 2008 BMW already had some in-car entertainments devices, such as DVD, TV and music. This game uses headphones (for the driver who can decide to join in or not) and two touch pads (for front and back seat passengers or children) to play and which can link with each other with a wireless connection. The game is a picture puzzle game; there are pictures in the application album, and the passenger can randomly choose one and begin the game. Each picture has a missing part and the player needs to solve the mini puzzle game to earn these missing parts, for instance, a music quiz, drawing pictures and a labyrinth game. Passengers can choose particular mini game which they are good at, and win the game as a team [11].

In 2013, "Mileys" was developed as an in-vehicle game designed particularly for families traveling long-distance journeys [7]. The aim of the game was to educationally enrich the driving time by providing geographical and location orientated information. The game employs AR technology, GPS, radar and mobile phone interaction in order for users to present and access the provided information. In the game, parents can see the location of the character Miley, which they can position de novo at a location or find Mileys already planted or dropped by other people. At the same time, children use a radar device to search different positions in order to find the exact location and in the process, talk to their parents. When the family reach the location, children can use the phones to pick up Miley and a secondary objective is to keep the character's health high with safe and steady driving. Once/if the health runs out, the children will get their final score and the character Miley will be dropped and wait for the next player to pick it [7].

3. Head-Up Display vs Head-Down Display

The HUD systems typically comprise a projection unit embedded on the vehicle dashboard. The projected image is in turn inverted with a set of mirrors directly positioned in front of the projector, and reflected on the windshield area. The area of the windscreen that receives the projection is covered with a transparent surface, namely the combiner, which enables the correct depiction of the projected colours and shapes. Depending on the system calibration, the image can be superimposed on the environment, seemingly appearing at approximately 1.5m to 2.5m ahead of the windscreen [12, 13]. This is deemed as an ideal projection distance in order to avoid the cognitive capture effect which forces the human eye to focus between two different layers, that in turn results in degradation of human attention and performance [14,15]. Evidently, the HUD systems adhering to the aforementioned design requirements can improve significantly the driver's response times in accident situations [1,2, 9, 13,15].

In contrast the information depicted on the vehicle's dashboard is known as Head-Down Display (HDD). The HDD by default, forces the user to take their eyes from the road, and focus momentarily on the lower section of the vehicle (dashboard) which accommodates multiple infotainment displays [15,16,17,18]. A plethora of studies have analysed the drivers' cognitive load in different scenarios involving vehicle instrumentation, navigation systems, radio, CD and mobile phones amongst other devices [16,17,18,19]. The conclusions of such studies suggest a high collision probability whilst the driver

operates any of the aforementioned in tandem to the main driving task [1,2,4,20, 21].

In our case, we opted for a HUD system in order to enable the younger passengers to look out of the window and experience the route. The HUD in this case can offer an Augmented Reality gaming environment that could mix the Computer-Generated Image (CGI) game characters with the real environment and create different gameplay experience on each drive [22, 23]. Additionally, the HUD offers the capability to project succinct educational information regarding the landmarks and areas viewed during the car-travel. The software and hardware system requirements are presented analytically on the following sections.

4. Proposed System Solution

Significant attempts to enhance the rear passengers commuting experience via in vehicle game solutions have been described in section 2, however, the issue of passenger infotainment in a vehicular environment remains largely unsolved [24, 25]. Previously suggested solutions and research projects have managed to identify various interactions and patterns that could possibly occupy the rear passengers, they were, however, not offering a holistic approach that could adapt to different ages and interest groups.

The proposed system aims to offer a combination of activities for the rear passengers in order to avoid in-vehicle distractions that could affect driver's performance and concentration. The system is designed to facilitate both educational and entertainment activities with particular focus on the younger passenger age group.

4.1. Software requirements:

The HUD system is comprised of a generic interface that allows the rear-passengers to access information in real-time through the navigation GPS. The navigation type of information can vary from maps and navigation, distance covered, estimated time of arrival and highlighted on-route landmarks (i.e monuments, churches, castles etc.). The latter can be accessed in real-time, through online connectivity.

The landmark related specifics can be superimposed through the side HUD to the landmarks approached en route providing succinct information. This aims to keep the rear passengers occupied throughout the travelling time whilst educating them with regards to the surrounding environment. Additionally, the system offers an overview of the vehicle functions such as speed, consumption, revs amongst other, that might be of interest to the passengers.

The second arm of the interface offers the entertainment suite which entails games, movies, internet, music and audiobooks. The audio related data can be provided by individualized headsets so as to avoid further in vehicle sound distractions. On the entertainment section we have primarily targeted the younger audience with a set of onboard games that could be augmented in the external window scenery. As such we have developed (a) a platform flying game and (b) a historic Augmented Reality (AR) game.

Fig. 1: Prototype HUD interface and offered activities.

In the first game the user commands a superhero flying over the scenery whilst avoiding and shooting back a flying villain. As the scenery constantly changes both rival characters' battle in different terrains and weather conditions (figure 2).

The second game aims to entice the upper age limit of the passengers as it utilizes the external scenery and landmarks as temporary objects for the game. The game employs the external scenery of a real castle (i.e. Stirling Castle) for the duration that the user can see the castle through his/her window. In turn the user has to complete a set of challenges, such as assaulting the castle with the use of a 3D catapult. If the catapult shots are successful, the user gathers some points and continues to the next landmark and task.

This can obviously be altered depending on the country, landmarks and activities that the user wishes to access and interact with. In this paper we examine the overall user experience in regards to the HUD interface and the first game.

4.2. Hardware requirements:

The projected interface is presented on a rear passenger side window HUD system. The selection of the particular surface was chosen as it is in close proximity to the passengers and can utilise the external environment and scenery for augmented reality and geotagged application. Also, the side windows on any vehicle offer a large area for any type of data projection without requiring additional monitors. The utilisation of this highly neglected surface within the vehicle additionally offers an interactive projection field within arm's distance.

In the Virtual Reality Driving Simulation laboratory (VRSD Lab) a dedicated High Definition (HD) mini projector was employed to project the HUD interface and context on a custom transparent film positioned on the top of the glass window. The mini projector superimposes both the external view and the HUD interface (game and generic interface) in a side window of a full scale vehicle as presented in Figure 2. This method offered a realistic projection of the HUD.

Due to cost limitations and time constrains, a suitable micro projector was not a viable option at this stage. An alternative system would have been the transparent screen by Samsung [26],

however for the same reasons as stated above it was not feasible to use it for this experiment.

Fig. 2: The figure presents the screenshot of the HUD interface and projected Augmented Reality en route to Stirling Castle in Scotland.

Prior to selecting the particular projection method alternative emerging technologies have been investigated. The Virtual Reality (VR) option was a simple and easier option yet it was isolating and depriving the user from experiencing the environment view and the route.

Furthermore, the bulky Head Mounted Display (HMD) is not ideal for younger users as they might suffer significant neck injuries in an abrupt braking situation. The more advanced Augmented Reality (AR) option of Microsoft HoloLens was also excluded for the same concerns related to cost and safety of the system in a vehicular environment by younger passengers.

A second side HD projector was mounted externally on the ceiling. The latter projector was projecting a video recorded footage from a 45 minutes' drive nearby the Stirling Castle. The interface was designed to provide geospatial information related to the en route landmarks.

Notably the Stirling Caste is visible on the background for the duration of each trial whilst the user can operate the interface. The interface provides related educational and historic information, local news and weather amongst other information.

The overall simulation is run by a custom server PC in a Cave Automatic Virtual Environment (CAVE) which uses five HD 3D projectors that create a full surround driving environment surrounding a real-life Mercedes A Class vehicle as presented in the Figure 3 & 4 below.

The vehicle has rebuilt and rewired steering wheel and pedals compatible with the Unity3D game engine. The vehicle has also a central vibration system which enhances the motion feeling. A 5.1 surround audio creates the vehicle and environment sounds. The internal mini projector provides the required audio feedback for the HUD interface and the related infotainment activities.

The physical interaction with the HUD interface is either through gesture recognition software or typical game pad device. The first is utilizing a Leap Motion device and the second an XBOX controller described succinctly below.

4.2.1 Leap Motion Interaction:

Prior to the main user trials, the Leap Motion device was tested in a desktop PC environment and received positive feedback in preliminary trials. The particular device was chosen as an ideal system to interact with the provided interface as it was not involving any handheld devices [27].

In the main trials the Leap Motion was positioned with an adhesive tape on the door inside plastic section, underneath the car window in close proximity to the user. Although the system performed well the users' experienced arm fatigue as their right hand was operating the system without any support. This was in contrast to the desktop environment that offered the table as arm rest for the focus group users.

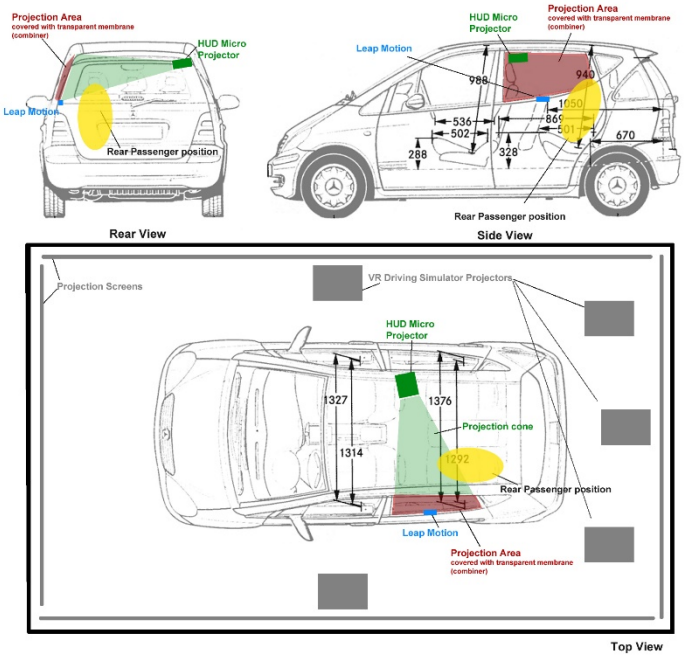


Fig. 3: Explanatory design of the HUD system and the VR Driving Simulator environment.

The game developed for the Leap Motion involved a cartoon style battle between a hero and a villain as depicted in Figure 2. The hero, controlled by the user was operated by pointing with one finger and moving the flying hero in four directions (up, down, forwards and backwards). The aim of the game was to fly the hero and avoid the fireball projectile shot by the villain. As the real environment background was changing en route their battle was taking place in different real locations. To activate a different app of the bottom menu the user had to point and push the button choice (without touching the window glass). The particular gestures were simple and efficient [28,29] however the duration of the actions and the position of the user within the vehicle, resulted in users' arm fatigue.

The game was developed in Unity3D game engine, exclusively for the evaluation of the proposed HUD system.

4.2.2 XBOX controller Interaction:

The typical XBOX controller was chosen as an alternative to the Leap Motion. The handheld controller aimed to counteract the fatigue issues and offer a familiar and faster interaction medium. The controller's drawback was mainly the handheld nature of the system. This could be a potential issue in an abrupt braking as the inertia of the device could result in injuries of the passengers.

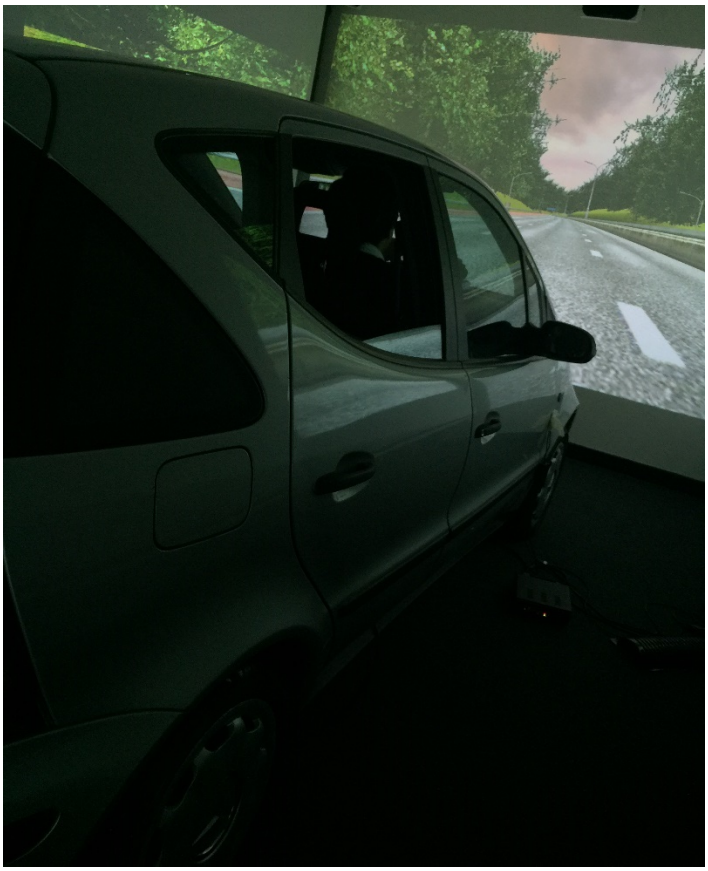


Fig. 4: Screenshot of the GCU VR Driving Simulator in action.

5. System Considerations

Adhering to the situations noted above, the proposed in-vehicle games should be able to entice the users to play for the duration of long distance runs whilst providing interesting information for the surroundings [4]. Due to the nature of these games, the in-vehicle activities should abide to a number of considerations in order to have the expected positive results according to Broy's research [10].

The preservation of driving safety, is a primary consideration, as the risk of the game distracting the driver by sound or virtual effects should be minimized.

Secondly, in-car games are not like traditional games or computer games, as they require a different playing environment, and should function steadily and effectively in a vehicular environment. As such the particular games should be simple and forgiving to minor movements that might occur due to road imperfections. Additionally, by utilizing the Augmented Reality (AR) provision of a HUD they could adapt to the external environment and create interesting and memorable game experiences [10, 11, 24, 30]. Interestingly this was highlighted by numerous remarks of our users in the simulation environment. Consequently, by utilizing the external environment the games do not require a full 3D gaming background in order to operate.

Additionally, motion sickness should also be taken into account, as some passengers feel uncomfortable when they read during travel, so it may be more comfortable and effective to use pictures and/or sound to present the information [9, 10]. Furthermore, Brunnberg provided evidence to suggest that, due to high speed travel, the view from the rear seat will pass by very quickly. It is therefore difficult for the device to identify and present the information, and this may be one of the most difficult technological challenges to be faced and overcome during the design process [9]. It should also be noted that passengers of different age groups (young children, teenagers, parents) have varying levels of ability in terms of reading and understanding; hence, the game's level of difficulty needs to be carefully considered [10].

Based on the aforementioned rear-passenger population differences the provided geospatial information should also be targeting the interest of the different age groups. As such we deemed essential to present local and international news as well as different educational data in different levels of detail and complexity for different ages. Additionally, the various movies or games which the HUD can project have been optimised for different age audiences. As such for the toddlers we incorporated simple platform games, popular kids' movies and indicative, simplified information for the external environment. For primary school pupils and teenagers, we introduced games and movies adhering to their preferences and age according to current top ten lists from relevant magazines and online publications. For adult passengers, we provided a collection of different online newspapers, news channels, weather channels mind games and contemporary movies. However, these were utilized as demonstration material of the system in a simulation environment. On a real-life commercial application, we would envisage the use of online applications to feed constantly the infotainment database and a small capacity Hard Disk Drive (HDD) that could maintain a selection of favourite movies and audio playlists.

Based on the aforementioned considerations, some in-car games have already been designed for short-distance drives situation as previously discussed. "Backseat Playground", "nICE" and "Mileys" use special equipment, phones or touchpads to play the games [7]. Similar studies have been conducted, in 2011 and 2012, by Toyota and General Motors (GM) respectively, utilizing enhanced versions of their backseat entertainment systems. Notably, GM used motion and optical sensor technology to transfer the rear seat window into a gesture touch panel [11,12,24].

However, there has been limited - if any at all - employment of HUD and gesture recognition technology, which could offer a more immersive and subtle way of interaction for the users. The ultimate purpose of this combination is to offer attention seeking infotainment to the passengers so as to reduce the level of distraction for the driver. All the current systems mentioned above have their own features and utilize their interfaces with multiple types of equipment, but without gesture recognition and typical console controls. Hence, the above, highlights the need for a new passenger infotainment and communication system that could alleviate the level of distraction of the driver.

6. Evaluation & Results

The evaluation process followed a threefold approach. Prior to commencing the main user trials we had to identify the nature of games that could be developed, projected and played in long duration within a vehicle environment. As mentioned above the idiosyncratic characteristics of these games offer new opportunities of Augmented Reality (AR) gameplay yet create a number of questions related with the type of games that could maintain passengers' attention and avoid motion sickness effect due to the vehicle's movement.

6.1. Focus Group

For this reason, a focus group of five users (3 male, 2 female) age 30 to 40 years of age was formed to test different popular games, aiming to identify the most suitable category for the vehicular environment.

The categories of games tested by a focus group prior to developing our own HUD games highlighted the difficulty of playing games that move in different directions to the vehicle's motion. The games tested were in categories of first person, action adventure, platform and table game.

The platform games following the vehicle's direction were perceived as easier to play without creating any motion sickness effect. In contrast, single person games that require from the user to move in 3D dimensions were confusing and tiring for the focus group users.

6.2. Main User Evaluation

The main evaluation was performed by 20 users varying from 4 to 13 years of age. The users tested the overall system. This paper focuses primarily on the game section of the overall interface evaluation. Two custom platform games were tested. Both games used the Leap Motion hardware in order to create gesture recognition interactions.

The results of the evaluation trials in this age group were promising and yielded a great deal of positive and useful feedback as presented in Figure 5. The first AR game received consistent approval and positive reactions from the participants. Yet certain issues were also pointed out, mainly regarding the gesture recognition system which required significant stamina from the users in order to operate it. This issue arose primarily from the demanding task of maintaining the arm lifted in order to operate the game on the side window. As an alternative, the system can be operated with a console wireless controller in order to avoid arm fatigue.

Notably, 90% of the participants rated the game highly. Furthermore, analyzing all the comments, gesture control is the focal point, with 80% of the participants enjoying the function of gesture control in the game, and 50% of those with no prior gesture control experience describing the gestures as hard initially, but acceptable and appropriate after a period of familiarization.

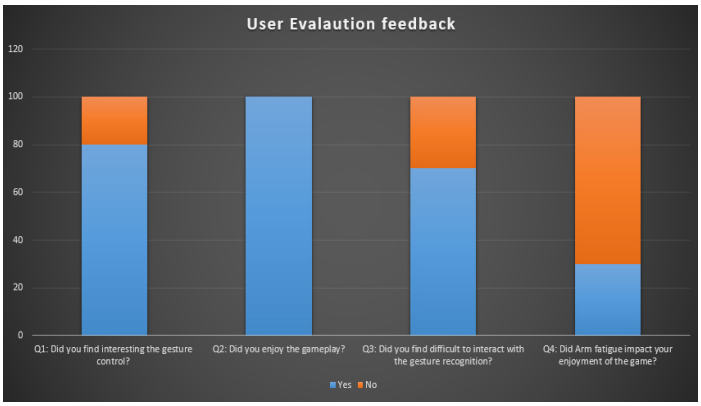


Figure 5: User evaluation feedback on main four usability questions.

Interestingly, despite expecting users to identify arm fatigue as the main problem caused by the hand gestures, the overwhelming response from the users was that arm fatigue did not significantly affect the participants' enjoyment of the game. 70% of the participants did not consider arm fatigue as an influencing factor in their enjoyment of the game. This result is not only contrary to the initial expectation but is also in contrast to the result of the test with the younger age group.

Although the participants did not think that arm fatigue affected their enjoyment in the game, they suggested that a pause button could be useful in extensive duration game-play. Finally, the HUD system and AR game managed to captivate the audience for the duration of the first level (10 minutes per level) and in some cases the users continued in the subsequent game-levels, achieving the 100% target for occupying the passengers for the duration of a typical long distance commute or trip [11].

6.3. Secondary User Evaluation and Driving Patterns

The aforementioned result had evidently a direct benefit to the driver's attention throughout the simulations. An indicative driver pattern driving the simulator with a toddler at the rear seat with and without the passenger HUD can be observed in Figure 6. This secondary trial aimed to identify the emerging driving patterns of the parent drivers during a short distance commute.

The simulation followed our previous work and challenged the driver to respond either by braking abruptly or performing collision avoidance maneuver, in a number of potential collision situations [1,2,4]. The simulation tested the parent/driver response times and collision occurrences with and without the use of the passenger HUD interface.

The simulation environment was a 3D photorealistic representation of an existing motorway section (M8) between Glasgow and Edinburgh. This smaller user trial involved 5 parents and their children. The driving patterns emerged from the five simulations which presented distinct similarities. The most interesting was that the use of the passenger HUD prevented any collisions as the driver was not concerned with the activities of the rear passengers.

In turn the children were preoccupied with the HUD interface for the duration of the simulation and were particularly interested in the custom platform games. In contrast the absence of the HUD activities on the simulation without the system, challenged the patience of the younger passengers which in turn distracted their driver/parent. However, this third part of the user trials was constrained by a number of factors.

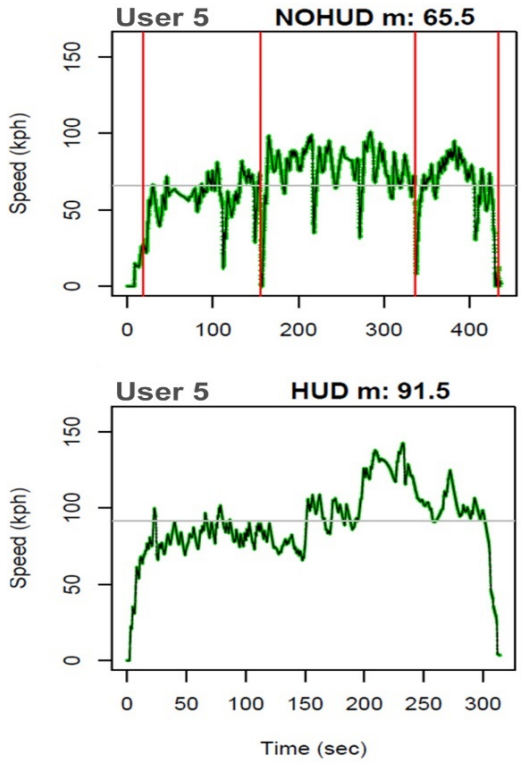


Figure 6: Indicative driving patterns recorded with and without the HUD interface. The Red vertical lines present the collisions occurred.

Primarily it was a difficult task to obtain both parents and toddler participants to run the simulation. Secondly it is very difficult to recreate in a short time the effect of boredom that can frustrate the younger age passengers. Furthermore, the VR Driving Simulator itself kept the young users occupied mitigating any boredom factor for substantially longer than in a typical everyday commuting environment.

Future multiple user trials will be required with the same users in order to create a familiarity with the VR Driving Simulator in order to avoid the third factor that affects the younger users' routine behaviour.

Additionally, users that are immune to motion sickness could potentially drive the simulator for longer trial times (i.e. two hours driving) for obtaining more accurate data. Acknowledging these three limitations, the HUD managed to maintain the rear seat vehicle occupants' attention and keep the parent-driver preoccupied with the driving task.

7. Conclusions & Future Work

This paper presented the design consideration, development and evaluation of a prototype Human Computer Interaction design for automotive HUD designed to entertain and inform the young passengers in the rear seat younger passengers. To facilitate an appraisal of the system, the experiment hosted two different games that utilized gesture recognition and typical console controls.

The results indicate that the participants were satisfied with the game's performance, with 80% of the participants having enjoyed the function of gesture control in the game, and half of the players who had no gesture control experience thinking that the gestures were hard at the start, but were acceptable and enjoyable after a period of familiarization. The problem of arm fatigue caused by hand gestures was highlighted in the experiments, even if not considered severe by the participants.

Yet, despite the positive outcomes of the presented work, it is apparent that the games need to be customized for the particular environment and take into consideration the duration of the trips, counteract the road-surface disturbances and adapt to the external weather conditions. Our tentative plan for future work entails the update of the games' and HUD interface's functionality in order to comply with the aforementioned observations and results. Consequently, further testing of the updated HUD system will be required in order to determine the optimal system parameters that will allow the system to facilitate lengthy interactions with minimal fatigue.

The motion sickness experienced by some users was not evaluated in this particular trial as this was not the aim of the particular experiment. However, we are interested to identify the motion sickness triggering mechanisms of this side window HUD interface in future trials explicitly designed to quantify this issue. In turn we plan to revise our interface in order to mitigate and reduce this side effect.

Additionally, we plan to introduce an interface indication for short breaks between HUD applications. This potentially will provide some time for the user's body to recover from the tilted position.

Finally, longer simulations and trials might be required in order to determine the passenger HUD's efficiency in quantitative trial with simulations of real life scenario.

Acknowledgment

The authors would like to acknowledge the contribution of Dr Warren Chan and Dr Soheeb Khan in the development of the Virtual Reality Driving Simulator.

References

[1] V. Charissis, and S. Papanastasiou, Human-Machine Collaboration Through Vehicle Head Up Display Interface, in International Journal of Cognition, Technology and Work, P. C. Cacciabue and E. Hollangel (eds.) Springer London Ltd Volume 12, Number 1(2010), pp 41-50, DOI: 10.1007/s10111-008-0117, 2010.

[2] V. Charissis, S. Papanastasiou, and G. Vlachos, Comparative Study of Prototype Automotive HUD vs. HDD: Collision Avoidance Simulation and Results, in Proceedings of the Society of Automotive Engineers World Congress 2008, 14-17 April, Detroit, Michigan, USA, 2008.

[3] T. Tsuyoshi, F. Junichi, O. Shigeru, S. Masao, and T. Hiroshi, Application of Head-up Displays for In-Vehicle Navigation/Route Guidance, 0-7803-2105-7/94 IEEE, 1994.

[4] V. Charissis, S. Papanastasiou, W. Chan and E. Peytchev, Evolution of a full-windshield HUD designed for current VANET communication standards, IEEE Intelligent Transportation Systems International Conference (IEEE ITS), The Hague, Netherlands, pp. 1637-1643. 2013.

[5] V.Charissis, W. Chan, S Khan, and R. Lagoo, Improving Human Responses with the use of prototype HUD interface, ACM SIGGRAPH Asia 2015, Kobe, Japan, 2015.

[6] Office of National Statistics UK, Commuting to Work, Travel and Transport Theme, Report, 2011.

[7] G. Hoffman, A. Gal-Oz, S. David, and O. Zuckerman, In-car Game Design for Children: Child vs. Parent Perspective, IDC,ACM, New York, NY, USA, pp. 112-119. 2013.

[8] J. Barker, Driven to Distraction: Children's Experiences of Car Travel,Brunel University, UK, DOI 10.1080/17450100802657962, 2009.

[9] V. Charissis, & M. Naef, Evaluation of Prototype Automotive Head-Up Display Interface, In Proceedings of the IEEE Intelligent Vehicle symposium, (IV '07), Istanbul, Turkey, pp.1–6. 2007.

[10] N. Broy, S. Goebl, M. Hauder, M, A Cooperative In-Car Game for Heterogeneous Players, AutomotiveUI 2011, Nov. 30th-Dec.2nd 2011, Salzburg, Austria, Copyright 2011 ACM 978-1-4503-1231-8/11/11, 2011.

[11] L. Brunnberg, O. Juhlin, and A. Gustafsson,"Games for passengers-Accounting for motion in location-based applications", ICFDG 2009, April 26-30, 2009, Orlando, FL, USA, Copyright 2009 ACM 978-1-60558-437-9, 2009.

[12] S. Okabayashi, M. Sakata, and T. Hatada, Driver's Ability to Recognize Objects in the Forward View With Superimposition of Head-Up Display Images, in Proceedings of the 1991 Society of Information Displays, pp 465-468, 1991.

[13] Y. Inzuka, Y. Osumi, and H. Shinkai,Visibility of Head-Up Display for Automobiles, in Proceedings of the 35th Annual Meeting of the Human Factors Society, San Francisco, Calif., September 2-6, Human Factors Society, Santa Monica, Calif., pp1574-1578, 1991.

[14] N.J. Ward, and A. M. Parkes, Head-Up Displays and Their Automotive Application: An overview of Human Factors Issues Affecting Safety, in International Journal of Accident Analysis and Prevention, 26 (6), Elsevier, pp 703-717, 1994.

[15] C. Dicke, On the Evaluation of Auditory and Head-up Displays While Driving, ACHI 2012, The Fifth International Conference on Advances in Computer-Human Interactions, 2012.

[16] K. Young, & M. Regan, Driver distraction: A review of the literature. In: I.J. Faulks, M. Regan, M. Stevenson, J. Brown, A. Porter & J.D. Irwin (Eds.). Distracted driving. Sydney, NSW: Australasian College of Road Safety, pp. 379-405, 2007.

[17] S. J. Kass, K. S. Cole, C.J. Stanny, Effects of Distraction and Experience on Situation Awareness and Simulated Driving, in International Journal in Transportation Research Part F, Elsevier Ltd. 2006.

[18] K.W. Gish, & L. Staplin, Human factors aspects of using head up displays in automobiles: A review of the literature (Report No. DOT HS 808 320). Washington, DC: National Highway Traffic Safety Administration, 1995.

[19] P. Green, Where do drivers look while driving (and for how long)? In R. Dewar, & P. Olson, (Eds.), Human Factors in Traffic Safety, 2nd ed., (pp. 77-110). Tucson, AZ: Lawyers & Judges Publishing Co, Inc, 2007

[20] A. M. Recarte, M.L. Nunes, Mental Workload While Driving: Effects on Visual Search, Discrimination, and Decision Making, Journal of Experimental Psychology, American Psychology Association, pp 119-137, 2003

[21] Siemens, VDO, Keeping Your Eyes on the Road at All Times; Siemens VDO's Color Head-Up Display Nears First Application, source Siemens VDO Automotive AG Web Available from: http://www.siemensauto.com, 2003.

[22] B.D. Wassom, Augmented Reality Law, Privacy, and Ethics: Law, Society, and Emerging AR, Elsevier, 2015.

[23] W. Wu, F. Blaicher, J. Yang, T. Seder, & D. Cui, A Prototype of Landmark-Based Car Navigation Using a Full-Windshield Head-Up Display System. Proceedings of the AMC'09, Beijing, China, 2009.

[24] C. Squatriglia, GM Opens a New window in Entertainment, WIRED, 2012. https://www.wired.com/2012/01/gm-windows-of-opportunity/

[25] J. Heikkinen, E. Makinen, J. Lylykangas, T. Panakkanen, KVV. Mattila, R. and Raisamo, Mobile devices as infotainment user interfaces in the car: contextual study and design implications. pp.137–146, 2013.

[26] J. Vincent, Samsung's Transparent and Mirrored Displays Look Like Optical Illusions, The Verge, 2015.
https://www.theverge.com/2015/6/10/8756355/samsung-transparent-mirror-oled-displays

[27] F. Weichert, D. Bachmann, B. Rudak, D. Fisseler, Analysis of the accuracy and robustness of the leap motion controller. Sensors, 14;13(5):6380-93. 2013.

[28] L.E. Potter, J. Araullo, & L. Carter L, The leap motion controller: a view on sign language. In Proceedings of the 25th Australian computer-human interaction conference: augmentation, application, innovation, collaboration, pp. 175-178. 2013.

[29] M.A. Rupp, P. Oppold, D.S. McConnell, Evaluating input device usability as a function of task difficulty in a tracking task. *Ergonomics*, 4;58(5):722-35. 2015.

[30] Y Takaki, Y. Urano, S. Kashiwada, H. Ando, K. Nakamura, Super multi-view windshield display for long-distance image information presentation, Optical Society of America, 2011.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 642-649 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ
ISSN: 2415-6698

Incorporating Spatial Information for Microaneurysm Detection in Retinal Images

Mohamed M. Habib[1], Roshan A. Welikala[1], Andreas Hoppe[1], Christopher G. Owen[2], Alicja R. Rudnicka[2], Adnan Tufail[3], Catherine Egan[3], Sarah A. Barman[*1]

[1] *School of Computer Science and Mathematics, Faculty of Science, Engineering and Computing, Kingston University, London, UK*

[2] *Population Health Research Institute, St. George's, University of London, United Kingdom*

[3] *Moorfields Biomedical Research Centre, Moorfields Eye Hospital, London, United Kingdom*

ARTICLE INFO

Article history:
Received : 07 April, 2017
Accepted : 12 May, 2017
Online: 04 June, 2017

Keywords:
Retinal Image Analysis
Microaneurysms
Image Registration
Image Segmentation

ABSTRACT

The presence of microaneurysms(MAs) in retinal images is a pathognomonic sign of Diabetic Retinopathy (DR). This is one of the leading causes of blindness in the working population worldwide. This paper introduces a novel algorithm that combines information from spatial views of the retina for the purpose of MA detection. Most published research in the literature has addressed the problem of detecting MAs from single retinal images. This work proposes the incorporation of information from two spatial views during the detection process. The algorithm is evaluated using 160 images from 40 patients seen as part of a UK diabetic eye screening programme which contained 207 MAs. An improvement in performance compared to detection from an algorithm that relies on a single image is shown as an increase of 2% ROC score, hence demonstrating the potential of this method.

1. Introduction

This paper is an extension of the algorithm originally presented in the International Conference on Signal Processing Theory and Applications (IPTA 2016) [1]. This work adds a combination of spatial information from two retinal images for microaneurysm (MA) detection. Diabetic Retinopathy (DR) is one of the most common causes of blindness among working-age adults [2]. Signs of DR can be detected from images of the retina which are captured using a fundus camera. Microaneurysms (MAs) are one of the early signs of DR. Several algorithms for automated MA detection from single 45 degree fundus images have been proposed in the literature. However, in many Diabetic Eye Screening Programmes, including the UK National Health Service Diabetic Eye Screening Programme (NHS DESP) [3], at least 2 views of the retina are captured including both optic disc centered view and the fovea centered images (Figure 1). These images overlap together and thus have common MAs that appear in both views (with variability in contrast). Despite the availability of both views, the algorithms that have been proposed have only taken into account the information contained in a single image. In this paper an increase in detection accuracy is achieved by fusing the information from two views of the retina.

Algorithms reported in the literature lie broadly in two categories: supervised and unsupervised techniques. Supervised techniques make use of a classifier to reduce the number of false detections. This classifier requires training on an additional training set in order to generate a classification model. Unsupervised methods do not require a classifier and hence no training step is needed.

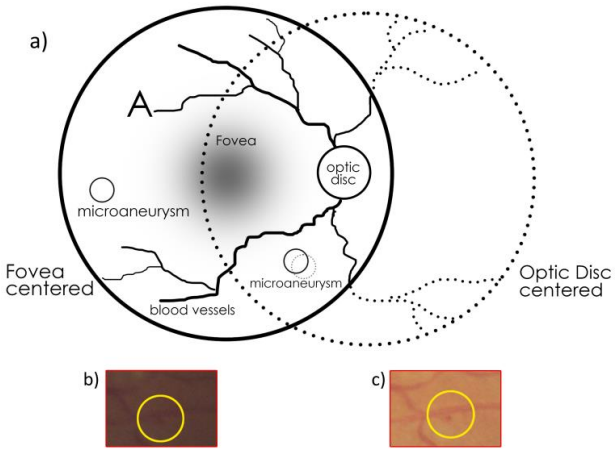


Figure 1. a) A conceptual diagram of the 2 spatial views of the retina (optic disc and fovea centered). The patches in b) and c) demonstrate how the same microaneurysm in different views of the retina can have a varying level of contrast.

The majority of the proposed methods in the literature fall under the supervised category. Most of the algorithms consist of three main stages: 1) Preprocessing 2) MA Candidate Detection and 3) Classification. The preprocessing phase corrects the image

[*]S.A. Barman, Kingston University, UK | E-mail: s.barman@kingston.ac.uk

with respect to non-uniform illumination and enhances MA contrast. MA Candidate Detection detects an initial set of candidates that are suspected to be MAs. While it is possible to stop here and report the detected results, a third phase is usually included in the algorithm to reduce the amount of false positives. This is the candidate classification phase and it classifies the detected candidates from phase 2 as either true or spurious.

A variety of classification techniques have been reported in the literature such as Linear Discriminant Analysis (LDA) [4] K-Nearest Neighbours (KNN) [5]–[8], Artificial Neural Networks [9], [10], Naive Bayes [11] and Logistic Regression [12]. There are a number of unsupervised methods that do not rely on a classifier include [13]–[16]. The obvious advantage of an unsupervised method is that they do not require a training phase. Some initial candidate detection methods that have been proposed are Gaussian filters [4]–[6] or their variants [8], [17], [18], simple thresholding [15], [16], [19], Moat operator [20], double ring filter [9], mixture model-based clustering [21] 1D scan lines [13], [14], extended minima transform [11], [22], Hessian matrix Eigenvalues [7], [23], Frangi-based filters [24] and hit-or-miss transform [10].

All the aforementioned methods were based on detection from a single colour image. Even though extra information may be available from another view of the retina, these algorithms are not designed to incorporate this extra information. A few methods have addressed the problem of detecting or measuring change from multiple images for the purpose of disease identification. Conor [25] performed vessel segmentation on a series of fundus images and measured both vessel tortuosity and width in these images. This was done in order to find a correlation between these measures and some signs including Diabetic Retinopathy. Arpenik [26] used fractal analysis to distinguish between normal and abnormal vascular structures in a human retina. Patterson [27] developed a statistical approach for quantifying change in the optic nerve head topography using a Heidelberg Retinal Tomograph (HRT). This was done for measuring disease progression in glaucoma patients. Artes [28] reported on the temporal relationship between visual field and optic disc changes in glaucoma patients. Bursell [29] investigated the difference in blood flow changes between insulin-dependent diabetes mellitus (IDDM) patients compared to healthy patients in video fluorescein angiography. Narasimha [30] used longitudinal change analysis to detect non-vascular anomalies such as exudates and microaneurysms. A Bayesian classifier is used to detect changes in image colour. A "redness" increase indicates the appearance of microaneurysms. Similarly, an increase in white or yellow indicates the appearance of exudates. While the problem of analysing "change" and "progression" of disease has been studied in the literature, to the best of our knowledge, the combination of a spatial pair of retinal images for the improvement of detection of MAs has not yet been explored.

The objectives of the present work are: 1) To present a novel method for combining information from two views of the retina (optic disc centered and fovea centered) and 2) Evaluate this method using a dataset of spatial image pairs. Following this introductory section, the methodology of the proposed method will be explained. Section 3 will discuss the details of the dataset and the methods employed for evaluating the proposed method. Results will be presented and discussed in Section 4. Final remarks and conclusions will be presented in Section 5.

2. Methodology

2.1. Method Overview

Figure 2 illustrates the overview of the proposed method. A way to compare MA candidate detection using the combined image versus the 2 singular images from the same patient was needed. A previous method [1] was used for the detection of candidates and measuring the probability of each candidate being true or spurious. The method was explained in detail in [1] and will be summarised in the following paragraph.

The previous method worked on the detection of microaneurysms from a single colour fundus image. The method was based on 3 stages: Preprocessing, MA candidate detection and classification. In the preprocessing phases, noise removal was performed and the image was corrected for non-uniform illumination by subtracting it from an estimate of the background. Salt and pepper noise was also removed during this stage. The vessel structure was removed from the image since vessel cross sections usually cause many false positive candidate detections. In the MA candidate detection phase a Gaussian filter response was thresholded in order to receive a set of potential candidates. Each candidate was region grown in order to enhance the shape of the candidates (to match the original shape in the image). Finally, during the classification stage, each region grown candidate was assigned a probability p between 0 and 1 representing the classifier's confidence in it being a true candidate or not. Each probability p can be thresholded at an operating point α to produce t such that:

$$t = \begin{cases} 1 \text{ if } p > \alpha \\ 0 \text{ otherwise} \end{cases}$$

Where $t = 1$ means that the corresponding candidate will be classified as true and $t = 0$ means it will be classified as spurious and removed from the candidates set.

The previous method has been adapted to work on 2 images of different spatial views. As shown in Figure 2, the 2 images are run through the algorithm as normal. This produces a set of 2 scores ($scores^{(1)}$ and $scores^{(2)}$). The intention is to combine these 2 scores together. However, before fusing the scores a way to find correspondences between candidates is needed. Therefore the images need to be aligned first (Image Registration) and then finding a match between corresponding candidates is needed (Candidates Matching). Each of the matched candidate's scores can then be combined to produce a single set of scores for both images. This combination should increase the confidence in some true MA candidates, and will hence improve the final algorithm after the final scores are thresholded with an operating point α.

In the following sections the Image Registration, Candidates, Matching and Fusion of Scores stages are described in greater detail.

2.2. Image Registration

Image registration is the process of aligning two images so that their corresponding pixels lie in the same space. One image is considered to be the *reference image* and the other image (known as the *moving image*) is transformed to be aligned to the reference image. A global transformation model was used which means that a single transformation was applied across the entire moving image. This has the advantage of simplicity and efficiency, but may not be as accurate as localised registration techniques. Since the goal of this paper was to introduce a proof of concept with regards to combining spatial information during microaneurysm detection, the accuracy achieved was sufficient for this purpose. More accurate registration techniques will be investigated in future work.

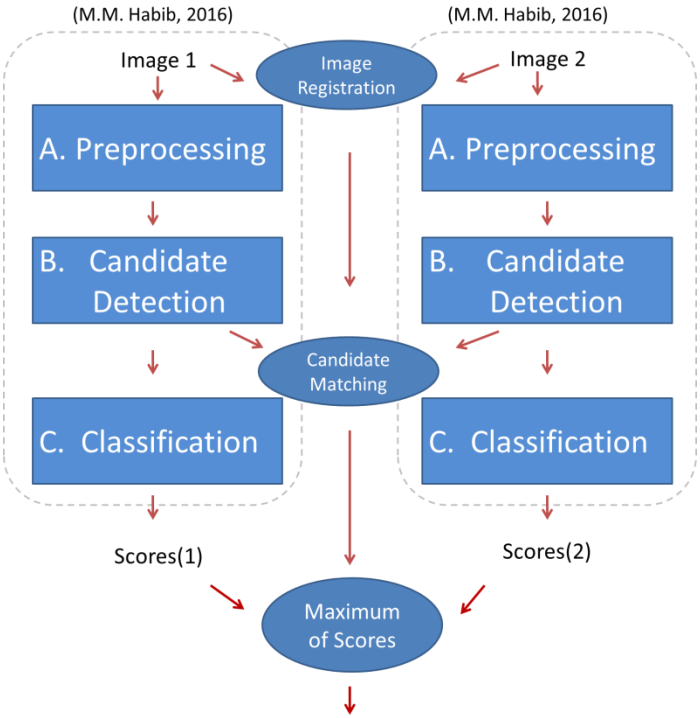


Figure 2. An overview of the proposed methodology.

A manual registration technique was employed where corresponding 'control points' were selected from each image. These control points were used to solve the global transformation model equations and find the transformation parameters (Figure 3). Corresponding points were annotated in each pair of images (as specified by Table 1).

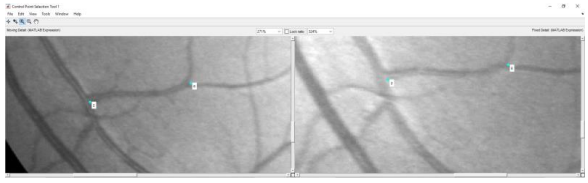


Figure 3. The manual control point selection process for the image registration phase.

Based on the literature, four transformation models were evaluated: These include Similarity [31], Affine [32], [33], Polynomial [34], [35] and RADIC [36], [37].

The transformation model parameters were estimated using the six control points that were manually selected on each pair of images. These control points were picked on each image's vessel cross sections since it was easiest to identify corresponding points at these areas. Figure 4 shows samples of checkerboard patches selected at random from the registered image pairs. In general it was difficult to identify the most accurately registered model by visual observation of the patches alone since there was an observed discrepancy in performance across rows.

In other words, none of the transformation models perfectly aligns the vessels in all four patches. Hence, a more objective method for selecting the model was needed. This will be discussed in Section 3.2.1.

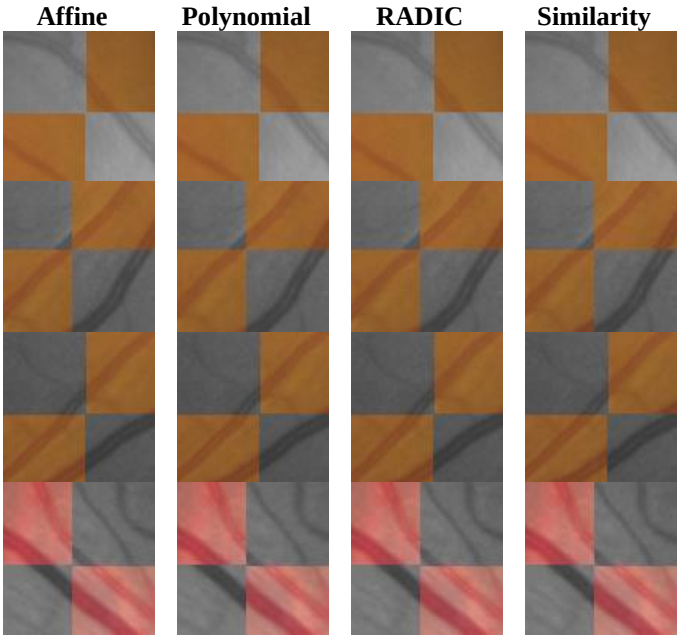


Figure 4. Sample checkerboard patches showing registration of multiple transformation models.

Table 1 The number of control points needed for each transformation model

Transformation model	No. of control points needed
Similarity	2
Affine	3
Polynomial	6
RADIC	3

2.3. Candidates Matching

Once both images were aligned the candidates detected from both images lie in the same coordinate space and hence can be matched by their location. In order to account for some inaccuracies in the registration, we used the following method to find matches between 2 candidates:

Given two aligned images I_1 and I_2, each candidate R detected in I_1 needs to be matched to one of the candidates detected in I_2. We start by finding the centre of candidate R and define a circular search region with radius r around R. A match is made with the candidate in I_2 whose center lies closest to R. If no candidate in I_2 is found in this region, no match will be made. This procedure is repeated for all candidates in I_1. In our case we defined r to be 15 pixels which is twice the size of an average candidate in our dataset

(Figure 5). This offers more tolerance to account for potential inaccuracies in the image registration.

More formally, let $P \in \{p_1, p_2, \ldots p_n\}$ be the set of candidates detected from the image I_1 and $Q \in \{q_1, q_2, \ldots q_m\}$ be the set of candidates detected from image I_2. Our goal is to find a set of correspondences $C = \{\{p_{r_1}, q_{s_1}\}, \{p_{r_2}, q_{s_2}\}, \ldots \{p_{r_l}, q_{s_l}\}\}$ (where $r_i \in [1..n]$ and $s_i \in [1..m]$) which represent the correspondences between these candidates. Note that some candidates would not have any correspondences and in this case they would not be a member of any set pair in C. In practice either P or Q are picked as a 'reference set' and matches are found from the other set. For instance, if we pick P as a reference then for each p_i we find a corresponding match from Q and add it to C if any exists. But we would not do vice versa – Q would not be used as a reference, and this is done for consistency, since we want to have consistent matches in order to make the fused scores consistent.

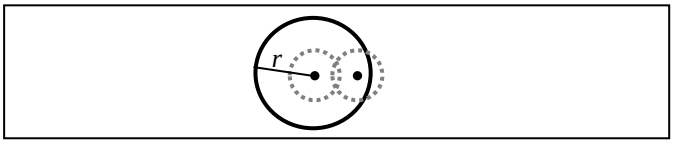


Figure 5. An illustration of the tolerance added while matching 2 candidates. The dashed circles represent two candidates from two views of the retina. It is assumed that they are misaligned due to registration inaccuracy. The solid circle represents a tolerance region around the first candidate and using this method a match is made.

Figure 6 shows an example of correspondences found after following the procedure above. The first row in the figure (a, b) shows a colour image pair while the second row in the figure (c,d) represents the green channel extracted from each image in the first row. Note that the candidates are matched from the image on the right column (b, d) to the image on the left column (a, c). The annotation numbers represent the matches from P to Q (A visual representation of C). Candidates annotated with "-1" in the right image represent a candidate that has no correspondence in the other image (no match found in C). The blue circle in the figure represents a true candidate. It can be seen that a match has been found between the true candidate in (b) and its corresponding candidate in (a). Furthermore, it is observed that the candidate has a much higher contrast in the right image than it does on the left one. The MA candidate is still visible in the left image but it much more subtle. Nevertheless, a combination of information from both candidates will give us higher confidence that is a true candidate.

In fact, human retinal graders often switch between both views of the retina when they have suspicions regarding one candidate. The existence of signs in both images would give graders more confidence about it being a microaneurysm. The process of matching in the proposed method attempts to replicate this.

2.4. Scores Fusion

Given that correspondences have been established between candidates and that each classifier has produced scores for each candidate we now need to find a final fused set of scores that represent a combination of information from both images. Suppose we have two matched candidates a and b from images I_1 and I_2 respectively. Furthermore, assume that I_2 is our reference image – i.e. we are currently interested to classify the candidates in I_2. We define the function fuse as follows (Figure 7):

$$\text{fuse}(a,b) = \begin{cases} \max(a,b) \; if \; \beta_1 < b < \beta_2 \\ b, \text{otherwise} \end{cases}$$

Where β_1 and β_2 are algorithm parameters specified between 0 and 1. In other words, given 2 matching candidates, the maximum of both their scores is taken only if b lies between the two threshold parameters (β_1 and β_2). The parameters β_1 and β_2 are used to limit the number of false candidates that get their scores maximized in the final set. This is because maximization of scores should only be done for true candidates

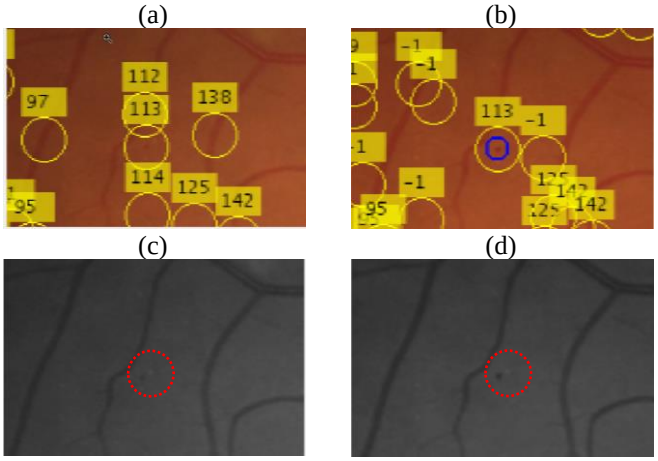


Figure 6. An example of the candidate detection result. a) A colour image from an optic disc centered image. b) The fovea centered view of (a). (c) The green channel image of (a). d) the green channel image of (b). The numeric annotations on (b) represent the result of the matching operation with a). '-1' represents no match. Candidate #113 is a true candidate and it has been correctly matched in both images (b). The candidate has variable contrasts in both images as can be seen in (c) and (d). The matching will hence improve the confidence regarding this candidate.

If a candidate in a reference image has no match in the other image then its score is simply copied over to the fused set (Figure 7). Once the fused score set is computed we can perform a final threshold at an operating point α to find the final set of classifications as described in Section 2.1.

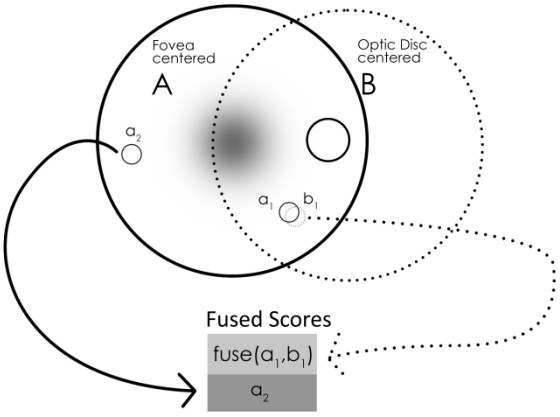


Figure 7. The method employed for the scores fusion phase. β_1 and β_2 are parameters set for the model.

3. Results

3.1. Dataset

The dataset used for evaluation consists of 40 patients imaged imaged as a part of UK NHS DESP. 4 images are available for each patient: fovea and optic disc centered images for 2 eyes. Hence there are 4 images per patient. The total amount of images is 160 images (Figure 8). The images were captured using different fundus cameras (including Canon 2CR, EOS and Topcon cameras) and hence were of different resolutions (Table 2) however they all had the same aspect ratio of 1.5. Moreover all the images had the same field of view (45 degrees).

The images in the dataset were split into 80 images for training the model and 80 images for validating it. The training set consisted of 112 MAs and the testing set consisted of 95 MAs.

Research Governance approval was obtained. Images were pseudonymized, and no change in the clinical pathway occurred.

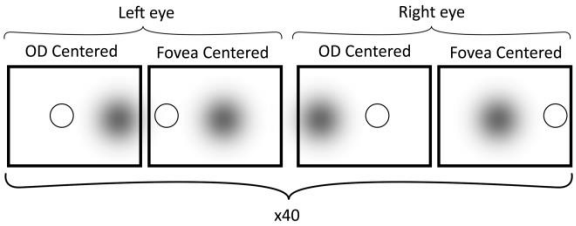


Figure 8. An illustration of the types of images available in the spatial dataset. OD centered – optic disc centered.

The same test set was used to validate both the proposed technique that takes into account spatial information and the normal technique. The classifier model was generated using the training set and the same model was used for both cases of with and without spatial information. To clarify, the same model was used to generate the scores for both the single-image method and the spatial-information method.

Table 2 The image resolutions of the dataset. All the images had an aspect ratio of 1.5.

Image Resolution	Count
3888 x 2592	24
2592 x 1728	48
3872 x 2592	12
4752 x 3168	32
4288 x 2848	20
3504 x 2336	24
TOTAL	160

3.2. Evaluation

In this section details regarding the evaluation of the registration transformation model and spatial information combination phases are presented. Section 3.2.1 details the selection of the appropriate transformation model objectively. Section 3.2.2 will describe the process for evaluating the spatial information combination and present its results.

3.2.1. Registration

As shown in Figure 4 it is difficult to decide which transformation model achieves best performance. Therefore we need a more objective measure of registration performance. During the registration we have a reference image I_r and a moving image I_m which is transformed to be in the coordinate space of I_r. After the image is transformed to the coordinate space of I_r we define an overlapping region as all the pixel coordinates where both I_r and I_m exist (overlap).

The Centreline Error Measure (CEM) [36] quantifies the mean of the minimum distance between each pixel along the centreline of the reference image and the closest pixel in the moving image. Given a set of coordinates in the reference image that lie on its vessel centreline (Figure 9) and are on the overlapping region of the two images: $V=\{v_1, v_2, \dots, v_N; v_i \in (x,y)\}$. Similarly, let $U=\{u_1, u_2, \dots, u_L; u_j \in (x,y)\}$ be the set of points on the moving image that lie on its vessel centreline and belong to the overlapping region. Let $t(p)$ be a transformation that transforms a point p from the moving image space to the reference space. We calculate the centreline error metric for a transformation $t(p)$ on the moving image as follows:

$$CEM = \frac{1}{L}\sum_{i=0}^{L-1} \left| M(\boldsymbol{u}_i) - t(\boldsymbol{u}_i) \right|$$

$$M(\boldsymbol{u}_i) = \underset{j=1..N}{\operatorname{argmin}}\ d(\boldsymbol{v}_j, t(\boldsymbol{u}_i))$$

Where $d(\boldsymbol{x}, \boldsymbol{y})$ represents the Euclidean distance between coordinates $\boldsymbol{x}$ and $\boldsymbol{y}$. Therefore the CEM calculates the average distance between each point on the reference image and the nearest points on the registered moving image (in the overlapping region).

A box and whisker plot of CEM values for each registration model in the same view was plotted in Figure 10. This plot is helpful to summarize the data since it shows the median value and the spread of the values (upper-quartile, lower quartile and the highest and lowest value). Based on this plot it is observed that while the polynomial model contains some of the lowest CEM values (highest accuracy) compared to other models, the distribution of its values also contain the highest spread (as can be seen by the height of the polynomial 'box', as well as the highest and lowest values). This can also be seen by the number of outliers in the polynomial plot. This high variance in values is also expressed by the standard deviation of the values as can be seen in Table 3. This suggests the undesirable "instability" of the polynomial model. We also see that the lowest standard deviation values are exhibited by both the affine and the similarity models (with the similarity model having a slightly lower standard deviation). Both the mean and standard deviations of these models are similar to each other which makes their performance comparable. The similarity model was selected since its values exhibited the lowest standard deviation.

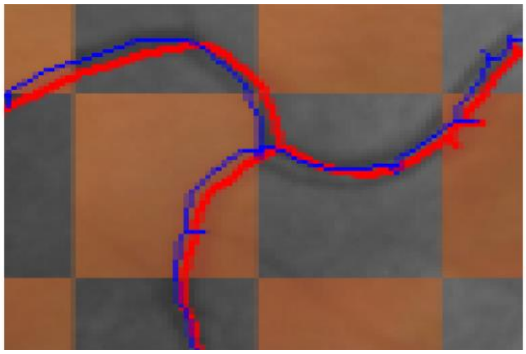


Figure 9 An example of overlapping vessel centerlines for computing Centerline error after registration. The distance between each red pixel and the nearest blue pixel is computed and the average of all distances is a measure of alignment accuracy.

Table 3 Centerline Error Metric (CEM) mean and standard deviation values for various transformation models. Std – Standard deviation.

	Similarity	Affine	Polynomial	RADIC
Mean	4.40	4.50	4.74	4.49
Std	1.51	1.536	3.82	1.536

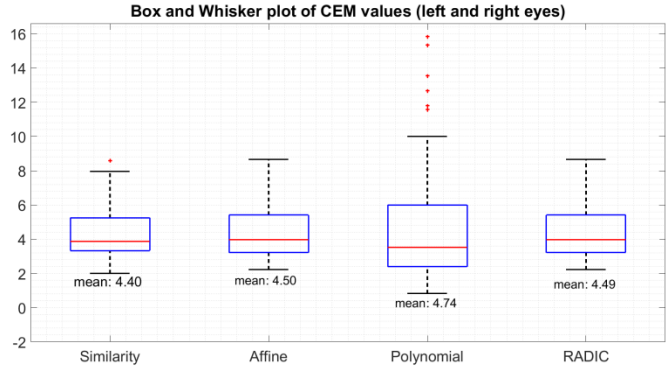


Figure 10. A Box and Whisker plot of the Centerline Error Measure (CEM) values

3.2.2. Spatial information combination

The method for fusing the scores has been discussed in Section 2.4. In order to evaluate the effectiveness of this method we need a baseline method to compare against. The baselines in our case would be the original proposed method that detects the candidates from a single image [1]. The goal was to provide a direct comparison between the performance when spatial information is accounted for and when it is not accounted for. The method used to compare [1] with the proposed method of this paper was as follows:

1- A tree ensemble model was generated using the training set.
2- Features were generated from the validation set and the ensemble model (decision tree ensemble) was used to assign scores to each candidate in the 80 images of the validation set.
3- In the case of the original image that used single images, an FROC (Free-Receiver operating) curve [38] was generated using these scores alone. This is the solid curve in Figure 11.
4- To incorporate spatial information, each image pair (optic disc centered and fovea centered) had their corresponding candidate scores fused as explained in the methodology section. The fused scores were then evaluated *collectively* for each image pair and this was used to generate the FROC curve in Figure 11.

The parameters for scores fusion were $\beta_1 = 0.4$ and $\beta_2 = 0.95$. Automating the process of selecting these parameters is left for future work. We see a slight increase after spatial information is incorporated. This increase can be captured quantitatively using the ROC score measure [39]. This score measures the sensitivity values at various x-axis intervals. The measured ROC score shows an increase of 0.02 after adding spatial information, which is a 2% increase. This increase can be explained intuitively as follows: Some candidates appear very subtle in the optic disc centered image, especially at the edge towards around the fovea region. This is because the retina is spherical and would get distorted during image acquisition. This distortion would affect the candidates towards the edge of the image. But these same candidates would appear more clearly in the Fovea centered image. This is because the same candidates now lie in the centre of the captured image, and hence their appearance will be obvious in the image. Intuitively, we expect the classifier to give higher scores to the more obvious candidates in terms of appearance. However, when the scores are fused, we take the maximum score of both candidates, and this will give us a higher score for the more subtle candidate that was originally given a lower score. This is why the FROC curve for the fused candidates shows a better performance.

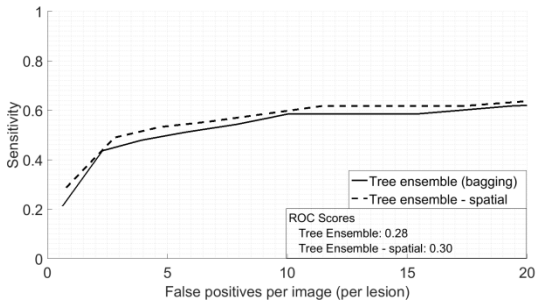


Figure 11 A comparison of performance between the technique applied to single images (solid) and after incorporating spatial information (dashed)

Figure 12 shows this intuitive concept by showing a plot of the corresponding pairs of scores. The figure shows a correspondence of scores for the test set where "score 1" and "score 2" on the axes refer to the methodology scores in Figure 2. Furthermore true candidates are labelled in blue while false detections are labelled in orange. Let us assume Score 2 is the reference image and that the candidate scores are being matched to score 1. Since in our case $\beta_1 = 0.4$ and $\beta_2 = 0.95$ we are only interested in this cross section from the score 2 axis. If we look to the extreme right of the graph we will see some candidates that have received scores within this range in the score 2. These receive higher scores along the x-axis. Hence, the maximum of both scores in the fused set will improve their scores and this will result in a higher FROC curve.

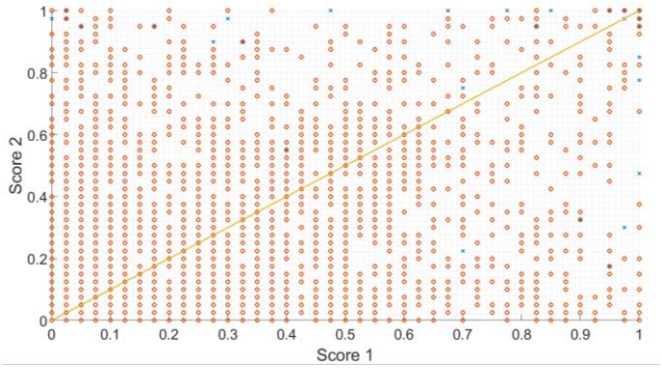


Figure 12 The scores from each corresponding spatial pair of images (these are the same scores as illustrated in Figure 2). True candidates are labelled in blue while false detections are labelled in orange. The orange line represents the linear line for the equation x=y.

The additional processing involved is the time for Image Registration, Candidates Matching and Score Fusion. These are the blue ellipse stages in Figure 2. The image registration is dependent on the algorithm of choice. In this work since it is based on manual control points the overhead is the control point selection and the time to solve a series of equations. An automated registration method may require an optimisation step and hence more time will be required for this step [40]. Given a spatial image pair, let m be the number of candidates detected in one image, and n be the number of candidates detected in the other image. The candidates matching step requires calculating the distance between each pairs of candidates detected in both images and hence has a complexity of $O(m \times n)$ (using Big O notation). The Score Fusion phase will fuse the scores together from one image (n fusions) and then fuse the scores from the other image (m fusions) and therefore will have a complexity of $O(m + n)$. In practice, on a core i5-4590 @ 3.30GHz CPU with 8GB RAM and an SSD hard drive, the time per image for each of the three stages was as follows:

1. Image Registration: 0.02 s
2. Candidates Matching: 1.30s
3. Scores fusion: 0.001s.

The sum of the above timings is 1.321s. The total time for test dataset was computed and then the average value per image pair was found. The average time per pair for the entire process was 4.141s. This means that the average overhead is about 32% of the time required per image (1.321s out of 4.141s). This does not take into account the time for manual control point selection, however, automated registration will be implemented in future work. Methods for optimising the speeds of the other stages are also left for future work.

4. Conclusions and Future Work

In this work a novel algorithm that combines spatial information from two views of the same retina for the purpose of microaneurysm (MA) detection is proposed. Most of the published work in the field of MA detection has been on detection from a single fundus image. The problem has been redefined to assume that two are available and the objective is to use the information from both to detect the microaneurysms in both images. The clinical application of this work would be its incorporation into diabetic eye screening programmes, thereby assisting the National Health Service in the detection of diabetic retinopathy. Screening in England has a recommendation of at least 2 images per eye so the assumption that two views are available would be valid in this context. ETDRS grading [41] is based on a categorical variable. Accurate counts of abnormalities could lead to a more refined grading/risk prediction using a continuous variable, or more accurate estimates of ma turnover that may better predict progression of disease. Accurate alignment of images would be an essential prerequisite to allow such scoring systems to develop. We propose a method for aligning the images, detecting candidates, matching candidates from both views and then combining the information from both views to perform microaneurysm detection from both views simultaneously. The proposed method was evaluated on a Diabetic Eye Screening Programme dataset of 160 images that contained 207 microaneurysms. The combination of information from multiple images was shown to increase the performance in comparison to depending on single images only. An account of the computational overhead required for the combination of information is also presented. Some of the limitations of the present work are: 1) The computational overhead required for the additional steps of the proposed method, 2) The accuracy of the registration method can be improved, 3) The control point selection of the image registration step can be automated to reduce manual labour work and 4) The scores fusion stage can be further improved to reduce more false positives and increase true positive detections. Future work will involve an exploration of other ways to combine information from spatial images, including multi-image features, or mosaicing images together to enhance the contrast of subtle MAs. Furthermore, this concept will be extended to include temporal images of the same patient as well.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This work is funded by an internal Scholarship, awarded by the Faculty of Science, Engineering and Computing, Kingston University.

References

[1] M. Habib, R. A. Welikala, A. Hoppe, C. G. Owen, A. R. Rudnicka, and S. A. Barman, "Microaneurysm Detection in Retinal Images Using an Ensemble Classifier," presented at the The Sixth International Conference on Image Processing Theory, Tools and Applications, Oulu, Finland, 2016.

[2] G. Liew, M. Michaelides, and C. Bunce, "A comparison of the causes of blindness certifications in England and Wales in working age adults (16–64 years), 1999–2000 with 2009–2010," *BMJ Open*, vol. 4, no. 2, p. e004015, Feb. 2014.

[3] "Diabetic Eye Screening Feature Based Grading Forms." [Online]. Available: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/402295/Feature_Based_Grading_Forms_V1_4_1Nov12_SSG.pdf. [Accessed: 02-Apr-2017].

[4] L. Streeter and M. J. Cree, "Microaneurysm detection in colour fundus images," *Image Vis. Comput N. Z.*, pp. 280–284, 2003.

[5] M. Niemeijer, B. van Ginneken, J. Staal, M. S. A. Suttorp-Schulten, and M. D. Abramoff, "Automatic detection of red lesions in digital color fundus

photographs," *IEEE Trans. Med. Imaging*, vol. 24, no. 5, pp. 584–592, May 2005.
[6] A. D. Fleming, S. Philip, K. A. Goatman, J. A. Olson, and P. F. Sharp, "Automated microaneurysm detection using local contrast normalization and local vessel detection," *IEEE Trans. Med. Imaging*, vol. 25, no. 9, pp. 1223–1232, Sep. 2006.
[7] K. M. Adal, D. Sidibé, S. Ali, E. Chaum, T. P. Karnowski, and F. Mériaudeau, "Automated detection of microaneurysms using scale-adapted blob analysis and semi-supervised learning," *Comput. Methods Programs Biomed.*, vol. 114, no. 1, pp. 1–10, Apr. 2014.
[8] J. Wu, J. Xin, L. Hong, J. You, and N. Zheng, "New hierarchical approach for microaneurysms detection with matched filter and machine learning," in *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*, 2015, pp. 4322–4325.
[9] A. Mizutani, C. Muramatsu, Y. Hatanaka, S. Suemori, T. Hara, and H. Fujita, "Automated microaneurysm detection method based on double ring filter in retinal fundus images," 2009, p. 72601N.
[10] R. Rosas-Romero, J. Martínez-Carballido, J. Hernández-Capistrán, and L. J. Uribe-Valencia, "A method to assist in the diagnosis of early diabetic retinopathy: Image processing applied to detection of microaneurysms in fundus images," *Comput. Med. Imaging Graph.*, vol. 44, pp. 41–53, Sep. 2015.
[11] A. Sopharak, B. Uyyanonvara, and S. Barman, "Simple hybrid method for fine microaneurysm detection from non-dilated diabetic retinopathy retinal images," *Comput. Med. Imaging Graph.*, vol. 37, no. 5, pp. 394–402, 2013.
[12] M. García, M. I. López, D. Álvarez, and R. Hornero, "Assessment of four neural network based classifiers to automatically detect red lesions in retinal images," *Med. Eng. Phys.*, vol. 32, no. 10, pp. 1085–1093, Dec. 2010.
[13] I. Lazar and A. Hajdu, "Microaneurysm detection in retinal images using a rotating cross-section based model," in *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, 2011, pp. 1405–1409.
[14] I. Lazar and A. Hajdu, "Retinal Microaneurysm Detection Through Local Rotating Cross-Section Profile Analysis," *IEEE Trans. Med. Imaging*, vol. 32, no. 2, pp. 400–407, Feb. 2013.
[15] L. Giancardo, F. Mériaudeau, T. P. Karnowski, K. W. Tobin, Y. Li, and E. Chaum, "Microaneurysms detection with the radon cliff operator in retinal fundus images," in *SPIE Medical Imaging*, 2010, p. 76230U–76230U.
[16] L. Giancardo, F. Meriaudeau, T. P. Karnowski, Y. Li, K. W. Tobin Jr, and E. Chaum, "Microaneurysm detection with radon transform-based classification on retina images," in *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*, 2011, pp. 5939–5942.
[17] B. Zhang, X. Wu, J. You, Q. Li, and F. Karray, "Detection of microaneurysms using multi-scale correlation coefficients," *Pattern Recognit.*, vol. 43, no. 6, pp. 2237–2248, Jun. 2010.
[18] Q. Li, R. Lu, S. Miao, and J. You, "Detection of microaneurysms in color retinal images using multi-orientation sum of matched filter," in *Proc. of the 3rd International Conference on Multimedia Technology*, 2013.
[19] J. H. Hipwell, F. Strachan, J. A. Olson, K. C. McHardy, P. F. Sharp, and J. V. Forrester, "Automated detection of microaneurysms in digital red-free photographs: a diabetic retinopathy screening tool," *Diabet. Med.*, vol. 17, no. 8, pp. 588–594, Aug. 2000.
[20] C. Sinthanayothin *et al.*, "Automated detection of diabetic retinopathy on digital fundus images," *Diabet. Med.*, vol. 19, no. 2, pp. 105–112, Feb. 2002.
[21] C. I. Sánchez, R. Hornero, A. Mayo, and M. García, "Mixture model-based clustering and logistic regression for automatic detection of microaneurysms in retinal images," in *SPIE medical imaging*, 2009, vol. 7260, p. 72601M–72601M–8.
[22] A. Sopharak, B. Uyyanonvara, and S. Barman, "Automatic microaneurysm detection from non-dilated diabetic retinopathy retinal images using mathematical morphology methods," *IAENG Int. J. Comput. Sci.*, vol. 38, no. 3, pp. 295–301, 2011.
[23] T. Inoue, Y. Hatanaka, S. Okumura, C. Muramatsu, and H. Fujita, "Automated microaneurysm detection method based on eigenvalue analysis using hessian matrix in retinal fundus images," in *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE*, 2013, pp. 5873–5876.
[24] R. Srivastava, D. W. Wong, L. Duan, J. Liu, and T. Y. Wong, "Red lesion detection in retinal fundus images using Frangi-based filters," in *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*, 2015, pp. 5663–5666.
[25] C. Heneghan, J. Flynn, M. O'Keefe, and M. Cahill, "Characterization of changes in blood vessel width and tortuosity in retinopathy of prematurity using image analysis," *Med. Image Anal.*, vol. 6, no. 4, pp. 407–429, 2002.
[26] A. Avakian *et al.*, "Fractal analysis of region-based vascular change in the normal and non-proliferative diabetic retina," *Curr. Eye Res.*, vol. 24, no. 4, pp. 274–280, 2002.
[27] A. J. Patterson, D. F. Garway-Heath, N. G. Strouthidis, and D. P. Crabb, "A new statistical approach for quantifying change in series of retinal and optic nerve head topography images," *Invest. Ophthalmol. Vis. Sci.*, vol. 46, no. 5, pp. 1659–1667, 2005.
[28] P. H. Artes and B. C. Chauhan, "Longitudinal changes in the visual field and optic disc in glaucoma," *Prog. Retin. Eye Res.*, vol. 24, no. 3, pp. 333–354, 2005.
[29] S.-E. Bursell, A. C. Clermont, B. T. Kinsley, D. C. Simonson, L. M. Aiello, and H. A. Wolpert, "Retinal blood flow changes in patients with insulin-dependent diabetes mellitus and no diabetic retinopathy.," *Invest. Ophthalmol. Vis. Sci.*, vol. 37, no. 5, pp. 886–897, 1996.
[30] H. Narasimha-Iyer *et al.*, "Robust Detection and Classification of Longitudinal Changes in Color Retinal Fundus Images for Monitoring Diabetic Retinopathy," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 6, pp. 1084–1098, Jun. 2006.
[31] S. S. Patankar and J. V. Kulkarni, "Directional gradient based registration of retinal images," in *Computational Intelligence & Computing Research (ICCIC), 2012 IEEE International Conference on*, 2012, pp. 1–5.
[32] G. K. Matsopoulos, N. A. Mouravliansky, K. K. Delibasis, and K. S. Nikita, "Automatic retinal image registration scheme using global optimization techniques," *Inf. Technol. Biomed. IEEE Trans. On*, vol. 3, no. 1, pp. 47–60, 1999.
[33] E. P. Ong *et al.*, "A robust multimodal retina image registration approach," 2014.
[34] A. Can, C. V. Stewart, B. Roysam, and H. L. Tanenbaum, "A feature-based, robust, hierarchical algorithm for registering pairs of images of the curved human retina," *Pattern Anal. Mach. Intell. IEEE Trans. On*, vol. 24, no. 3, pp. 347–364, 2002.
[35] P. C. Cattin, H. Bay, L. Van Gool, and G. Székely, "Retina mosaicing using local features," *Med. Image Comput. Comput.-Assist. Interv. 2006*, pp. 185–192, 2006.
[36] S. Lee, M. D. Abràmoff, and J. M. Reinhardt, "Feature-based pairwise retinal image registration by radial distortion correction," *Med. Imaging*, pp. 651220–651220, 2007.
[37] S. Lee, M. D. Abràmoff, and J. M. Reinhardt, "Retinal image mosaicing using the radial distortion correction model," *Med. Imaging*, pp. 691435–691435, 2008.
[38] T. Spencer, J. A. Olson, K. C. McHardy, P. F. Sharp, and J. V. Forrester, "An Image-Processing Strategy for the Segmentation and Quantification of Microaneurysms in Fluorescein Angiograms of the Ocular Fundus," *Comput. Biomed. Res.*, vol. 29, no. 4, pp. 284–302, Aug. 1996.
[39] M. Niemeijer *et al.*, "Retinopathy Online Challenge: Automatic Detection of Microaneurysms in Digital Color Fundus Photographs," *IEEE Trans. Med. Imaging*, vol. 29, no. 1, pp. 185–195, Jan. 2010.
[40] C. V. Stewart, C.-L. Tsai, and B. Roysam, "The dual-bootstrap iterative closest point algorithm with application to retinal image registration," *Med. Imaging IEEE Trans. On*, vol. 22, no. 11, pp. 1379–1394, 2003.
[41] B. Antal and A. Hajdu, "An ensemble-based system for microaneurysm detection and diabetic retinopathy grading," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 6, pp. 1720–1726, 2012.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 650-657 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Human Robot Interaction for Hybrid Collision Avoidance System for Indoor Mobile Robots

Mazen Ghandour*[1], Hui Liu[1], Norbert Stoll[2], Kerstin Thurow[1]

[1]*Center for Life Science Automation, University of Rostock, D-18119, Rostock, Germany*

[2]*Institute for Automation, University of Rostock, D-18119, Rostock,, Germany*

ARTICLE INFO

Article history:
Received: 07 April, 2017
Accepted: 12 May, 2017
Online: 04 June, 2017

Keywords :
Collision Avoidance
Human-Robot Interaction
Local Path Planning
Neural Networks
Support Vector Machine
Kinect

ABSTRACT

In this paper, a novel approach for collision avoidance for indoor mobile robots based on human-robot interaction is realized. The main contribution of this work is a new technique for collision avoidance by engaging the human and the robot in generating new collision-free paths. In mobile robotics, collision avoidance is critical for the success of the robots in implementing their tasks, especially when the robots navigate in crowded and dynamic environments, which include humans. Traditional collision avoidance methods deal with the human as a dynamic obstacle, without taking into consideration that the human will also try to avoid the robot, and this causes the people and the robot to get confused, especially in crowded social places such as restaurants, hospitals, and laboratories. To avoid such scenarios, a reactive-supervised collision avoidance system for mobile robots based on human-robot interaction is implemented. In this method, both the robot and the human will collaborate in generating the collision avoidance via interaction. The person will notify the robot about the avoidance direction via interaction, and the robot will search for the optimal collision-free path on the selected direction. In case that no people interacted with the robot, it will select the navigation path autonomously and select the path that is closest to the goal location. The humans will interact with the robot using gesture recognition and Kinect sensor. To build the gesture recognition system, two models were used to classify these gestures, the first model is Back-Propagation Neural Network (BPNN), and the second model is Support Vector Machine (SVM). Furthermore, a novel collision avoidance system for avoiding the obstacles is implemented and integrated with the HRI system. The system is tested on H20 robot from DrRobot Company (Canada) and a set of experiments were implemented to report the performance of the system in interacting with the human and avoiding collisions.

1. Introduction

In future work environments, robots will work alongside to human, and this raises big challenges related to the robustness of these robots in detecting people, interacting with them and avoiding physical accidents. Different collision avoidance techniques have been implemented for mobile robotics. In Bug algorithms [1], when the robot meets an obstacle in the navigation path, it will implement either a complete rotation around the obstacle (Bug1), or it will rotate around the obstacle till it sees again the goal location (Bug 2). The robot will then continue its movement to the goal. Other methods takes into consideration the robot's dynamics, such as dynamic windows approach (DWA) [2]. In DWA, after specifying the distances of each obstacle from the robot, it will consider only the velocities that the robot can use and decelerate before colliding with the obstacles. The robot will then select the velocity vector that is closest to the goal location. Other methods uses reactive techniques, in which the robot relies merely on the

*Corresponding Author: Mazen Ghandour, Center for Life Science Automation, Rostock University, Germany.
Address: Friedrich-Barnewitz-Straße 8, D-18119 Rostock
Tel: +4917657805747,
Email: mazen.ghandourr@gmail.com

sensor's information and the motion update is related to the new measurements that are obtained from the sensors. An example of these methods is potential field [3], vector field histogram [4], and nearness diagram [5]. Nearness Diagram simplifies the collision avoidance task by using a divide and conquer strategy; the robot splits the navigation area into sectors, and combines the sectors that don't include obstacles together to define the navigable regions. Based on this information, the robot will consider only the regions that are wide enough for the robot to navigate, and it will select one of five scenarios for selecting the region and avoiding the obstacles; the robot will then calculate the angular and linear velocities for passing the selected region.

Recently, many human-robot interaction systems have been realized in robotics. Wearable sensors are fixed on certain parts of the people to provide the robot with the motion of their bodies, and it will then execute physical reactions as a result of the interaction. Cifuentes et al. implemented a human-tracking system that allows the robot from following the human [6]. In this system, the robot will use LRF sensor to track the leg's motion of the human. Furthermore, an inertial measuring unit IMU is fixed on the trunk of the user to extract its motion and rotation. The robot moves in front of the person and adjusts its direction and velocity based on the sensory data from LRF and IMU.

3D vision is another common method to obtain the interaction with the humans. As an example of these sensors is the Kinect. Boubou implemented an interaction system to define nine actions for the humans (wave, sit, walk, stand, pick up, stretch, forward punishing, use hammer, and draw a circle) [7]. To define the gestures, a histogram of oriented velocity vectors HOVV algorithm is implemented. The HOVV will analyze the velocity and orientation of each joint of the body and build a velocity vector for each of these joints. The vectors will be then represented in a spatial histogram. The implemented histogram will be used to describe the activity of the person. Other human-robot interaction systems could be done by voice recognition [8], brain-signals [9], and face recognition [10].

In the literature, it is found that several collision avoidance systems show good performance in avoiding obstacles. Still, in these systems, when the robot meets humans in the path, it will consider them as dynamic obstacles. When the robot moves in social environments such as laboratories, hospitals and restaurants, these systems won't work efficiently since the people and the robot will get confused regarding the motion direction that each of them has to follow. Furthermore, these systems can't solve the bottleneck problem in which the robot and humans meet in narrow areas such as corridors and small rooms, so none of them is able to generate the collision-avoidance path due to the limited navigation area. Out of this, a collision avoidance system has been implemented. The system provides a mutual responsibility for both the robot and the people in avoiding each other. When the robot meets humans, it will send a voice message asking the people to interact. If a person interacted with it, the robot will move based on the gestures that were provided by the user. If no people interacted with the robot within a certain time, it will generate the collision avoidance path autonomously, taking into consideration generating the collision-free path that is closest to its original path.

Since this system is based on the interaction between the human and the robot, it will be called Cooperative Collision Avoidance system based on Interaction (CCAI).

This paper will be organized as follow: chapter 2 will show the human-robot interaction system. In chapter 3, the collision avoidance system will be explained, while chapter 4 will show the experimental results for the system.

This paper is an extension for the work originally presented in Mechatronika conference [11]. In this paper, deep experiments have been implemented to check the performance of the human-robot interaction and the collision avoidance systems and the results have been reported.

2. Human-Robot Interaction System

Human-Robot interaction (HRI) is defined as the ability of the robot from recognizing the humans, understanding their activities and implementing certain responses as a result of this interaction [12]. In the proposed system, the HRI will serve the task of collision avoidance. Thus the HRI will be based on gesture recognition. Kinect 2.0 sensor (Microsoft, USA) will be used for detecting the humans and extracting their joints' coordination. This sensor has a RGB camera and an infrared camera. The combination of these two cameras provides the robot with a 3D vision of the work environment. The sensor uses time-of-flight to measure the distance between the sensor and the objects. The Kinect 2.0 provides a skeleton frame, which includes the skeletal data for up to 6 people to the robot. 3D dimensions for 25 joints express each skeleton. Fig 1. shows the skeletal representation for the used gestures in this work. To implement the human-robot interaction, seven gestures are used in this system:

- Stop: the robot will stop as long as the stop gesture is raised.
- Move right: The robot will generate the collision-free path to the right side of the person.
- Move left: The robot will generate the collision-free path to the left side of the person.
- Move forwards: The robot will move toward the person as long as the "move forward" gesture is raised.
- Move backwards: The robot will move backwards as long as the master person raises the "move backwards" gesture.
- Master select: when several people are in the path of the robot, the robot will not know to which person it has to interact, since

two or more people might give different orders to the robot. To overcome this problem, "Master Select" gesture is assigned. When a person in a group raises the right arm 180°, the robot will know that it has to interact with this person and will ignore the gestures of other people in the work area.

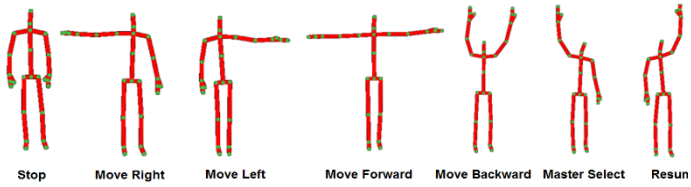


Figure 1: The gestures that are used for interaction with the robot

The Kinect sensor provides only the dimensions of the human's joints, without describing the gestures. Thus, to build the gesture recognition system, the "y" values for the right & left elbows, the right & left wrists and the "y, z" values of the neck joints were extracted. Furthermore, machine learning (SVM) and artificial intelligence (BPNN) models were trained to define the gestures. To train the models, a training set of 448 samples is constructed, and the k-fold cross validation algorithm is used to train the models. Additionally, it is worthy to mention that the HRI system took into consideration that in real environments, the robot and the human might not be located face-to-face. Thus, the training set is extracted with different deviation angles between [-40°, 40°].

3. Cooperative Collision Avoidance

In previous collision avoidance systems, the robot will search for the closest collision-free path to the goal location, without taking into consideration the motion direction of the human. Since the human will also try to avoid the robot, both participants will get disordered about the correct direction that each of them has to go. This situation can cause problems when the robot navigates in social areas which include several people leading to a collision when both the human and the robot decide to use the same collision-avoidance direction. To overcome this situation, a new method for the collision avoidance is proposed which is called *Cooperative Collision Avoidance System based on Interaction (CCAI)* since both the robot and the human will cooperate in generating the collision-free path via interaction. Fig. 2 shows the flow chart of the proposed system.

In the proposed collision avoidance system after identifying people in the path the robot will send a voice message "Robot in the path", alarming the people that the robot is moving in the same area. The people can move away from the robot's path, or be prepared for interaction. If the human remain in the path, the robot will stop keeping a certain distance from the people (2m), asking them for interaction by sending a voice message "Interact". Then the following situations can occur.

- If a user raised the right arm activating the "master gesture", the robot will execute the orders provided by the user.
- If no human interacted with the robot, it will plan a collision-free path autonomously, taking into consideration the goal location.
- If the human moved away from the path of the robot, it will complete its path to the goal without changing its direction.
- If there is no available path that the robot can go through, it will send an alarm message "no free path" to inform the people that they have to keep a free path for the robot so it can generate a new collision-free path.

Fig. 3 shows three human in front of the robot; non of them interacted with it. The robot will then implement the collision avoidance path by searching for the regions that the robot can pass through. Then, it will select the region that is closest to its original path obtained from the global navigation system [13]. Since the robot has two available regions (R1, R2), it will select the region R2 since it is the closest to the original orientation to the goal.

In contrary, when the user interacts with the robot, it will follow the orders provided by the master via gestures. The red colored person in fig. 4 activated the "master person", and asked the robot to move to the left. The robot will then select the region R1 since there is no free space between the persons P1, P2, and the region R2 is not located to the right of the master person.

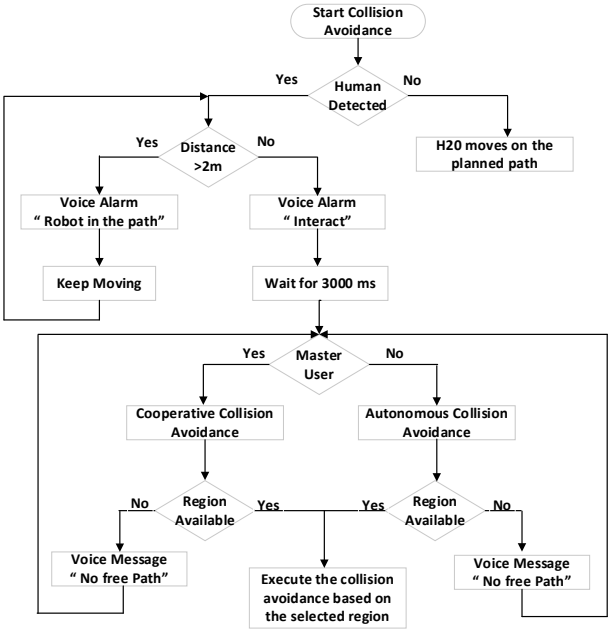


Figure 2: The flow chart of the collision avoidance system

To calculate the collision-free path, the robot has to search for all available regions that the robot can pass safely. Here the robot distinguishes between the terminal regions that are located to the

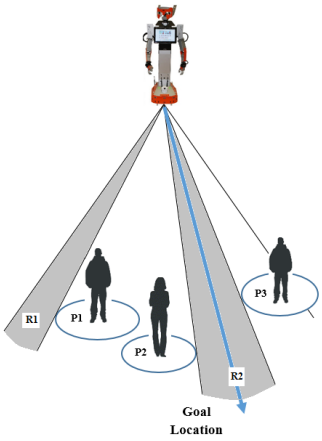


Figure 3: Selecting the CA region based on the goal location

right/left of the last person in the group, and the regions that exist between the people.

In the first step, the robot detects the positions of the people, and calculates the relative distances between them. For the middle regions between the people, the robot calculates the width of the available distance between them:

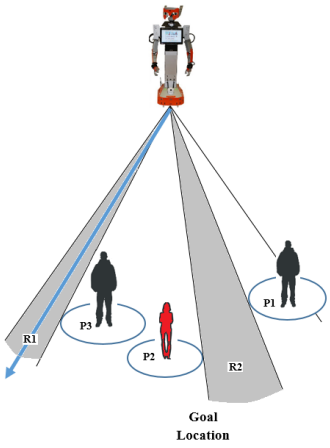


Figure 4. Selecting the CA path based on HRI

$$R_{av}^{i} = X_{p}^{i+1} - X_{p}^{i} \; where: i = 1, .., n-1 \qquad (1)$$

where n represents the number of people detected by the Kinect sensor.

Furthermore, the robot estimates the minimum region width required for the robot to pass without collision:

$$R_{\min\,(mid)}^{i} = 2 \times r_r + r_p^{i+1} + r_p^{i} + d \qquad (2)$$

Where:

r_r robot's radius.

r_p^i the distance between the shoulder and the neck of the user i obtained from Kinect sensor.

d the safety distance around the robot.

In the next step, the robot checks the possibility of generating the avoidance path to the right and left of the most right/left people in the group. This calculation is done by considering the maximum detection angle of the sensor (70°), so the robot will compare the width of the region between the terminal person and the last point that the sensor can detect for the given depth, as it could be found in fig. 5, which shows the calculation of the region for the right-terminal person.

The following equations are used to get the width of the right-terminal region:

$$\delta_n = \sin^{-1} \frac{X_p^n}{Z_p^n} \qquad (3)$$

$$Z_{temp}^{n} = Z_p^n . \cos \delta_n \qquad (4)$$

$$X_{term}^{n} = Z_{temp}^{n} . \tan 35 \qquad (5)$$

where

Z_p^n: The distance of the person n from the robot

X_{term}^n: The last point that the Kinect sensor can detect for the given depth Z_p^n

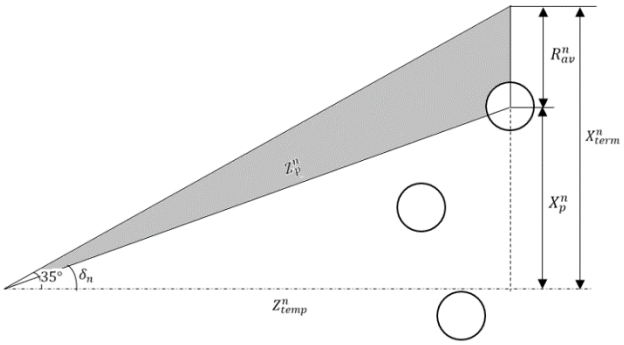


Figure 5. Calculation of available terminal region

The available region is given as:

$$R_{av}^{n} = \left| X_{term}^{n} - X_{n}^{p} \right| \qquad (6)$$

The same calculations are done to get the width of the left terminal region. The robot will then calculate the minimum width of the terminal regions that the robot can pass through:

$$R_{\min}^{n} = r_r + r_p^n + d/2 \qquad (7)$$

$$R_{\min}^{0} = r_r + r_p^0 + d/2 \qquad (8)$$

The candidate valleys are the regions that are wider than the minimum required width:

$$V = \sum_{i=1}^{n-1} R_{av}^{i} + R_{av}^{0} + R_{av}^{n} \quad if \; R_{av}^{i} > R_{min}^{i}, R_{av}^{0} > R_{min}^{0}, R_{av}^{n} > R_{min}^{n} \qquad (9)$$

If $V = \emptyset$, the robot couldn't find a free walking area between the group of people, and it will send the voice message "no free path" to alarm the people to move and keep enough area for the robot to pass as it could be found in fig. 2.

If $V \neq \emptyset$, the robot will select the valley depending on whether the master user asked the robot to go to a certain direction, or in case that no people interacted it will select the valley that is closest to its path. To calculate the collision-free path for the selected valley, the robot distinguishes between the terminal valley and the intermediate valley. Fig. 6 shows the calculation of the collision avoidance path for an intermediate valley using the following equations:

$$x_{ca} = \frac{x_p^{i+1} - x_p^i}{2} \quad (10)$$

$$Z_f = Max\{Z_i, Z_{i+1}\} \quad (11)$$

$$\psi = \sin^{-1}\frac{x_f}{z_f} \quad (12)$$

$$Z_{temp} = Z_f.\cos\psi\ \theta_{ca} = \tan^{-1}\frac{x_{ca}}{Z_{temp}} \quad (13)$$

$$\theta_{ca} = \tan^{-1}\frac{x_{ca}}{Z_{temp}} \quad (14)$$

$$Z_{ca} = \frac{X_{ca}}{\sin\theta_{ca}} \quad (15)$$

$$d_{ca} = Z_{ca} + r_r + r_f + d \quad (16)$$

where:

x_{ca}: The middle of the selected valley.

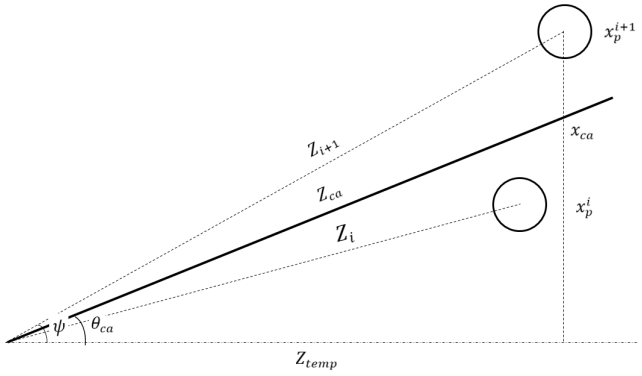


Figure 6. Calculating the collision-free path

Z_f: The distance of the farthermost person from the robot in the selected region.

θ_{ca}: The robot's orientation toward the valley.

ψ: The angle between the robot's horizontal axis and the farthermost person in the selected region.

Z_{temp}: The horizontal distance between the robot and farthermost person.

Z_{ca}: The distance between the robot and middle of the Region.

d_{ca}: The travelled distance to pass the people.

For the terminal valleys, the robot uses the following equations for the most right person:

$$x_{ca} = X_p^n + r_p^n + r_r + d/2 \quad (17)$$

$$Z_{ca} = Z_p^n \quad (18)$$

$$\theta_{ca} = \sin^{-1}\frac{x_{ca}}{Z_{ca}} \quad (19)$$

$$d_{ca} = \frac{x_{ca}}{\tan\theta_{ca}} + r_p^n + r_r \quad (20)$$

And for the most left person:

$$x_{ca} = X_p^0 - r_p^0 - r_r - d/2 \quad (21)$$

$$Z_{ca} = Z_p^0 \quad (22)$$

$$\theta_{ca} = \sin^{-1}\frac{x_{ca}}{Z_{ca}} \quad (23)$$

$$d_{ca} = \frac{x_{ca}}{\tan\theta_{ca}} - r_p^n - r_r \quad (24)$$

Calculating the angular and linear velocities is based on the width of the selected valley:

$$v = v_{max}.\frac{\log_{10}\theta_r}{\log_{10}\theta_{max}} \quad (25)$$

$$\omega = \omega_{max}.\frac{\log_{10}\theta_r}{\log_{10}\theta_{max}} \quad (26)$$

Where θ_{max} the maximum vision angle of the Kinect sensor and equal 70°, θ_r the width of the selected region in degrees.

4. Experimental Results

The experiments have been implemented to check the performance of the HRI and the collision avoidance systems.

4.1 The experiments for the human-robot interaction

The experiments over the HRI were implemented on two steps. The first step, the experiments were applied to compare the performance of the SVM and BPNN and find out the best model that could fit the HRI system. After finding out the optimal model that can classify the training data, the HRI is tested for different gestures and different people to find the success rate of the implemented HRI system.

The goal of using the L_SVM and BPNN is to find the optimal model that could be used to recognize the gestures of different people with different heights and deviations from the Kinect level. To do so, a training set is implemented to train the models. Each training vector is composed of the "y" values of the right and left elbows, the "y" values of the right and left wrists, and the "y, z" values for the neck joint. The training data is collected from four people with different heights [152-187] cm. Each person is asked to do 112 gestures with distances [1.7, 4] m from the Kinect, with deviation angle [-40, 40]. Thus the total training set is 448. The L_SVM is trained for different penalty values, while the BPNN is trained for different number of hidden neurons.

To implement the training, the K-fold cross validation model is used with k=8.

4.1.1 Training the L_SVM model

The L_SVM is trained to find out the best penalty value that provides minimum error over the test set. The model is trained for penalty values $c = e^n\ where\ n \in [-10,70]$. Table I shows the training results.

Table I. The training results of L_SVM model

C	Total Error	Average Error	Success Rate (%)	Train time (ms)	Test Time (ms)	No. Support Vectors
$4.5 \times 10^{-5}, 0.0067$	379	47.3	15.53	[103, 112]	<1	21
0.0183	335	44.3	20.9	102	<1	21
0.0497	156	19.5	65.17	88	<1	21
0.135	33	4.12	92.64	71	<1	21
0.367	8	1	98.2	68	<1	21
1, 2.718	9	1.125	97.99	49.56	<1	21
7.389	**8**	**1**	**98.2**	**45**	**<1**	**21**
$20.08, 9.2 \times 10^{29}$	16	2	96.4	[41,53]	<1	21

From table I it can be seen, that the best penalty value that shows minimum error is at c=7.389 with 8 misclassified training samples for the 8 folds.

4.1.2 Training the BPNN model

The BPNN model is trained to find the best number of hidden neurons. The learning rate is set to $\eta = 0.15$. The model has five input neurons, three output neurons, while it is trained and tested to find the best number of hidden neurons with the range [3, 15]. Table II shows the experimental results for training the BPNN.

Table II. The training results of BPNN model

No. Hidden neurons	No. Errors	Training Time (ms)	Test Time (ms)
3	0	94	<1
4	0	137	<1
5	0	92	<1
6	**0**	**126**	**<1**
7	0	135	<1
8	0	113	<1
9	0	117	<1
10	0	119	<1
11	0	122	<1
12	0	133	<1
13	0	122	<1
14	0	162	<1
15	0	161	<1

From table II, it could be concluded that despite of the number of hidden neurons, the model could correctly classify the whole test data successfully with no errors. Furthermore, in comparison with table I, it could be found that despite of the number of hidden neurons, the BPNN outperforms the implemented SVM model. Thus, the BPNN model is selected in the HRI system with a number of hidden neurons equal to 6.

4.1.3 Test the HRI system

After selecting the optimal model for the HRI system, further experiments have been implemented in real work conditions to check the performance of the HRI. To do so, five people with different heights and physical shapes were asked to do tests. Each person is asked to implement 18 gestures with different deviation angles [-40°, 40°]. Table III shows the experimental results for the HRI system.

Table III. The test results for the HRI

Person	No. Experiments	Height	False Gesture	State Transition
1	18	165	1	0
2	18	170	1	3
3	18	177	0	1
4	18	179	0	1
5	18	181	0	1

Out of 90 experiments, two gestures were misclassified. These misclassifications occur due to false measurements for the joints' locations by the Kinect sensor. Furthermore, state transition means the Kinect couldn't detect the person that is located in front of it, and it required from him to move or shake their bodies to be detected from the sensor.

4.2 The experiments for the collision avoidance system

As mentioned in chapter 3, the robot uses the equations (25), and (26) to adjust its velocity based on the width of the selected region. Fig. 7 shows simulation for the velocity controller for different maximum linear and angular velocities taking into consideration that the maximum angular view for the sensor is 70°. Furthermore, to check the validity of the implemented system, several experiments were implemented for each function. H20 humanoid robot from DrRobot Company is used in the experiments.

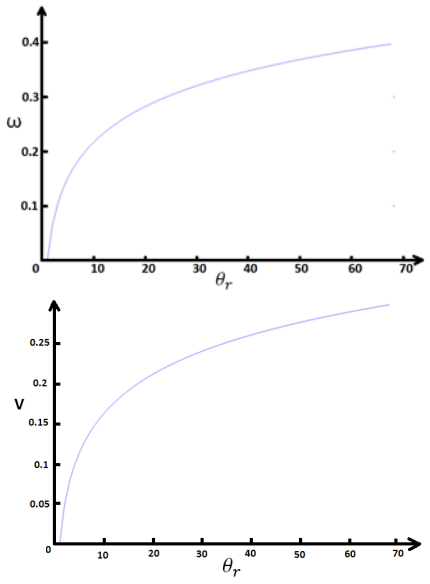


Figure 7: The simulation result for the velocity controller

4.2.1 Tests for the move forward/backward

The experiments for the move forward/backward gesture were implemented for different velocities $V_{max} = 0.1, 0.2, 0.3$. Two experiments were implemented for each velocity, one for each function. The experiments show that the robot could detect people, interact with the master and execute the orders based on the raised function. Fig 7. Shows an example of the cooperative collision avoidance in which the master asked the robot to move forward, and then he asked it to move to the right.

4.2.1 Tests for the move right/left

Fig 7. shows an example of the "move right" function after using the "move forward" gesture. To test the move right/left functions, 6 experiments were implemented. Table IV summarizes these experiments. V_{max}, ω_{max} represent the maximum linear and angular velocities that the robot uses in case that no obstacles in its path. Region candidates represent the possible steering angles that the robot can use to pass the detected regions. Heading angle represents the steering angle that the robot uses after selecting the region that the robot will go through it. (d) represents the distance that the robot will travel to pass the region. θ_r is the width of selected region. v, ω are the linear and angular velocities for the robot obtained from the velocity controller. Time represents the time that the robot needs to pass the people.

As it could be found, the robot will detect the whole navigable regions that are located at the right or left of the master person, depending on the direction that he/she ordered the robot to move. Then, it will calculate the steering angle and the distance that it has to move to pass the region. The robot will also consider the width of the selected region to calculate the linear and angular velocities.

4.2.2 Tests for the autonomous collision avoidance

As discussed before, in case of no human interaction the robot will generate the collision avoidance autonomously without any interaction. To test the function, additional six experiments were implemented to check the performance of the autonomous collision avoidance system. Table IV shows the experimental results for the system. Waypoint represents the original heading angle that the robot had to move if there are no people in the path.

The robot could successfully define the navigable regions, and select the region that is closest to its original path, by comparing the regions' angles with the angle of the waypoint. Furthermore, it could be seen that the robot could successfully define the steering angle and the distance that it has to move to pass the region. Furthermore, the velocity controller will adjust the robot's linear and angular velocities based on the width of the selected region.

4.3 Discussion and comparison with other systems

As it could be found in the experiments over the HRI system, the system is able to classify the whole gestures with a success rate 100%. Furthermore, the implemented system is able to detect and define the gestures even when the user is deviated with [-40°, 40°] from the straight line between the user and the robot. Other systems which are based on analyzing the geometrical displacements of the joints as in [14] and [15] will fail in such situations.

Although many collision avoidance systems proved a good performance in avoiding dynamic and static obstacles, all of these systems considered the humans as dynamic obstacles. This is a short come if we consider that in cluttered social environments, robots will navigate between many people who will share the location, and this could lead for both of them to get confused due to the lack of the interaction between them. Furthermore, the previous collision avoidance systems didn't solve the bottleneck problem which could occur in narrow corridors and rooms, so the robot can't generate collision avoidance paths due to the lack of space. Additionally, it is possible in certain situations that the human needs to interrupt the motion of the robot, and control it when needed.

The implemented system solved the bottleneck problem, by using the move forward/backward gestures. Thus, the master can lead the robot to another area which is wide enough to avoid each other. Furthermore, when there are several people in the path of the robot, a person can raise the "master select" gesture, and provide the robot the direction that it has to move, so the other group of people can go to another direction, and no motion conflict will occur. Besides to the interactive collision-avoidance features, the system is also able to generate the collision-avoidance path autonomously in similar concepts that other collision avoidance systems follow.

Future work will focus on detecting the static obstacles and integrate it to the implemented collision avoidance system, so the robot can avoid both the static, dynamic and human obstacles.

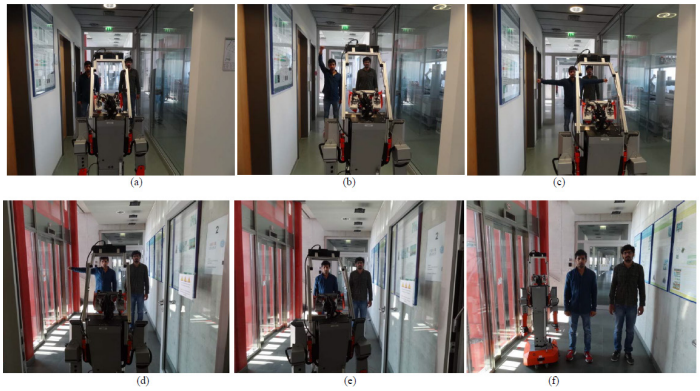


Figure 8: The cooperative collision avoidance, (a) the robot meets people in narrow path and ask them to interact, (b) the user raises his right arm to inform the robot that he will be the master, (c) the master person orders the robot to move forward, (d) the robot moves forward till it reach to a wide area and the user asks the robot to implement the collision avoidance to the right, (e, f) the robot adjusts its motion and moves to the free region to the right of the master

(a) (b) (c)

Figure 9: the autonomous collision avoidance, (a) the robot detects two people in the path and defines three navigable regions, (b) since no people interacted with the robot, it will select the region that is closest to its original path, (c) the robot calculates the path, pass the people and keep moving to the goal location.

Table IV: The experimental results for the move right/left functions

No.	Function	V_{max} m/s	ω_{max} rad/s	Region Candidate	Region Candidate	Heading Angle	d(m)	θ_r (°)	V (m/s)	ω (rad/s)	Time (s)
1	Right	0.2	0.4	25.4	-3.7	-3.7	3.03	34.9	0.167	0.335	18.3
2	Right	02.5	0.3	1.16	31.1	1.16	2.95	38.7	0.215	0.256	13.8
3	Right	0.3	0.5	24.1	----	24.1	3.14	24.6	0.226	0.379	14.9
4	Left	0.15	0.3	-13.9	----	-13.9	2.97	36.4	0.126	0.252	24.4
5	Left	0.2	0.4	-25.9	----	-25.9	3.08	22.0	0.145	0.29	22.7
6	Left	0.3	0.25	-34.3	-8.9	-8.9	2.95	36.4	0.253	0.212	12.3

Table V: The experimental results for the autonomous collision avoidance function

No.	V_{max} m/s	ω_{max} rad/s	Region Cand	Region Cand	Region Cand	Waypoint	Heading Angle	d(m)	θ_r (°)	V (m/s)	ω (rad/s)	Time (s)
1	0.15	0.2	-29.9	5.2	---	12.7	5.2	2.66	37.4	0.127	0.171	21.3
2	0.15	0.2	-16.77	13.32	---	8.82	13.32	2.93	36.8	0.127	0.169	24.4
3	0.2	0.5	-30.5	3.2	---	8.3	3.2	2.69	37.9	0.171	0.425	15.8
4	0.2	0.5	-25.3	9.2	33.6	84.4	33.6	3.54	12.3	0.118	0.295	32.0
5	0.3	0.25	-30.6	3.3	34.7	71.6	34.7	3.25	12.3	0.177	0.147	22.4
6	0.3	0.25	-12.6	18.3	---	-27.4	-12.6	2.89	38.0	0.256	0.215	12.3

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The study is funded by the Federal Ministry of Education and Research (FKZ: 03Z1KN11, 03A1KI1) and the German Academic Exchange Service (Ph.D. stipend M. Ghandour). The authors would also like to thank the Canadian DrRobot Company for the technical support for the H20 mobile robots in this study.

References

[1] V. Lumelsky and A. Stepanov, "Dynamic path planning for a mobile automaton with limited information on the environment," *IEEE Trans. Autom. Control*, vol. 31, no. 11, pp. 1058–1063, Nov. 1986.

[2] D. Fox, W. Burgard, and S. Thrun, "The dynamic window approach to collision avoidance," *IEEE Robot. Autom. Mag.*, vol. 4, no. 1, pp. 23–33, Mar. 1997.

[3] O. Khatib, "Real-time obstacle avoidance for manipulators and mobile robots," in *1985 IEEE International Conference on Robotics and Automation. Proceedings*, 1985, vol. 2, pp. 500–505.

[4] J. Borenstein and Y. Koren, "The vector field histogram-fast obstacle avoidance for mobile robots," *IEEE Trans. Robot. Autom.*, vol. 7, no. 3, pp. 278–288, Jun. 1991.

[5] J. Minguez and L. Montano, "Nearness diagram (ND) navigation: collision avoidance in troublesome scenarios," *IEEE Trans. Robot. Autom.*, vol. 20, no. 1, pp. 45–59, Feb. 2004.

[6] C. A. Cifuentes, A. Frizera, R. Carelli, and T. Bastos, "Human–robot interaction based on wearable IMU sensor and laser range finder," *Robot. Auton. Syst.*, vol. 62, no. 10, pp. 1425–1439, Oct. 2014.

[7] S. Boubou and E. Suzuki, "Classifying actions based on histogram of oriented velocity vectors," *J. Intell. Inf. Syst.*, vol. 44, no. 1, pp. 49–65, Jul. 2014.

[8] I.-J. Ding and J.-Y. Shi, "Kinect microphone array-based speech and speaker recognition for the exhibition control of humanoid robots," *Comput. Electr. Eng.*

[9] E. Hortal *et al.*, "SVM-based Brain–Machine Interface for controlling a robot arm through four mental tasks," *Neurocomputing*, vol. 151, Part 1, pp. 116–121, Mar. 2015.

[10] C. C. Tsai, Y. Z. Chen, and C. W. Liao, "Interactive emotion recognition using Support Vector Machine for human-robot interaction," in *IEEE International Conference on Systems, Man and Cybernetics, 2009. SMC 2009*, 2009, pp. 407–412.

[11] M. Ghandour, H. Liu, N. Stoll, and K. Thurow, "A hybrid collision avoidance system for indoor mobile robots based on human-robot interaction," in *2016 17th International Conference on Mechatronics - Mechatronika (ME)*, 2016, pp. 1–7.

[12] M. Ghandour, H. Liu, N. Stoll, and K. Thurow, "Interactive collision avoidance system for indoor mobile robots based on human-robot interaction," in *2016 9th International Conference on Human System Interactions (HSI)*, 2016, pp. 209–215.

[13] A. A. Abdulla, H. Liu, N. Stoll, and K. Thurow, "Multi-floor navigation method for mobile robot transportation based on StarGazer sensors in life science automation," in *Instrumentation and Measurement Technology Conference (I2MTC), 2015 IEEE International*, 2015, pp. 428–433.

[14] G. Canal, S. Escalera, and C. Angulo, "A Real-time Human-Robot Interaction system based on gestures for assistive scenarios," *Comput. Vis. Image Underst.*

[15] N. Yang *et al.*, "A study of the human-robot synchronous control system based on skeletal tracking technology," in *2013 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, 2013, pp. 2191–2196.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 658-668 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Multi Level Integrity Management in LTE/LTE-A Networks

Ouafae Kasmi*, Amine Baina, Mostafa Bellafkih

STRS Lab., National Institute of Posts and Telecommunications –INPT, Department of Telecommunications Systems, Network and Services, Rabat, Morocco

ARTICLE INFO

Article history:
Received: 05 April, 2017
Accepted: 24 May, 2017
Online: 05 June, 2017

Keywords:
Integrity
Criticality
QoS
Fuzzy Logic

ABSTRACT

The sector of telecommunications is experiencing a large growth and rapid change in developing applications due to the variety of technologies, to increase the number of users and to arrival services varied from the high demand of the service quality such as streaming, web navigation. In this regard, the 3rd Generation Partnership Project (3GPP) organization has greatly motivated to develop Next Generation Network (NGN). Indeed, telecom operators are forced to make a portfolio of available services more efficient to ensure customer loyalty. This article aims to propose an incorporation a layer above the quality of service (QoS) controller based on Fuzzy Logic to determine the guaranteed levels of QoS and levels of criticality. It describes a new approach of multi criticality in order to ensure the integrity in IP Multimedia Subsystem (IMS) networks. This approach is based on the recent model of integrity Totel. The proposed approach is tested and evaluated using Fuzzy Inference System to simulate levels of criticality for determining an adequate QoS level that ensures the customer's needs. The simulation results describe the comparison between three membership function Triangle, Trapezoidal, Gaussian also describes the comparison between the methods of defuzzification to find whose gives better results and has more efficient in the aspect of chance value for levels of QoS and levels of criticality.

1. Introduction

The sector of telecommunications is undergoing a profound and rapid change in developing applications due to convergence towards Telecom, Internet, Multimedia and continual technological innovations. The IMS architecture towards all-IP domain supporting at all services voice, video and data. Thus, according to IMS network convergence, the issue of maximizing the resource use and the continuity of QoS and Quality of Experience (QoE) for satisfying the user's requirements has gained importance to find success or failure of applications and services. This quality shows performance metrics in terms of perspectives for the service user. Indeed, the IMS architecture is composed of several entities types of session initiation protocol (SIP) servers to process SIP signaling packets of the IMS domain: Proxy Call Session Control Function (P-CSCF), Interrogating CSCF (I-CSCF), and Serving CSCF (S-CSCF).

Thus, IMS architecture also defined entities for QoS management, which are Policy Control Rule Function (PCRF) and Policy Charging Enforcement Function (PCEF). Although, due to the high-quality IMS services, it is not possible to guarantee a QoS for IMS services at anytime, anywhere due to the increase of user numbers and to their mobility.

This paper is an extension of work originally presented in conference on Intelligent Systems Theories and Applications (SITA) [1]. The goal of this article is to propose a policy of integrity implementation that allows a security policy integrated on IMS architecture. This integration is based on Totel model. Our contribution described a new approach for multi criticality on IMS architecture to ensure the integrity and to verify the exchange between different software entities. Furthermore, to dynamically find the level of criticality of customers for offering QoS appropriated to their needs by using Fuzzy Logic for decision-making according to multi criteria such as environment, equipment and the importance of customers. This approach is introduced by simulation using Matlab.

*Corresponding Author: Ouafae Kasmi, STRS, National Institute of Posts and Telecommunications, Rabat, Morocco | Email: kasmi.ouafae@gmail.com

This paper is organized as follows. In section 2, we briefly review related work. We summarize basic standards and mechanisms, integrity models, Fuzzy Logic in Section 3. The proposed contribution and discussion architecture presented in Section 4. Section 5 concludes this paper, sums up and explains the next steps.

2. Related works And Motivation

2.1. Related works

Recently, the increase in mobile usage data and the new application emergence has greatly motivated the 3GPP organization for working on the IMS network. Various studies made to enhance the QoS in IMS networks [2, 3, 4, 5, 6, 7] due to the network convergence on account to several issues including security, mobility, and QoS management. Some works, focus on detecting problems of degradation of service. Others provide methods for QoS Management [2]. Basic mechanisms for management of QoS in the IMS are presented by 3GPP specification [8, 9]. The 3GPP approaches based on the phase of provisioning services by offering stable and dynamic solutions for the resources according to the type of service. The goal of this approach is to find the needs of each service in terms of QoS requirements. Although, these approaches focus on the phase of providing service, without taking into consideration the customer type, which influenced on the overload of network [10].

In paper [11], the authors propose a technique of causal map for monitoring the resources to identify the fault cause via the technique of map causal [11]. The aims of this solution is to detect a failure by identifying the probable cause. This approach allows monitoring resources and not services flows [5].

The paper [13], propose a platform SLM&M (service Level Management & Monitoring) for services monitoring, it is based on indicators collected directly over the network. However, before deducting the requirements provisioning service to users, these indicators are analyzed using thresholds defined in the Service Level Agreement (SLA) of each client. The aim of this solution is to detect only the degradations without including the corrective actions in real-time.

In paper [2], authors propose a new approach INQA (IMS Network QoS Architecture) that is presented at SLM&M platform to supervise and monitor a service in the real-time according to key performance (KPI) and quality indicators (KQI). The aim of this solution is to project enhanced Telecom Operation Map (eTOM) in the context of IMS network by integrating entities in a multi-technologies context using web technology; more particularly service oriented architecture SOA and simple object access protocol SOAP [12].

In [14], authors discuss three types of SLA, which are SLAs Customer, SLAs Supplier/Partner, and SLAs Interns. A service provider for offering products to the customer intends SLAs Customer. This type of SLA is generally subject to a formal contract between the customer and the service provider. SLAs Supplier/Partner are intended for suppliers who are designed to ensure whether performance targets met for the service components provided by a supplier and for derived from a partial or total contractual agreement between suppliers. An agreement for all customers using the services delivered by the service provider. SLAs Interns are designed primarily to target services and working resources. They are intended to service elements or groups of service elements and are described as the agreements between organizational functions within the business operators.

Paper [13], proposes the verification of SLA using eTOM process [15] for monitoring services. The aim of this verification is to identify the appropriated QoS, which is determined in the SLA contract from the information collected by the network entities. Four steps, structures this verification: Performance Indicators Collection [13], Mapping quality indicators, Comparison between quality indicators and SLA, and Report overall service and resource. In the performance indicators collection step, a key performance indicator KPIs is identified by Packet Loss (%), Delay (ms), and Bandwidth (kb/s). Unlike indicators of performance, quality indicators are correlated to the service, these indicators are MOS-V (Mean Opinion Scores Video) [16] and MOS-A (Mean Opinion Scores Audio) [17]. The following step is a comparison between quality indicators and SLA to verify the negotiation of SLA. Lastly, a report on the resource during service provisioning and its status are created for detecting the type of SLA execution.

2.2. Motivation

In IMS network, SLA execution takes two forms: Normal SLA Execution and SLA execution with a violation. The Normal SLA Execution detects the exceeded thresholds or failure of a component and the manager intervenes. However, the SLA execution with violation restores services and QoS operation of Video service on demand (VoD).

To enhance an SLA security policy and to control the exchange between different entities, the proposed approach is to integrate a layer called integrity verification communication in IMS architecture for applying integrity policy [1] in terms of criticality [1]. To achieve this goal, the Totel model of integrity is adopted to improve a new approach of multi criticality using multi-criteria for decision-making and to guarantee QoS needs according to the level of criticality of users. This proposition is integrated into INQA and SLM&M architecture as illustrated in Figure 1. It is composed of two layers: assurance and monitoring. The assurance layer is dedicated for verification process of SLA according to several entities of eTOM process. The monitoring layer is dedicated for supervising; it is composed of a set of functional processes for ordering services and other execution scenarios of a resource. It is divided into two parts: A distributed part that is a set of agents and probes that are able to recover all data in real-time. Moreover, a centralized part is dedicated to synchronization.

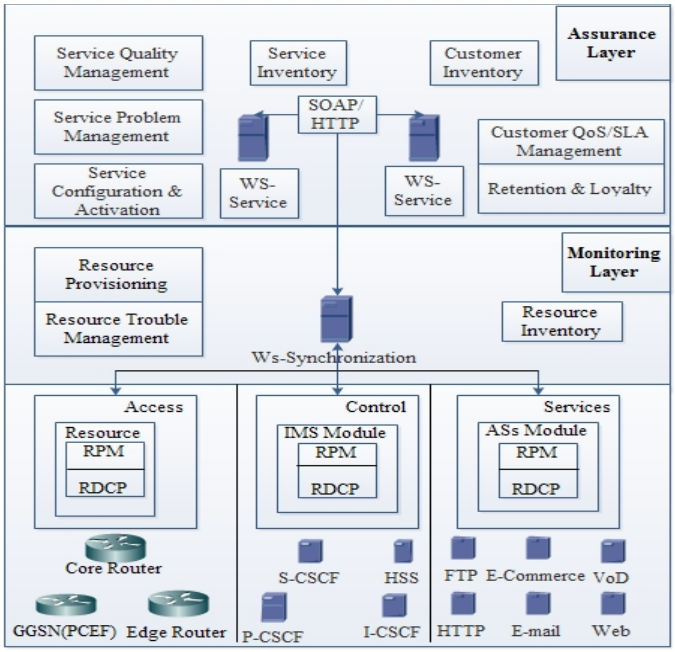


Figure 1: *Architecture SLM&M and INQA [2]*

3. Overview

An access control model intended to abstract the policy and handle its complexity, to verify the consistency of the security policy and detect possible conflicts. It is based on essential model and mechanisms, integrity policy, IMS, Fuzzy Logic, which will be outlined in this section.

3.1. Survey of integrity

Computer security in critical infrastructure is designed to control access and manipulate data by using the mechanisms of authentication, integrity, authorization, in order to protect a system against the vulnerabilities and failures. One basic element of security is integrity. It is a key for checking the trust change of data by using verification mechanisms such as integrity models. These models keep data safe and trustworthy by protecting system data from any intentional or accidental changes. For a long time in computer systems, this concept was used for three goals: to prohibit unauthorized users from making any changes to data or programs, to prohibit authorized users from making unauthorized modifications and finally to prevent internal and external consistency of data and programs [18]. Integrity models allow quantifying the integrity in term of credibility or criticality [19]. Credibility can be considered as an integrity fact, since it is defined as the character of something that can be believed, on what trust can be placed. The measure indicator of credibility is based on reliability, internal actions, and reputation. However, criticality is the characteristic of a system that does not tolerate failures and must secure its properties even through internal and external events. This sensitivity is related to system importance on the socio-economic development of a nation, and the grave consequences that could emerge from a service interruption caused by any inadvertence or malicious source [19].

3.2. Integrity Models

The models of integrity policy are:

- The policy defined by Bell & Lapadula that addresses confidentiality of data systems. It is based on two properties: Simple Integrity (No read down) and Integrity* property (No Write up) [21]. For simple integrity property, the subject can only read objects at its own integrity level or above. On the other side the Integrity* property, the subject can only write objects at its own integrity level or below.
- The Chinese Wall policies aim is to protect the financial system from illicit manipulation. It consists of three levels of abstraction object: Objects that contain information about only one company, Company data set (CD), and the conflict of interest (COI). CD is an object collection related to a certain company. COI is a class containing the data sets of competing companies [22].
- The policy defined by Clark & Wilson is designed for commercial aims and focuses on data integrity preservation. Indeed, each data item can only change using a certified procedure, which keeps its integrity. Furthermore, it is dedicated to verifying data items and certify their integrity. This model can be interpreted as two integrity levels: data are classified into Unconstrained Data Items (UDI) and Constrained Data Items (CDI). This model defines two notions: the transactions and separation of duties. The first notion states that information may only be treated through Transformation Procedures (TPs) which support the involved data integrity. Additionally, integrity Verification Procedures (IVPs) can be verified modification integrity data. The system keeps lists of a relation between these entities which allowing UDIs into CDIs transformation according to TPs and IVPs [23]. Unlike Biba policy, two ways influenced the policy defined by Clark & Wilson: a data who gives integrity level should keep its integrity level only if it is manipulated by a procedure validated to at least the same level. Moreover, the information flow passage from a lower to the highest level is possible only if it is validated to certify that the data has a high integrity level.
- Totel model is based on the object-oriented approach. It aims to offer a usable and flexible generalization of the other models mentioned above for providing kinds for objects with defined properties. Figure 2 presents the software architecture proposed in Totel's Model. It is based on objects classification in order to give a strict information flow to check integrity policy and to manipulate the mechanisms of information integrity upgrade. Authors in [24, 25] defined three object types: **Single Level Objects (SLOs):** are single constant integrity level. This king contains information and creates flows using its internal data at different levels of integrity. **Multi Level Objects (MLOs):** are intended for providing more flexibility by being able to offer some services at any integrity level. They are validated to a given integrity level and can accept data from any level. They are modified dynamically and restoring it at the beginning of the next invocation. To achieve this aim, the authors prove that these objects have no memory. At each invocation, these objects will be created from scratch. **Validation Object (OV):** provides information with a sufficient certified integrity level to

allow flow from low to high integrity level. Furthermore, it runs fault tolerance mechanisms to produce high integrity outputs. The software architecture described a Trust Computing Base (TCB), which is composed of operating system micro and kernel integrity. The goal of TCB is to control all communications in the system, prevent unauthorized rising flows by authorizing only ascending flows, and sending it to a validation object.

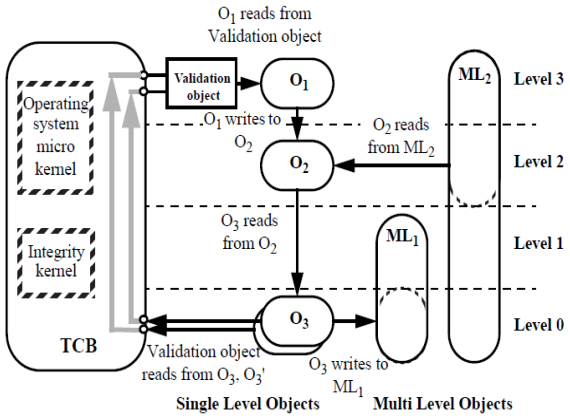


Figure 2: Illustrations of the integrity policy rules [25]

3.3. Integrity Based IMS architecture

In IMS networks, the integrity protection is applied between user equipment (UE) and the P-CSCF for protecting the SIP signaling according to several steps: Firstly, UE and P-CSCF negotiate the integrity algorithm for the session. Secondly, these entities agree on key of security, which is based on AKA algorithm and they verify that data originates received. Finally, the replay and reflection attacks must be mitigated. The integrity algorithm defined in IMS architecture is HMAC-MD5-96 or HMAC-SHA-1-96. This algorithm should be supported by UE and the P-CSCF. The protection indicator of integrity to decide the registration is as follows: (1) REGISTER request from the UE needs to be authenticated. (2) The S-CSCF should be known about the integrity protection applied to a message. (3) The p-CSCF lies an indication in the REGISTER request to inform the S-CSCF that the message was integrity protected for verifying if the message of integrity is the same or not. (4) For all other REGISTER requests, the P-CSCF attached an indication in order to verify that was not integrity protected or ensures that there is no indication about integrity protection in the message [5].

3.4. Fuzzy Logic

Fuzzy Logic is a technique that provides mechanisms to handle nonlinear uncertainties that exist in physical systems. It introduced by Lotfi Zadeh [26]. It is applicable in several applications [27]. Fuzzy sets represent objects collection, which is characterized by a membership function. The membership function is a type variable, which uses words instead of numbers to represent its knowledge in fuzzy systems that express a conditional rules statement [28]. The linguistic variables are parameters that are used to control the system and to show their performance. It is divided into levels. Fuzzy operations are the same operations in classical sets, which are union, intersection, containment, and complement. Fuzzy rules consist of three parts antecedent, fuzzy proposition, and consequence. Fuzzy inference system (FIS) is the regrouping of fuzzy rules, the linguistic variables of membership functions and fuzzy reasoning [27]. The FIS is designed for constructing complex and nonlinear relationships between the input and output. There are many FIS used in applications such as Mamdani, Tsukamoto, and Sugeno [30]. The difference between these types of FIS is in the aggregation and defuzzification processes. The Mamdani FIS method is used in our case because it has widespread acceptance and it has the most commonly used of the three FIS [30]. To implement an FIS, four steps to follow fuzzification, rule evaluation, aggregate output, and defuzzification.

4. Contribution And Discussion

The approach of multi criticality on IMS networks aims applying of a policy of security in terms of criticality. The proposed contribution is to integrate a layer called Integrity verification communication in Test Bench [2] as shown in Figure 3. This integration is based on Totel model. Even though, in the IMS network, the entities of the controller have much knowledge of the network. Although, the administration of real-time services is a difficult task given the fluidity of the network to be ensured at any time. That is why, the contribution approach is to integrate a layer of security policy to monitor business processes [2] responsible for the implementation regarding verification and correction scenarios to ensure the integrity of the exchanged messages. Using Totel Model, for integrity policy presents an advanced innovation in terms of QoS management in IMS network compared with the other models [20, 21, 22, 23]. These models prohibit the passage flow from a lower level of criticality to the highest level of criticality. On the other hand, in IMS network; the users are mobile devices, have different preferences such as environment, equipment …

For this effect, the communications should be circulated in both ways from less critical tasks to more critical tasks. The proposed contribution is based on an object-oriented, the SLO objects described levels of QoS and the MLO objects described levels of criticality of customers. In [24], the authors have specified that the MLO allow changing their level of criticality at any invocation method to control and accept data from lower or equal level. These objects have no memory, and are created at each invocation, which influences execution duration and reduces network performance; also, the identification of the exacted values for these thresholds is very difficult. For this reason, the concept of Fuzzy logic is used to determine dynamically criticality levels for users according to several criteria: customers, equipment, and environment. This technique allows a significant gain in terms of determining appropriate QoS to the users' levels of criticality.

4.1. Integration of fuzzy logic in the multi criticality approach

There are many main components of soft computing such as Fuzzy Logic, Neural Networks, Probabilistic reasoning [29] and Genetic algorithms [31]. The most important one is Fuzzy Logic.

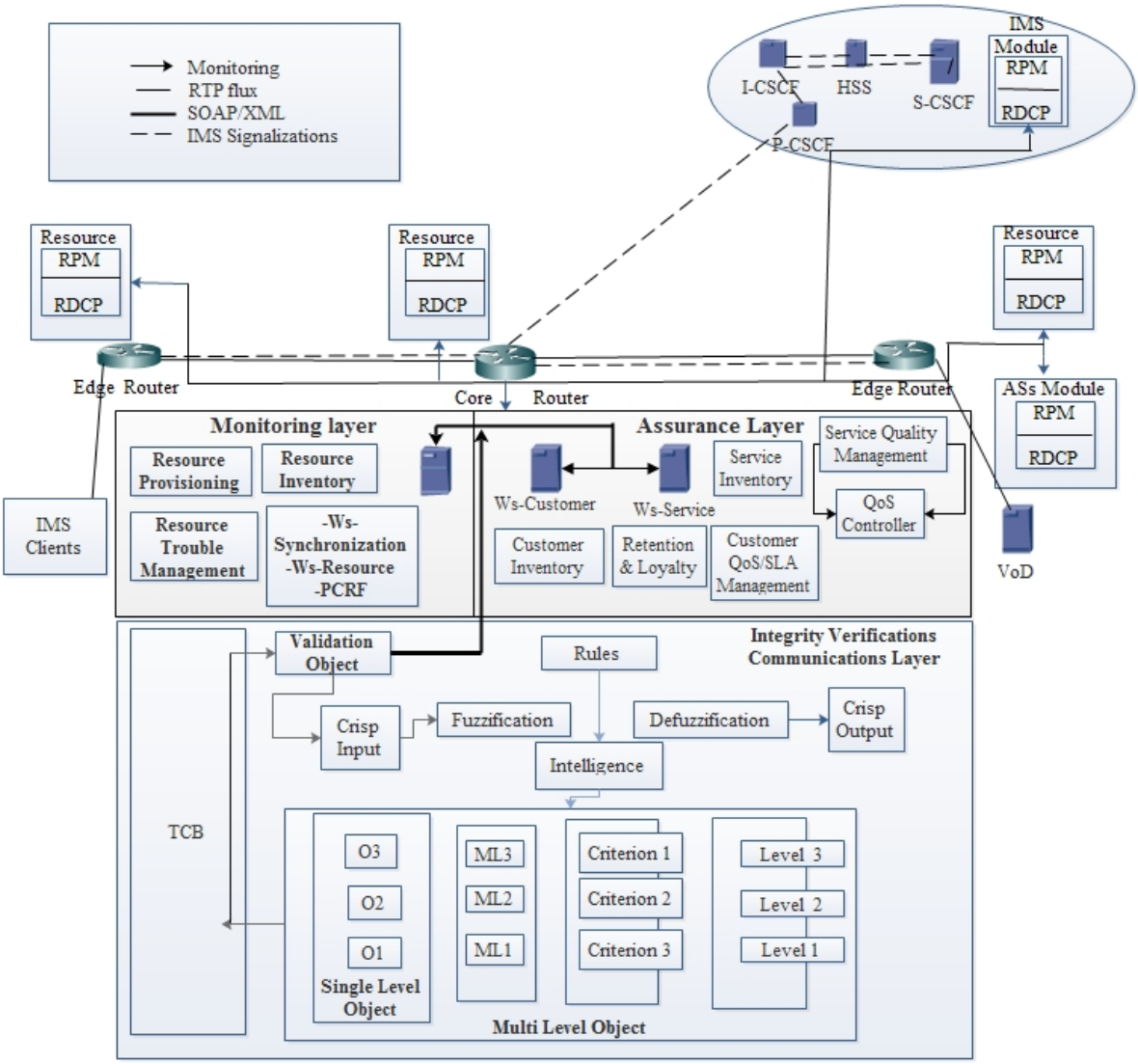


Figure 3: Architecture proposed [1]

It incorporates modeling, control management, forecasting and performance estimation. Also, it used on several applications. The integration of multi criticality approach, based on fuzzy logic as shown in Figure 3 has several advantages while estimating a QoS considering the increasing or decreasing levels of criticality of customers, levels of QoS services and optimizing costs. However, this approach is a key to satisfaction both QoS and QoE of customers' needs.

4.2. Use Case of Fuzzy Logic

The simulation results are based on Matlab tools. The fuzzy logic tool allows users to create Fuzzy inference for estimating conclusions problems. To simulate a system, the system is built through the Graphical User Interface (GUI) Tools. Five GUI tools are being executed to have simulations for the input and output of our system: building, editing FIS, Membership function, Rules, and reviewing Rule and surface. The viewer Rule and surface are being used for a survey with FIS editor.

FIS Design

Fuzzy Logic is based on Fuzzification, rule evaluation, aggregation, and defuzzification as mentioned previously. The simulation is based on Mamdani method, which is the most frequently used method [30]. This method allows a system to work in an easy and simplified mathematical way as described in Figure 4.

For modeling the control system, the FIS design considers three parameters: Environment, Equipment, and Customer to define the chance value of SLOService and levelOfCriticality MLO. Customer input presents a degree of importance of users according to their activities in the IMS network. Equipment input presents categories of mobile use. Environment input presents areas when users placed. SLOService describes a QoS and LevelOfCriticality describes levels of criticality for the customer.

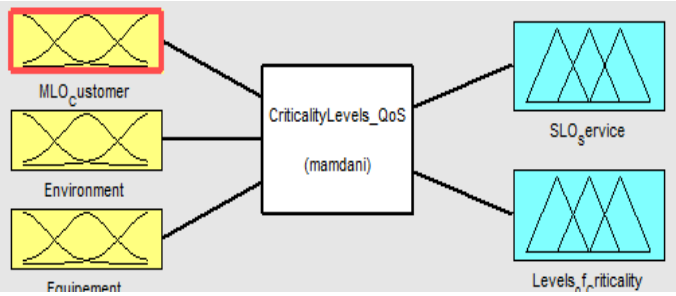


Figure 4: Fuzzy Inference System for CriticalyLevels_QoS

Each parameter is divided into three linguistic value levels: Average, Good, and very Good for Customer input, and Dense Urban, Urban, Suburban for Environment, and category 1, category 3, category 5 which present bit rate for Equipment. Moreover, Matlab Fuzzy Logic tools included several types of membership functions such as triangle, Trapezoidal, Gaussian, sigmoidal and S-shape. In the next step, some simulations are found to determine the important membership functions between Gaussian, trapezoidal and Triangle whose gives a good result in the aspect of chance value for SLOService and MLO.

The variable linguistic for input parameters is presented by a Gaussian membership function as illustrated in Figure 5, Figure 6 and Figure 7.

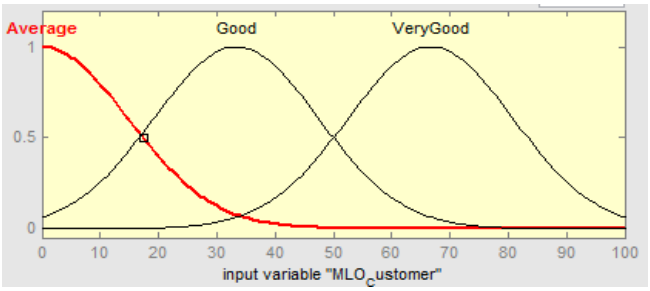


Figure 5: Membership function for Customer input

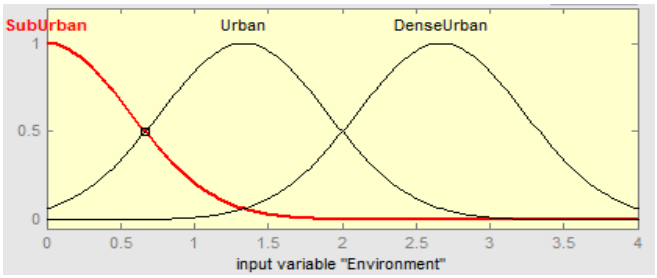
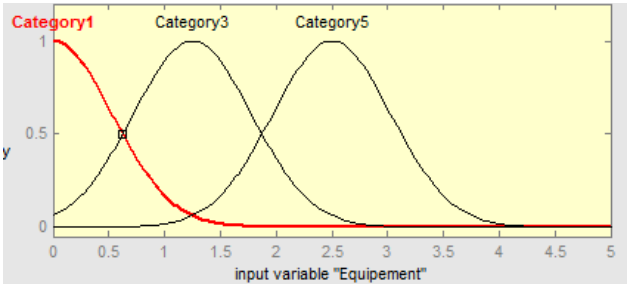
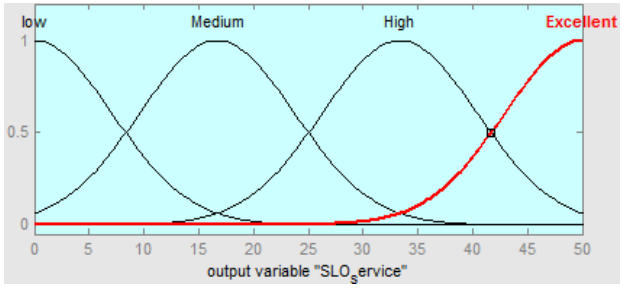


Figure 6: Membership function for Environment input





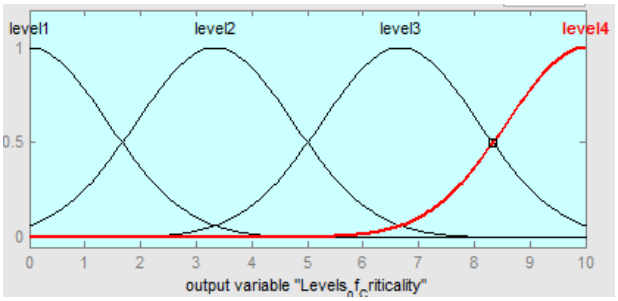


Figure 9: Membership function for LevelsOfCriticality output

Table 1: Fuzzy Inference System IF-THEN rules

Rule NO	Customer	Environment	Equipment	SLO	MLO
1	Average	DenseUrban	Category5	M	L2
2	Average	DenseUrban	Category 3	L	L1
3	Average	DenseUrban	Category1	L	L1
4	Average	Urban	Category5	M	L2
5	Average	Urban	Category3	L	L1
6	Average	Urban	Category1	L	L1
7	Average	SubUrban	Category5	M	L2
8	Average	SubUrban	Category3	L	L1
9	Average	SubUrban	Category1	L	L1
10	Good	DenseUrban	Category5	E	L4
11	Good	DenseUrban	Category 3	H	L3
12	Good	DenseUrban	Category1	M	L2
13	Good	Urban	Category5	H	L3
14	Good	Urban	Category3	H	L3
15	Good	Urban	Category1	M	L2
16	Good	SubUrban	Category5	H	L3
17	Good	SubUrban	Category3	H	L3
18	Good	SubUrban	Category1	M	L2
19	VeryGood	DenseUrban	Category5	E	L4
20	VeryGood	DenseUrban	Category 3	H	L3
21	VeryGood	DenseUrban	Category1	M	L2
22	VeryGood	Urban	Category5	E	L4
23	VeryGood	Urban	Category3	H	L3
24	VeryGood	Urban	Category1	M	L2
25	VeryGood	SubUrban	Category5	E	L4
26	VeryGood	SubUrban	Category3	H	L3
27	VeryGood	SubUrban	Category1	M	L2

M=Medium, L=Low, E= Excellent, H= High, L1=Level 1, L2=Level2, Level 3= L3, L4= Level 4, SLO: SLOService, MLO: LevelsOfCriticality.

Table 2: Evaluation of Fuzzy IF-THEN rules

Rule NO	Customer	Environment	Equipment	SLO	MLO
1	Average (=0.7)	DenseUrban (=1)	Category5 (=1)	M (=0.7)	L2 (=0.7)
6	Average (=0.7)	Urban (=0.48)	Category1 (=0.55)	L (=0.48)	L1 (=0.48)
9	Average (=0.7)	SubUrban (=0.58)	Category3 (=0.49)	L (=0.49)	L1 (=0.49)
11	Good (=0.4)	DenseUrban (=1)	Category3 (=0.49)	H (=0.4)	L3 (=0.4)
14	Good (=0.4)	Urban (=0.48)	Category3 (=0.49)	H (=0.4)	L3 (=0.4)
18	Good (=0.4)	SubUrban (=0.5)	Category1 (=0.55)	M (=0.4)	L2 (=0.4)
20	VeryGood (=0)	DenseUrban (=1)	Category3 (=0.49)	H (= 0)	L3 (= 0)
22	VeryGood (=0)	Urban (= 0.48)	Category5 (=1)	E (= 0)	L4 (= 0)
26	VeryGood (=0)	SubUrban (=0.58)	Category3 (=0.49)	H (= 0)	L3 (= 0)

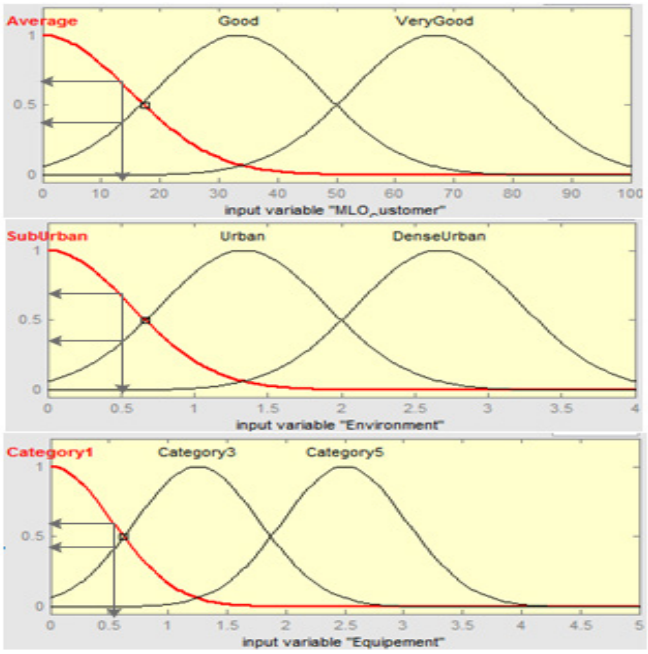


Figure 10: Fuzzification of crisp value input

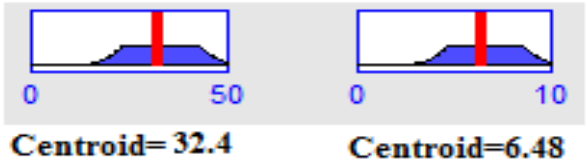


Figure 11: Centroid Point for Defuzzification

Analysis and examples for trapezoidal and Triangle membership function

The following simulation describes a comparison between Gaussian, triangle and trapezoidal membership function. The linguistic variable for the middle level (good, Urban, category3) is presented by the triangle membership function, while the other levels are presented by Trapezoidal membership function as illustrated in Figure 12, Figure 13 and Figure 14.

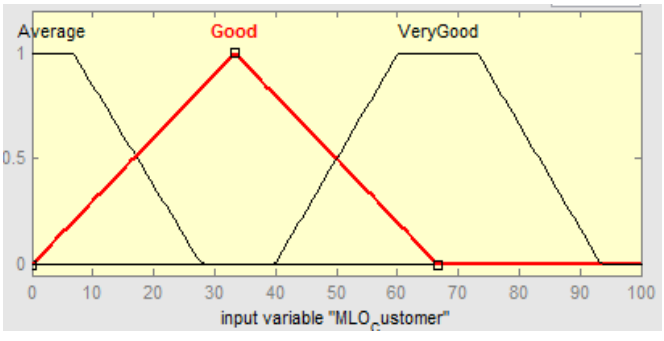


Figure 12: Fuzzy set of Customer

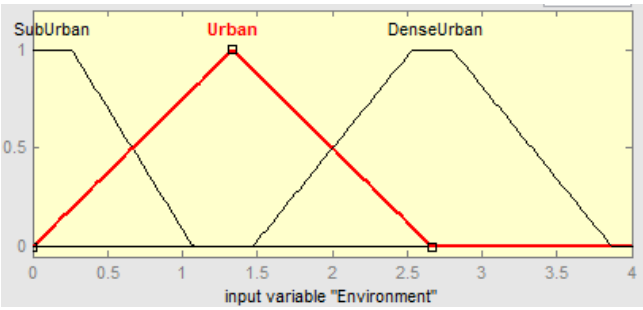


Figure 13: Fuzzy set of Environment

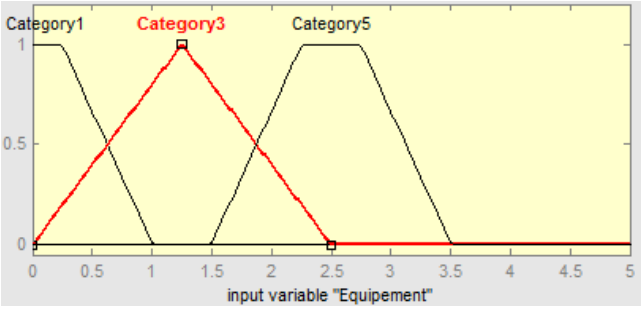


Figure 14: Fuzzy set of Equipment

To give our incidence feature of flexibility, the linguistic variable for chance value is divided into 9 levels following as VeryLow, Low, ratherLow, LittleMedium, Medium, HighMedium, ratherHigh, High, VeryHigh for levels of criticality and veryWeak, Weak, littleLow, LittleMedium, Medium, HighMedium, littleHigh, High, VeryHigh for levels of QoS as illustrated in Figure 15 and Figure 16.

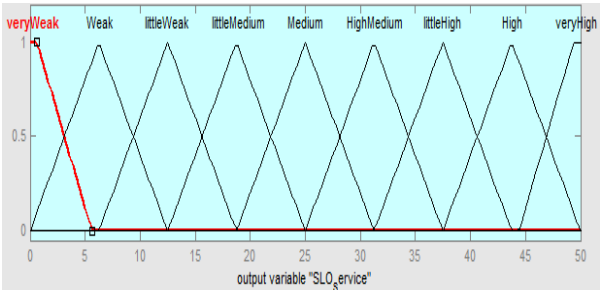


Figure 15: Fuzzy set of SLOsService

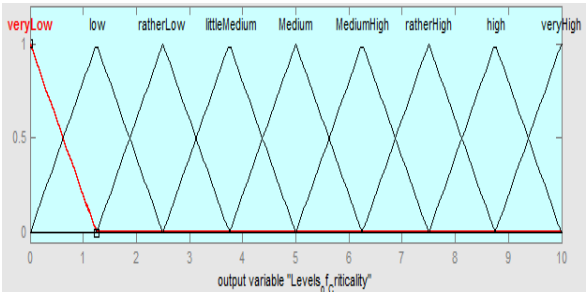


Figure 16: Fuzzy set of SLOsService

According to this result, the degree that is determined by a Triangle and Trapezoidal membership functions is more specific than a membership function of Gaussian. In addition, Triangle and Trapezoidal membership functions are more useful than the other types because their degree is more easily determined. Table 3 illustrates IF-THRN rules. The following Figure 17 described a result obtained for defuzzification by triangle and trapezoidal membership function.

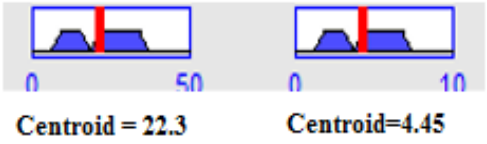


Figure 17: The centroid point

Table 3: Fuzzy Inference System IF-THEN rules

Rule No	Customer	Environment	Equipment	SLO	MLO
1	Average	DenseUrban	Category5	LW	rL
2	Average	DenseUrban	Category 3	weak	Low
3	Average	DenseUrban	Category1	VW	vL
4	Average	Urban	Category5	LW	rL
5	Average	Urban	Category3	weak	Low
6	Average	Urban	Category1	VW	vL
7	Average	SubUrban	Category5	LM	rL
8	Average	SubUrban	Category3	LW	Low
9	Average	SubUrban	Category1	Weak	vL
10	Good	DenseUrban	Category5	HM	HM
11	Good	DenseUrban	Category 3	M	M
12	Good	DenseUrban	Category1	LM	LM
13	Good	Urban	Category5	HM	HM
14	Good	Urban	Category3	M	M
15	Good	Urban	Category1	LM	LM
16	Good	SubUrban	Category5	HM	HM
17	Good	SubUrban	Category3	M	M
18	Good	SubUrban	Category1	LM	LM
19	VeryGood	DenseUrban	Category5	VH	VH
20	VeryGood	DenseUrban	Category 3	High	High
21	VeryGood	DenseUrban	Category1	LH	rH
22	VeryGood	Urban	Category5	VH	VH
23	VeryGood	Urban	Category3	High	High
24	VeryGood	Urban	Category1	LH	rH
25	VeryGood	SubUrban	Category5	VH	VH
26	VeryGood	SubUrban	Category3	High	High
27	VeryGood	SubUrban	Category1	LH	rH

For SLOService QoS: Low: veryWeak(VW)-weak-LittleWeak (LW), Medium: LittleMedium (LM)-Medium (M)-HighMedium (HM), High: LittleHigh (LH)-High-HighMedium (HM). For LevelsOfCriticality MLO: Level1 (low level of criticality): veryLow (vL)-Low-ratherLow (rL), Level2-3(Medium Level of criticality): LittleMedium (LM)-Medium (M) - MediumHigh (MH), Level4 (High level of criticality): ratherHigh (rH) - high- veryHigh (vH).

Analysis and examples for different defuzzification methods:

Our analysis result takes into consideration two cases:

Case 1: Low quality of service and Low Level of Criticality.

Case 2: Excellent quality of service and High Level of Criticality.

Case1: if (customer is average) and (Environment is SubUrban) and (Equipment is Category1) then (LevelofCriticality is veryLow) and (SLOService is veryWeak), which is described in Rule viewer and illustrated in Figure 18.

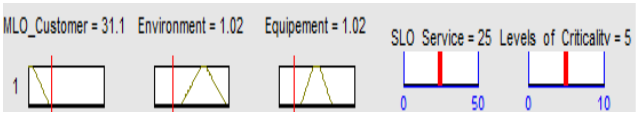


Figure 18: Rule Viewer for case 1

While, the values for MLO_Customer as average (31.1), Environment 1.02 and Equipment 1.02, then SLOService is 25 and Level of criticality is 5. In this case, the Level of criticality is low and the level of QoS also is Low. Therefore, the activity of users is average, the environment is suburban and the equipment is category1, which bit rate is low. Thus, the level of criticality and QoS are low. According to this result of defuzzification, the service of VoD is delivered. This value is obtained by using COG method. The proposed approach of multi criticality is based on several criteria for making decision problems. For this reason, the FIS incorporates three inputs and two outputs. In this case, a 3D surface cannot show the relations between all output criteria: SLOService and levelofCriticality. In such a case, several 3D fuzzy surfaces are illustrated in different examples in Figure 19. Each surface corresponds to one scenario and allows the sensibility of the SLO and LevelsOfCriticality output with Customer and Environment input. Three main methods of defuzzification methods such as COG (Centroid of Gravity), MOM (Mean value of maximum), and LOM (Largest value of Maximum) are simulated by using 3D surface viewer as illustrated in Figure 19. A method COG returns a maximum of aggregation solutions. MOM method takes as a solution the average of the most likely values output. In addition, LOM method returns the maximum value, which is determined from membership functions.

Table 4 defines the results of comparative for three methods of defuzzification such as COG, MOM and LOM. Rule number 9 describes case 1 that represents a low quality of service (SLO) which value of COG method is 13.6, MOM is 6.25 and LOM is 9. In addition, the value of low Level of Criticality (MLO) using COG method is 2.92, MOM is 0.25 and LOM is 0.5.

From the result above, The COG method returns the maximum chance values for providing the more stable results among the other methods (MOM and LOM), considering the variations in the set conclusion which are graphically illustrated in Figure 19. From these figures, also the same results are concluded, the COG defuzzification methods take into a consideration several values for input that represented by different color as compared to MOM and LOM method. Thus, the concluded result is that the COG method gives a better result than MOM and LOM methods.

Table 4: Output values obtained for different defuzzification method

S. No	Input			Output SLO			Output MLO		
	Ct	**Ev**	**Eq**	**C**	**M**	**L**	**C**	**M**	**L**
1	8.3	3.54	4.43	25	25	25	5	5	5
2	8.3	3.54	2.54	21.1	12.5	16.5	4.21	2.5	3.3
3	8.3	3.54	0.56	18.3	6.25	10.5	3.66	1.25	2.1
4	8.3	2.37	4.43	25	25	25	5	5	5
5	8.3	2.37	2.54	18.3	12.5	13	3.66	2.5	2.6
6	8.3	2.37	0.56	17	6.25	9.5	3.4	1.25	1.9
7	8.3	0.64	4.43	25	25	25	5	5	5
8	8.3	0.64	2.54	19.5	12.5	15.5	3.9	2.5	3.1
9	8.3	0.64	0.56	13.6	6.25	9	2.92	0.25	0.5
10	50	3.54	4.43	25	25	25	5	5	5
11	50	3.54	2.54	21.1	36.8	50	4.21	7.37	10
12	50	3.54	0.56	18.3	31.3	48	3.66	6.25	9.6
13	50	2.37	4.43	25	25	25	5	5	5
14	50	2.37	2.54	37.1	48.5	50	7.45	9.7	10
15	50	2.37	0.56	31.3	37.5	40.5	6.27	7.5	8.1
16	50	0.64	4.43	25	25	25	5	5	5
17	50	0.64	2.54	37	48.5	50	7.44	9.7	10
18	50	0.64	0.56	31.3	37.5	40.5	6.26	7.5	8.1
19	85	3.54	4.43	25	25	25	5	5	5
20	85	3.54	2.54	47.7	48	50	9.49	9.6	10
21	85	3.54	0.56	40.6	40.8	48	8.12	8.15	9.6
22	85	2.37	4.43	25	25	25	5	5	5
23	85	2.37	2.54	47.8	48.3	50	9.51	9.65	10
24	85	2.37	0.56	40.6	40.8	47.5	8.12	8.15	9.5
25	85	0.64	4.43	25	25	25	5	5	5
26	85	0.64	2.54	47.8	48.3	50	9.51	9.65	10
27	85	0.64	0.56	40.6	40.8	47.5	8.12	8.15	9.5

Ct (Customer), Ev (Environment), Eq (Equipment), C (COG), M (MOM), L (LOM).

Figure 19 shows that the environment and type of customer are low, thus resulting of SLO and MLO is Low, this is why the result is different for three defuzzification methods. A service VoD can be transmitted with low quality for a customer whose have a low level of criticality.

Case2: if (customer is VeryGood) and (Environment is DenseUrban) and (Equipment is Category5) then (LevelofCriticality is veryHigh) and (SLOService is veryHigh) as described in Rule viewer and illustrated in Figure 20.

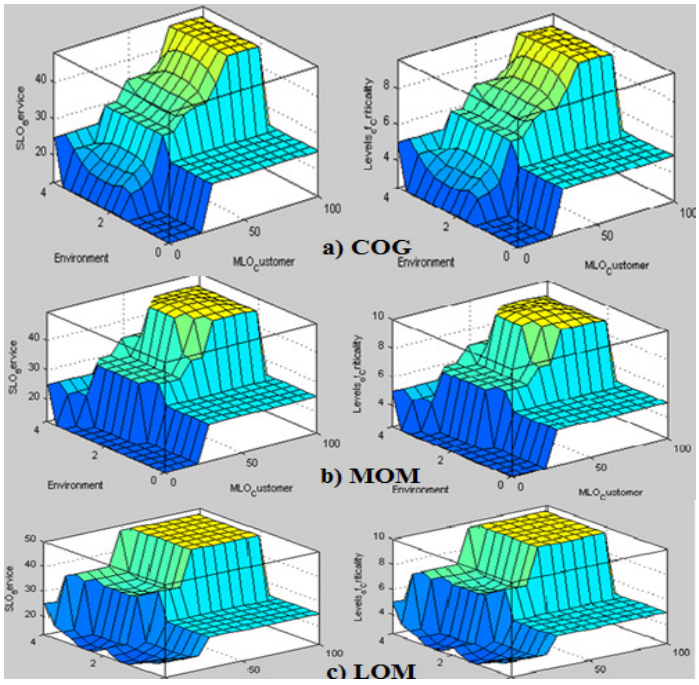


Figure 19: Surface viewer with method: a) COG, b) MOM and c) LOM

Figure 20: Rule Viewer for case 2

In case 2, the value of the Customer is 82.6, Environment is 3.23 and Equipment is 3.14. For SLOServices the value is 47.5 and LevelOfCriticality is 9.53. The analysis determined that the QoS is Excellent and the level of criticality is high. This means that using the parameters above, the service VoD can be transmitted with excellent quality for a customer whose have a high level of criticality. The case 2 is illustrated with rule 19 in Table 4, which describes DenseUrban environment, veryGood type of customer and high category of mobile, thus resulting in high SLO and MLO this is why the result is approximately the same for three defuzzification methods: GOC, MOM, and LOM, which are 25 for SLO output and 5 for MLO output. A service VoD can be transmitted with high quality for a customer whose have a high level of criticality.

From the above examples in case 1 and case 2, the COG defuzzification method gives better results in the aspect of SLOService and LevelsOfCriticality. This can be seen with the surface viewer and the rule viewer as presented in the above examples using different rules. In this section, the proposition aims are to design some control mechanisms, for solving problems of QoS during multimedia service transmission by using an intelligent controller, in order to determine dynamically levels of criticality of customers. This controller is analyzed by using Fuzzy Controller Language (FCL), the input of FCL is determined and the most appropriate membership functions and defuzzification method are founded for determining better crisp values in the aspect of SLOService and LevelsOfCriticality in IMS network. In addition, the methods for membership function (Gaussian, Triangle, and Trapezoidal) and methods for defuzzification: GOC, LOM, and MOM are compared. The 3D surface viewer is used for describing the comparison offered by three methods. Through these analyses, the COG method, triangle, and trapezoidal membership functions give better results in all analyzed cases.

5. Conclusion

In this paper, the proposed approach of multi-criticality is based on Fuzzy Logic to ensure integrity on IMS architecture, in order to determine levels of criticality dynamically by using multi-criteria: Environment, Equipment and Type of customer for making the decision to offer a level of QoS of VoD. It is based on Totel model. From the simulation results, the triangle and trapezoidal membership functions give better results than Gaussian membership also COG method of defuzzification is more efficient than MOM and LOM methods. Future works, will be focused on the technical approach of Totel model using technologies web SOA, make a VoD by using openIMS, and will be focused on the prediction of the indicators for storing the antecedents' cases of degradation, furthermore another aspect of QoE will be discussed.

References

[1] O. Kasmi, A. Baina, M. Bellafkih, "Multi level integrity management in critical infrastructure," International Conference on Intelligent Systems: Theories and Applications, IEEE, 1-6, 2016.

[2] M. Errais, M. Bellafkih, R. Daniel, "Autonomous System for network monitoring and service correction, in IMS Architecture," International Journal of Computer Science and Applications, 12 (1), 80-93, 2015.

[3] R. Rajavelsamy, P. Basavaraj, C. Suresh, "Efficient registration procedure for multi-domain authentication for mission critical communication services," Standards for Communications and Networking, IEEE, 282-287, 2015.

[4] M. Navarro, Y. Donoso, V. Rodriguez, "An IMS architecture with QoS parameters for flexible convergent services," Symposium on: Computers and Communications, IEEE, 640-645, 2010.

[5] M. Poikselka, A. Niemi, H. Khartabil, G. Mayer: The IMS: IP Multimedia Concepts and Services, John Wiley & Sons, 2013.

[6] U. Baniel, K. C Jackson, T. Abou-Assali, M. Mercurio, D. M. Sprague, "Methods, systems, and computer readable media for policy and charging rules function (PCRF) node selection," U.S. Patent No. 8,615,237, 2013.

[7] 3GPP TS 23.228: Service Requirements for the Internet Protocol (IP) Multimedia Core Network Subsystem (IMS), Stage 1, (Release 7), 2007.

[8] 3GPP TS 29.207 : Policy control over Go interface, (Release 6), 2011.

[9] J. Cao, M. Ma, H. Li, Y. Zhang, Z. Luo, "A survey on security aspects for LTE and LTE-A networks," Communications Surveys & Tutorials 16.1, IEEE , 283-302, 2014.

[10] J. Balcerzak, T. Grzegorz, S. Stephane, "QoS monitoring model of registration procedure for IMS platform," 16th International Conference on Telecommunications Network Strategy and Planning Symposium, IEEE, 1-6, 2014.

[11] J. Lu, D. Christoph, K. Francine, "Distributed self-diagnosis algorithm for VoIP service over IMS network," International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, IEEE, 186-191, 2015.

[12] V.K Tiwari, K.G Amit, D. Saurabh, "Web service discovery and integration with QOS parameter using SOA based repository," International Conference on Advances in Computer Engineering and Applications, IEEE, 1014-1009, 2015.

[13] M. Raouyane, M. Bellafkih, M. Errais, "SLA Management & Monitoring based-eTOM and WS-composite for IMS networks," 4th IFIP International Conference on New Technologies, Mobility and Security, IEEE, 1-6, 2011.

[14] Application Note to the SLA Management Handbook - Voice over IP, TMF document reference GB 934, 2005.

[15] SERIES M: Telecommunications management network Enhanced Telecom Operations Map (eTOM) –Representative process flows, ITU-T Recommendation M.3050.

[16] L. Anegekuh, L. Sun, E. Jammeh, I. Mkwawa, E. Ifeachor, "Content-based video quality prediction for HEVC encoded videos streamed over packet networks," Transactions on Multimedia, 17(8), IEEE, 1323-1334, 2015.

[17] S. Lingfen, E. Ifeachor, C. Emmanuel "Voice Quality Prediction Models and Their Application in VoIP Networks, IEEE transaction on Multimedia, 8 (4), 809-820, 2006.

[18] J. Breithaupt, S. Merkow, "Information Security Priciples of Success," part of the certification/training series, 3-15, 2014.

[19] A. Ameziane El Hassani, A. Abou El kalam, A. Ait Ouahman, A. Bouhoula, R. Abassi, "Integrty-OrBAC: a new model to preserve Critical Infrastructure integrity," International Journal of Information Security, 14 (4), 367-385, 2015.

[20] L.J. LaPadula, D.E. Bell, "Secure Computer Systems: A Mathematical Model," Volume II (No. MTR-2547-Vol-2), MITRE CORP BEDFORD MASS, 1973.

[21] K. Biba, "Integrity considerations for secure computer systems," (No.MTR-3153-REV-1), MITRE CORP BEDFORD MA, 1997.

[22] M.B. Baig, C. Fitzsimons, S. Balasubramanian, R. Sion, D. E. Porter, "CloudFlow: Cloud-wide policy enforcement using fast VM introspection," International Conference on Cloud Engineering, IEEE, 159-164, 2014.

[23] K. Li, H. Feng, Y. Li, Z. Zhang, "Information flow control model and method in distribute MILS," International Conference on Computational Intelligence and Security, IEEE, 598-602, 2014.

[24] E. Totel, J. Blanquart, Y.Deswarte, D. Powell, "Supporting multiple levels of criticality," Fault-Tolerant Computing, Twenty-Eighth Annual International Symposium, IEEE, 70-79, 1998.

[25] A. Baina, Y. Laarouchi, "MultiLevel-OrBAC: Multi-Level Integrity management in organization based access control framework," International Conference on Multimedia Comuting and Systems, IEEE, 933-938, 2012.

[26] B. Bouchon-Meunier "la logique floue et ses applications," Vie artificielle. Addistion-Wesley, 1995.

[27] J.T .Ross "Fuzzy Logic with Engineering Applications," John Wiley & Sons Inc.Chichester, England, 2004.

[28] S. Chaudhari, M. Patil, J. Bambhori, "Study and review of fuzzy inference systems for decision making and control," American International Journal of Research in Science Technology Engineering and Mathematics, 5 (1), 88-92, 2014.

[29] S. Montani, C. J Lakhmi, "Case-Based Reasoning Systems," Successful Case-based Reasoning Applications*, 2*. Springer Berlin Heidelberg, 1-6, 2014.

[30] V. Kamboj,A. Kaur, "Comparison of constant SUGENO-type and MAMDANI-type fuzzy inference system for load sensor," International Journal of Soft Computing and Engineering, 3(2), 204-207, 2013.

[31] V. Chandwani, V. Agrawal, R. Nagar, "Modeling slump of ready mix concrete using genetic algorithms assisted training of Artificial Neural Networks," Expert Systems with Applications, 42(2), 885-893, 2015.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 669-673 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Development of an automated vehicle stop system for cardiac emergencies

Tung T. Nguyen[1], Issey Takahashi[2], Takahiro Tanaka[2], Eijiro Takeuchi[3], Shinpei Kato[2,4], Tatsuya Suzuki[1], Hitoshi Kanamori[2], Yoshiki Ninomiya[2], Hirofumi Aoki[2*]

[1]*Graduate School of Engineering, Nagoya University, Nagoya, Japan*

[2]*Institute of Innovation for Future Society, Nagoya University, Nagoya, Japan*

[3]*Gradute School of Information Science, Nagoya University, Nagoya, Japan*

[4]*Graduate School of Information Science and Technology, University of Tokyo, Japan*

ARTICLE INFO

Article history:
Received: 05 April, 2017
Accepted: 12 May, 2017
Online: 05 June, 2017

Keywords:
Vehicle safety
Cardiac emergency
Automated driving
Bio-signal

ABSTRACT

This paper describes the concept and configuration of a novel automated safety vehicle stop system, and a future prospect of the study. Intrinsic sudden death may cause traffic accident since such accidents sometimes involve not only the driver but also other traffic users such as passengers and pedestrians. Cardiovascular disease (CVD) is considered as a serious driving risk factor. The pain and others effects of cardiac events degrade driver's performance, and CVD causes ischemia brought by the CVD induces incapacity of driving. In the automated safety vehicle stop system, which our research group has developed, steer-sensors collects bio-signals and a camera captures the driver's posture to monitor driver's incapability. When the driver loses his or her driving capability, the system takes over the maneuver of the vehicle and automatically drives to a safety spot by observing the traffic environment. An emergency scenario was used to demonstrate the system verifying its potential.

1. Introduction

The association of driving and physical disorder has been researched for many years. Driving requires both physical and cognitive functions. Beside the effects of long time driving and many cognition stressors, deterioration of driving performance by physical disorder increases the risk of accidents [1]. Although multiple technologies such as adaptive cruise control and emergency braking embedded in the vehicle have decreased the stress of driving on the driver, the number of intrinsic sudden death in traffic accident is still high over the years (Table 1).

Table 1. Autopsy report of intrinsic sudden death of driver

Author	Location	Period	Driver reported as intrinsic sudden death	
			Number	Percentage
West et al. (1968)[2]	USA (California)	1963-1965	155/1026	15%
Osawa et al. (1998)[3]	Japan (Western Kanagawa)	1992-1997	15/188	8%
Oliva et al. (2011)[4]	Canada (Quebec)	2002-2006	123/1260	9.7%
Tervo et al (2013)[5]	Finland	2008-2009	55/488	11.3%

Cardiovascular disease (CVD) and cerebrovascular disease are the most common causes of sudden natural death while driving[6][7]. The effects of cardiovascular disease, such as myocardial infraction, are not immediately but gradual, so that the driver can somehow step on the brake and/or maneuver the vehicle to the side of the road. The pain of cardiac events distracts driver's concentration, and ischemia brought by CVD induces impairment of awareness, recognition, eyesight, physical strength, and others functions.. Syncope is the sudden and transient loss of consciousness. The aftereffect of some serious arrhythmias is syncope that may lead to the lost control of the vehicle. The association of driving and CVD has been investigated for many years [2][8][9]. In Japan, Motozawa et al [6] have reported that among the autopsies results of Transportation Bureau of National Police Agency (Japan), the main cause of death is ischemic heart disease. In European countries, meta-analysis has shown that CVD patients have 23% higher risk of accident involvement than drivers without CVD [10]. Many other researches also pointed out that CVD drivers have higher risk of accidents than healthy drivers [11][12][13][14].

*Corresponding Author: Hirofumi Aoki, Nagoya University, Nagoya, Japan Email: hiro.aoki@nagoya-u.jp

Table 2. Percentage of CVD death among intrinsic sudden death at wheel

Author	Location	Period	Intrinsic death caused by CVD (%)
West et al. (1968)[2]	USA (California)	1963-1965	94%
Copeland (1987)[18]	USA (Florida)	1980-1984	82%
Oliva et al. (2011)[4]	Canada (Quebec)	2002-2006	36.8%
Tervo et al (2013)[5]	Finland (Uusimaa)	2008-2009	17.8%

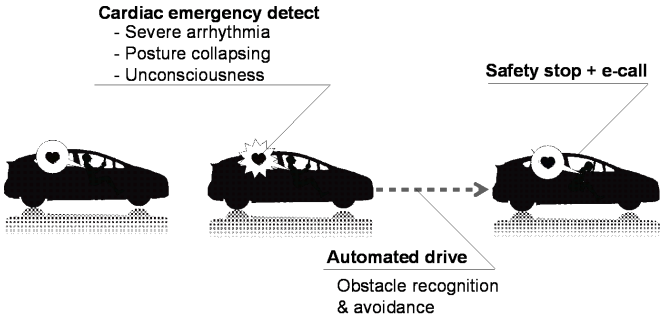


Figure 1. System objective

Various authorities have issued guidelines to detail the criteria for declaring an individual with CVD safe to drive [15][16][17]. Although these methods help decrease the number of death at wheel caused by CVD, drivers still have high potential to get cardiac emergencies behind the wheel, as shown in Table 2. Development of the system that prevents from the CVD-derived accident is essential besides developing the regulations.

This study aims to develop a system that prevents traffic accidents caused by physical disorders. The system detects driver's incapacity by collecting various types of bio-signals and driver's ergonomic factors: electrocardiogram (ECG), photoplethysmogram (PPG), and blood pressure (BP), by using steer-sensors. A driver agent using a camera in front of the driver monitors the driver's posture. The system will take over the maneuver of the vehicle when the driver lost his capacity to drive, and automatically drive to a safety spot by sensing the traffic environment. Figure 1 shows the objective of the research. The following sections describe the concept and the configuration of our first prototype system, The prototype system was designed focusing on cardiac emergencies, and a future prospect of the study.

2. System configuration

The system is composed by the following units: Steering wheel-type bio-signal sensor (Steer-sensor), which is for the cardiac emergencies, Driver's ergonomic factors and driver agent assist to confirm situation, and Automated driving system will take over the vehicle to maneuver to safety spot.

2.1. Bio-signal sensing and cardiac emergency detection

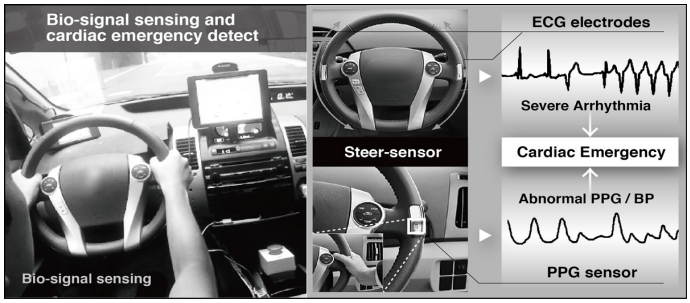


Figure 2. Bio-signal sensing by the steer-sensor and cardiac emergency detection

The collected bio-signals are essential data to detect CVD-event. The electrodes are placed on the outer side of the wheel. ECG and PPG signal are collected by gripping the steering wheel with both hands. Figure 2 shows the position of the both sensors on the wheel. The PPG sensor is located at the place where the base of right thumb touches when a driver grips the wheel in the usual manner. BP is estimated from the pulse transit time (PTT) measured by the ECG and PPG. Further detail is referred to [19][20].

Combining the characteristics of following patterns: 1) tachycardia or bradycardia, 2) abnormality of QRS complex, 3) irregularity of R-R or P-Q intervals, 4) abnormality of BP; the type of cardiac emergency is estimated. Depending on the type of arrhythmias, the normal heart rate will be affected and heart rate might be faster or slower or erratic. These abnormal states of heart rate will affect the control capacity of the driver. When the life-threatening arrhythmias such as ventricular tachycardia (VT), ventricular fibrillation (VF), Torsade de Pointes, or cardiac arrest with the abnormal PPG and BP are detected, the system triggers the emergency signal to the later unit.

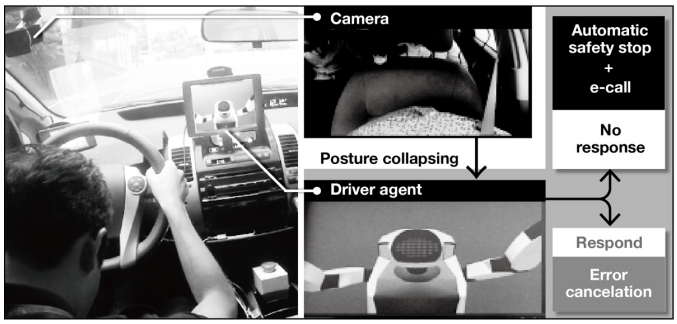


Figure 3. Posture collapsing detection

2.2. Driver's ergonomic factor and driver agent assist

When the system detects an abnormality in driver's bio-state sent from the prior unit, the driver's posture is captured by the driver-monitor camera settled in front of the driver (Figure 3). Some arrhythmias cause pain and/or numb in the chest and shoulder, some make driver feel dizzy or faint. In most of the cases, drivers with serious trouble feel uncomfortable in their chest first and then eventually lose their normal driving posture. A driver agent is installed inside the vehicle to give instruction and alerts while driving.

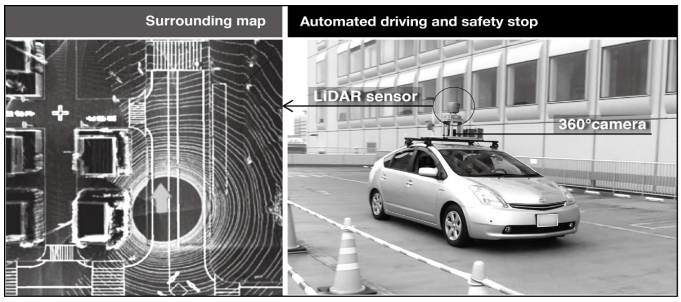


Figure 4. **Automated driving and safety stop**

By using driving instruction model, the driver agent handles cameras data and driving records to monitor driver's driving performance and give instructions if needed to improve driving behavior. If the driver's abnormal posture is detected, the driver agent alerts the driver. The alerting is settled in order to avoid the misdetection of the arrhythmia and posture collapsing. Autonomous drive is activated by the absence of response to the alert for certain duration, and then emergency call (e-call) is taken. Further detail of the driver agent is referred to [21].

2.3. Automated driving system

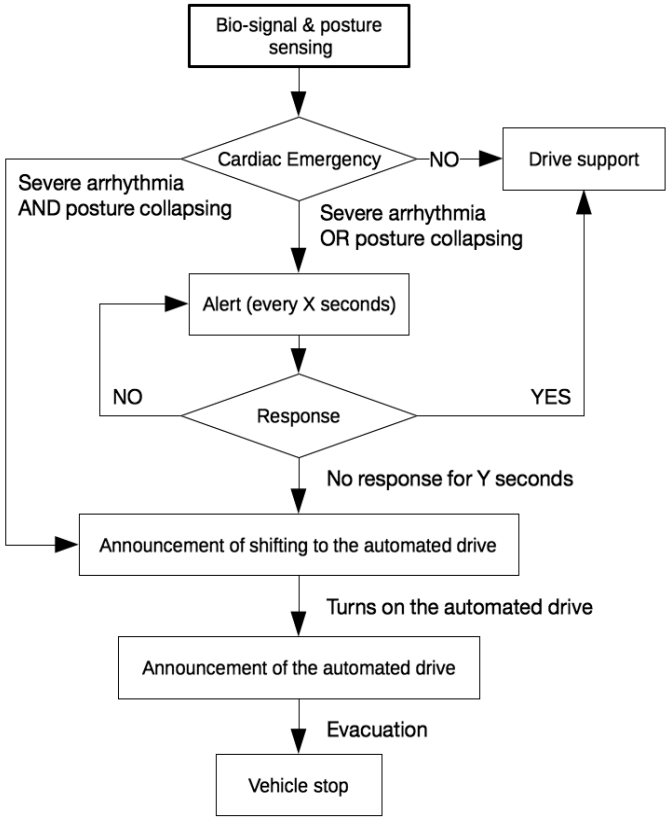


Figure 5. Decision-making procedure of the manual-to automated shift

Various sub-systems in the automated driving system enable the vehicle capture to the surrounding environment: by a LiDAR sensor (HDL-64E, Velodyne LiDAR, Inc.), 360 degrees camera (Ladybug 5, Point Grey Research, Inc.), Grasshoper 3 cameras and Javad RTK sensors receive global positioning information from satellites as shown in the Figure 4. By combining Velodyne LiDAR sensor and Ladybug 5, the system obtains 3D point-cloud data for localization and mapping, and is also used to measure the distance to surrounding objects and moving objects. The Grasshoper can be used to detect traffic lights. The system enables to estimate the position with the precision of 0.1m. This configuration allows the vehicle to drive and avoid obstacles on the road after taking control from the driver. Further detail is referred to [22].

When the driver's critical state is confirmed and trigger signal is sent from prior units, the automated driving mode is activated. The system recognizes the surrounding of the vehicle and automatically finds a safety spot to park the vehicle. Figure 5 shows the decision-making procedure of the manual-to-automated shift.

3. Demonstration

The prototype of the system has been demonstrated at TOC Ariake, Tokyo, Japan in 28th August 2015. A well-prepared scenario was set to demonstrate the working prototype: The driver is driving with normal state then the abnormal state is injected to the system to simulate the critical situation.

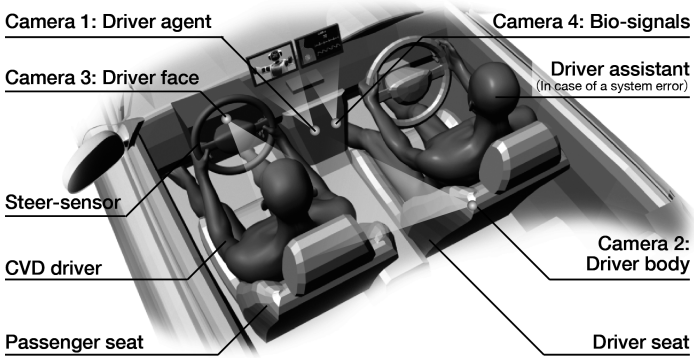


Figure 5. Interior layout of the demonstration

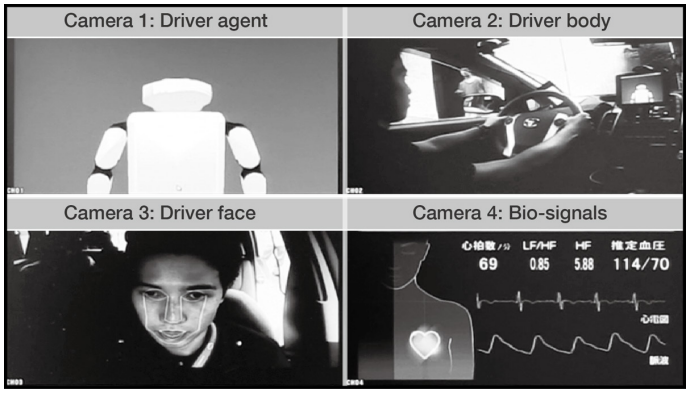


Figure 6. Captured images by the cameras at the demonstration

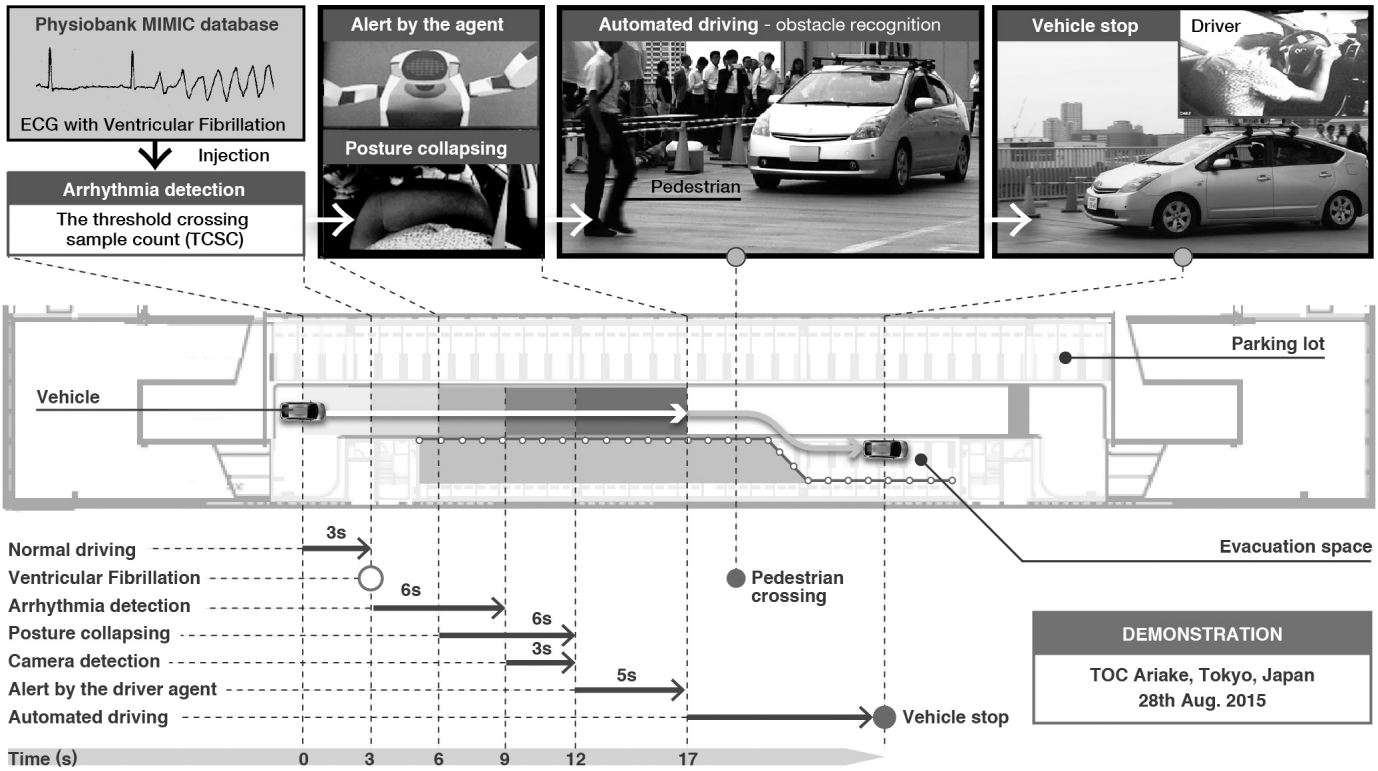


Figure 7. Scenario of the demonstration conducted at TOC Ariake, Tokyo, Japan in 28th August 2015

Normal ECG data was collected by the sensor, then the ECG with ventricular fibrillation (VF), obtained from PhysioBank MIMIC database [23], was used for simulation of the emergency situation.

The threshold crossing sample count (TCSC) [24] was implemented for the VF detection. Considering the safeness of the demonstration, the steer-sensor and the driver-monitoring camera were settled on the passenger seat as shown on the Figure 5. The assistant driver on the driver seat will take countermeasures in case of system errors during the transition of manual-to-autonomous driving and while automated driving. Four cameras were settled to monitor the conditions of inside the vehicle: two for capturing driver's face and posture, one for driver agent and one for driver's bio-signals.

The captured images (Figure 6) were sent to the monitor placed in front of the audience wirelessly. The scenario of the demonstration is illustrated in the Figure 8. It started from normal driving with the speed of 20 km/h, it took about 3 seconds to let the vehicle reached the designated velocity and all systems worked in normal state. Soon after the simulated VF was injected, the arrhythmia detection and cameras worked simultaneously to catch the emergency situation. The posture collapsing of the driver were detected by the face recognition. In the meantime, the agent gave an alert to the driver to confirm the situation. Then the system automatically turned on the automated driving when it confirms the driver's critical state. In the scenario, a pedestrian crossed in front of the vehicle while the system turned into the automated mode. It successfully recognized the pedestrian and stopped while he was crossing the road. Soon after it confirmed no obstacles ahead; it drove to the safety spot (Evacuation space on the Figure 8) and stopped the vehicle. All the procedures after the VF injection were done automatically.

4. Discussion

European automobile manufacturers have developed the related system in recent years [25]. Their system prevents frontal crash and lane departures by an adaptive cruise control and a lane-keep assist when a driver lost his capacity to drive by a physical disorder. It also alerts the driver by an indicator and warning brakes, and evacuates the vehicle with hazard lumps when there is no response from the driver. The similar demonstration has shown at the 22nd ITS world congress, Bordeaux, France, in October 2015 [26]. Driver's incapacity was detected by a camera and automatically stopped the vehicle on the shoulder of a road at the demonstration. Although the related system is well designed for the countermeasures for incapacitated drivers, the future studies require detection methodologies of the driver's incapacity at earlier stage. Even a few seconds of beforehand detection decreases the risk of crash. Since the bio-signals indicate the abnormality of the driver before he/she gets incapacity, our system has a potential for minimizing the crash risk. The prototype of the system using bio-signal detection was successfully demonstrated its function at the prepared scenario.

The demonstration, however, also revealed its limitations and problems. Tests with conventional ECG sensor has showed that the driving condition actively adds noise to the signal. Moving artifacts due to body motion were sometimes detected as severe arrhythmias, resulting in misdetections. Since the steer-sensor generates more noises on the ECG than conventional sensors, that uses adhesive electrodes, the waveform of the PQRST complex easily gets distorted. Therefore, it requires additional methodologies on the conventional arrhythmia detection. Moreover, gripping the wheel by both hands all the time is not practical in the real driving scenes, so that it requires

accompanying sensors such as capacitive ECG sensor or wearable PPG sensor to compensate the signal distortions or the signal losses or combine with other technologies to compensate signal when the driver drives with one hand. As shown in the Figure 7, the duration of uncontrolled vehicle is 14 seconds and the initial velocity when the arrhythmia started was 20 km/h (5.6 m/s). Thus, the distance the vehicle kept moving is 78 meters. In unconsciousness state, the driver can step on the gas pedal or steer the vehicle out of the lane. Both of above situations will eventually lead to accident. It is needed to shorten the detection time of both arrhythmia and posture collapse. Normally, if the driver feels any discomfort while driving he/she will apply brake and slow down. But others might suffer more serious situation and collapse before the accident happen. In that case, it is important to develop methodology to predict arrhythmia before it happen. In addition, it is not always safe to stop the vehicle at a certain spot. Further investigation of coping behavior in the emergency situation is essential for the development.

5. Conclusion

The study demonstrated the concept of the automated safety vehicle stop system for cardiac emergencies. The prototype system has shown potential benefits to both patients and normal healthy people. But the system contains many limitations requiring their solutions: the sensor quality and investigation of emergency management on the various driving situations. Moreover, detecting driving incapacity requires not only bio-signals but also other risk factors such as driving performance, driver's state, driver's medical history etc. Currently, we are focusing on the effects of the driving situations, the effects of physical stress and cognitive stress on CVD patients and healthy subjects. These topics are dedicated to our further studies.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The authors would like to acknowledge to Sadayoshi Tateishi, Prof. Nishiyama Kiyohisa in Nagoya University, Minoru Makiguchi in Toyota Motor Corporation, and Tsuyoshi Nakagawa with Denso. This research is in part supported by the Center of Innovation Program (Nagoya COI) from Japan Science and Technology Agency and Toyota Motor Corporation.

References

[1] M. T. Schultheis, J. DeLuca, and D. Chute, *Handbook for the assessment of driving capacity*. Academic Press, 2011.

[2] I. West, G. L. Nielsen, A. E. Gilmore, and J. R. Ryan, "Natural Death at the Wheel," *JAMA J. Am. Med. Assoc.*, vol. 205, no. 5, p. 266, Jul. 1968.

[3] M. Osawa *et al.*, "Sudden natural death in driving: case studies in the western area of Kanagawa.," *Nihon Hoigaku Zasshi*, vol. 52, no. 5, pp. 315–8, Oct. 1998.

[4] A. Oliva *et al.*, "Autopsy investigation and Bayesian approach to coronary artery disease in victims of motor-vehicle accidents," *Atherosclerosis*, vol. 218, no. 1, pp. 28–32, 2011.

[5] T. Tervo *et al.*, "Sudden Death at the Wheel Due to a Disease Attack Sudden Death at the Wheel Due to a Disease Attack," *Traffic Inj. Prev.*, vol. 14, no. 14, pp. 138–144, 2013.

[6] Y. Motozawa, J. M. Hitosugi, and S. Tokudome, "Analysis of Sudden natural deaths while driving with forensic autopsy findings."

[7] P. Schmidt, K. Haarhoff, and W. Bonte, "Sudden natural death at the wheel - A particular problem of the elderly?," *Forensic Sci. Int.*, vol. 48, no. 2, pp. 155–162, 1990.

[8] S. P. Baker and W. U. Spitz, "An evaluation of the hazard created by natural death at the wheel," *N. Engl. J. Med.*, vol. 283, no. 8, pp. 405–409, 1970.

[9] T. Tervo, E. Räty, P. Sulander, J. M. Holopainen, T. Jaakkola, and K. Parkkari, "Sudden Death at the Wheel Due to a Disease Attack," *Traffic Inj. Prev.*, vol. 14, no. 2, pp. 138–144, Jan. 2013.

[10] C. N. O. Gma, T. Vaa, T. Economics, J. Alvarez, and B. Hockey, *Impairments, Diseases, Age and their Relative Risks of Accident Involvement: results from meta-analysis, IMMORTAL*, no. August. 2003.

[11] J. A. Waller, "Chronic medical conditions and traffic safety: review of the California experience," *N. Engl. J. Med.*, vol. 273, no. 26, pp. 1413–1420, 1965.

[12] J. A. Waller, "Cardiovascular disease, aging, and traffic accidents," *J. Chronic Dis.*, vol. 20, no. 8, pp. 615–620, 1967.

[13] a. Crancer and L. McMurray, "Accident and Violation Rates of Washington's Medically Restricted Drivers," *JAMA J. Am. Med. Assoc.*, vol. 205, no. 5, pp. 272–276, 1968.

[14] Y. Lerman, M. Matar, B. Lavie, and Y. L. Danon, "Effect of Valvular Heart Diseases, Migraine Headaches, and Perianal Diseases on the Risk of Involvement in Motor Vehicle Crashes," *J. Trauma Inj. Infect. Crit. Care*, vol. 39, no. 6, pp. 1058–1062, 1995.

[15] D. G. K. Theodoros A. Zografos, "Review Article Guidelines and Regulations for Driving in Heart Disease," *Hell. J. Cardiol.*, vol. 51, pp. 226–234, 2010.

[16] A. E. Epstein *et al.*, "Personal and Public Safety Issues Related to Arrhythmias That May Affect Consciousness : Implications for Regulation and Physician Recommendations," *Circulation*, vol. 94, pp. 1147–1166, 1996.

[17] M. C. Petch, "Driving and heart disease, Prepared on behalf of the Task Force by M. C. Petch," *Eur. Heart J.*, vol. 19, no. 8, pp. 1165–1177, Aug. 1998.

[18] A. R. Copeland, "Sudden Natural Death 'At the Wheel'—Revisited," *Med. Sci. Law*, vol. 27, no. 2, pp. 106–113, Apr. 1987.

[19] K. Futatsuyama, N. Mitsumoto, T. Kawachi, and T. Nakagawa, "Noise Robust Optical Sensor for Driver's Vital Signs," in *SAE Technical Paper*, 2011, p. No. 2011-01-1024.

[20] M. Osaka, H. Murata, Y. Fuwamoto, S. Nanba, K. Sakai, and T. Katoh, "Application of heart rate variability analysis to electrocardiogram recorded outside the driver's awareness from an automobile steering wheel," *Circ. J.*, vol. 72, no. 11, pp. 1867–1873, 2008.

[21] T. Tanaka *et al.*, "Driver Agent for Supporting Elderly Driver," in *Human-Agent Interaction Symposium 2015*, 2015, pp. 115–118.

[22] S. Kato, E. Takeuchi, Y. Ishiguro, Y. Ninomiya, K. Takeda, and T. Hamada, "An Open Approach to Autonomous Vehicles," *IEEE Micro*, vol. 35, no. 6, pp. 60–68, 2015.

[23] G. B. Moody and R. G. Mark, "A database to support development and evaluation of intelligent intensive care monitoring," *Comput. Cardiol.*, vol. 23, pp. 657–660, 1996.

[24] M. A. Arafat, A. W. Chowdhury, and M. K. Hasan, "A simple time domain algorithm for the detection of ventricular fibrillation in electrocardiogram," *Signal, Image Video Process.*, vol. 5, no. 1, pp. 1–10, 2011.

[25] Volkswagen, "Emergency Assist," 2016. .

[26] AISIN SEIKI Co.Ltd., "Aisin Group to exhibit at 22nd ITS World Congress Bordeaux 2015," 2015. .

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 674-677 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Computer Aided Classification using Support Vector Machines in Detecting Cysts of Jaws

Veena Divya K[1*], Anand Jatti[1], Revan Joshi P[2], Sabah Meharaj[1]

[1]*Department of Electronics and Instrumentation Engineering, RVCE, Bengaluru, 560059, INDIA*

[2]*Department of Radiology, D.A.Pandu Memeorial R.V.Dental College and Hospital, Bengaluru, 560078, INDIA*

ARTICLE INFO

Article history:
Received: 09 April, 2017
Accepted: 11 May, 2017
Online: 05 June, 2017

Keywords:
Oral cyst
Gray level co-occurrence matrix
Support vector machines

ABSTRACT

Jaw cyst is one of the most common pathology observed in the field of dentistry. Early detection of the cystic lesion will help the surgeons to take appropriate therapeutic measures after a thorough diagnostic procedure. One of the challenging task for surgeons is to differentiate the cysts from the other pathologies. The appearance of these pathologies on a radiograph is a complex and confusing task due to the close similarity between the cysts and tumors which is a difficult to differentiate just by its appearance. Hence to resolve this problem, a computer aided classification algorithm is needed for accurate classification of cysts. The work presents a new approach for the determination of the presence or severity of the jaw bone disease aiding the diagnosis and radiotherapy planning. This paper presents texture characterization for the dental panoramic images. The transposed images are analyzed using Gray level co-occurrence matrix(GLCM). The textural properties such as entropy, contrast, correlation, energy and homogeneity are determined for both cyst and non-cystic images. The results obtained are fed to the classification model to classify the given image into normal or abnormal images containing cyst. Support vector machines are chosen for image classification. Image dataset of 30 were used in training and validation. The image set consists of 20 abnormal images and 10 normal images used in image classification.

1. Introduction

Digital dental x-rays are the most preliminary step in imaging various dental anomalies. Dental x-rays also referred as orthopantomograph is the radiographic examination of jaws and the supporting structure across the teeths [12]. Using this equipment, the area of interest in the jaw are scanned in few seconds, often happen to be distorted by the noise and have inhomogeneous distribution, due to this, few regions are blurred making it difficult to interpret certain features from the raw images. Orthopantomography being a two dimensional imaging modality is riddled with problem of superimposition of anatomical structures. In this approach, the limitations are overcome by using specific image processing algorithm to retrieve more information from the original image.

A cyst in the jaws occurs most frequently than any other part of the body. The most frequent manifestation of jaw cysts is swelling followed by the combination of swelling and pain. The swelling associated with the cysts causes facial deformation adversely affecting the patient's psycho-social status. A definitive diagnosis can be made by age gender and prevalence of cyst. If the cysts are not detected at an early stage, it may lead to tumors**.** Jawbone cysts are classified based on the odontogenic or nonodontogenic, epithelial or non-epithelial, inflammatory or developmental in origin. The distribution of jaw cysts in the ordinary population according to the diagnosis are radicular cysts, dentigerous cysts, odontogenic keratocyst(OKC), eruption cysts, nasopalatine duct and traumatic bone cysts to name a few. Diverse cysts or cysts like lesions appear in maxillofacial region. According to the study, radicular cysts (RC) and dentigerous cysts(DC) are the commonest of all the cysts [1][2].

*Corresponding Author: Veena Divya K, Department of Electronics and Instrumentation Engineering, RVCE, Bengaluru, 560059, INDIA
Email: veenadivya@gmail.com

2. Image Processing Chain

Image processing chain consists of preprocessing, segmentation, feature extraction and classification [9]. First the dental x-ray images are preprocessed based on image enhancement technique namely contrast stretching. Next, it involves segmentation of cysts and lesion using level sets method. In addition to this, certain features are extracted based on texture statistical distribution using gray level co-occurrence matrix (GLCM). In this step the texture properties such as energy correlation, contrast, entropy and homogeneity are analyzed and given as an input for classification. Image classification is made using support vector machines (SVM) to classify the presence or absence of cysts for the given image set. Figure 1 show the block diagram of the image processing system.

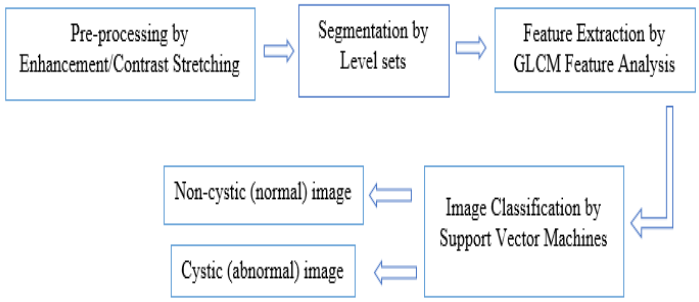


Figure 1: Block diagram for the classification of Cysts.

2.1. Preprocessing

Pre-processing represents the lowest level of abstraction at the images where both input and output are gray scale intensity images. Pre-processing deliberately reduces the information content of the image [13]. If the entropy of the images is measured, the preprocessed image typically has low information content compare to the original image. Therefore, the aim of preprocessing is to reduce the noise and improve the quality of image highlighting few important features which is necessary for further processing in segmentation and recognition tasks [4]. Image pre-processing is classified into two categories such as image enhancement and image restoration techniques. In this paper the raw dental images are scanned by digital dental x-rays are preprocessed based on one of the enhancement techniques called contrast stretching. Contrasting the image, possibly increase the brightness so that the transformed image can be better interpreted. Usually digital dental x-rays images have limited band of gray levels. So the gray level transformation corrects the brightness by setting up the two image contrast between two brightness values p1 and p2. Using the proper brightness threshold, the contrast of the image can be improved and significantly recognizing low level features which is distorted due to poor illuminance properties [3].

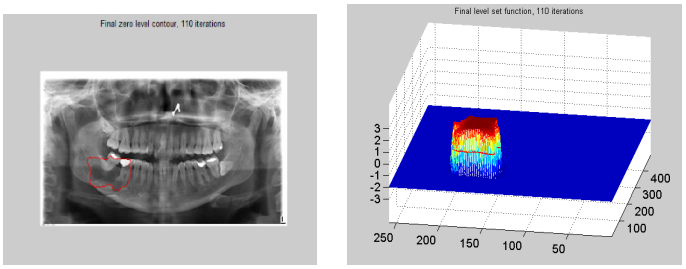
2.2. Segmentation by Level Sets

The main idea behind this level sets segmentation is to represent the initial contour of the object as the zero level set function which is defined by the level set function defined in the higher dimensions [12]. The level sets function is evolved using partial differential equations (PDF) [10]. This method presents many advantages in which the contour described by the extent level set function has the flexibleness to interrupt and merge naturally during the evolution. [5].

The level set method is a curve evolution technique for detecting interfaces and shapes in the images. In level set method initially the contours are set as zero level set of higher dimensional function called as level set function. firstly, geometrical shaped contour and iterations are defined and after given number of iterations the contour spreads over the region of interest for the object boundaries to be extracted. [5].

The level set evolution function is defined as the gradient flow minimizing energy functional based on the term distance regularization and basically it has energy functional that tends to move the zero level set function to obtain the desired object boundaries and its location. This level set function(LSF) created with the signed distance profile near the zero-level set function give rise to the new technique called as distance regularized level set function(DRLSE). The numerical errors accounted induced during computation of the level sets are reduced with the effect of distance regularization term.

In traditional level sets methods, level sets functions create instability and causes numerical errors with the irregularities developed during its evolution. Hence to avoid such errors, a mathematical approach called initialization is used [6]. Figure 2 illustrates the segmentation of cystic region carried out by the level sets algorithm for 110 number of iterations.



(a) (b)

Figure 2: (a)Segmentation by level sets for extraction region of cysts from the entire image (b)contour showing final level sets function.

2.3. Texture Analysis

Feature extraction is a method to extract the desired features and representing the data for further processing. This plays a crucial part in machine vision in understanding and describing the objects extracted after segmentation [11]. The resulting representation can be used as an input for classification techniques to classify the medical data into normal or abnormal. In this step, the features pertaining to the dental images are extracted using statistical texture analysis. Texture is represented as the spatial distribution of gray levels in the neighborhood pixels. Each pixel is characterized by its position and tonal properties referred as texture elements or texels. Thus image texture is described by the spatial relationship and its primitives. Texture constitute the repeated pattern of local deviations in the image intensity finely distinguishing the objects subjectively at the given resolution. Hence the general rule is to group this connected or similar set of pixels occurring repeatedly in the image region satisfying certain the gray level properties is assigned as textured region. Typically,

texture deals with the properties representing the surface or shape of an object. In this phase the dental x-ray images are analyzed based on one of the statistical texture description using co-occurrence matrix called gray level co-occurrence matrix(GLCM) [3].

Texture is spatial quantity which is represented by the two dimensional matrix known as gray level co-occurrence matrix. The occurrence matrix gives the interdependencies of the gray level intensities to extensively analyze texture [3].

The gray level co-occurrence matrix P[i, j] is defined as how frequently the pixels with the gray level i and j are occurring in the window separated by the distance vector[6]. Distinct features are extracted by GLCM are entropy, energy, contrast correlation and homogeneity for the dental panoramic images. This GLCM estimates the similarity measures among these co-occurrence matrices and five features are selected for this course of study which are discussed in this paper.

Entropy defines the randomness present in an image. It is a measure of disorder of an image. It is inversely proportional to the energy of GLCM. If the image is inhomogeneous, the GLCM matrix P will have small values. Hence entropy will be low if energy is high in the GLCM matrix. It is difficult to understand subjective wise as mathematically it given by

$$E_{entropy} = \sum_i \sum_j p(i,j) \log (p(i,j)) \qquad (1)[8]$$

Energy cost for the uniformity of the image representing second angular moment. A monochrome or homogeneous image usually has less gray level distributions hence a co-occurrence matrix, P will predominantly have high values for energy feature. In this context, Energy is, the opposite of entropy. Image exhibits its distinct energy forces acting internally and externally on the images[8]. In this sense, Energy is useful in representing orderliness. That is the reason for finding Energy in texture measures.

$$E_{energy} = \sum_i \sum_j p(i,j)^2 \qquad (2)[8]$$

Contrast is accounted for difference moment of the co-occurrence matrix P representing the local variations and brightness distribution in the image.

$$E_{contrast} = \sum_{n=0}^{N_g-1} n^2 \{\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j)\} | i, j | \ (3)[8]$$

Homogeneity represents inverse difference moment. This term attains the highest values when most of the GLCM elements are concentrated towards the main diagonal of the GLCM matrix P. Hence it is inversely proportional to the GLCM contrast.

$$E_{homogeneity} = \sum_i \sum_j \frac{1}{1=(i,j)^2} p(i,j) \qquad (4)[8]$$

The correlation feature is a measure of gray-level linear dependencies in the image [7].

$$E_{correlation} = \sum_i \sum_j \frac{(i,j)p(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y} \qquad (5)[8]$$

Table 1: Feature extracted for cystic images

Features /Images	Image1	Image2	Image3	Image4	Image5
Energy	0.0872	0.783	0.1	0.087	0.0987
Entropy	7.7096	7.5304	7.6366	7.7872	7.6889
Contrast	0.4477	0.4529	0.5305	0.5208	0.4895
Correlation	0.9466	0.9267	0.9208	0.932	0.9161
Homogeneity	0.8693	0.8824	0.8688	0.8581	0.8513

Table 1: Feature extracted for non-cystic images

Features /Images	Image 1	Image2	Image3	Image4	Image5
Energy	0.1173	0.1046	0.1119	0.0999	0.0991
Entropy	7.5469	7.4637	7.4786	7.6337	7.6255
Contrast	0.53	0.555	0.5686	0.5572	0.5833
Correlation	0.9216	0.9249	0.9174	0.9342	0.9307
Homogeneity	0.8729	0.8585	0.8614	0.8614	0.8601

2.4. Image Classification

Image classification is one of the application of medical image processing and has a great significance in the field of pattern recognition. Most of the times, the jaw bone cysts and lesions have close similarity which is hard to differentiate just viewing the images where the patient is allowed to undergo deferential diagnosis. Hence this common problem can be taken as classification problem where the main purpose of classifier is to differentiate the normal and abnormal pixels based on the texture features. Support vector machines is one of the most successful method proved for pixel classification

Support vector classification is one of the advance approach used in statistical recognition which is computationally faster than other classifiers. SVM designs a model which group the pixels into normal and abnormal pixels yielding binary classification. In this approach, the optimal classification can be obtained by increasing the margin or the empty space between the two classes. The width of the margin is defined as the distance between the discrimination hypersurface in n-dimensional feature space and the closest training patterns which are known as support vectors. Thus support vectors specify the discrimination function in hyperplane. Earlier SVM is employed for dealing only linear classification yielding two separate classes. Later this limitation was overcome by supporting non linearity classification where combing two separate classes yielding multi class classification [3].

Support vector classification involves designing a model for two separable classes. Hence training is required for the classifier where the kernel function is chosen appropriately to yield better performance of the classifier [7]. A large amount of training data set is required for the learning algorithm to obtain high accuracy. Figure 3 shows the basic steps involved in generic classification algorithm.

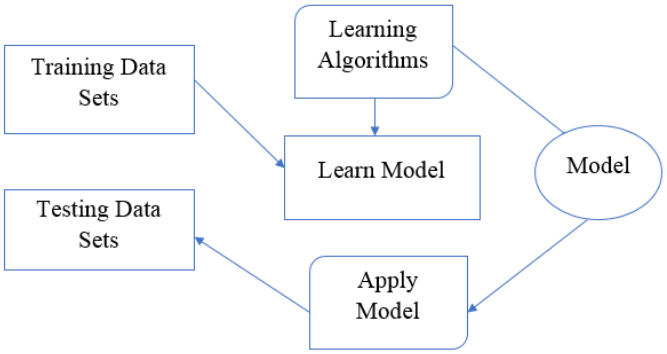


Figure 3: Model for building Generalized Classification

3. Conclusion

Dental x-ray images exhibit complex nature with respect to distribution of pixel intensities. This paper presents a new algorithm for determination of texture using on co-occurrence matrix(GLCM). The selection of feature plays a vital role in any classification. Five distinct features are extracted using GLCM namely energy, contrast, entropy, correlation and homogeneity. These features are given to the classifier to classify into presence or absence of cysts for the given dental image. The classification is made by support vector machines. This yield a binary classification which group the given image set into cystic(abnormal) image or non-cystic(normal) image. Totally 30 images were used for training and testing the classifier, in which 20 images are abnormal (or cystic images) and 10 (non-cystic) images are normal. The proposed classification algorithm accounted for 68% of accuracy. Hence to improve the accuracy large dataset are used to train the network which constitute the future scope of this proposed work.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The authors of this paper would like to thank R.V.College of Engineering, Bengaluru for supporting the research work and D.A.Pandu Memorial R.V.Dental College & Hospitals, Bengaluru for their constant support towards the research work.

References

[1] Johnson, Nigel R., Orla M. Gannon, Neil W. Savage, and Martin D. Batstone, “Frequency of odontogenic cysts and tumors: a systematic review” Journal of investigative and clinical dentistry 5, no. 1, pp. 9-14,2014.

[2] Cyrus S, Micha, Chindia Mark L, Awange David O, Dimba Elizabeth, Osundwa Tom M, and Otwoma Josiah G, “Pattern of occurrence of jaw cysts and cyst-like lesions at the University of Nairobi Dental Hospital: A 10-year histopathologic audit” Open Journal of Stomatology 2012.

[3] Sonka, Milan, Vaclav Hlavac, and Roger Boyle. Image processing, analysis, and machine vision. Cengage Learning, 2014.

[4] Divya, K. Veena, Anand Jatti, P. Sabah Meharaj, and Revan Joshi. "Appending active contour model on digital panoramic dental X-rays images for segmentation of maxillofacial region." In Conference on Biomedical Engineering and Sciences (IECBES), pp. 450-453, IEEE EMBS, 2016.

[5] Choong, Mei Yeen, May Chin Seng, Aroland Kiring, Soo Siang Yang, and Kenneth Tze Kin Teo. "Variational Level Set Algorithm in Image Segmentation for Foetus Ultrasound Imaging System", IEEE International conference, 2012.

[6] Li, Chunming, Chenyang Xu, Changfeng Gui, and Martin D. Fox, “Distance regularized level set evolution and its application to image segmentation”, IEEE transactions on image processing 19, no. 12 pp. 3243-3254, 2010.

[7] Partio, Mari, Bogdan Cramariuc, Moncef Gabbouj, and Ari Visa. "Rock texture retrieval using gray level co-occurrence matrix." In Proc. of 5th Nordic Signal Processing Symposium, vol. 75. 2002.

[8] Abdolvahab Ehsani, Mohd Shafry Mohd Rahim, and Alireza Norouzi. "Digital dental X-ray image segmentation and feature extraction." Indonesian Journal of Electrical Engineering and Computer Science 11, no. 6, pp. 3109-3114,2013.

[9] Ayachi, Raouia, and Nahla Ben Amor. "Brain tumor segmentation using support vector machines" European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty, pp. 736-747. Springer Berlin Heidelberg, 2009.

[10] Nurtanio, Ingrid, “Classifying Cyst and Tumor Lesion Using Support Vector Machine Based on Dental Panoramic Images Texture Features” IAENG International Journal of Computer Science,vol no .40:1:4 , 2013.

[11] Vijayakumari, B., G. Ulaganathan, A. Banumathi, A. Farzana Shahar Banu, and M. Kayalvizhi. "Dental cyst diagnosis using texture analysis." In Machine Vision and Image Processing (MVIP), 2012 International Conference on, pp. 117-120. IEEE, 2012.

[12] Veena Divya.K, Jatti, Anand, and Revan Joshi, "Image processing and parameter extraction of digital panoramic dental X-rays with ImageJ." Innternational Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS), International Conference on, pp. 450-454. IEEE, 2016.

[13] R. Gonzalez and R. Woods, Digital Image Processing, Prentice Hall, 3rd edition, Jan. 2008.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 678-686 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Spatiotemporal Traffic State Prediction Based on Discriminatively Pre-trained Deep Neural Networks

Mohammed Elhenawy[1], Hesham Rakha [2,*]

[1]Postdoctoral associate with the Center for Sustainable Mobility, Virginia Tech Transportation Institute, VA, 24061, USA

[2]Director of the Center for Sustainable Mobility (CSM) at the Virginia Tech Transportation Institute (VTTI), USA

ARTICLE INFO

Article history:
Received: 13 April 2017
Accepted: 12 May, 2017
Online: 05 June, 2017

Keywords:
Traffic state prediction
Proactive decision support systems
Deep Neural Network

ABSTRACT

The availability of traffic data and computational advances now make it possible to build data-driven models that capture the evolution of the state of traffic along modeled stretches of road. These models are used for short-time prediction so that transportation facilities can be operated in an efficient way that guarantees a high level of service. In this paper, we adopted a state-of-the-art machine learning deep neural network and the divide-and-conquer approach to model large road stretches. The proposed approach is expected to be a tool used in daily routines to enhance proactive decision support systems. The proposed approach maintains spatiotemporal correlations between contiguous road segments and is suitable for practical applications because it divides the large prediction problem into a set of smaller overlapping problems. These smaller problems can be solved in a reasonable time using a medium configuration PC. The proposed approach was used to model 21.1- and 30.7-mile stretches of highway along I-15 and I-66, respectively. The resulting predictions were better than predictions obtained using partial least squares regression.

1. Introduction

The growing economy and an increasing number of vehicles on the road have brought about traffic problems that increase travelers' inconvenience. Traffic congestion features prominently among these problems, and has become an everyday concern in many urban areas, bringing with it negative environmental effects. During periods of congestion, cars cannot run efficiently, resulting in air pollution, carbon dioxide (CO_2) emissions, and increased fuel use. In 2007, wasted fuel and lost productivity cost Americans $87.2 billion. This number reached $115 billion in 2009 [1]. Congestion also increases travel time. In 1993, driving under congested conditions caused a delay of about 1.2 minutes per kilometer of travel on arterial roads [2]. Since then, this problem has grown worse, as evidenced by a report from the Texas Transportation Institute that indicates that the number of wasted hours due to traffic congestion increased fivefold between 1982 and 2005 for Americans.

In the past, the solution to these problems was constructing new roads, but that era of build-out ended due to its social and economic effects. Subsequently, Transportation Management Centers (TMCs) were deployed to solve traffic problems. Most of these TMCs operate in reactive mode, applying management strategies only after conditions in their regions change. Today, due to advances in technology, applying predictive traffic control strategies offers a cost-effective solution to improving traffic flow. Predictive traffic control is a proactive approach where traffic problems, such as congestion, are anticipated and countermeasures are applied prior to their occurrence.

Data-driven, short-term prediction of traffic characteristics, such as flow, density, and speed, is essential for applying predictive traffic control strategies. However, because of unstable traffic conditions and complex road settings, short-term prediction is not a straightforward task [3]. Applying mathematical models that are based on macroscopic and microscopic theories of traffic flow is difficult so data-driven modeling is considered a good approach to model complex traffic characteristics. Thanks to computational advances that make data collection and processing easy and rapid, a promising research area based on data-driven algorithms has emerged.

* Corresponding author: Hesham Rakha, phone: 540-231-1505,
E-mail: hrakha@vt.edu

The use of deep learning neural networks for traffic state prediction was demonstrated to be promising by Elhenawy and Rakha [4]. In this paper, we describe new research on the prediction of traffic speeds and flows for basic freeway segments using a data-driven approach. Predictions are based on the current traffic state data, weather conditions, visibility levels, and seasonal predictors (i.e., day of the week and time of the day). State-of-the-art deep neural networks were trained and tested to predict traffic states for a larger spatial and temporal horizon. The proposed approach allows for identification of traffic problems up to 2 hours in advance, and implementation of solutions up to several miles away. In order to make the proposed approach suitable for practical applications at TMCs, a principal component analysis (PCA) was used to reduce the dimensionality of the inputs to the deep neural network. Moreover, we used the divide-and-conquer approach to divide the large prediction problem into a set of smaller overlapping problems.

Following the introduction, in Section 2 related work is discussed. Section 3 defines the short-term prediction problem. Section 4 briefly presents a background of the methods used in this paper. The model calibration is described in Section 5. The study sites and the preprocessing of the traffic state data are discussed in Section 6. In Section 7, the results of the experimental work are outlined. Finally, the paper's conclusions and recommendations for future work are described in Section 8.

2. Related Work

Over the past few decades, traffic characteristic prediction models have been developed by researchers from different areas, such as transportation engineering, control engineering, and economics. The developed prediction approaches can be classified into three broad categories: parametric models, nonparametric models, and simulations. Parametric models include time-series models, stochastic differential equations, etc. Nonparametric models include a variety of techniques that range from simple methods like k-nearest neighbor (k-NN) to complex support vector regressions (SVR) and artificial neural networks (ANNs).

The time-series technique is a parametric model used widely in traffic flow prediction. The autoregressive integrated moving average (ARIMA) model was used very early to predict short-term freeway traffic flow [5]. After that, different advanced versions of ARIMA were used to develop more-accurate prediction models. Voort et al. integrated the Kohonen self-organizing map and ARIMA into a new method called KARIMA, which uses a Kohonen self-organizing map to cluster the data and then models each cluster using ARIMA [6]. Lee et al. used a subset ARIMA model for the one-step-ahead forecasting task, which returned more stable and accurate results than the full ARIMA model [7]. Williams and Hoel used seasonal ARIMA (SARIMA) to analyze data from two freeways, and the results showed that one-step seasonal ARIMA predictions outperformed heuristic forecast benchmarks [8].

Due to both the highly nonlinear nature of traffic characteristics and the availability of data, nonparametric methods have also attracted researchers' attention. For traffic flow prediction, there are many versions of the k-NN algorithm that show a good prediction accuracy. Davis and Nihan argued that k-NN can capture linear and nonlinear relationships and is therefore able to model the nonlinear transition between free-flow and congested traffic [9]. However, the results of their empirical study showed that k-NN was no better than a simple univariate time-series forecast. Sun et al. considered the traffic prediction model as a nonlinear system with historical and current traffic characteristics as inputs and future traffic characteristics as outputs [10]. They used the local linear regression model to approximate the nonlinear relationship between system inputs and outputs to predict future traffic characteristics. Young-Seon et al. proposed a short-term traffic flow prediction algorithm combining the online-based SVR with the weighted learning method for short-term traffic flow predictions [1]. ANN is considered one of the best tools to model highly nonlinear relationships between inputs and outputs, and many papers have adopted different ANN models, such as the Bayesian neural network [11] and radial basis function neural network [12], for predicting traffic flow. For more information, Vlahogianni et al. [13] provide a good review of the proposed techniques as well as the challenges of short-term prediction.

Recently, a number of algorithms have been developed to train neural networks with many hidden layers. These deep learning neural networks are state-of-the-art machine learning techniques used to solve complex problems. However, this technique requires a huge amount of data for the training to return good results. In recent years, deep neural networks have attracted the attention of researchers in the transportation field as a super-tool to model traffic evolution. Huang et al. used a deep belief network to learn the important features for flow prediction and then passed those features to the regression output layer [14]. In another paper, the problem of traffic forecasting at peak hour and after an accident is approached using a generic deep learning framework based on long short term memory units [15]. Polson and Sokolov used a deep neural network to predict traffic flow, demonstrating that the deep neural network is capable of giving precise short-term prediction at the sharp traffic flow regime [16]. Lv et al. used a stacked autoencoder model to learn features that capture the nonlinear spatial and temporal correlations from the traffic data, then forwarded these features to the output layer to predict the traffic flow [17]. In an another recent study by Ma et al., spatiotemporal speed matrices were considered as images and used to train a convolution neural network, and then this network was used to predict large-scale, network-wide traffic speed with high accuracy [18].

3. Problem Definition

In this research, we are interested in predicting the traffic states of several segments of a freeway for up to a 2-hour horizon. This is not an easy task, since the evolution of traffic states is a complex spatiotemporal process. In order to define our prediction problem, we first define the spatiotemporal speed/flow matrix, S_{IxT}, where I is the number of the stretch's segments and T is the day's time intervals. The spatiotemporal traffic state prediction problem can be stated as follows. Let s_i^t be the observed elements of the spatiotemporal state matrix at the time interval $t = 1, 2, \dots, t_0$ and segment $i = 1, 2, \dots, h$ of the studied stretch of road. Given the spatiotemporal observed traffic state submatrix that ends at time t_0, the forecasted weather conditions, and the visibility level, our goal is to predict the spatiotemporal traffic state submatrix that spans the time interval $[t_0 + 1,\ t_0 + \Delta]$ for some prediction horizon Δ.

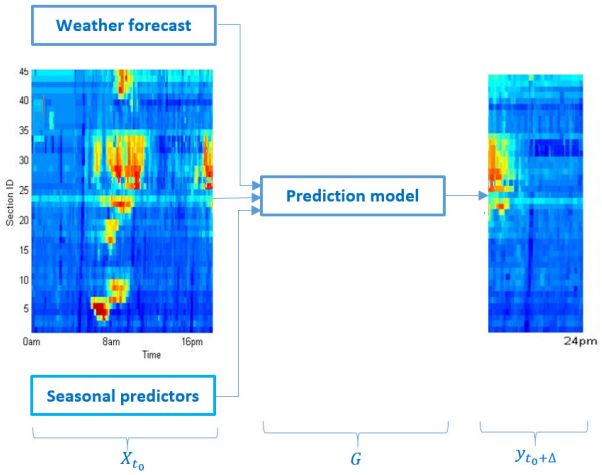


Figure 1. Illustration of the problem where a model is needed to predict the speed evolution in time (x-axis) and space (y-axis).

The general form of the solution for the problem above is shown in Equation 1 [19];

$$y_{t_0+\Delta} = G\left(X_{t_0}, e_{t_0}, \Theta\right) + e_{t_0+\Delta} \quad (1)$$

where:

G	the chosen model
$y_{t_0+\Delta}$	the prediction at horizon Δ
X_{t_0}	the inputs' predictors, including the observed elements of the spatiotemporal state submatrix at the time interval t_0, time of day, day of week, and the forecasted weather conditions and visibility
Θ	estimated model parameters
$e_{t_0+\Delta}$	errors (unexplained variability) due to the absence of factors we cannot observe

We propose using a deep neural network for this prediction problem for the following three reasons. First, the stochastic nature of the input-output data makes it possible to find two different responses for the same input. In other words, the response corresponding to any input predictor is a distribution rather than a single point in the response space. Second, the spatiotemporal response is multivariate and the input-output relationship is nonlinear. Third, there is no closed mathematical form (model) that can be used to explain the relationship between the input predictors and the response.

4. Methods

4.1. Discriminatively Pre-Trained Deep Neural Networks

In machine learning, ANNs are powerful tools used to estimate or approximate unknown linear and nonlinear functions, and are used for both classification and regression problems.

An ANN consists of an input layer, hidden layers, and an output layer. Each layer consists of several processing units, called neurons. In this paper, we used a multi-layered, feed-forward ANN, which is commonly used for classification/regression analysis. In a multi-layered, feed-forward ANN, the neurons make use of directed connections, which allow information flow in the direction from the input layer to output layer. A neuron k at layer m receives an input x_j from each neuron j at layer $m-1$. The neuron adds the weighted sum of its inputs to a bias term, then applies the whole thing to a transfer function and passes the result to its output toward the downstream layer. Given the training dataset, the ANN can use a learning algorithm such as back propagation to learn the weights and biases for each single neuron [20]. In general, training a deep neural network that has more than one hidden layer is challenging because the gradient vanishes as it propagates back. Over the past 10 years or so, many algorithms have been proposed to train deep neural networks, the discriminative pre-training method being one of the most simple [21]. The discriminative pre-training algorithm starts by training a neural network that has only one hidden layer. The input predictors and the responses are used in the supervised pre-training. Once the one-hidden-layer network has been trained, the algorithm inserts a new hidden layer before the output layer. The expanded whole network keeps the weights and biases of the input layer and the previously pre-trained hidden layer as initial weights for the next pre-training cycle. This algorithm continues growing the network until the desired depth is reached. Finally, the whole network is fine-tuned using a backpropagation algorithm.

4.2. Partial Least Squares Regression

Multiple linear regression (MLR) is a good tool to model the relationship between predictors and responses. MLR is effective when the number of predictors is small, there is no significant multicollinearity, and there is a well-understood relationship between predictors and responses [22]. In many scientific problems, the relationship between predictors and responses are poorly understood, and the main goal is to construct a good predictive model using a large number of predictors. In this case, MLR is not a suitable tool. If the number of predictors becomes too large, an MLR model will over-fit the sampled data perfectly but fail to predict new data well.

When the number of the observations is less than the number of predictors, the chance of multicollinearity increases and ordinary MLR fails. Several approaches have been proposed to overcome this problem. For instance, principal component regression is used to remove the multicollinearity by projecting X into the orthogonal component and then regressing X's scores on Y. These orthogonal components explain X but may not be relevant for Y.

Partial least squares regression (PLSR) is a more recent technique that generalizes and combines features from principal component analysis and MLR. It is used to predict Y from X and to describe their common structure. PLSR assumes that there are a few latent factors that account for most of the variation in the response. The general idea of PLSR is to try to extract those latent factors, accounting for as much of the predictors' X variation as possible, while at the same time modeling the responses well. A number of variants of PLSR algorithms exist, most of which estimate the coefficients of the linear regression between X and Y as $Y = X\tilde{\beta} + \widetilde{\beta_0}$. Some PLSR algorithms are designed for the case where Y is a column vector, while others are suitable for

the general case, in which Y is a matrix. One example of a simple PLSR algorithm is the nonlinear iterative partial least squares (NIPALS) algorithm [23].

4.3. Principal Component Analysis (PCA)

The stretch-wide prediction problem is a multivariate problem that may involve a considerable number of correlated predictors. PCA is a popular technique for dimensionality reduction that linearly transforms possibly correlated variables into uncorrelated variables called principal components.

PCA is usually used to reduce the number of predictors involved in the downstream analysis; however, the smaller set of transformed predictors still contains most of the information (variance) in the large set. The principal components are the Eigenvectors of the dataset covariance matrix. The following steps show how to find k principle components given N data vectors in the space R^n where $k \leq n$.

1. Normalize input data vectors.
2. Compute the covariance matrix S for the normalized dataset.
3. Compute the Eigenvectors and the associated Eigenvalue of the covariance matrix S.
4. Sort the Eigenvectors in order of decreasing Eigenvalues.
5. Reduce the size of the data by choosing the first k Eigenvectors.
6. Construct a good approximation of the original dataset by multiplying the transpose of the matrix consisting of the k Eigenvectors with the original data matrix.

The first principal component is the normalized Eigenvector, which is associated with the highest Eigenvalue. The first principal component represents the direction in the space that has the most variability in the data, and each succeeding component accounts for as much of the remaining variability as possible.

5. Model Calibration

5.1. Divide-and-Conquer Approach

One major challenge in traffic state short-term prediction using the spatiotemporal state matrix is the dimension of the predictor and response vectors and the large number of required parameters that must be estimated. As the stretch of road grows, the deep neural network needs a great deal of time for training or suffers from memory problems. To deal with these issues, this study adopts a divide-and-conquer approach model.

The divide-and-conquer paradigm suggests that if the problem cannot be solved as is, it should be decomposed into two or more small problems of the same type, which are then solved. The final solution to the difficult big problem is the combination of solutions to the smaller problems.

We applied divide and conquer in a straightforward manner to our prediction problem by dividing the input predictors of the spatial-temporal state matrix into smaller overlapping windows. The same method was applied to the responses, as shown in Figure 2.

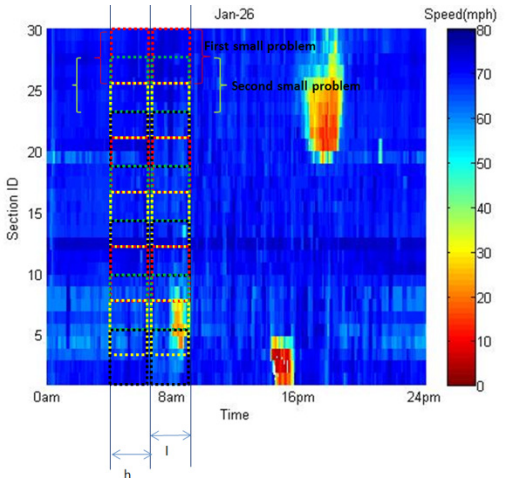


Figure 2. Illustration of divide-and-conquer approach, where l is the prediction horizon and h is the width of the predictors' windows.

The overlap of the windows is important for determining a smooth predicted speed. In the experimental work done in this paper, we set the window length equal to four segments and the overlap between contiguous windows equal to two segments. Because of this overlap between windows, each segment has two predicted traffic states at the testing phase, and the final predicted traffic state for overlapped segments is the average.

5.2. Training and Testing Phase

The model calibration process consists of a training phase and a testing phase. In the training phase, the deep neural network's weights are estimated using the training dataset. In the testing phase, the constructed deep neural network's accuracy is tested using an unseen dataset called the testing dataset. The training phase in our approach includes the following steps:

1. Partitioning (dividing) the whole stretch into small windows, each with four segments.
2. Preparing the X and Y matrices for each window by reshaping the traffic state, weather, and visibility inside the windows, with widths of h and l respectively.
3. Shifting the window to the right and repeating Step 2 to get another raw X and Y, as shown in Figure 3.
4. Using PCA to find $\bar{X}$, which is the scores of the matrix X.
5. Using $\bar{X}$ and Y matrices as the inputs and outputs, respectively, to train the one-hidden-layer neural network. Once the one-hidden layer network is trained, the algorithm discards the output layer and keeps the first hidden layer with its trained weights. Then, it stacks another hidden layer with a new output layer on top of the first layer and randomly initializes the weights of the new layers. The grown network is retrained again using the same $\bar{X}$ and Y matrices. This algorithm continues growing the network until the desired depth is reached. Finally, the whole network is fine-tuned using the backpropagation algorithm.

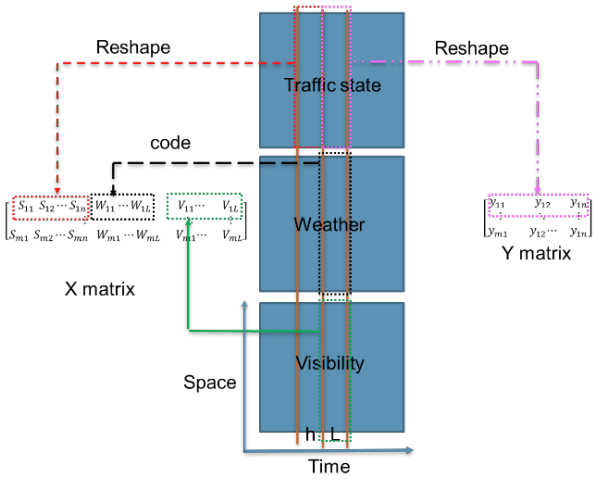


Figure 3. Illustration of preparing *X* and *Y* matrices for one window.

The testing phase is always simpler and does not require a great deal of time. This phase includes passing the testing data through the neural network after reducing it using the same principal components. The last step in testing is putting the predicted pieces together to find the prediction for the whole stretch.

6. Case Study

The performance of the proposed algorithm was tested on two real datasets collected from Interstate 66 (I-66) in Virginia and Interstate 15 (I-55) in California. This section describes the study sites and the preparation of the data.

6.1. Virginia Study Site

The selected I-66 eastbound study site is located between Gainesville and Arlington in Northern Virginia. There are 36 segments along the selected 30.7-mile stretch of freeway. This corridor has four lanes in each direction, with the leftmost lane designated as an HOV-2 lane during morning and afternoon peak hours. This is the major commuting corridor between Northern Virginia and Washington, DC. INRIX probe data from 2011 to 2013 were used to develop the prediction models, and data reduction was conducted to extract the daily traffic speed matrices. One-minute average speeds (or travel times) are available in the raw data for each roadway segment. Initially, the daily speed data were sorted by time and location from the raw data into a two-dimensional matrix. In order to reduce the stochastic noise and measurement error, the speed matrices were aggregated by 15-minute intervals. The missing data in the aggregated speed matrices were estimated using the moving average from a 3 × 3 window.

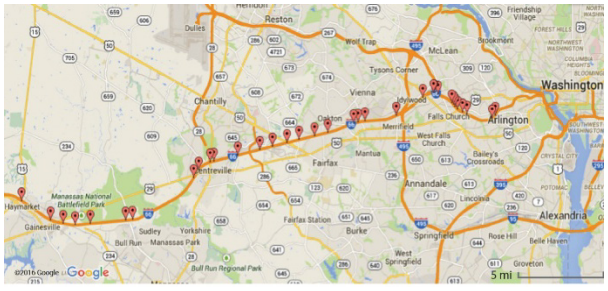


Figure 4. Virginia study site (I-66) (source: Google Maps).

6.2. California Study Site

The 2013 and 2014 loop detector data from I-15 southbound were used to develop traffic prediction models for the California study site. The layout of the I-15 freeway corridor and the locations of loop detectors are presented in Figure 4. There are 43 loop detectors along the 21.1-mile stretch of the corridor. Data reduction was conducted on the loop detector data collected from the test site to extract the daily traffic data matrices of speed and flow. The speed and flow matrices were 15-minute aggregated matrices.

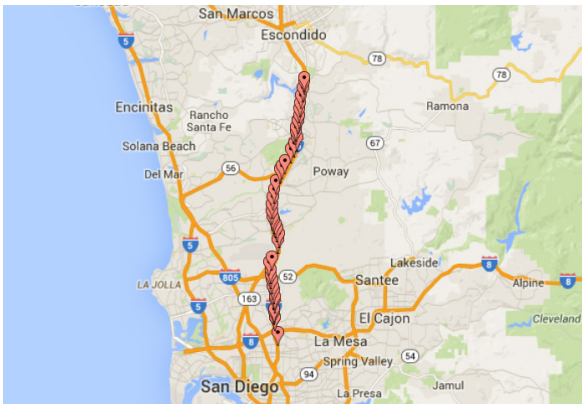


Figure 5. California study site (I-15) (source: Google Maps)

7. Prediction Experimental Results

This section describes the evaluation criteria and then shows the experimental results of the proposed traffic state prediction for the Virginia and California study sites.

7.1. Evaluation Criteria

Relative and absolute prediction errors were calculated to compare the models using shallow networks, those using deep neural networks, and those using PLSR. The relative error was computed as the Mean Absolute Percentage Error (MAPE) using (2). This error is the average absolute percentage change between the predicted and the true values. The corresponding absolute error is presented by the mean absolute error (MAE) using (3). This error is the absolute difference between the predicted and the true values.

$$\mathrm{MAPE} = \frac{100}{\mathrm{IxJ}} \sum_{j=1}^{J} \sum_{i=1}^{I} \frac{\left|y_i^j - \hat{y}_i^j\right|}{y_i^j} \tag{2}$$

$$\mathrm{MAE} = \frac{1}{\mathrm{IxJ}} \sum_{j=1}^{J} \sum_{i=1}^{I} \left|y_i^j - \hat{y}_i^j\right| \tag{3}$$

where:

J = total number of observations in the testing data set

I = total number of elements in each observation

y = ground truth traffic state

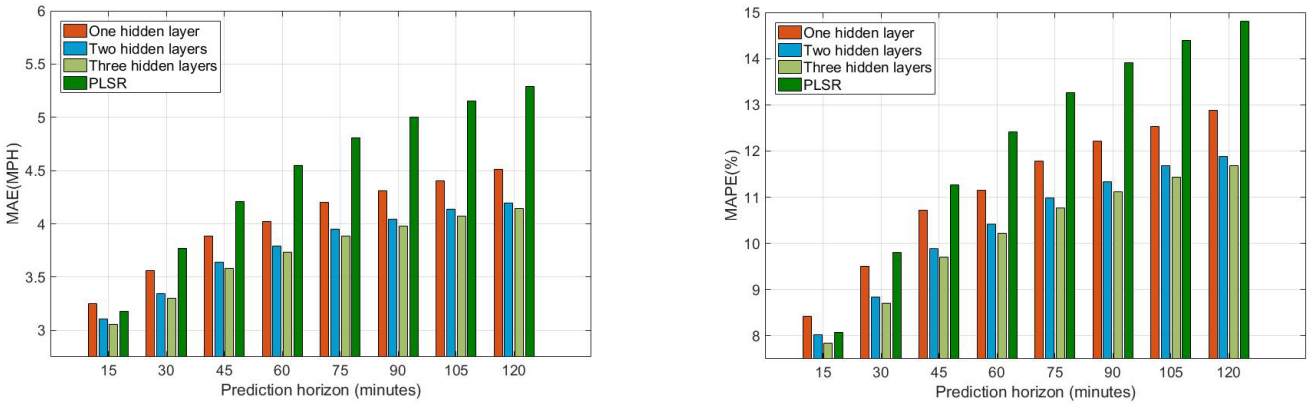
$\hat{y}$ = predicted traffic state

In the following experimental work, leave-one-out validation was used to cross-validate the prediction errors. During each fold

of the leave-one-out validation, one month was kept out for testing and the other months were used to train the model such that each month was used only once for testing. The leave-one-out validation subsequently reported the final prediction errors as the average errors of all folds.

7.2. I-66 Results

In this section, we present the results of applying the PLSR model, shallow network model, and deep neural network model to predict the speed for up to 120 minutes in the future for the I-66 stretch. The dataset for the I-66 road stretch contained only speed data. The input to the neural network is the score matrix $\bar{X}$, which is obtained from the predictors' matrix X using PCA. In our previous study [24], we trained different one-hidden-layer networks with different numbers of neurons, ranging from 3 to 31, in the hidden layer. The different networks were compared using the MAPE and MAE, and the results showed that nine neurons was a good choice. Accordingly, in this experimental work, the number of neurons in each hidden layer was set equal to nine. The MAPE and MAE for the speed at different prediction horizon are shown in Figure 6.



(a) (b)

Figure 6. Speed prediction errors for I-66 in Virginia: (a) MAE; (b) MAPE.

expected to be high, and the variability in the response could be explained using a linear model such as PLSR. However, as the prediction horizon increased, the relationship between the predictors and responses became nonlinear and the PLSR did not predict well. The neural network, on the other hand, had the ability to estimate a nonlinear function, and predicted well. It should also be noted that, as the number of hidden layers increased, the average errors decreased. In order to visually inspect differences among the four models, we plotted the average of the errors of all folds and estimated the 95% confidence interval, as shown in Figure 7.

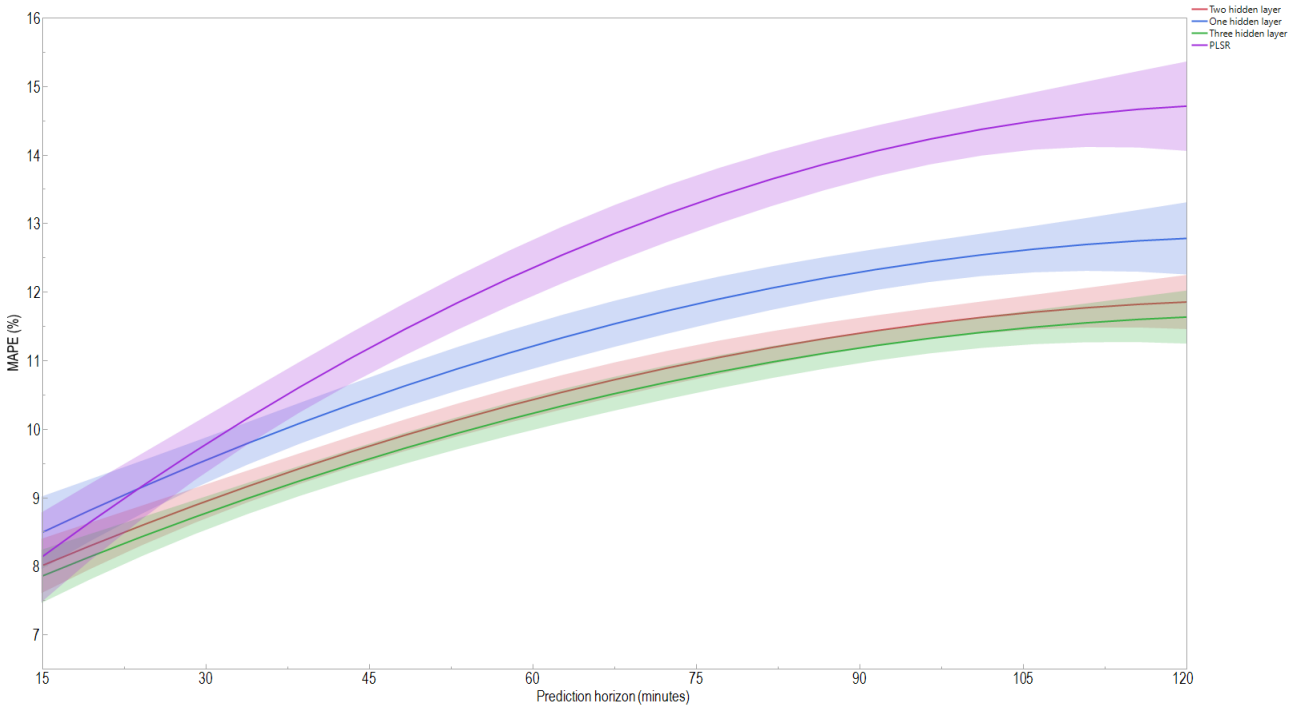
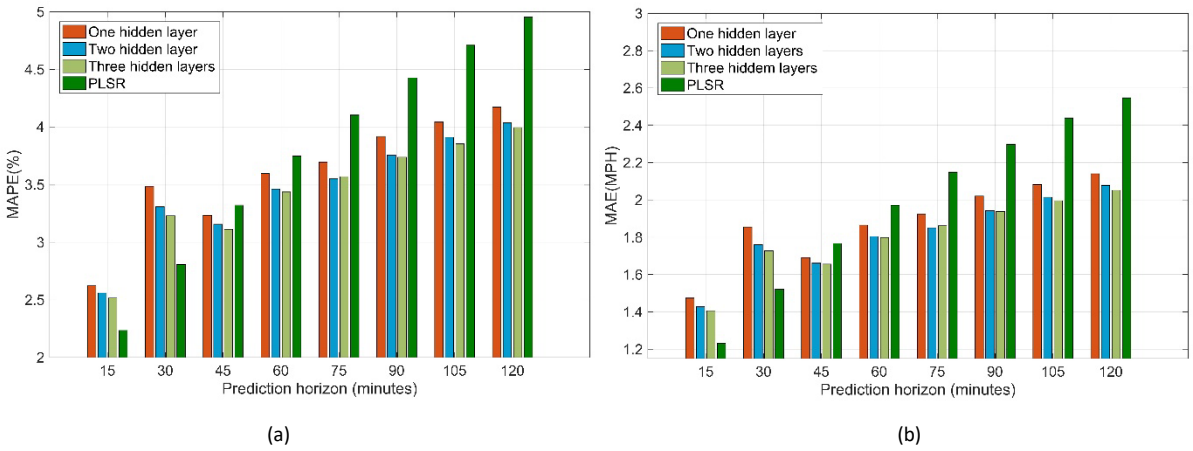
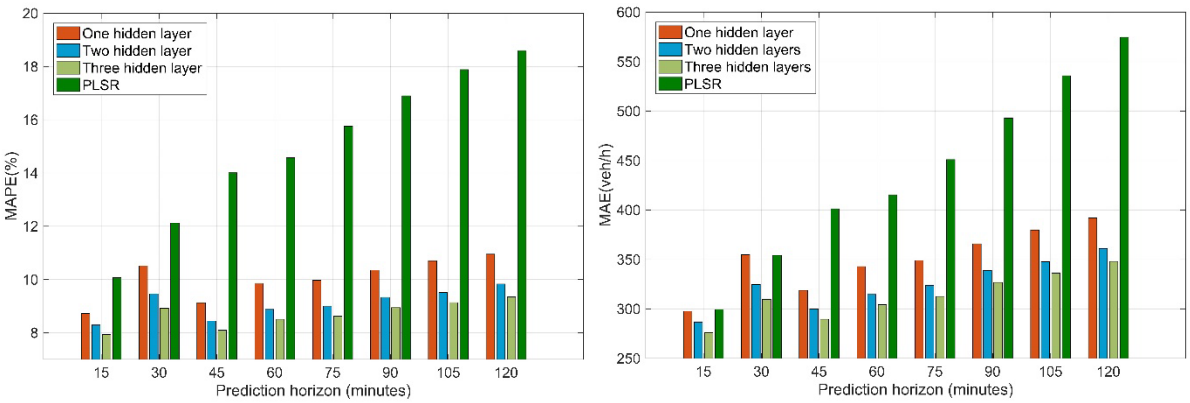


Figure 7. The speed MAPE and 95% confidence interval for the four developed models I-66 in Virginia).



(a) (b)

Figure 8. Speed prediction errors for I-15 in California; (a) MAPE; (b) MAE.



(a) (b)

Figure 9. Flow prediction errors for I-15 in California; (a) MAPE; (b) MAE

As shown in Figure 7, at the 15-minute prediction horizon, all models are quite similar where their four confidence intervals intersect. At a larger prediction horizon, we can group the four models into three different groups. The first group includes the PLSR model; the second group includes the one-hidden-layer neural network model; the third group includes both the two- and three-hidden-layers neural network models. This figure suggests that the three-hidden-layers neural network is the best model, particularly at the 120-minute prediction horizon, as its confidence interval does not intersect with the one-hidden-layer neural network's confidence interval.

7.3. I-15 Results

In this section, we discuss the results of applying the proposed approach to a different dataset to predict the traffic states (speed/flow) for up to 120 minutes in the future for the studied stretch if I-15. We modeled the traffic state using shallow network models, the PLSR model, and deep network models. The prediction errors for both speed and flow at different prediction horizons are shown in Figure 8 and Figure 9, respectively.

Figure 8 compares the shallow network and the two deep networks for the speed prediction, with the top panel showing the MAPE and the bottom panel showing the MAE. The PLSR model is shown as well for the sake of comparison. As the figure illustrates, the deep neural networks had smaller prediction errors compared to the shallow neural network. Moreover, the three-hidden-layers network did not remarkably reduce the average prediction error. However, the confidence intervals of the various models still need to be considered before deciding on a preferred model. Figure 8 also shows that at the 15-minute and 30-minute prediction horizon PLSR performed quite well, as there was high correlation between the input predictors and the responses. However, for the longer prediction horizons, neural networks returned better results than PLSR.

The flow prediction errors shown in Figure 9 indicate that the deep neural network had smaller prediction errors compared to the shallow neural network. Moreover, the three-hidden-layers network had the lowest prediction errors. Figure 9 also shows that the prediction errors for PLSR were the worst among all the models, especially for the long prediction horizons.

differentiate between the models with close average errors. In this figure, we can visually inspect whether the four models are, in fact, different by plotting the average of the errors of all folds and estimating the 95% confidence interval.

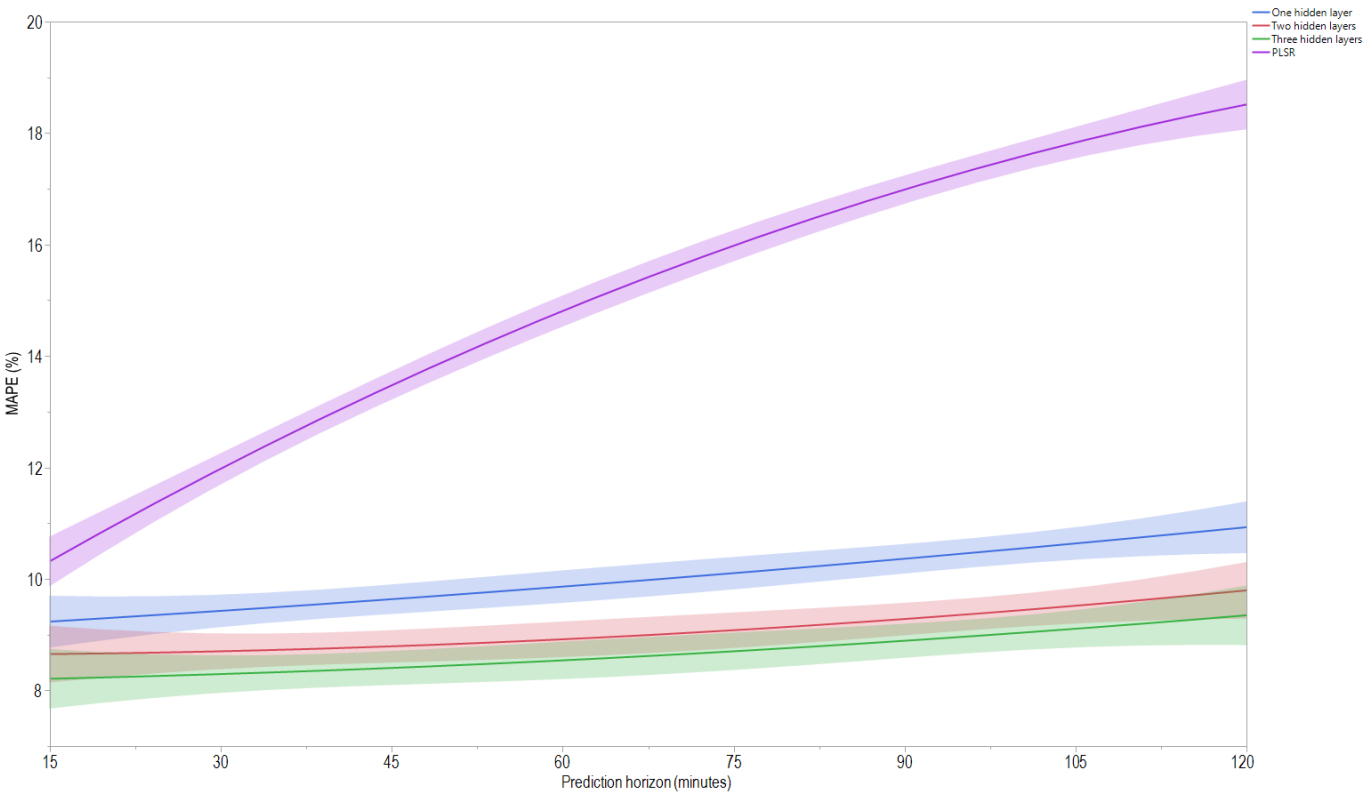


Figure 10. The flow MAPE and 95% confidence interval for the four developed models (I-15 in California).

As shown in Figure 10, the confidence intervals of the two- and three-hidden-layers neural networks intersect, indicating that they are not significantly different. However, the three-hidden-layers model is preferred because it does not intersect with the one-hidden-layer model at the 15-minute prediction horizon, and its confidence interval is far from that of the one-hidden-layer model's confidence interval.

8. Conclusions and Future work

In this paper, we adopted discriminatively pre-trained deep neural networks to build a traffic state short-term prediction model. We compared a shallow neural network, a two-hidden-layers neural network, a three-hidden-layers neural network, and PLSR. The four models were used to predict traffic states for two different freeway road stretches. We used the divide-and-conquer approach to overcome the central processing unit time and memory requirements for long roadway stretches and large prediction horizons. The models were compared using MAE and MAPE measures. Experimental results showed that the three-hidden-layers neural network was the best model for traffic state short-term prediction and that PLSR was the worst among the four models. However, PLSR performed well at the 15-minute prediction horizon, as evidenced by its high correlation between the responses and predictors at that horizon.

Adding other factors, such as work zones and incident information, to the prediction models is planned in future studies. Adding these factors may explain some of the variability that cannot be accounted for using the current prediction models.

Note that the models proposed in this paper do not consider the response of travelers if they are receiving predicted traffic information disseminated by the agencies operating the network. Studying the interaction of informed travelers and including travelers' responses as an input factor to the prediction models is an area for further work. Another area for future work is network-wide traffic prediction, for which models to predict the traffic on different roadway segments in a network will be developed. For network-wide prediction, we can make use of newly available technologies along with new big data techniques to integrate travel behavior and enhance traffic predictions. Moreover, online learning algorithms that continue to learn from each new incoming observation are needed to capture the dynamics of traffic patterns inside cities, which change over time.

References

[1] J. Young-Seon, B. Young-Ji, M. Mendonca Castro-Neto, and S. M. Easa, "Supervised Weighting-Online Learning Algorithm for Short-Term Traffic Flow Prediction," *Intelligent Transportation Systems, IEEE Transactions on,* vol. 14, pp. 1700-1707, 2013.

[2] R. Arnott and K. Small, "The Economics of Traffic Congestion," *American Scientist,* vol. 82, pp. 446-455, 1994.

[3] E. I. Vlahogianni, M. G. Karlaftis, and J. C. Golias, "Short-term traffic forecasting: Where we are and where we're going," *Transportation Research Part C: Emerging Technologies,* vol. 43, pp. 3-19, 2014.

[4] M. Elhenawy and H. Rakha, "Stretch-wide traffic state prediction using discriminatively pre-trained deep neural networks," in *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, 2016, pp. 1065-1070.

[5] M. S. Ahmed and A. R. Cook, *Analysis of freeway traffic time-series data by using Box-Jenkins techniques*, 1979.

[6] M. Van Der Voort, M. Dougherty, and S. Watson, "Combining Kohonen maps with ARIMA time series models to forecast traffic flow," *Transportation Research Part C: Emerging Technologies,* vol. 4, pp. 307-318, 1996.

[7] S. Lee and D. Fambro, "Application of subset autoregressive integrated moving average model for short-term freeway traffic volume forecasting,"

Transportation Research Record: Journal of the Transportation Research Board, pp. 179-188, 1999.
[8] B. M. Williams and L. A. Hoel, "Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: Theoretical basis and empirical results," *Journal of transportation engineering,* vol. 129, pp. 664-672, 2003.
[9] G. A. Davis and N. L. Nihan, "Nonparametric Regression and Short-Term Freeway Traffic Forecasting," *Journal of Transportation Engineering,* 1991.
[10] H. Sun, H. Liu, H. Xiao, R. He, and B. Ran, "Use of local linear regression model for short-term traffic forecasting," *Transportation Research Record: Journal of the Transportation Research Board,* pp. 143-150, 2003.
[11] W. Zheng, D.-H. Lee, and Q. Shi, "Short-term freeway traffic flow prediction: Bayesian combined neural network approach," *Journal of transportation engineering,* vol. 132, pp. 114-121, 2006.
[12] B. Park, C. Messer, and T. Urbanik II, "Short-term freeway traffic volume forecasting using radial basis function neural network," *Transportation Research Record: Journal of the Transportation Research Board,* pp. 39-47, 1998.
[13] E. I. Vlahogianni, M. G. Karlaftis, and J. C. Golias, "Short-term traffic forecasting: Where we are and where we're going," *Transportation Research Part C: Emerging Technologies,* vol. 43, Part 1, pp. 3-19, 6// 2014.
[14] W. Huang, G. Song, H. Hong, and K. Xie, "Deep architecture for traffic flow prediction: Deep belief networks with multitask learning," *IEEE Transactions on Intelligent Transportation Systems,* vol. 15, pp. 2191-2201, 2014.
[15] R. Yu, Y. Li, C. Shahabi, U. Demiryurek, and Y. Liu, "Deep Learning: A Generic Approach for Extreme Condition Traffic Forecasting."
[16] N. G. Polson and V. O. Sokolov, "Deep learning for short-term traffic flow prediction," *Transportation Research Part C: Emerging Technologies,* vol. 79, pp. 1-17, 6// 2017.
[17] Y. Lv, Y. Duan, W. Kang, Z. Li, and F.-Y. Wang, "Traffic flow prediction with big data: a deep learning approach," *IEEE Transactions on Intelligent Transportation Systems,* vol. 16, pp. 865-873, 2015.
[18] X. Ma, Z. Dai, Z. He, and Y. Wang, "Learning Traffic as Images: A Deep Convolution Neural Network for Large-scale Transportation Network Speed Prediction," *arXiv preprint arXiv:1701.04245,* 2017.
[19] J. Van Lint and C. Van Hinsbergen, "Short-term traffic and travel time prediction models," *Artificial Intelligence Applications to Critical Transportation Issues,* vol. 22, 2012.
[20] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Cognitive modeling,* vol. 5.
[21] D. Yu, L. Deng, F. T. B. Seide, and G. Li, "Discriminative pretraining of deep neural networks," ed: Google Patents, 2011.
[22] H. Abdi, "Partial least square regression (PLS regression)," ed.
[23] S. Wold, A. Ruhe, H. Wold, and I. Dunn, W., "The Collinearity Problem in Linear Regression. The Partial Least Squares (PLS) Approach to Generalized Inverses," *SIAM Journal on Scientific and Statistical Computing,* vol. 5, pp. 735-743, 1984.
[24] M. Elhenawy and H. Rakha, "Traffic Stream Speed Short-term Prediction using Machine Learning Techniques: I-66 Case Study," presented at the the 95th Transportation Research Board Annual Meeting, Washington DC, 2016.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 687-694 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Low-cost Hybrid Wheelchair Controller for Quadriplegias and Paralysis Patients

Mohammed Faeik Ruzaij Al-Okby[*, 1, 2], Sebastian Neubert[3], Norbert Stoll[3], Kerstin Thurow[1]

[1]*Center for Life Science Automation (celisca), University of Rostock, Rostock 18119, Germany*

[2]*Technical Institute of Babylon, Al-Furat Al-Awsat Technical University (ATU), Babylon, Iraq*

[3]*Institute of Automation, University of Rostock, Rostock 18119, Germany*

ARTICLE INFO

Article history:
Received: 04 May, 2017
Accepted: 22 May, 2017
Online: 05 June, 2017

Keywords:
Intelligent Wheelchair
Voice Recognition
Quadriplegia
Orientation Detection
Embedded System

ABSTRACT

Wheelchair controller design is very important for handicapped users such as quadriplegia, amputee, paralyzed and elderly. The safe use, easiness, and comfort are an important factor that can directly affect the user's social efficiency and quality of life. In this paper, the design, implementation, and test of a new low-cost hybrid wheelchair controller for quadriplegics and paralysis patient have been proposed. Two sub-controllers are combined together to build the hybrid controller which are the voice controller and the head tilt controller. The voice controller consists of two voice recognition (VR) modules. The first module uses the dynamic time warping (DTW) and the second used both the hidden Markov model (HMM) and the DTW to process the user's voice and recognize the required commands. This controller has been tested with three global languages English, German, and Chinese at two different noise levels 42 and 72 dB. The voice controllers can be used for motion commands as well as for controlling other parameters like lights, light signals, sound alarm, etc. The second sub-controller is the head tilt controller. It consists of two sophisticated orientation detection modules BNO055 to detect and track the user's head and the wheelchair-robot orientation. The controller has an auto-calibrated algorithm to calibrate the user head orientation with reference wheelchair orientation to adjust the control commands and the speed of the system motors in case of ascending or descending a hill or when passing non-straight roads. The head tilt controller was tested successfully for indoor and outdoor applications.

1. Introduction

Quadriplegia, paralyzed, hand amputees and elderly users lost the ability to control their hands and they need a special control system to use an electrical wheelchair which is normally controlled via joystick. Thus, they still need the help of others to move and realize their daily activities. Nowadays the population of this type of users increases dramatically due to rapid increases in accidents and elderly people. The main goal of the presented work is to help this type of patients to control the electrical wheelchair without the need of others help. The active body regions for quadriplegia patients include the head, neck, shoulders and some chest muscles. One of the effective control signals in this body regions is the user's voice. It can be used as a simple solution to control the intelligent application using voice recognition (VR) technology. The VR technology means the processing and converting of the user audio signal to an electrical signal, which can be used by a computer or microcontroller as a control signal. Two main strategies have been used to build VR-based systems. The first uses a computer program to implement the voice recognition and then the computer controls the target system. The second uses embedded modules to implement the VR tasks. In the proposed work the second approach is preferred due to small size, low cost, and less power consumption [1-6].

Other signals which can be used for quadriplegia patients can be acquired from the user's skin surface in the active body regions that are still under the user control such as shoulders, face and head

[*]Mohammed Faeik Ruzaij Al-Okby, Center for Life Science Automation (celisca), University of Rostock, Germany. Phone: +49 381-4987823
Email: Mohammed.Faeik.Ruzaij@celisca.de, abnalfratain@yahoo.com.

scalp using different bio-signal acquisition techniques, like electromyography (EMG) [7-8], electroencephalography (EEG) [9-10], and electrooculography (EOG) [11]. These methods have many drawbacks. The bio-signals are highly affected by the interferences and noises from the power sources and also from other body bio-signals [12-14]. The acquisition of the EEG, EMG, and EOG need to use and connect the user's body with special electrodes which may be uncomfortable and limit the user motion.

Head tilts and body movements have been used by other researchers. User's head tilts around basic axes x, y and z-axes were used to control the motion of the wheelchair. Several scenarios have been used to achieve this goal. Ultrasound and infrared sensors and detectors have been used to track the head motion with reference to the sensor array [15-18]. Cameras and optical sensors have been used to detect the head orientation and used it as a control signal [19]. Nowadays, accelerometer, micro electrical mechanical system (MEMS) based inertial and orientation sensors are used to detect the objects motion and gesture recognition [20-24]. Most of the above work used open loop control systems with thresholds for the control commands. Some applications use fixed speed for each motion command which is not sufficient for a real-time application.

In this paper, the design, implementation, and test of a hybrid low-cost controller for wheelchairs have been proposed. It has two sub-controllers which are the voice and head tilts to give the user more flexibility in the selection of the comfortable control method. The presented work is innovative in the sense that it uses two different voice recognition modules that are combined together in the voice controller. At the same time, the head tilts controller uses two orientation detection modules to calibrate the control commands of the user's head with the wheelchair reference orientation in case of ascending or descending a ramp. The proposed solution is safe, comfortable, and easy to use compared with previous other solutions which used only the head orientation without taking in consideration the wheelchair orientation and the road slope.

The system has a simple design. A headset with microphone is used to pick up the voice commands and send them to the voice recognition modules VR1 and VR2. Also, the first orientation module is fixed on the headset to detect the user's head tilts and send the orientation data to the microcontroller to take the required control commands. The user can simply control the wheelchair by his voice using specific commands or he can use his head tilt to the front, back, left, right and middle for forward, backward, left, right and stop commands respectively. The system uses the powerful ARM EFM32GG990F1024 microcontroller (Silicon Laboratories Inc., USA) which makes the system work standalone, without the need of using any computers and also enables a low cost, small size, and low energy consumption design.

2. Methodology

The system consists of two input units and two output units connected to the core microcontroller. The input units are the two VR modules for voice controller and also the two orientation detection modules for the head tilts controllers. The output control signals will be sent to the motor driver unit of the wheelchair to implement the motion commands and to the system LCD display to show the system status and the active controller. The microcontroller is responsible for receiving the data from the input unit, processing the received data, taking the required control commands and sending them to the output units. Fig. 1 shows the block diagram of the system. All system units will be explained in detail in next sub-section.

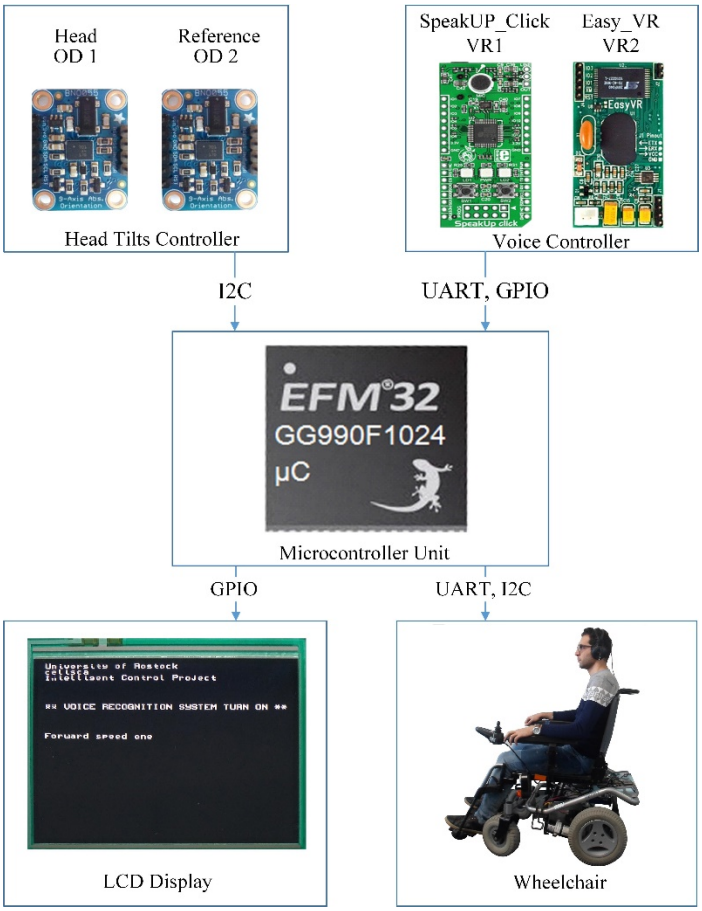


Fig. 1: Block Diagram

2.1. Microcontroller Unit (MCU)

The Silicon Labs EFM32GG990F1024 ARM microcontroller, has been selected as the core processor for the presented system. The low energy consumption (200 µA/MHz in running mode; 50 µA/MHz in sleep mode), makes it one of the most energy efficient microcontrollers. The EFM32GG990F1024 has a 32-bit central processing unit (CPU) and can operate at the speed of up to 48 MHz. The main reason and embedded features for selecting this microcontroller are:

- 5 universal synchronous/asynchronous receiver/transmitter UART/USART ports are available and can be used separately. Tow UART ports have been used for communication with the voice recognition modules. Also, another UART port is used for the communication with the motor driver unit. The remained UART/USART ports can be used for future system requirements.

- The controller has Two I²C bus interfaces: The proposed system requires only one I²C bus for communication with the two BNO055 orientation modules. The remaining I²C interface can be reserved for future works.
- 3 pulse counters: the system requires two pulse counters to communicate with the motor encoders to calculate the speed feedback of the system motors.
- Wake up time: It has a short wake-up time of 2 µs. The wake-up time is the time required to change the energy mode from different sleep modes to active mode.

Several other unique embedded features have been built in the EFM32GG990F1024 microcontroller found elsewhere [25].

2.2. Voice Controller

The voice controller consists of two voice recognition modules which are the VR1: SpeakUp Click module (MikroElektronika D.O.O., Serbia) and the VR2: Easy VR module (Veear, Italy). The two modules VR1 and VR2 connected together with the main MCU and build a strong voice controller. The voice controller uses both, dynamic time warping (DTW) and hidden Markov model (HMM) algorithms to perform the voice processing and it can work with speaker dependent SD or speaker independent SI voice recognition modes. This controller has already been tested with Jaguar lite robot. In this paper, the same controller will be integrated with a wheelchair. More details about the previous work and the used module can be found elsewhere [4, 6, 26-28].

2.3. Head Tilts Controller

The head tilts controller consists of two BNO055 orientation detection modules (Bosch Sensortec, Germany). The BNO055 module consists of 3 triaxial MEMS sensors which are an accelerometer, a gyroscope, and a magnetometer. All 3 sensors are combined with a 32-bit microcontroller in a single package to produce a nine degree of freedom (9DOF) orientation detection module. The module output has different data formats and measuring units such as Euler angles, quaternion vector, linear acceleration, gravity vector, and magnetic field strength [29]. The Euler angles have been used in the proposed system to measure the user's head and the reference wheelchair orientation.

2.4. Motors Driver Unit and Motors

Meyra Ortopedia Smart 9.906 (MEYRA GmbH, Germany) has been used for system tests. The Meyra wheelchair uses CAN bus for communication between the joystick and the motion controller. In the presented work, the original joystick control system has been replaced with the voice and head tilt controller. Sabertooth 2x32 Amp (Dimension Engineering LLC., USA; see Fig. 2) has been used for the wheelchair motor driving. It communicates with the main ARM microcontroller using a UART bus (settings: 115,200 kbps, N, 8, 1).

Two DC motors SRG0531 (AMT Schmid GmbH & Co. KG, Germany) have been used to drive the wheelchair. Each motor has a gearbox including a helical gearset on the input side and spur gear set on the output side. The operation voltage of the motor is 24 V with a rated power from 220 to 400 Watt. The total load capacity is 160 kg and it can run with an output speed from 105 to 200 RPM (revolutions/minute). The motor operates at low-level noise which is less than 58 dB. It has an electromagnetic brake work with 12/24 V [30].

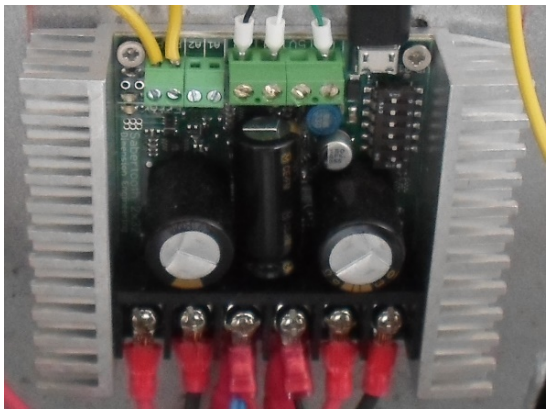


Fig. 2: Sabertooth 2x32 motor driver unit

3. System Description

The microcontroller communicates with the voice controller modules VR1 and VR2 via the general-purpose input/output GPIO pins and UART communication ports respectively. The voice command is picked up and feeding to the VR1 and VR2 using a sensitive microphone with an internal impedance of 1 k Ohm at 1 kHz, frequency range of 50-13000 Hz and a sensitivity of -34 dB. The VR modules process the voice commands and send the recognition results to the microcontroller. “OR” function has been used to combine the recognition results of the VR modules which allow the system to response when any one of the both VR modules recognizes correctly. A false positive (FP) cancellation function has been programmed and added to the “OR” function to cancel the voice recognition results if the two modules show a different command recognition. The FP cancellation function converts the FP errors to a false negative FN error which mean the system cannot recognize the voice command.

The head tilts controller can be activated using the voice command “Orientation”. Two BNO055 orientation detection modules have been used to track and read the orientation data of the user's head and the wheelchair. The first module is fixed on a headset for reading the head orientation; the second is fixed on the wheelchair chassis to read the reference wheelchair orientation. The user's head tilts around the principal axes x and y activates the control commands of the head tilt controller. The orientation detection modules have been programmed to send orientation information in Euler angles format (Pitch, Roll, and Yaw) to the microcontroller using inter-integrated circuit I²C bus.

The use of the second module for reference orientation detection plays an important role in the control commands calibration in the case of ascending or descending a hill. The microcontroller will compare the tilt angles (Euler angles) of the

user's head with the wheelchair tilt angles instead of comparing with ground original (x, y, and z) axes. This increases the system safety, and easy use when the system passes through ramps or non-straight roads. When only the head orientation module is used for controlling the wheelchair without using the reference wheelchair orientation, the user needs to increase the head tilt angle proportional to the value of hill slope angle in the ascent and vice versa the descent to compensate the hill slope. The same procedure happens in case of a side slope in the road. The user needs to increase the head tilt more to the right when the slope of the right side is greater than left and vice versa. In the presented system only two Euler angles have been used which are the Pitch and Roll for motion control. The Yaw angle has been neglected to allow the user turn his head free around the z-axes when he is driving the wheelchair.

3.1. System motor controlling

The wheelchair speed and direction control have been implemented through controlling the left (M2) and the right (M1) motor rotation speed and direction. The head tilts controller controls the speed of the wheelchair motors depending on the x, and y position of the user's head with reference to the origin point (x=0°, y=0°). If the user increases his Pitch tilt angle in the x-axis and keeps the Roll at -5°<y<5°, the both wheelchair motors will rotate in the forward direction at the same speed. If the user tilts his head in the forward direction with right side tilt (see Fig. 3) to move the wheelchair forward with a right curve direction. In this case, the left motor (M2) speed will increase (M2= x + y) and the right motor (M1) speed will decrease (M1= x - y) depending on the value of the head tilt in the x and y-axes. The same procedure applies when the user needs to go forward with a left curve where the speed of the right motor will increase and the speed of the left motor will be decreased. A similar situation will be applied for backward commands.

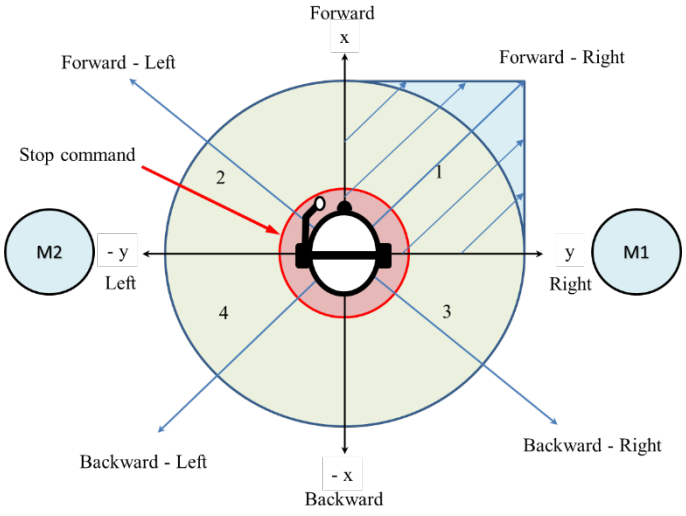


Fig. 3: Head tilts control motor speed

The user can rotate the wheelchair left or right at the same point by tilting his head left or right and keeping the value of the x-axis at -10°<x<0°. The forward command is activated when the Pitch head tilt angle become at 0°<x<25°. This tilt range will keep acceptable user viewing in the front direction. The backward command is activated when the Pitch tilt angle become at -45°<x<-10°.

3.2. Sub-Controller selection

The system design allows the user easily to select and change between the voice and head tilt as a motion controller. The system will start by listening for the user voice command to activate the required controller. If the user wants to select the voice controller hen need to give the voice command “voice”. The voice controller controls the motion of the wheelchair using a group of motion commands which are forward, left, right, stop, backward, speed one and speed two. Another group of voice commands has been used for activating the control of lights, signals and sound alarms which are light-on, light-off, signal-on, and signal-off. When the user needs to activate the head tilts controller he must give the voice command “orientation”. Then the wheelchair motion is guided by the user’s head tilts and the voice controller will still be used to turn on the lights, light signals, and sound alarms. The user can change the motion controller at any time by giving a voice command with the name of the required controller (“voice” or “orientation”). Fig. 4 shows the flowchart of the presented system.

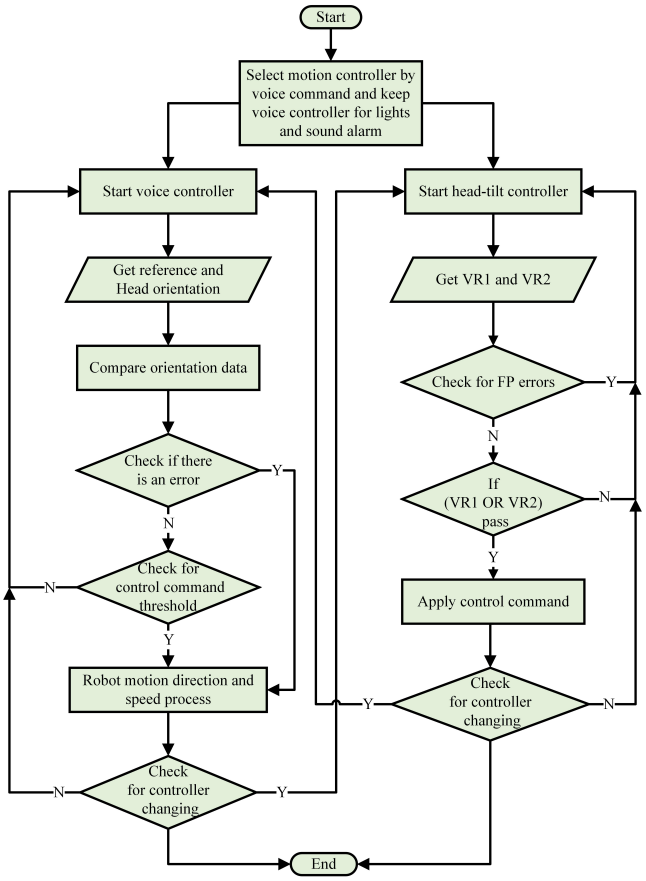


Fig. 4: System flowchart

4. Experimental Results

4.1. Voice Controller

In the previous work, the voice recognition modules VR1 and VR2 have been tested individually and also combined with "AND" and "OR" algorithms [6] in two different noise level environments with 42 dB and 72 dB. Seven voice commands to control the motion of the wheelchair have been selected for the test which are: forward, left, right, stop, backward, speed one and speed two (more detailed results can be found elsewhere [4, 6]).

In the presented work, 8 new voice commands have been used to 1) select the motion controller, 2) control the maximum speed of the wheelchair and 3) control the lights. The used commands are: voice, orientation, speed-one, speed-two, light-on, light-off, signal-on, and signal-off. Since the voice controller was tested before [4, 6], the SD voice recognition mode has been selected to control the wheelchair, which is the best choice for medical and rehabilitation application. It makes the wheelchair respond only to a specific user and not respond to any person saying the same trained commands.

"OR" with FP cancellation function has been used in the voice controller tests. The function has been tested before and the results show that the VR accuracy has been increased ≈ 3% and the FP error has been reduced to ≈ 0% with FN error rate of 5.2 % in 72 dB noisy environments [4, 6].

The voice commands "forward", "speed one" and "speed two" have been used when the head tilt controller is activated to control the user's head tilts to speed ratio. This is important to let the user increase or decrease the system speed sensitivity proportional to the head tilts angle value. The "forward" voice command sets the tilt to speed ratio as 0°-25° tilts angles have a 0-2 km/h speed with ≈ 80 m/h speed for each individual tilt degree. The command "speed one" increases the speed ratio to (0°-25° / 0-4 km/h) with ≈ 160 m/h speed for each tilts degree. The command "speed two" increases the tilts to speed ratio to (0°-25° / 0-6 km/h) with ≈ 240 m/h speed for each tilt degree. Fig. 5 shows the tilt to speed ratio of the tested voice commands.

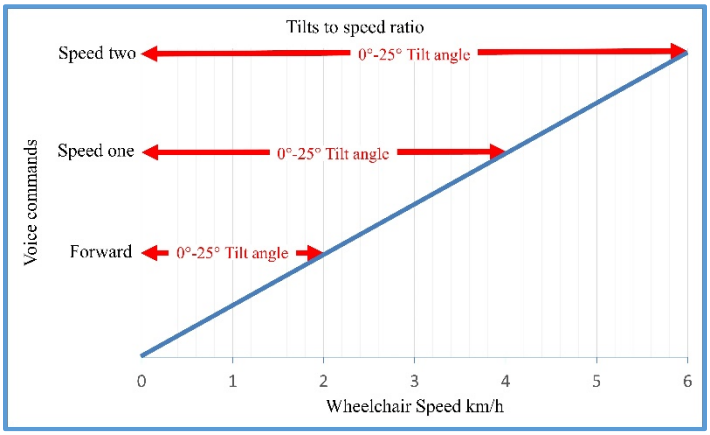


Fig. 5: System tilts to speed ratio

4.2. Head Tilts Controller

The head tilts controller has been tested for indoor as well as for outdoor environments.

4.2.1. Indoor Head Tilts Controller Tests

Several different tasks have been executed to test the system for the indoor motion control. The first task is used to test each control command for the head tilts controller. The users performed the test two times. The first is without previous training and the result titled NTU, and the second is after training the user for approximately 30 minutes and titled TU. Seven control commands have been tested by the users using head tilts around the x and y-axis. The tested commands are: forward, backward, left, right, stop, forward-left, forward-right. Each user tests each control command 10 times. The system speed to tilts ratio has been fixed at 2 km/h which is the primary system speed to tilt ratio. The test results before and after 30 mints training for the control command tests are explained in Fig. 6.

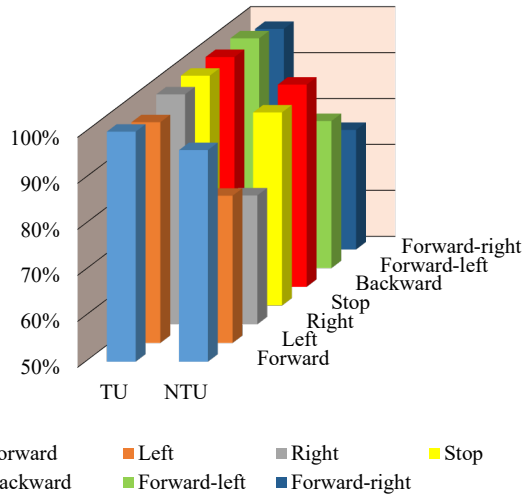
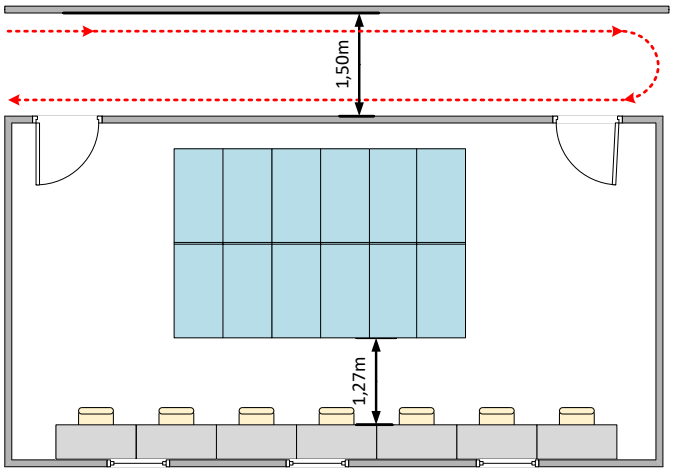


Fig. 6: Commands success rate of the head tilt controller for trained user TU and non-trained users NTU

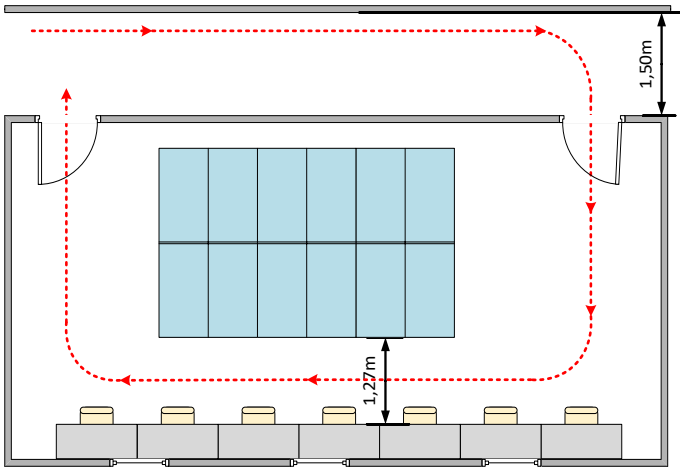
The results in Fig. 6 show the control commands success rate. The success rate is the correct implementation of the tested command in speed and direction from the start to the end point. The test results revealed that the success rate is increased when the user is trained before using the system.

The second test for indoor environments is used to check the flexibility of driving the wheelchair at a narrow area. This test includes implementing two tasks which are: 1) driving the wheelchair through a narrow corridor and 2) driving the wheelchair through a furnished hall. The test has been performed in a laboratory environment at the University of Rostock. Five trained users implemented the two tasks of the test 5 times. The results show a successful implementation of the two tasks using

the head tilts controller. All 5 users completed the tasks 1 and 2 without collisions and followed the required paths correctly. Fig. 7 shows the test paths.



1- Passing a narrow corridor



2- Passing a pathway in furnished hall

Fig. 7: Tasks for system indoor tests

4.2.2. Outdoor Head Tilts Controller Tests

The auto-calibrated algorithm of the head tilts controller has been tested for outdoor environments. The test was designed to check the performance of the algorithm, especially in the case of ascending or descending a hill. The algorithm is used to cancel the effect of road slope on the user's head tilt angle. The tests was performed on a small narrow and long ramp with a height of 80 cm and a slope angle of 7 degrees. The test included two tasks, the first is to pass the ramp without activating the algorithm while the second is to pass the ramp with the auto-calibrated algorithm. Five users performed the test using fixed tilt to speed ratio of 25° tilt angle to 2 km/h for both tasks.

The test results revealed that the auto-calibrated algorithm enhances the performance of the head tilts controller for driving the wheelchair while passing the ramp and makes it similar to driving the wheelchair with normal head tilt angles on a straight road. The test of the system without activating the auto-calibrated algorithm for ascending a hill show that the control of the wheelchair is more difficult. In the ascending of the hill the user needs to tilt his head 7 degrees more down (same ramp slope) compared to the motion with the auto-calibrated algorithm to compensate the hill slope (7°) and vice versa. Fig. 8 shows the tests of the system with and without the auto-calibrated algorithm.

a) Without auto-calibrated algorithm, 7° more in the head tilts

b) With auto-calibrated algorithm

Fig. 8: System tests for passing an 80cm height and 7° slope angle ramp with and without the auto-calibrated algorithm.

4.2.3. Hysteresis and Wrong Head Orientation Function

During the test of the system, the head tilt controller is faced with a harmful situation with hysteresis around the stop command area. This problem occurs when the user brings his head into the stop area and moves his head involuntarily out of the stop area for less than 50 ms without intending the system to move. This problem makes continuous discrete motion-stop commands that can cause concern and bother the user and also unwanted motion may occur. A new function has been added to the head tilt controller to avoid this hysteresis by increasing the number of measurements for the desired control command from the head orientation sensor. Instead of using only one readout in the range of 0º<tilt angle<25º for the forward command, the system uses 10 continuous measurements at the region of 0º<tilt angle<25º for the forward command tilt area in order to implement the forward command. This procedure cancels the effect of the involuntary head movement and makes the start of the motion more smoothly.

Another function has been built and embedded in the head tilt controller to stop the system and protect the user if the head orientation goes out of the programmed range of the control commands tilts angles. The function stops the wheelchair and suspends the control commands of the head tilt controller until the headset returns to the center stop region (0°>x>-10°, 5°>y>-5°). This function protects the user when the headset falls down or the user loses consciousness.

The two functions have been tested with the system and the results revealed a successful performance for both of them. The hysteresis function increased the reaction time of the system ≈ 50 ms in the initial motion only which do not really affect the system performance.

4.2.4. Orientation Sensor Error Handling

Orientation sensor error handling is one of the most important tasks in the present study since the system's motion can cause the loss of the instantaneous updating of orientation data. This can affect the microcontroller performance and may cause restarting or hanging in the system. An error handling function has been programmed to avoid missing orientation data. The error handling function tracks changes in the orientation data and sends alert information to the user if there are any communication problems with orientation modules. It also resets and recalibrates the orientation sensors after a limited number of unchanged readings.

The function has been tested with the system and the test results show a success rate of 100% for tracking the information data, resetting and calibrating the orientation detection modules. Fig. 9 shows the flowchart of the system's error handling function.

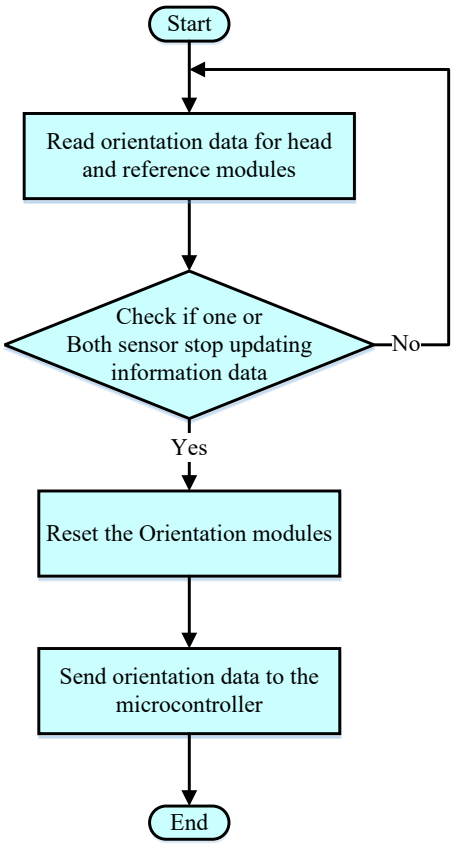


Fig. 9: Error handling function of the head tilts controller

5. Conclusions

In this paper, the design, implementation, and testing of a hybrid wheelchair controller for quadriplegic and paralysis users has been proposed. Two sub-controllers, which are the voice and head tits controller, have been used to allow the user to select and change the active controller for more comfortable wheelchair driving. The test results of the voice controller revealed that it has a better performance using "OR" and FP cancellation function compared with "AND" and individual tests. The results revealed a ≈ 3% increase in the VR accuracy which reached 98.2% in 42 dB noise environments with a 0% FP errors and 1.8% FN errors which are important features for biomedical and rehabilitation application. The voice controller is ideal for the control of lights, light signals, sound alarms, rather than control the wheelchair motion, especially in narrow indoor areas.

The head tilts controller tests revealed that it can be used for both indoor as well as outdoor environments. The indoor tests showed accurate control and good performance in a narrow area with a successful implementation of task 1 and task 2. The outdoor tests showed that the system is ideal for non-straight roads and especially for ascending and descending a hill with the auto-calibrated algorithm.

Hysteresis, wrong head orientation function and orientation sensor error handling function have been successfully added to the head tilts controller. All the functions have been tested and the results show that they enhance the performance of the controller and make the driving of the wheelchair more safe and comfortable.

The head tilts controller tests revealed that previous training of the user is required before using it to have a better human system interaction, good adaptation, and enhancing the system performance.

The voice and the head tilt controllers have been combined together successfully, the user can select or change the wanted controller easily using voice commands.

Acknowledgment

This work is supported by the German Academic Exchange Service (DAAD, Germany), and Al-Furat Al-Awsat Technical University (ATU, Iraq). The authors would like to thank and appreciate their support and scholar funding. Also, the author would like to thank Dr.-Ing. Thomas Roddelkopf, Dr.-Ing. Steffen Junginger and Mr. Heiko Engelhardt for help and technical support.

References

[1] Mohammed Faeik Ruzaij, Sebastian Neubert, Norbert Stoll, Kerstin Thurow, "Design and implementation of low-cost intelligent wheelchair controller for quadriplegias and paralysis patient", in Proc. 2017 IEEE 15th International Symposium on Applied Machine Intelligence and Informatics (SAMI), Herl'any, Slovakia, 26-28 January 2017, pp. 399-404.

[2] Akira Murai, Masaharu Mizuguchi, Takeshi Saitoh, Member, IEEE, Tomoyuki Osakiand, Ryosuke Konishi, "Elevator Available Voice Activated Wheelchair", in Proc. 18th IEEE International Symposium on Robot and Human Interactive Communication, Toyama, Japan, 27 September- 2 October 2009, pp. 730-735.

[3] Xiaoling Lv, Minglu Zhang and Hui Li, "Robot Control Based on Voice Command", in Proc. IEEE International Conference on Automation and Logistics, Qingdao, September 2008, pp. 2490-2494.

[4] Mohammed Faeik Ruzaij, Sebastian Neubert, Norbert Stoll, Kerstin Thurow, "Design and Testing of Low Cost Three-Modes of Operation Voice

Controller for Wheelchairs and Rehabilitation Robotics", in Proc. 2015 IEEE 9th International Symposium on Intelligent Signal Processing (WISP), Siena, Italy, 15-17 May 2015, pp. 114-119.

[5] Mohammed Faeik Ruzaij, S. Poonguzhali, "Design and Implementation of Low Cost Intelligent Wheelchair", in Proc. Second International Conference on Recent Trends in Information Technology, Chennai, 19-21 April 2012, pp. 468-471.

[6] Mohammed Faeik Ruzaij, Sebastian Neubert, Norbert Stoll, Kerstin Thurow, "Hybrid Voice Controller for Intelligent Wheelchair and Rehabilitation Robot Using Voice Recognition and Embedded Technologies", Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol. 20 No. 4-2016, Fuji Technology Press Ltd, Tokyo, Japan, pp. 615-622.

[7] Inhyuk Moon, Myungjoon Lee, Jeicheong Ryu, and Museong Mun, "Intelligent Robotic Wheelchair with EMG-, Gesture-,and Voice-based Interfaces", in Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems, Las Vegas. Nevada, 27-31October 2003, Vol. 4, pp. 3453 – 3458.

[8] Satoshi Ohishi and Toshiyuki Kondo, "A Proposal of EMG-based Wheelchair for Preventing Disuse of Lower Motor Function", in Proc. Annual Conference of Society of Instrument and Control Engineers (SICE), 20-23 August 2012, Akita University, Akita, Japan, pp. 236-239.

[9] Kazuo Tanaka, Kazuyuki Matsunaga, and Hua O. Wang, "Electroencephalogram-Based Control of an Electric Wheelchair", IEEE Transactions on Robotics, Vol. 21, No. 4, August 2005.

[10] I. Iturrate, J. Antelis and J. Minguez, "Synchronous EEG Brain-Actuated Wheelchair with Automated Navigation", in Proc. 2009 IEEE International Conference on Robotics and Automation, International Conference Center, Kobe, Japan, May 12-17, 2009, pp. 2318-2327.

[11] Biswajeet Champaty, Jobin Jose, Kunal Pal, Thirugnanam A., "Development of EOG Based Human Machine Interface control System for Motorized Wheelchair", in Proc. International Conference on Magnetics, Machines & Drives (AICERA-2014 iCMMD), Kottayam, 24-26 July 2014, pp. 1-7.

[12] Samuel Akwei-Sekyere, "Powerline noise elimination in biomedical signals via blind source separation and wavelet analysis," PeerJ, Vol. 3, p.e1086, July 2015, pp.1-24.

[13] Wilfried Philips, "Adaptive noise removal from biomedical signals using warped polynomials," IEEE Transactions On Biomedical Engineering, Vol. 43, no. 5, May 1996. pp. 480–492.

[14] Kayvan Najarian, Robert Splinter, Biomedical Signal, and Image Processing. CRC Press, 2005.

[15] O. Partaatmadja, B. Benhabib, A. Sun, and A. A. Goldenberg, "An Electrooptical Orientation Sensor for Robotics", IEEE Transactions On Robotics And Automation, Vol. 8, No. 1, February 1992, pp. 111-119.

[16] D. J. Kupetz, S. A. Wentzell, B. F. BuSha, "Head Motion Controlled Power Wheelchair", in Proc. 2010 IEEE 36th Annual Northeast Bioengineering Conference, New York, USA, 26-28 March 2010, pp. 1-2.

[17] Henrik Vie Christensen a and Juan Carlos Garcia b,"Infrared Non-Contact Head Sensor, for Control of Wheelchair Movements," Book Title Book Editors IOS Press, 2003.

[18] James M Ford, Saleem J.Sheredos, " Ultrasonic Head Controller For Powered Wheelchair," Journal of Rehabilitation Research and Development, Vol. 32 No. 3, October 1995, pp. 280-284.

[19] Farid Abedan Kondori, Shahrouz Yousefi, Li Liu, Haibo Li, "Head Operated Electric Wheelchair", in Proc. 2014 IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI), San Diego, CA, USA, 6-8 April 2014, pp. 53-56.

[20] Mohammed Faeik Ruzaij, Sebastian Neubert, Norbert Stoll, Kerstin Thurow, "Auto Calibrated Head Orientation Controller for Robotic-Wheelchair Using MEMS Sensors and Embedded Technologies", in Proc. 2016 IEEE Sensors Applications Symposium (SAS 2016), Catania, Italy, 20-22 April 2016, pp. 433-438.

[21] Mohammed Faeik Ruzaij, Sebastian Neubert, Norbert Stoll, Kerstin Thurow, "Multi-Sensor Robotic-Wheelchair Controller for Handicap and Quadriplegia Patients Using Embedded Technologies", ", in Proc. 2016-9th International Conference on Human System Interactions (HSI), Portsmouth, United Kingdom, 6-8 July 2016. pp.103-109.

[22] Deepesh K Rathore, Pulkit Srivastava, Sankalp Pandey and Sudhanshu Jaiswal, "A Novel Multipurpose Smart Wheelchair", in Proc. 2014 IEEE Students' Conference on Electrical, Electronics and Computer Science, Bhopal, India, 1-2 March 2014, pp. 1-4.

[23] Mohammed Faeik Ruzaij, Sebastian Neubert, Norbert Stoll, Kerstin Thurow, "A speed compensation algorithm for a head tilts controller used for wheelchairs and rehabilitation applications", in Proc. 2017 IEEE 15th International Symposium on Applied Machine Intelligence and Informatics (SAMI), Herl'any, Slovakia, 26-28 January 2017, pp. 497-502.

[24] Tao Lu, "A Motion Control Method of Intelligent Wheelchair Based on Hand Gesture Recognition", in Proc. 2013 8th IEEE Conference on Industrial Electronics and Applications (ICIEA), Melbourne, VIC, 19-21 June 2013, pp. 957-962.

[25] EFM32GG990 DATASHEET, [Online]. Available: http://www.silabs.com. [Accessed: 20 December 2015].

[26] Speak Up Click User Manual Ver.101, MikroElektronika, Belgrade, Serbia, 2014.

[27] EasyVR 2.0 User Manual R.3.6.6., TIGAL KG, Vienna, Austria, 2014.

[28] RSC-4128 Speech Recognition Processor data sheet, [Online]. Available: http://www.sensory.com/wp-content/uploads/80-0206-W.pdf. [Accessed: 04 May 2016].

[29] BNO055 data sheet, [online]. Available: https://ae-bst.resource.bosch.com/media/_tech/media/datasheets/BST_BNO055_DS000_14.pdf. [Accessed: 15 Jun 2016].

[30] "AMT - Auto Move Technologies > Home > Products > Geared Motors > SRG 05." [Online]. Available: http://www.amt-schmid.com/en/products/right_angled_drives/srg_05/srg_05.php. [Accessed: 10 October 2016].

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 695-701 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Preparation of Ni-C Ultrafine Composite from Waste Material

Mahmoud A. Rabah*, S.M. Abdelbasir

Electrochemical Treatment Dept., Mineral Processing Div. Central metallurgical Research and Development Institute (CMRDI) Egypt.

ARTICLE INFO

Article history:
Received: 02 April, 2017
Accepted: 15 May, 2017
Online: 09 June, 2017

Keywords:
Carbon Ultrafine Scale
Carbon-Nickel Ultrafine Composite
Plasma Technology
Spent Engine Oil
Plasma Heat Decomposition

ABSTRACT

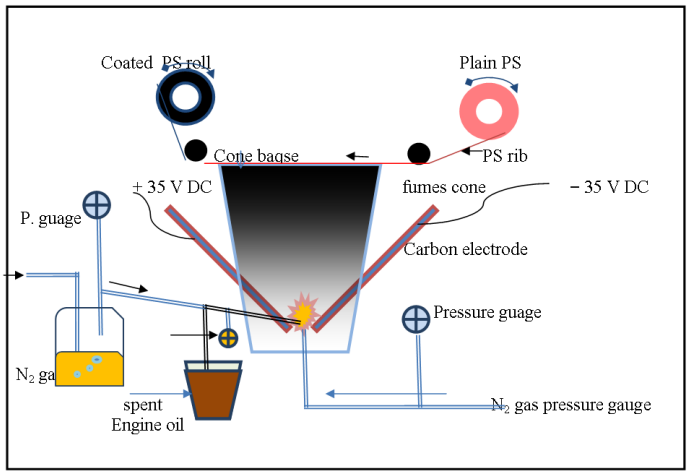
This work depicts the preparation of Ni-C ultrafine composite from used engine oil. The used oil was emulsified with detergent loaded with Ni $(OH)_2$. The loaded emulsion was sprayed on electric plasma generated between two C electrodes to a DC main 28 V and 70-80 A. The purged Ni-doped carbon fume was trapped on a polymer film moistened with synthetic adhesive to fix the trapped smoke. Characterization of the deposit was made using SEM. XRD examined the crystal morphology. Carbon density in the cloud was calculated. The average size and thickness of the deposited composite is 120-160 nm. Aliphatic hydrocarbons readily decompose to gaseous products. Solid carbon smoke originates from aromatic compounds. Plasma heat blasts the oil in short time to decompose in one step.

1. Introduction

Teng [1] reported the use of plasma arc technology to produce fluid of carbon ultrafine particles dispersed in distilled water. Carbon is simultaneously heated and vaporized in a chamber furnace. Carbon vapor was cooled to prepare a carbon/water ultrafine slurry. The problem of heat dissipation from electrical appliances became a significant issue. To ameliorate this problem, there are four approaches commonly taken [2] enlarged the heat exchanger area and structure, [3] fabricated the heat exchanger using materials with higher thermal conductivity, [4] increased the working fluid flow rate to the heat exchanger, and [5] improved the heat transfer performance of the heat exchange working fluid. *Akoh* [6] prepared superfine fluids using the VEROS method. Teng [7] improved the VEROS technique and developed successful and effective preparation of Ag, Fe fine fluids. Bo [8] showed that plasma-enhanced chemical vapour deposition (PEVCD) has emerged as a key method to generate vapor-gas; and determine the development of VG with required properties. The paper summarized the state-of-the-art research on PECVD growth of VG ultrafine sheets. It provided guidelines on the design of plasma sources and operation parameters. Also to offer a future good challenges that would help commercial applications of VG. The authors studied the influence of feedstock gas temperature and pressure on VG growth. The effect of pretreatment and the growth of VG patterns on cylindrical and carbon ultrafine tube (CNT) substrates was shown. After discussion conclusion showed challenges and future directions for PEVCD growth of VG.

In the year 2014, defects-free graphene and with highest electron mobility was successfully prepared. Mayer [9] used adhesive tape to separate graphene sheets from graphite. It was reported that hydrazine followed by annealing in argon/hydrogen gases was successful to reduce graphite oxide. It was shown that almost intact carbon framework allowed efficient removal of functional groups. Some molten salts attack graphite particles to form a variety of carbon ultrafine structures including graphene [10]. Hydrogen cations, dissolved in molten lithium chloride, were discharged and intercalated and peeled to graphene sheets [11]. The graphene ultrafine sheets produced displayed a single-crystalline structure with a lateral size of several hundred ultrafine meters and a high degree of crystalline and thermal stability [12]. In 2014 defect-free, not oxidized graphene-containing liquids were made from graphite using mixers that produce local shear rates greater than 10×10^4 [13]. Mubarak [14] reported that the chemical vapor deposition (CVD) had been found a potential way to produce high quality and at high yield of CNTs. El Mel et al., [15]

*Corresponding Author: Mahmoud A. Rabah, Electrochemical Treatment Dept., Mineral Processing Div. Central metallurgical Research and Development Institute (CMRDI), Egypt | Email: mrabah010@gmail.com



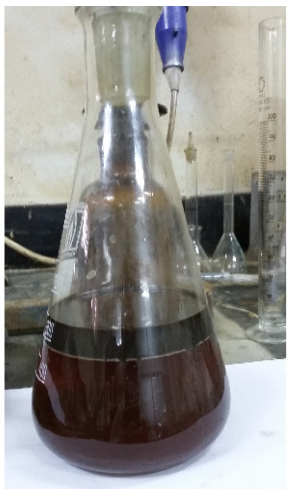
Industrial nitrogen gas was used as a propellant to spray the emulsified oil on the plasma. The gas was supplied in cylinder with the following specifications

Nitrogen and trace inserts	99.998%
Oxygen	< 0.001%
Density (g cm^{-3})	1.2506
Sp. Gravity (g.cm^{-3})	0.9737
Viscosity (g $cm^{-1}s^{-1}$)	$1.809x10^{-4}$

The chemicals used are technical pure grade. Nickel hydroxide salt was synthetically prepared by precipitation by adding 10% ammonium hydroxide to nickel sulfate solution. The precipitated $Ni(OH)_2$ was filtered, washed and dried.

2. Emulsification of the used engine oil

About 100 g of the used engine oil was placed in a conical flask of a thermostated agitator maintained at 80°C. The flask was fitted with an electrical mechanical stirrer. After the oil has acquired the required temperature, sodium sulphonate detergent containing nickel hydroxide suspended in distilled water was added step wise while stirring for 4 hours. The emulsion was then transferred to a separating funnel and left over night.



3. Method of preparation of the Ni-doped carbon ultrafine particles.

Determination of the extent of engine oil emulsification was determined by weighing the separated portion of the not emulsified oil. Fig. 2 shows partial (2a) and full (2b) emulsification of the used oil. Fig. 2 her

4. Methods of measurement of the physic-chemical properties

Determination of nickel in the oil suspension or in the prepared composite was carried out by dissolving 1 g of the oil or 250 mg of the catalyst in 1 N nitric acid. Nickel content was determined using spectrophotometer Unicom 7400 Fig.3 shows a schematic diagram of the Electric Plasma device used for thermal decomposition of the engine oil. Particle size and shape of the product were determined using the light-scattering size analyzer, SEM. Crystal morphology was examined by X ray diffraction (XRD).

5. Results

The specifications of the engine oil are given in Table 1. It can be seen that the spent oil has lower viscosity, foam and gelation index and higher phosphorus content.

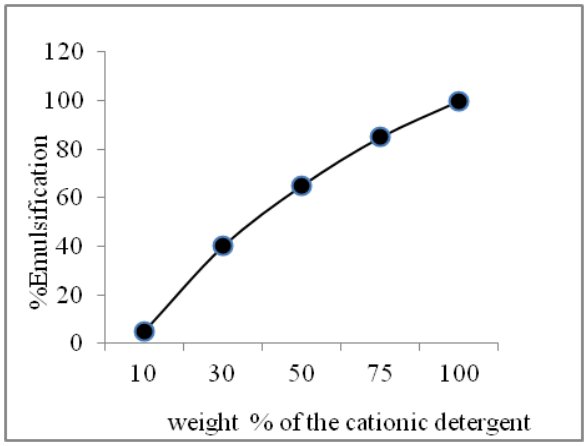


Fig. 4 shows the extent of emulsification of the spent engine oil (SEO) as affected by the weight of commercial anionic detergents containing 10 % Ni as $Ni(OH)_2$. It is seen that the extent of emulsification of the spent engine oil increases linearly with the corresponding increase of the detergent. Anionic detergent is far efficient to emulsify engine oil as compared to the cationic detergent to emulsify the spent engine oil.

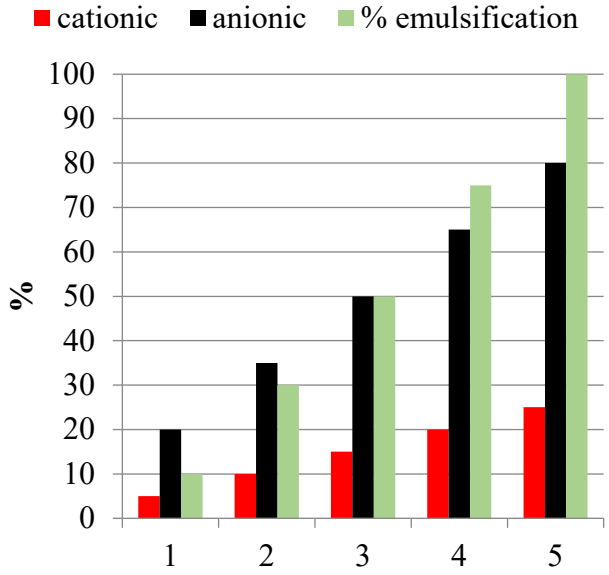


Fig 5 Effect of detergent type on the extent of emulsification of spent engine oil.

Table 1. The specifications of the engine oil used in this study

	API. SN		API.SN + RC
Viscosity Grade	0W20, 0W30, **5W20, 5W30**, 10W30 **0W8** **(spent)**	Others	Multigrade variations of 0W, 5W and 10W
Sculpture, max %	0.5 **0.45**	NA	0.5
High temp. deposits, max mg	35 **46**	45	35
Foam	1 min settling **<1 min**	10 min. settling	1 min. settling
Gelation Index	12 **6**	NR	12
Emulsion Retention	NA NA	NA	No Water Separation
P, max% / min%	0.08/0.06 **0.06**	NA / 0.06	0.08 / 0.06

Table 2 shows a comparative data on the effect of mass of anionic and cationic polymers as a detergent for emulsifying the used engine oil. It is shown that anionic polymer is more power

Table 2 The effect of mass of the polymer on the emulsification extent of Em. at room temperature

% emulsification	Anionic polymer	Cationic polymer
20	18	20
40	30	30
60	38	50
80	50	72
100	60	100

Fig. 6 shows the rate of discharge of oil from the spraying nozzle as affected by the pressure of propellant nitrogen gas. It is seen that convenient rate of discharge amounting to 2.5-3.0 g/min is achieved with 3 – 3.5 KPa. Such rate was found suitable to generate and decompose the Em. together with the Ns solutioFig. 5 Effect of the pressure of the propellant nitrogen gas on the rate of discharge of the emulsified engine oil

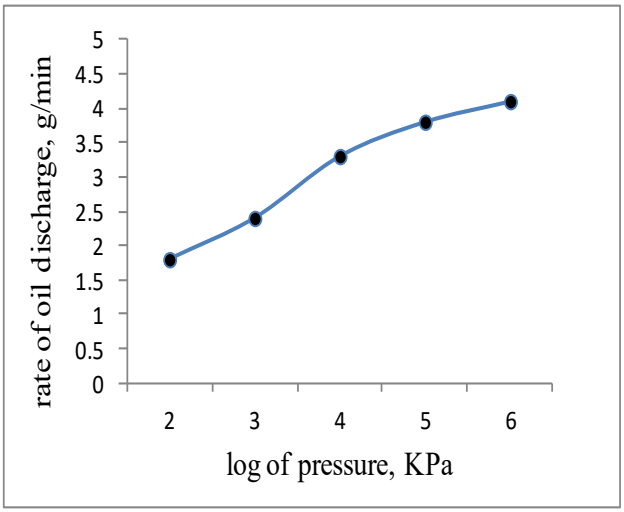


Fig. 6 The rate of the oil discharge from the spraying nozzle as affected by the propellant nitrogen gas

Fig. 7 shows a photograph of the electric plasma generated between the carbon electrodes.

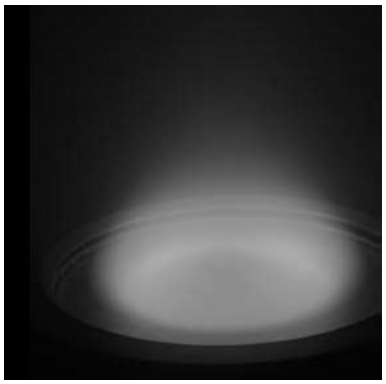


Fig. 7 A photograph of t electric plasma exerted between two carbon electrodes

Table 3 The carbon density in the plasma hood

No.	Oilg/ min.	∑C, mg/min	H, mm	r, mm at H, mm			$\hat{r}^2$ mm^2	V, cm^3	C_c, mg/ cm^3/min	Cc, ug/ cm^2/min
				H_o	H_1	$\hat{r}$				
1	1	275	100	10	50	30	900	94.247	2.917	297
				20	60	40	1600	167.5	1.642	164
				20	70	45	2025	212.0	1.297	129
				25	80	55	3025	318.2	0.864	86.4
				30	90	60	3600	375.04	**0.7332**	**73.3**
2	1.5	412.5	175	10	50	30	900	164.0	2.515	251
				20	70	45	2025	369.2	1.1172	111.7
				30	90	60	3600	656.3	0.6285	**62.85**
3	2	550	200	10	50	30	900	187.5	2.933	293.3
				20	70	45	2025	421.9	1.303	130.3
				30	90	60	3600	750.0	**0.7333**	**73.33**
4	2.5	687.5	250	10	50	30	900	234.4	2.933	293.3
				20	70	45	2025	527.4	1.3035	130.35
				30	90	60	3600	937.6	**0.7339**	**73.39**
5	3	825	300	10	50	30	900	281.3	2.933	293.3
				20	70	45	2025	632.9	1303.5	130.3
				30	90	60	3600	1125.13	**0.7332**	**73.32**

Table 3 shows the mean value of the weight of carbon particles disseminated in the cone of a plasma heat as a function of the weight of inserted oil and the dimensions of the cone.

Fig.8 shows the XRD diffraction patterns of nickel-carbon composite species It is seen that the diffraction peaks of Ni confirm the existence of well crystalline Ni metal. Raman spectrum of Ni doping C appeared with I_D/I_G=0.63

Fig. 9 shows the weight of the deposited carbon -Ni composite on the polymer film as a function of the weight of the sprayed engine oil as a function of the weight of the sprayed emulsified engine oil.

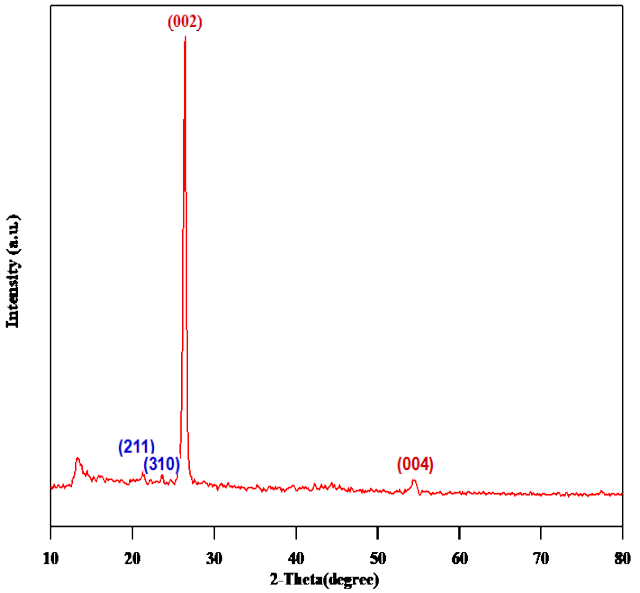


Fig. 8 The XRD diffraction patterns of nickel-carbon composite species It

It can be seen that the weight of the deposited C-Ni amounts to 10 ug/cm2 and increases gradually with the increase of the weight of sprayed engine oil up to 350 mg whereby the maximum WCNIO is attained (45.5 ug/cm2). Fig. 10 shows the SEM of the deposited C-Ni composite material.

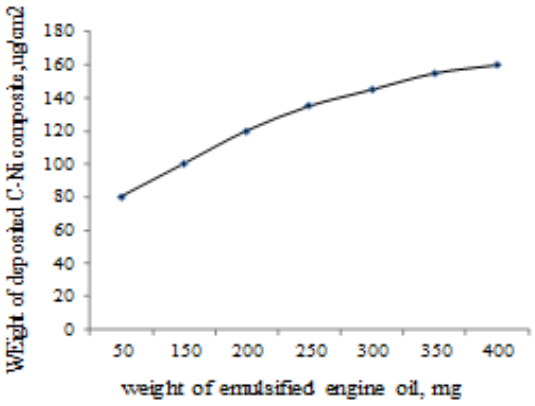


Fig. 9 The weight of the deposited C-NiO (W_{CNiO})

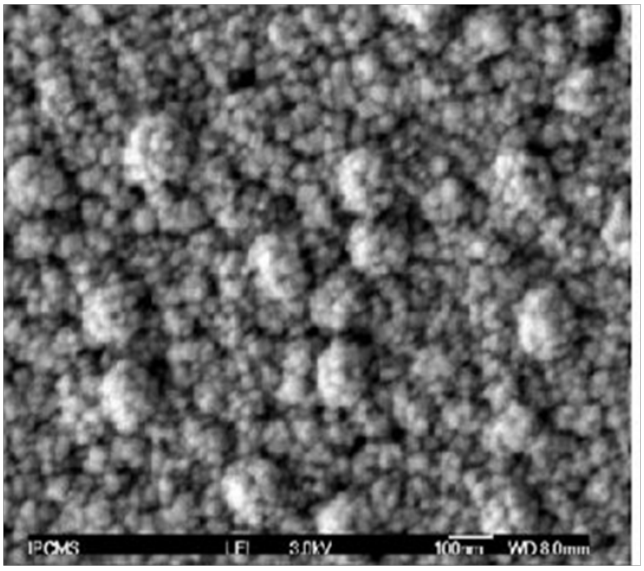


Fig. 10 SEM photograph of the deposited C—Ni composite

6. Mathematical Model

A mathematical model has been proposed to simulate the formation of C-Ni particles in the hood of the plasma heat vicinity based on the following assumptions:

1. The chemical analysis of the used engine oil (O_u) and the projected plasma heat at the hood vicinity (hv) are kept constant throughout the experiments.
2. Solid carbon particles generate at a constant weight percentage of 27.5 % of the weight of the oil as determined experimentally.
3. The weight of carbon generates in nearly no time; $Wt_c = 0.275\ Wt_o$.
4. Ejection of the emulsified engine oil takes place in the plasma core at a constant discharge rate for a given driving pressure.
5. No extra air was fed into the hood vicinity of the plasma arc
6. The carbon particles dissemination is distributed symmetrically at a given horizontal level in the hood.

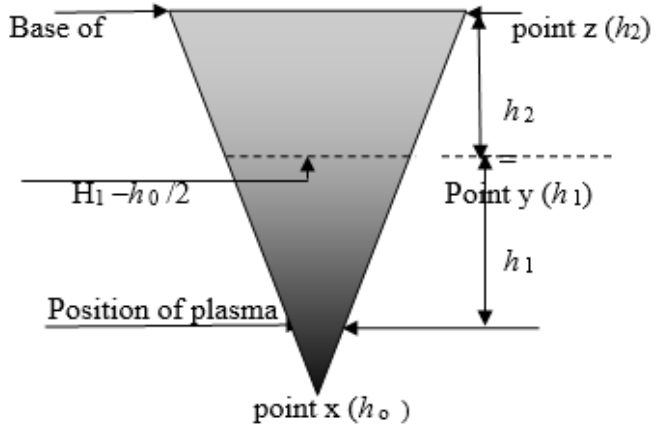


Fig. 11 A Schematic proposal of the caron-Ni smoke in the blasma hook.

The average concentration of C-Ni / cm^3 in the hood at a given horizontal level H_x =

$$\frac{0.275\ (Wt_o)_p}{V_{Hx}} \qquad (1)$$

Volume V of a right cone (hood) with a height H,

$$V = 1/3\ \pi\ r^2 H_x \qquad (2)$$

Wt_o is the weight of discharged oil at constant pressure p.
The average C-Ni concentration/ cm3 in the hood:

$$\frac{0.275\ (Wt_o)_p}{1/3\ \pi\ r\ h_x} \qquad (3)$$

Carbon particles disseminate in the hood volume under a convection model of heat transfer. Convection above a hot surface occurs because hot air expands and becomes less dense and rises upwards. Convective transfer involves the combined processes of diffusion (conduction) and advection transfer by bulk fluid flow. Natural buoyancy forces alone are entirely responsible for the C-Ni fluid motion upon heating the mass transfer of carbon.

Where k is a transfer coefficient, A is the total surface area of the carbon particles, T_p is the temperature at the plasma core and T_{Hx} is the temperature at the level Hx in the hood. The convective transfer coefficient is dependent upon the physical properties of the fluid (here air). Values of h have been measured and tabulated in the literature for commonly encountered fluids and flow of situations [18]. The mathematical presentation model of the release, development and change of air-born solid fluid (C-Ni particles) as a function of variables such as time, temperature, pressure, distance, and concentration has been processed using "the experimental design" program. Results are under publication elsewhere

$$\bar{Q}c = kA\ (TT_{Hx}) \ldots \ldots (4)$$

Temperature gradient of the plasma heat vicinity comply the curve given in Fig. 12

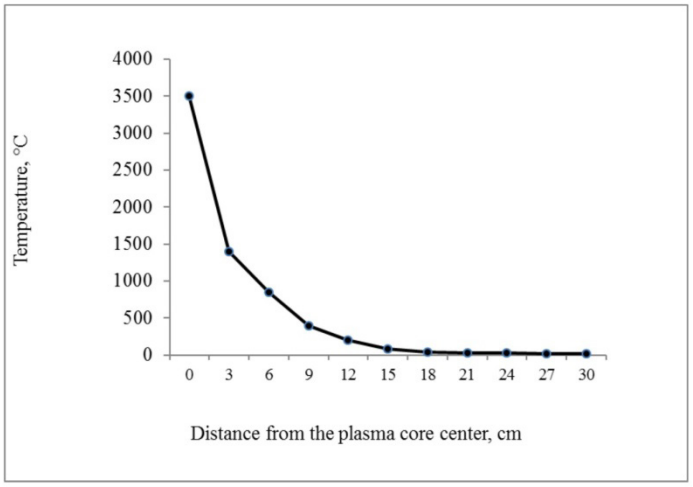


Fig. 12 Temperature gradient of blasma heat as a function of distance

7. Discussion

Engine oil is composed of hydrocarbons, poly alpha olefins (PAO), and poly internal olefins(PIO). The organic compounds consisting entirely of carbon and hydrogen [17]. The oil contains 73-80% weight/weight aliphatic hydrocarbons (primarily alkanes and cycloalkanes with l-6 rings); 11-15% monoaromatic hydrocarbons; 2-5% diaromatic hydrocarbons; and 4-8% polyaromatic hydrocarbons. Additives for lubrication, (≈ 20% of the oil), consist primarily of zinc diaryl, molybdenum disulfide, zinc dithiophosphate, metal soaps, and other organo metallic compounds. The Canadian Environmental Protection Agency reported that up to 15% of the additives are detergents and dispersants [18]. Other used mineral-based crankcase oil may contain compounds of barium, phosphorus, zinc, and some chlorine and bromine as a result of oil additives [18]. Inorganic compounds of sulfur, aluminum, arsenic, calcium, chromium, copper, iron, magnesium, manganese, potassium, silicon, sodium, tin are detected. Organic solvents of toluene, benzene, xylene,and ethylbenzene are present although most of the organic compounds may be destroyed during combustion of crankcase oil as fuel [19]. In general, sources of dirt, sand and dust from the air, soot, contaminate unburned fuel in the motor oil. Water if detected is formed from condensation of the combustion products. Particulates of metal wear that the oil filter

cannot trap together with the corrosion by-products and degraded additive elements [20, 21]. All these contaminants enter the engine oil and creat more wear debris and cause more damage than they would separately.

Heating engine oil at high temperatures undergoes a decrease in viscosity grade and degradation reactions of multi sequence order. Decomposition begins with the formation of compounds having more simple molecular chemical formulae. According to the infrared spectra, the paraffinic-naphthenic fractions of new and used oils are substantially similar to each other. The aromatic fractions of the used oil contain oxygen groups that can be removed only by catalytic hydro treating [22]. The aromatic fractions contain oxygen groups detrimental to the thermal-oxidative stability of the oil. [23]. The decomposition process proceeds in a multi sequence order to finally produce the free elements of the engine oil (C, H, O). In 1959, Chevron developed a new method of refining base oils (Hydro cracking) where many of the paraffin and wax molecules are broken up into mineral oil molecules. The process removed the aromatics, sulphur and nitrogen compounds from the oil [22,23].

In 1993, Chevron invented the Hydro-Isomerization process, where wax and paraffin molecules are reshaped into useful lubricants instead of simply being broken up into smaller molecules. In 2004 accounts for almost half of all base oils [22]. Iso-dewaxing produces much higher-performance oil and allowed to start with low quality oil. In addition, oil manufacturers add what's called an additive package [23]. Additive packages contain several different chemicals with different purposes.

Results of weight loss upon charring of oil emulsion revealed that NiO may catalyze one or more of the decomposition reactions. The temperature of the electric plasma exceeds some thousands and it promotes decomposition reactions to complete to end in one step in nearly no time. Generated carbon fumes, doped with nickel, deposit on a cold surface of polyurethane film positioned over the spark zone. Frequent deposition of C-Ni composite on the polymer surface would form a layer with measurable thickness. Thickness development is a time dependent process. The particle size and shape are critical for to build up the composite layer. Characterization of the crystal structure of the deposit was carried out by SEM. Crystal morphology was examined by XRD. Results revealed that the thickness of the deposited composite film was few microns. The average size of the crystals of composite ultrafine material ranged from 50-90 nanometer whereas thickness of the layer was 120 – 160 nm depending on the experimental conditions. Finer size was obtained with a rate of oil discharge of 0.5 g/min. It increases with the increase of the oil discharged rate ≥ 5 g oil/min. Results were explained in the premise of a model which assumes that engine oil decomposed in a multistep reactions ending with the formation of ultrafine carbon fumes doped with nickel vapor. The weight % of the solid fumes generated from decomposition process of the engine oil amounts to 65%. The mathematical model given in page 3 estimates the hypothetical weight of carbon disseminated in the experimental test rig. It was studied based on multi sequence reactions. Results will be published later on.

8. Conclusion

The output results conclude the following.

1. Used engine oil is a useful material that can be utilized to prepare C-Ni composite.
2. The plasma heat is an effective heat source to decompose the used engine oil.
3. Decomposition process is a multi-sequential reaction involving generation of solid carbon particles and gaseous products.
4. Plasma decomposition blasts oil to undergo decomposition in nearly no time and in less decomposition reactions.
5. The particle size of the produced carbon -Ni was 50-90 nm whereas the thickness of the formed layer amounts to 120-160 nm
6. Doping of carbon ultrafine particles with Ni has been successfully carried out by loading the oil emulsion with Ni Hydroxide.

Acknowledgment

The authors are thankful to Dr. Morsi Ameen and Eng. Usama Khedr, Welding Department, CMRDI, for their great help and support.

References

[1] T.P.Teng, C.M. Cheng and F.Y. Pai,. Preparation and characterization of carbon ultrafine fluid by a plasma arc ultrafine particles synthesis system". Nanoscalescale Res. Lett.,6(1), 293.2011

[2] G. Corsico, L. Mattei, A. Roselli and C. Gommellini. "Poly(internal olefins)- Synthetic Lubricants and high-performance functional fluids". Marcel Dekker, Chapter 2, p. 53-62, C. SUS, Enhancing thermal conductivity of fluids with ultrafine particles. 1999.

[3] L. D. Chang, C.M. Mou. "Nanomaterials and Nanostructure". Peking: Science Press.2001

[4] Y. Xuan and Q. Li "Heat transfer enhancement of ultrafine fluids". Int. J. Heat Fluid Flow, 21: 58. 2000.

[5] I.A. Eastman,,C. SUS, S. Li, W.Yu and L.J. Thompson. "Anomalously increased effective thermal conductivities of ethylene glycol-based ultrafine fluids containing copper ultrafine particles". Appl. Phy. Lett., 78: 718, 2001.

[6] H. Akoh, Y. Tsukasaki, S. Yatsuya and A. Tasaki "Magnetic properties of ferromagnetic ultrafine particles prepared by vacuum evaporation on running oil substrate". J. Cryst. Growth, 45: 495. 10.1016/0022-0248(78)90482-7 1978

[7] T.P.Teng, C.M. Cheng and F.Y. Pai."Preparation and characterization of carbon nanofluid by a plasma arc nanoparticles synthesis system". Nanoscale Res Lett.; 6(1): 293. 2011.

[8] Bo. Z., Yong Y., Junhong C., Yu K., Jianhua Y. and Kefa C., 2013. Plasma-enhanced chemical vapour deposition synthesis of vertically oriented graphene penosheets. Ultrafinescale, 5, 5180-5204.

[9] J. Meyer, A.K. Geim, M. Katsnelson, M.I. Novoselov, K.S. Booth and T. J. Roth." The structure of suspended graphene sheets". Nature 446(7131):60–63., 2007.

[10] Eigler S., Enzelberger H. M., Grimm S., Hofmann P., Kroener W., Geworski A., Dotzer C., Röckert M., Xiao J., Papp C., Lytken O., Steinrück H. P., Müller F., Hirsch A., 2013. Wet Chemical Synthesis of Graphene. Advanced Materials 25 (26): 3583–3587

[11] A.R. Kamali and D.I. Fray., "Molten 0salt corrosion of graphite as a possible way to make carbon ultrafine structures". Carbon, 56, 121–131. 2013.

[12] A.R. Kamali and D.I.Fray, "Large-scale preparation of graphene by high temperature insertion of hydrogen into graphite". Nanoscale, 7, 11310–11320. 2015.

[13] R. Keith R., "Scalable production of large quantities of defect-free few-layer graphene by shear exfoliation in liquids" Nature Materials, 13 (6): 624–630. 2014.

[14] N.M. Mubarak, E.C. Abdullah, N.S. Jayakumar and J.N. Sahu J.N., An overview on methods for the production of carbon ultrafine tubes. Journal of Industrial and Engineering Chemistry 20, (4) 25, 1186–1197 2014.

[15] A.A. El Mel, E. Gautron, B. Angleraud, A. Granie and P.V. Tessier "Synthesis of nickel-filled carbon ultrafinetubes at 350 °C". Carbon Volume 49, Issue 13, Pages 4595–4598, 2011.

[16] S.,Abu Bakar, S. Alfarisa, A. Mohamed, I.M. Isa, A. Kamari, N. Hashim,, M.H. Mamat, M. Ruson. and A.R. Mohamed, "Quasi-aligned carbon nanotubes synthesized from waste engine oil". Material Letters, 139. pp 220-223. 2015.

[17] D.R. Vazquez "Environmental impact of used motor oil. Sci" Total Environment, 79, 1-23. 1989.

[18] Canadian Environmental Protection Agency 1994; Vermont Agency of Natural Resources 1994.

[19] A.Shenoy, M. Sheremet, P. Ioan,."Convective Flow and Heat Transfer from Wavy Surfaces: Viscous Fluids, Porous Media, and Nanofluids", CRC Press, Taylor & Francis Group, Florida ISBN 978-1-498-76090-4, 2016.

[20] K. Dieter, R.R. Rost,."Lubricants and Related Products", Verlag Chemie. ISBN 0-89573-177-0. American Petroleum Institute 2016. 1984

[21] P.L. Fan, L. Fei, Y. Xuecheng, I. Gu, Y.Z. Zhong. S. H. G., Qiu. and Y. Xiangdong,."Atomically isolated nickel species anchored on graphitized carbon for efficient hydrogen evolution electro catalysis". Nature Communication **7**, Article number:10667, 2016.

[22] V.A. Litvishkova,, A.I. Bukhter, A.V.Nepogod'ev and A.M. Bezhanidze, "Chemistry and Technology " Fuels and Oils. Vol. 10, Issue 12, pp 962–965, 1974.

[23] M. Lawrence, "All About Oil". http://www.calsci.com/motorcycleinfo/Oils1.html.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 2, No. 3, 702-710 (2017)
www.astesj.com
Special Issue on Recent Advances in Engineering Systems

ASTES Journal
ISSN: 2415-6698

Schizophrenia Prediction Using Integrated Imaging Genomic Networks

Su-Ping Deng[1], Wenxing Hu[1], Vince D. Calhoun[2], Yu-Ping Wang[*,1],

[1]*Tulane University, Department of Biomedical Engineering, School of Science and Engineering, 70118, USA*

[2]*Mind Research Network, 87106, USA*

ARTICLEINFO

Article history:
Received: 06 April, 2017
Accepted: 12 May, 2017
Online: 14 June, 2017

Keywords :
Classification
Data Integration
Network Fusion
Imaging Genomic Data

ABSTRACT

In order to increase the diagnosis accuracy of schizophrenia (SCZ) disease, it is essential to comprehensively employ complementary information from multiple types of data. It is well known that a network is a general method for analyzing relationships between patients, with its nodes representing patients and its edges showing relationships between them. In this study, we constructed a fused network using three types of data including genetic, epigenetic and neuroimaging data from a study of schizophrenia patients. We developed a network-based prediction approach taking advantage of the whole network of patients rather than just individual clusters in the network. The majority neighborhood of a node in the network was exploited as a feature for discriminating SCZ from healthy controls. Compared with other 9 graph-based label prediction methods, our network-fusion based label prediction method shows the best performance according to the prediction accuracy. The prediction power of our proposed method was also tested under different parameters settings and an optimal parameter was found for achieving the best performance. The method is also computationally efficient and can be extended to identify other clinical outcomes.

1 Introduction

This paper is an extension of work originally presented in the IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2016) [1].

Currently, there is a lack of objective ways for the diagnosis of schizophrenia (SCZ) [2]. It is generally diagnosed according to criteria set in either the American Psychiatric Associations fifth edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM 5), or the World Health Organizations International Statistical Classification of Diseases and Related Health Problems (ICD-10). These criteria use the self-reported experiences of the patients and report abnormalities in their behaviors, followed by a clinical assessment by a mental health professional. However, symptoms associated with schizophrenia occur along a continuum in the population and must reach a certain severity before a diagnosis is made [3]. Although in psychiatry as in all of general medicine there is an irreducible element of the subjective [4], we should try our best to reduce the subjective part of medical and psychiatry practice. This goal is well recognized by the National Institutes of Mental Health as well [5].

Like many other complex diseases, SCZ is caused by a combination of genetic, biological, and environmental factors, creating a difficult challenge for diagnosis and defining subtypes. In addition, SCZ is a highly heritable disorder [6]. The current belief is that there are a number of genes that contribute to susceptibility or pathology of SCZ, but none of them exhibits full responsibility for the disease. It is believed that SCZ is caused by a number of genetic and environmental factors [7].

Recently, high-throughput studies are rapidly accumulating a wealth of 'omics'-scale data, which enable comprehensive monitoring of a biological system. Meanwhile, there are remarkable medical images accumulated in hospitals, such as images from functional magnetic resonance imaging (fMRI). fMRI

[*]Yu-Ping Wang, 535 Lindy Boggs Bldg., New Orleans, LA, +1(504) 865-5867 & wyp@tulane.edu

https://dx.doi.org/10.25046/aj020390

is a fairly new tool that has been used to measure brain activation utilising the dependency of the magnetic properties of haemoglobin on the amount of oxygen it carries. Blood-oxygen-level-dependent (BOLD) signals measure the alterations in cerebral blood flow that mark functional brain activity [8]. The intrinsic BOLD contrast makes fMRI an invaluable non-invasive instrument for the investigation of the underlying physiological disturbances that lead to manifest psychiatric disorders. It is anticipated that successful application and analysis of fMRI in neurological disorders can be used to characterise and diagnose mental illnesses such as Alzheimers disease, SCZ, bipolar disorder, mild traumatic brain injury and addiction.
It is a valuable way to combine genetic, epigenetic and imaging data for the diagnosis of SCZ disease. Many studies have demonstrated that the integration of fMRI and SNP information will give a more comprehensive analysis of schizophrenia. For example, Yang et al. [9] proposed a hybrid machine learning method to identify schizophrenia patients from healthy controls by combining genetic and fMRI data, and this method achieved better classification accuracy than using either data alone. Castro et al. [10] developed a multiple kernel learning approach that employed both the phase and the magnitude of complex-valued fMRI data for the classification, showing improved classification accuracy compared with using only the magnitude of fMRI data. Cao et al. proposed a sparse representation clustering (SRC) model [11, 12] to simultaneously select SNPs and fMRI voxels as biomarkers for schizophrenia followed by a classifier such as support vector machine (SVM). The integrative model was tested on 20 SCZ patients and 20 healthy controls, demonstrating that the complementary information from both fMRI and SNP data can be utilized to improve diagnosis [13].

The above mentioned SCZ studies, however, only used two types of data: SNPs and fMRI data. DNA methylation is one of several epigenetic mechanisms and can be used by cells to control gene expression. There are a number of mechanisms controlling gene expression in eukaryotes, but DNA methylation is a frequently used epigenetic signaling tool, which can fix genes in the "off " position [14]. In addition, the integration methods used in the above mentioned research directly used features extracted from data as predictors. Alternatively, network-based classification methods mainly employ the topological features as predictors. With the rapid development of systems biology, network-based classification or prediction methods have been widely used [15, 16, 17, 18, 19, 20, 21].

In our study, we combined three types of data including SNPs, DNA methylation and fMRI to construct networks, which were subsequently be fused for diagnosis of SCZ. Specifically, for each type of data, we first constructed a patient network, called a single network. Thus for three types of data, we constructed three single networks. Then we fused two or three single patient networks into one subject network. As a result, we created four fused networks: three fused networks from the combinations of two data types and one fused network from three data types. Finally, we employed the neighborhood majority of the nodes in the network to predict unknown labels, that is, identifying whether a subject is a schizophrenia patient.

2 Data Preparation

In this study, participant recruitment and data collection were conducted at the Mind Research Network. Three types of data (SNP, DNA methylation and fMRI) were collected from 208 subjects including 96 schizophrenia patients (age: 34 ± 11, 22 females) and 112 healthy controls (age: 32 ± 11, 44 females). All of them were provided written informed consents. Healthy participants were free of any medical, neurological or psychiatric illnesses and had no history of substance abuse. By the clinical interview of patients for DSM IV-TR Disorders or the Comprehensive Assessment of Symptoms and History, patients met criteria for DSM-IV-TR schizophrenia. Antipsychotic history was collected as part of the psychiatric assessment. After a series of quality controls, we selected 184 subjects, including 80 SZ cases (age: 34 ± 11, 20 females and 60 males) and 104 healthy controls (age:32 ± 11, 38 females and 66 males). After pre-processing, 27,508 DNA methylation sites, 41,s236 fMRI voxels and 722,177 SNP loci were obtained for the subsequent biomarker selections [22, 23, 24, 25, 26, 27].

2.1 SNPs data collection and preprocessing

A blood sample was obtained for each participant and DNA was extracted. Genotyping for all participants was performed at the Mind Research Network using the Illumina Infinium HumanOmni1-Quad assay covering 1,140,419 SNP loci. Bead Studio was used to make the final genotype calls. Next, the PLINK software package(http://pngu.mgh.harvard.edu/~purcell/plink) was used to perform a series of standard quality control procedures, resulting in the final dataset spanning 722,177 SNP loci. Each SNP was categorized into three clusters based on their genotype and was represented with discrete numbers: 0 for 'BB'(no minor allele), 1 for 'AB' (one minor allele) and 2 for 'AA' (two minor alleles).

2.2 DNA methylation data collection and preprocessing

DNA from blood samples was assessed by the Illumina Infinium Methylation27 Assay. A methylation value, beta (β), represents the ratio of the methylated

probe intensity to the total probe intensity. A series of quality controls (QC) on the beta values were applied to remove bad samples and probes, such as 1) Beta value QC: Change any beta value to NaN, if p>0.05. 2) Bad sample/ bad marker removing: Samples with >5% of missing (NaN) values; markers with >5% of missing (NaN) values. This resulted in the identification of good methylation data from 224 subjects, 27,508 markers (some have missing values <5%). After QC, we used the K nearest neighbor (KNN) method [28]to impute for the missing values.

2.3 fMRI data collection and preprocessing

The fMRI data were collected during a sensorimotor task, a block-design motor response to auditory stimulation. During the on-block, 200msec tones were presented with a 500 msec stimulus onset asynchrony (SOA). A total of 16 different tones were presented in each on-block, with frequency ranging from 236 Hz to 1318 Hz. The fMRI images were acquired on Siemens 3T Trio Scanners and a 1.5T Sonata with echo-planar imaging (EPI) sequences using the following parameters (TR = 2000 msec, TE = 30 msec (3.0T)/40 msec (1.5T), field of view = 22cm, slice thickness = 4mm, 1mm skip, 27 slices, acquisition matrix = 64×64, flip angle =90°.). Data were pre-processed in SPM5 (http://www.fil.ion.ucl.ac.uk/spm) and were realigned, spatially normalized and resliced to $3 \times 3 \times 3mm^3$, smoothed with a $10 \times 10 \times 10mm^3$ Gaussian kernel to reduce spatial noise, and analyzed by multiple regression considering the stimulus and their temporal derivatives plus an intercept term as regressors. Finally the stimulus-on versus stimulus-off contrast images were extracted with $53 \times 63 \times 46$ voxels and all the voxels with missing measurements were excluded.

3 Methods

3.1 Overview

We employed a network-based semi-supervised learning (SSL) scheme, which improves the predictive power by using unlabeled data [29, 30, 31, 32]. It is compuationally efficient and the learning time depends nearly linearly on the number of network edges. Also, the accuracy is comparable with other methods such as kernel-based methods with a longer learning time [18, 33]. Moreover, the network structure can facilitate the interpretation of gene-gene interactions and/or region- region connections in the brain [34, 35, 36], which is one of the advantages of network-based SSL.

Based on a fused network, our label prediction method mainly consists of three procedures: network construction, network fusion and label prediction using network-derived features.

3.2 Network Construction

Based on the features selected from SNPs, DNA methylation and fMRI data, a subject network can be constructed. Suppose we have n samples and m measurements (e.g., DNA methylation). A subject similarity network is represented as a graph $G = (V, E)$. The vertices V correspond to the subjects $\{x_1, x_2, ..., x_n\}$ and the strengths of the edges E are the weighted value of the similarity between subjects. The edge weights are represented by an $n \times n$ similarity matrix$\mathbf{W}$, with each W_{ij} indicating the similarity between subjects x_i and x_j. $\rho_(x_i, x_j)$ is represented as the Euclidean distance between subjects x_i and x_j. In order to calculate edge strength, a Gaussian function was taken on the Euclidean distance between subjects:

$$W_{ij} = \begin{cases} exp(-\frac{\rho(x_i,x_j)^2}{\sigma^2}), & \text{if } i \sim j \\ 0. & \text{otherwise} \end{cases} \quad (1)$$

If i is in $j^{'}$ s K-nearest-neighborhood and vice versa, nodes i, j can be connected by an edge. In our study, we constructed three individual networks for SNPs, DNA methylation, and fMRI data.

3.3 Network Fusion

We fuse multiple networks into one network, using mean fusion in our study. Multiple networks, denoted by $G^m = (V^m, E^m, w^m), m = 1, 2, 3, \ldots, N$, are constructed from multiple types of data. They have the same network nodes and the edge strength in the fused network is the mean value of edge strength of all individual networks.

$$G = (V, E, w),\ with\ V = \cup_m V^m, E = \cup_m E^m,$$
$$and\ w(i,j) = \frac{1}{N} \sum_{m=1}^{N} w^m(i,j). \quad (2)$$

Using this graph fusion method, we fused two or three single patient networks into one subject network. As a result, we created four fused networks: three fused networks from the combinations of two data types and one fused network for three data types.

3.4 Label Prediction Based on Neighborhood Majority

A network is a map of interactions, with the links measuring the association between objects or subunits. A labeled subject is marked either by 0 or 1, representing two possible labels: healthy or disease, respectively. The edges have an essential role in influencing the propagation between the subjects to predict the true label of the unknown subject. We assumed that the labels of two subjects are more likely to be the same if the two subjects are more closely related to each other. Therefore, the labels can be predicted based on the similarities between subjects

with their genomic profiles and/or imaging features. Edges in network-based SSL represent the similarities between subjects that can be extracted from different data including SNPs, DNA methylation and fMRI data (Figure).

Network-based classification methods can be classified into two types [21]: direct annotation schemes, which infer the class label of a node based on its connections in the network, and module-assisted schemes, which first identify modules of related nodes and then annotate each module based on the known labels of its members. Neighborhood counting is a simple and direct method for label prediction, which determines the label of a subject based on the known label of subjects lying in its immediate neighborhood. Here, we used neighborhood majority to predict the unknown labels. We first calculated the neighborhood majority of a query subject. Then, the subject is assigned the same type of label as the one of its immediate neighbors with greatest neighborhood majority (Figure).

For a query subject i, the majority of its immediate neighbors can be defined as:

$$M_k = \sum_{j=1}^{N_k} W_{ij} \tag{3}$$

where k represents subject class index and N_k is the number of subjects in class k. Then the label of subject i is defined as:

$$y(i) = y(arg\max_k M_k) \tag{4}$$

4 Results and Discussions

We employed a network-based classification method to predict whether new subjects are schizophrenia patients. After preprocessing the three types of collected data: SNPs, DNA methylation and function MRI data, we chose 184 subjects with all three types of data. We used 5-fold cross validation to evaluate our prediction method. A two-step feature selection method was applied to the labeled data sets. At first, for SNPs and DNA methylation data sets, we preselected the features based on the genes in KEGG pathways and this procedure yielded 14,875 SNPs and 6,935 DNA methylation sites. Then, for three types of data, we utilized the t-test to select the significantly differential features ($p \leq 0.01$) between normal and disease status. The preselected features were used to construct a subject-subject network. A single-type network was built for each data type. That is, three single-type networks were generated. Next, a network fusion method was used to build fused networks from two or three single-type networks. Thus we can get four fused networks: three fused networks from the pairwise combinations of two data types and one fused network from three data types. In our previous study [37], it has been shown that the technique of data fusion will improve the classification performance of network based classifiers. Therefore, in this paper, we focus on the study of the fused network from three types of data. Based on the fused network, we utilized nodes' features in a network to predict the label of a testing subject. This process is illustrated in Figure. Our prediction method is called MMN, which is a majority-neighborhood-based classification method by mean fusion.

We also compared the performance of our model with that of several other network based classification methods, with fixed parameter settings ($\sigma^2 = 510, and K = 20$). We compared 10 network based methods, whose differences consist in the procedure of network fusion and the algorithm of classification. For network fusion, two alternative methods are the mean fusion as used by our model, and Similarity Network Fusion (SNF) [38]. For classification procedure, we used five different classification methods: majority-neighborhood based classification (MN), which is adopted by our model, spectral clustering (SPC) [39, 40], support vector machine (SVM) [41], local and global consistency (LGC) [32], and label propagation (LP) [42]. SPC is a module-assisted classification approach based on spectral clustering, which is efficient in capturing global structure of a graph [43]. SVM utilizes 11 node features extracted from the network for label classification, and these 11 node features are: betweenness centrality, closeness centrality, degree centrality, Bonacich Power centrality scores, the (Harary) graph centrality, information centrality, the load centrality, the vertex prestige scores, the stress centrality, and clustering coefficients [44]. LGC utilizes both the local and the global consistency of a new patient in a patient-patient network for classification. LP predicts the clinical outcome of the new patient based on a network structure called label propagation. Thus, we have 10 different models, which are MMN (our model), SMN, MSPC, SSPC, MSVM, SSVM, MLGC, SLGC, MLP, and SLP, respectively. To evaluate the performances of these 10 network-based label prediction methods, we computed some widely used metrics, e.g., true positive rate (TPR), true negative rate (TNR), negative predictive value (NPV), false positive rate (FPR), false negative rate (FNR), false discovery rate (FDR), accuracy (ACC), and F1 (F1 = 2TP/2TP+FP+FN) of each method. The classification performances of all these models are listed in Table 1.

From these results it is evident that our model, MMN performs significantly better than other network based classification methods, in terms of some key metrics, e.g. ACC, FDR, etc. Also, it seems that those mean fusion based methods (MLP, MLGC, MMN, MSPC, MSVM) works slightly better than thef corresponding SNF based methods (SLP, SLGC, SMN, SSPC, SSVM), which demonstrates the advantage of mean fusion over SNF. In summary, mean fusion is more suitable than SNF for classification, and MMN works significantly better than other graph-based

Table 1: The performances of different graph-based label prediction methods

Methods	TPR	TNR	NPV	FPR	FNR	FDR	ACC
SLP	1	0	NA	1	0	0.568	0.432
MLP	1	0	NA	1	0	0.568	0.432
SLGC	1	0	NA	1	0	0.568	0.432
MLGC	1	0	NA	1	0	0.568	0.432
SMN	0	1	0.568	0	1	NA	0.568
MMN	0.588	0.756	0.698	0.244	0.412	0.332	0.686
SSPC	0	1	0.568	0	1	NA	0.568
MSPC	0.338	0.726	0.6	0.274	0.662	NA	0.556
SSVM	0	1	0.56	0	1	NA	0.56
MSVM	0.388	0.754	0.614	0.246	0.612	0.424	0.598

Table 2: The performances of our label prediction method MMN with different $\sigma^2 (K = 20)$

σ^2	TPR	SPC	NPV	FPR	FNR	FDR	ACC	F1
100	0.54	0.688	0.65	0.312	0.464	0.408	0.624	0.566
150	0.54	0.688	0.65	0.312	0.464	0.408	0.624	0.576
200	0.564	0.716	0.672	0.284	0.438	0.374	0.65	0.582
250	0.552	0.726	0.674	0.274	0.45	0.378	0.652	0.592
300	0.564	0.746	0.684	0.254	0.438	0.352	0.668	0.603
350	0.564	0.746	0.684	0.254	0.438	0.352	0.668	0.598
400	0.564	0.756	0.686	0.244	0.438	0.342	0.674	0.607
450	0.564	0.756	0.686	0.244	0.438	0.342	0.674	0.602
500	0.588	0.756	0.698	0.244	0.412	0.332	0.686	0.625
510	0.603	0.758	0.707	0.242	0.399	0.329	0.694	0.630
520	0.612	0.746	0.71	0.254	0.388	0.336	0.692	0.632
530	0.612	0.746	0.71	0.254	0.388	0.336	0.69	0.632
540	0.612	0.738	0.708	0.282	0.388	0.366	0.686	0.628
550	0.612	0.738	0.708	0.262	0.388	0.346	0.684	0.628
560	0.612	0.738	0.708	0.262	0.388	0.346	0.686	0.628
570	0.612	0.738	0.728	0.262	0.388	0.344	0.686	0.628
580	0.612	0.738	0.708	0.262	0.388	0.346	0.686	0.628
590	0.612	0.738	0.708	0.262	0.388	0.346	0.686	0.628
600	0.612	0.738	0.708	0.262	0.388	0.346	0.686	0.631
650	0.6	0.738	0.702	0.262	0.402	0.35	0.68	0.620
700	0.6	0.738	0.702	0.262	0.394	0.346	0.68	0.624
750	0.6	0.738	0.702	0.262	0.402	0.35	0.68	0.620
800	0.588	0.738	0.694	0.262	0.414	0.352	0.674	0.617
850	0.588	0.738	0.694	0.262	0.414	0.352	0.674	0.612
900	0.588	0.728	0.692	0.272	0.414	0.364	0.668	0.611
950	0.588	0.728	0.692	0.272	0.414	0.364	0.67	0.608

methods.

At first, we want to compare two different network fusion methods: mean fusion and SNF [38]. For each classification approach, we used both network fusion methods. So according to the network fusion, the 10 prediction methods implemented in our study are all grouped into pairwise comparisons. For example, in SLP and MLP, the first letter S in the method symbol SLP denotes SNF fusion method and the first M in the method symbol MLP means mean fusion approach. It is the same for other pairs of prediction methods: SLGC and MLGC, SMN and MMN, SSPC and MSPC, SSVM and MSVM. Table 1 shows the performances of different prediction algorithms between two network fusion methods. As can be seen in Table 2, for the prediction methods based on LP and LGC, there are no differences in performances between two network

Table 3: The performances of our label prediction method MMN with different $K(\sigma^2 = 510)$

K	TPR	SPC	NPV	FPR	FNR	FDR	ACC	F1
37	0.513	0.708	0.648	0.292	0.488	0.41	0.625	0.546
20	0.603	0.758	0.707	0.242	0.399	0.329	0.694	0.630
18	0.563	0.727	0.678	0.273	0.438	0.371	0.657	0.592
12	0.538	0.728	0.67	0.272	0.463	0.389	0.646	0.57
9	0.525	0.752	0.665	0.248	0.475	0.347	0.654	0.578
7	0.538	0.742	0.669	0.258	0.463	0.34	0.656	0.581

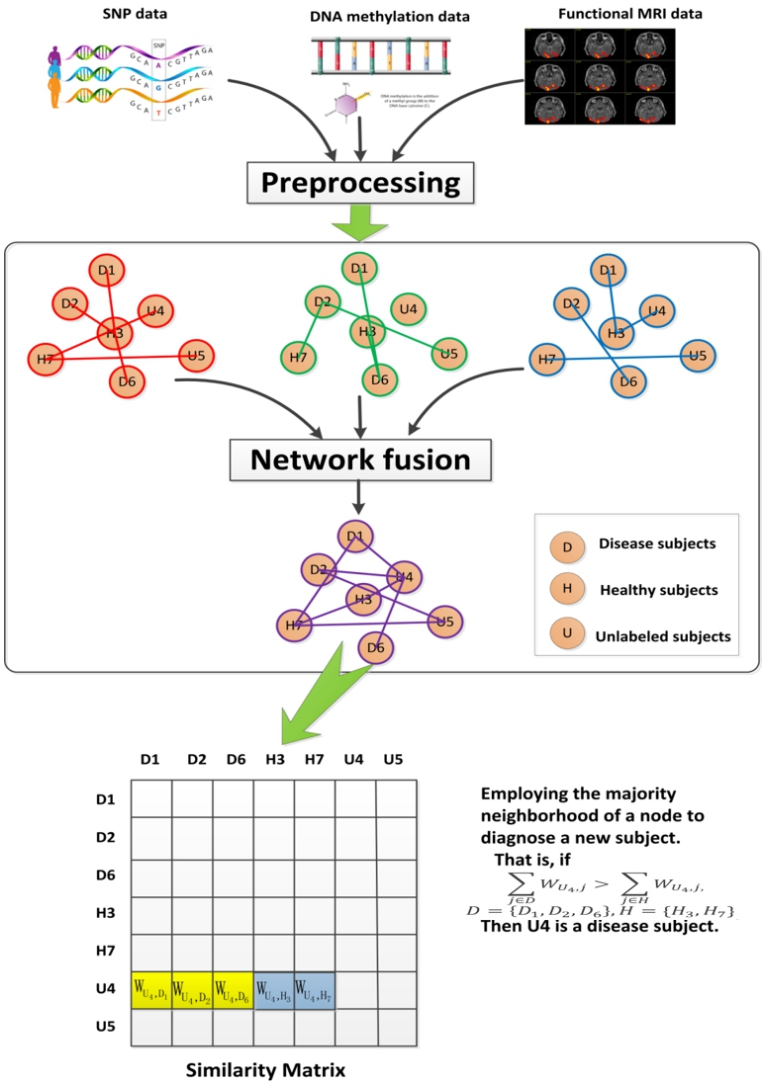


Figure 1 The Schematic of Schizophrenia Diagnosis

fusion approaches. For SSPC and MSPC, the only one group of module-assisted prediction methods, their accuracies are close to each other and neither got good performance. The results are similar to another pair of prediction methods: SSVM and MSVM. The major difference between SPC-based pair is that the mean-fusion-based method MSVM obtained higher accuracy than SNF-based method SSVM. However, the comparison results between SMN and MMN differ greatly from other graph-based prediction method pairs. For each of the performance metrics, MMN performs better than SMN, especially the ACC. Actually, the network fusion method used in this study uses uniform weight, i.e., equal weight for each of multiple individual networks. In the future, we will try to develop a weighted mean fusion method, which can assign an optimal weight for each individual network.

Above all, our prediction method MMN performs the best among all those graph-based label prediction methods and it is computationally efficient.

When computing the edge strength for constructing networks, there are two parameters σ, K related to

the prediction accuracy. When selecting an optimal σ, we fixed the parameter K (K=20) and evaluated the performances with σ^2 ranging from 100 to 950. The results are shown in Table 2 and we can see that when σ^2 is too small or too large, the prediction accuracy decreases and MMN performs best with $\sigma^2 = 510$. When selecting an optimal K, σ^2 was fixed at 510 with K changing in the range 7, 9, 12, 18, 20, 37 corresponding to 1/25, 1/20, 1/15, 1/10, 1/9, 1/5 of the number of the samples. As shown in Table 3, our methods achieves the best accuracy with $K = 20$, which is about 1/10 of whole sample (nodes) size.

5 Conclusions

We combined SNPs, DNA methylation and fMRI data into a single comprehensive network using a simple network fusion approach. A network-based label prediction method was applied to the network for predicting schizophrenia patient. Compared with other 9 graph-based label prediction approaches, our prediction method shows the best performance. However, our network fusion method used a uniform weight based combination of each network and in the future we will find an optimal weight based method to improve the prediction accuracy. In addition, Kim et al.[19] incorporated genomic knowledge when integrating multi-omics data for predicting the clinical outcome of cancer, which improved the predictive power. Therefore, we will also incorporate other prior knowledge into the construction of patient connection networks.

Conflict of Interest The authors declare no conflict of interest.

Acknowledgment The work are partly supported by the grants: the NIH (R01 GM109068, R01 MH104680, and R01MH107354) and NSF (#1539067). China Postdoctoral Science Foundation Grant, No. 2015M580352.

References

1. S.-P. Deng, D. Lin, V. D. Calhoun, and Y.-P. Wang, "Diagnosing schizophrenia by integrating genomic and imaging data through network fusion," in *Bioinformatics and Biomedicine (BIBM), 2016 IEEE International Conference on*. IEEE, 2016, pp. 1307–1313.
2. A. P. Association *et al.*, *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub, 2013.
3. M. M. Picchioni and R. M. Murray, "Schizophrenia," *BMJ*, vol. 335, no. 7610, 2007.
4. R. Pies, "How objective are psychiatric diagnoses," *Psychiatry*, vol. 4, no. 10, 2007.
5. T. Insel, B. Cuthbert, M. Garvey, R. Heinssen, D. S. Pine, K. Quinn, C. Sanislow, and P. Wang, "Research domain criteria (rdoc): toward a new classification framework for research on mental disorders," *American Journal of Psychiatry*, vol. 167, no. 7, pp. 748–751, 2010.
6. S. Ripke, B. M. Neale, A. Corvin, J. T. Walters, K.-H. Farh, P. A. Holmans, P. Lee, B. Bulik-Sullivan, D. A. Collier, H. Huang *et al.*, "Biological insights from 108 schizophrenia-associated genetic loci," *Nature*, vol. 511, no. 7510, p. 421, 2014.
7. A. P. M. M. S. Owen, *Schizophrenia*. London, England: Lancet, 2016.
8. S. Ogawa, T.-M. Lee, A. R. Kay, and D. W. Tank, "Brain magnetic resonance imaging with contrast dependent on blood oxygenation," *Proceedings of the National Academy of Sciences*, vol. 87, no. 24, pp. 9868–9872, 1990.
9. H. Yang, J. Liu, J. Sui, G. Pearlson, and V. D. Calhoun, "A hybrid machine learning method for fusing fmri and genetic data: combining both improves classification of schizophrenia," *Frontiers in human neuroscience*, vol. 4, p. 192, 2010.
10. E. Castro, V. Gómez-Verdejo, M. Martínez-Ramón, K. A. Kiehl, and V. D. Calhoun, "A multiple kernel learning approach to perform classification of groups from complex-valued fmri data analysis: Application to schizophrenia," *NeuroImage*, vol. 87, pp. 1–17, 2014.
11. H. Cao, J. Duan, D. Lin, V. Calhoun, and Y.-P. Wang, "Integrating fmri and snp data for biomarker identification for schizophrenia with a sparse representation based variable selection method," *BMC medical genomics*, vol. 6, no. Suppl 3, p. S2, 2013.
12. H. Cao, D. Lin, J. Duan, Y.-P. Wang, and V. Calhoun, "Bio marker identification for diagnosis of schizophrenia with integrated analysis of fmri and snps," in *Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference on*. IEEE, 2012, pp. 1–6.
13. D. Lin, H. Cao, Y.-P. Wang, and V. Calhoun, "Classification of schizophrenia patients with combined analysis of snp and fmri data based on sparse representation," in *Bioinformatics and Biomedicine (BIBM), 2011 IEEE International Conference on*. IEEE, 2011, pp. 394–397.
14. S. Robertson, "What is dna methylation?" new Medical. [Online]. Available: http://www.news-medical.net/life-sciences/What-is-DNA-Methylation.aspx
15. T. Shlomi, M. N. Cabili, M. J. Herrgrd, B. Ø. Palsson, and E. Ruppin, "Network-based prediction of human tissue-specific metabolism," *Nature biotechnology*, vol. 26, no. 9, pp. 1003–1010, 2008.

16. H.-Y. Chuang, E. Lee, Y.-T. Liu, D. Lee, and T. Ideker, "Network-based classification of breast cancer metastasis," *Molecular systems biology*, vol. 3, no. 1, p. 140, 2007.

17. L. Q. Uddin, K. Supekar, C. J. Lynch, A. Khouzam, J. Phillips, C. Feinstein, S. Ryali, and V. Menon, "Salience network–based classification and prediction of symptom severity in children with autism," *JAMA psychiatry*, vol. 70, no. 8, pp. 869–879, 2013.

18. K. Tsuda, H. Shin, and B. Schölkopf, "Fast protein classification with multiple networks," *Bioinformatics*, vol. 21, no. suppl 2, pp. ii59–ii65, 2005.

19. D. Kim, J.-G. Joung, K.-A. Sohn, H. Shin, Y. R. Park, M. D. Ritchie, and J. H. Kim, "Knowledge boosting: a graph-based integration approach with multi-omics data and genomic knowledge for cancer clinical outcome prediction," *Journal of the American Medical Informatics Association*, vol. 22, no. 1, pp. 109–120, 2015.

20. D. Kim, H. Shin, Y. S. Song, and J. H. Kim, "Synergistic effect of different levels of genomic data for cancer clinical outcome prediction," *Journal of biomedical informatics*, vol. 45, no. 6, pp. 1191–1198, 2012.

21. R. Sharan, I. Ulitsky, and R. Shamir, "Network-based prediction of protein function," *Molecular systems biology*, vol. 3, no. 1, p. 88, 2007.

22. H. Cao, J. Duan, D. Lin, Y. Y. Shugart, V. Calhoun, and Y.-P. Wang, "Sparse representation based biomarker selection for schizophrenia with integrated analysis of fmri and snps," *Neuroimage*, vol. 102, pp. 220–228, 2014.

23. D. Lin, J. Zhang, J. Li, H. He, H.-W. Deng, and Y.-P. Wang, "Integrative analysis of multiple diverse omics datasets by sparse group multitask regression," *Multi-omic Data Integration*, p. 126, 2015.

24. R. L. Gollub, J. M. Shoemaker, M. D. King, T. White, S. Ehrlich, S. R. Sponheim, V. P. Clark, J. A. Turner, B. A. Mueller, V. Magnotta *et al.*, "The mcic collection: a shared repository of multimodal, multi-site brain image data from a clinical investigation of schizophrenia," *Neuroinformatics*, vol. 11, no. 3, pp. 367–388, 2013.

25. J. Liu, J. Chen, S. Ehrlich, E. Walton, T. White, N. Perrone-Bizzozero, J. Bustillo, J. A. Turner, and V. D. Calhoun, "Methylation patterns in whole blood correlate with symptoms in schizophrenia patients," *Schizophrenia bulletin*, vol. 40, no. 4, pp. 769–776, 2014.

26. J. Liu, M. Morgan, K. Hutchison, and V. D. Calhoun, "A study of the influence of sex on genome wide methylation," *PloS one*, vol. 5, no. 4, p. e10028, 2010.

27. E. Walton, J. Liu, J. Hass, T. White, M. Scholz, V. Roessner, R. Gollub, V. D. Calhoun, and S. Ehrlich, "Mb-comt promoter dna methylation is associated with working-memory processing in schizophrenia patients and healthy controls," *epigenetics*, vol. 9, no. 8, pp. 1101–1107, 2014.

28. N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *The American Statistician*, vol. 46, no. 3, pp. 175–185, 1992.

29. O. Chapelle, J. Weston, and B. Schölkopf, "Cluster kernels for semi-supervised learning," in *Advances in neural information processing systems*, 2002, pp. 585–592.

30. X. Zhu, Z. Ghahramani, J. Lafferty *et al.*, "Semi-supervised learning using gaussian fields and harmonic functions," in *ICML*, vol. 3, 2003, pp. 912–919.

31. M. Belkin, I. Matveeva, and P. Niyogi, "Regularization and semi-supervised learning on large graphs," in *International Conference on Computational Learning Theory*. Springer, 2004, pp. 624–638.

32. D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," *Advances in neural information processing systems*, vol. 16, no. 16, pp. 321–328, 2004.

33. H. Shin, K. Tsuda, B. Schölkopf, A. Zien *et al.*, "Prediction of protein function from networks," in *Semi-supervised learning*. MIT press, 2006, pp. 361–376.

34. P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher, "Comprehensive identification of cell cycle–regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization," *Molecular biology of the cell*, vol. 9, no. 12, pp. 3273–3297, 1998.

35. E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman, "Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data," *Nature genetics*, vol. 34, no. 2, pp. 166–176, 2003.

36. J. H. Ohn, J. Kim, and J. H. Kim, "Genomic characterization of perturbation sensitivity," *Bioinformatics*, vol. 23, no. 13, pp. i354–i358, 2007.

37. S.-P. Deng, D. Lin, V. D. Calhoun, and Y.-P. Wang, "Predicting schizophrenia by fusing networks from snps, dna methylation and fmri data," in *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the*. IEEE, 2016, pp. 1447–1450.

38. B. Wang, A. M. Mezlini, F. Demir, M. Fiume, Z. Tu, M. Brudno, B. Haibe-Kains, and A. Goldenberg, "Similarity network fusion for aggregating data types on a genomic scale," *Nature methods*, vol. 11, no. 3, pp. 333–337, 2014.

39. A. Y. Ng, M. I. Jordan, Y. Weiss *et al.*, "On spectral clustering: Analysis and an algorithm," *Advances in neural information processing systems*, vol. 2, pp. 849–856, 2002.

40. Y.-C. Wei and C.-K. Cheng, "Towards efficient hierarchical designs by ratio cut partitioning," in *Computer-Aided Design, 1989. ICCAD-89. Digest of Technical Papers., 1989 IEEE International Conference on*. IEEE, 1989, pp. 298–301.

41. C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

42. X. Zhu and Z. Ghahramani, "Learning from labeled and unlabeled data with label propagation," Citeseer, Tech. Rep., 2002.

43. U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.

44. A. W. Wolfe, "Social network analysis: Methods and applications," *American Ethnologist*, vol. 24, no. 1, pp. 219–220, 1997.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 2, No. 3, 711-721 (2017)
www.astesj.com
Special Issue on Recent Advances in Engineering Systems

ASTES Journal
ISSN: 2415-6698

IGS: The Novel Fast IC Power Ground Network Optimization Flow Based on Improved Gauss-Seidel Method

Qinghao Ye, Xiaoxiao Wang*, Jiansong Gong, Songhao Ru, Donglin Su

School of Electronics and Information Engineering, Beihang Univeristy, 100191, China

ARTICLEINFO

Article history:
Received: 04 April, 2017
Accepted: 02 June, 2017
Online: 15 June, 2017

Keywords:
Integrated Circuit
Power Ground Network
IR-drop

ABSTRACT

With silicon technology further scaling, the switching activities together with GHz operation frequency greatly affects the power integrity by generating large IR-drop noises. Excessive IR-drop causes functional failures such as timing failure, abnormal reset and SRAM flipping. The PGN needs to be optimized to reduce IR-drop. The traditional EDA optimization routine repeats the steps such as generating layout, extracting parameter and simulation, which greatly increases the cost of integrated circuit design. This paper proposes a novel fast optimization flow optimizing the PGN based on the improved Gauss-Seidel method (IGS), which calculates the accurate IR-drop distribution according to the initial voltage distribution and the 3-D parasitic resistance distribution of PGN. The proposed optimization flow compares the calculated IR-drop distribution with the IR-drop distribution constraints, and changes the parameters of power straps until the performance of PGN meets the requirement of constraints without full chip simulation, which reduces the time cost of PGN optimization. This paper also provides the comparison between the IGS-based optimization flow and traditional EDA optimization routine. The difference of IR-drop distributions calculated by the proposed optimization flow and traditional EDA optimization routine is less than 4.7%. And the PGN optimization time has been reduced by 96.2% on average for ITC99 benchmark s9234, s13207, s35932 and b19. According to the results and analysis, the IGS-based optimization flow is reliable to improve the performance of PGNs.

1 Introduction

For advanced technologies like 28nm and below, billions of CMOS gates are integrated into a modern SOCs. As a result, the density of power ground network (PGN) has been greatly increased, which causes extra IR-drop. Due to the parasitic resistance and inductance of PGN, as well as the high current consumption, IR-drop can reach multiple hundreds millivolts, which becomes a significant challenge for IC design and test [1]. Excessive IR-drop causes functional failures such as timing failure, abnormal reset, and SRAM flipping. Hence, it is necessary to optimize the PGN and reduce IR-drop. The details of PGN optimization routine, which improves its performance, consists of two parts, including the calculation of the IR-drop distribution for different PGN parameters, and checking whether the performance of PGN meets the IR-drop constraints.

The traditional optimization routine of PGN is based on EDA tools such as Synopsys IC Compiler™ and Cadence EDI™, which uses full chip simulation to calculate the IR-drop distribution. However, when the parameters of PGN are changed, the traditional EDA optimization routine repeats the steps such as generating layout, extracting parameter and simulation, which greatly increases the cost of integrated circuit design. Therefore, it is difficult to calculate the IR-drop distribution for the PGN in a short time. Also, EDA tools only understand 1-D IR-drop constraints,

*Xiaoxiao Wang, Beihang University, XueYuan Road No.37,HaiDian District,BeiJing,China, 8601082338721. Email: wangxiaoxiao@buaa.edu.cn

The work of Xiaoxiao Wang was supported by National Science Foundation of China (NSFC) under grant 61427803, 61504007 and 61631002.

https://dx.doi.org/10.25046/aj020391

which is the maximum allowable IR-drop value. For example, when the IR-drop constraint is a 2-D IR-drop distributions, EDA tools cannot understand the constraint and determine whether the calculated IR-drop distribution meets the requirement of the constraint. So the traditional optimization routine cannot provide efficient and high quality PGN optimization results. Therefore, the optimization flow with ability efficiently calculating IR-drop distribution and comparing the performance of PGN with complex IR-drop constraints is of great value.

1.1 Previous Work

Several types of optimization methods to reduce IR-drop have been presented in literatures. The optimization methods proposed in [2] and [3] reduce the IR-drop by adjusting power straps such as changing the place of power straps, increasing the density of power straps where current noise is high, and amplifying the power strap area. However, the excess power straps lead to serious placement and routing problem for chips. Also it is technically difficult to increase power straps in the tape-out process. [4] presents the analysis method of Graphene-Based power distribution networks. A Walking pads method is presented in [5] and [6], which reduces IR-drop by changing the position of power supply pads according to critical path. But the changed power pads lead to distribution problem. Also, the optimization methods based on TSVs have been analysed. A thermal-aware power network design method for the power pad optimization is presented in [7]. [8] and [9] present the optimization method to reduce IR-drop by modifying the parameters and position of TSVs, which highly increases the area of chips. The optimization methods presented in [10] and [11] change the size of TSVs to reduce IR-drop. However, the size of TSVs is constant in the 28nm technologies. The optimization method presented in [12] inserts compensation cells when the delay of critical path exceeds the allowed value to reduce IR-drop. All the researches mentioned above reduce IR-drop and optimize the PGN of chips finally. But there are a few disadvantages of these methods as follows.

- These methods reduce IR-drop based on other parts of integrated circuits rather than the PGN, which leads to appended problems such as clock delay and increasing area power dissipation.
- These methods ignore the effect of power straps in PGN, which have a non-negligible impact on the IR-drop. Power straps take a large area of PGN, whose width and density are adjustable.
- These methods are used to reduce IR-drop instead of optimizing PGN, so they cannot optimize the PGN according to the IR-drop constraints.

Previous work refers to the Random Walk algorithm in [13], which is widely used to calculate the voltage distribution. And the Representative Random Walk algorithm presented in [14] establishes a relatively simple conductance model according to the width and density of the real PGN, then calculates the IR-drop distribution of the model, which is efficient to estimate the IR-drop distribution. But this algorithm cannot get accurate IR-drop distribution and calculate the voltage of each node in PGN. [15] presents the multiple triple method, which costs too much computing resource due to the high integration of the PGN. The Gauss-Seidel Method is presented in [16] to calculate the IR-drop distribution using recursive calculation. But this method only calculates the IR-drop distribution of the 2-D PGN and ignores the conductance of power rails and TSVs. Thus it cannot calculate the IR-drop distribution of the real PGN accurately. An IR-drop measurement system on chips is presented in [17]. This method adds a sensor on the basic circuit to measure the power supply noise. The unit level sensor provides a accurate IR-drop distribution of power rails, which produces extra noise. Also, the method adding a fully digital on-chip distributed sensor network is presented in [18]. The extra sensor network monitors the power supply noise across the chip continuously and generates a trace for diagnosis of the noise-induced failure. These methods spend a lot of time to repeat the full chip simulation, and change the circuit configuration. And they cannot provide a accurate voltage distribution of PGN to meet the requirement of IR-drop constraints automatically. So they cannot be used in the actual PGN.

1.2 Contributions and Paper Organization

In this paper, a novel fast optimization flow based on the improved Gauss-Seidel (IGS) method is proposed. The IGS method takes the conductance of power rail and TSV into consideration compared with the Gauss-Seidel method presented in [16]. The proposed IGS-based optimization flow has the following advantages.

- It optimizes the PGN by adjusting the parameters of power straps.
- It uses improved Gauss-Seidel method to calculate the IR-drop distributions for different power straps without repeating the steps such as generating layout, extracting parameter and simulation.
- It calculates the voltage of each node in the PGN accurately.
- It automatically optimizes the PGN with less manual operation and full chip simulation compared with the traditional EDA optimization routine, which saves a lot of time.
- It can provide satisfactory parameters of power straps according to 2-D IR-drop constraints.

This paper is organized as follows. In Section 2, the IGS method is presented to calculate the IR-drop of the PGN. The novel fast optimization flow of the PGN is presented in Section 3. The analysis of the results is shown in Section 4. Finally, the conclusion is presented in Section 5.

2 The Improved Gauss-Seisel Method

2.1 The Gauss-Seidel Method for the 2-D Conductance Network

In this part, the Gauss-Seidel method, which calculates the IR-drop distribution for the 2-D conductance network, is presented. Figure 1 shows a representative node i and the neighboring nodes j in a 2-D conductance network. The conductance between the node i and j is defined as $g_{i,j}$, and the voltage of node i is defined as V_i. The leakage current of node i is defined as I_i. According to the Kirchhoff Voltage and Current Law, the following equation is educed, in which N_i represents the gather of neighboring nodes of node i:

$$V_i = \sum_{j \in N_i} \frac{g_{i,j}}{\sum_{j \in N_i} g_{i,j}} V_j - \frac{I_i}{\sum_{j \in N_i} g_{i,j}} \tag{1}$$

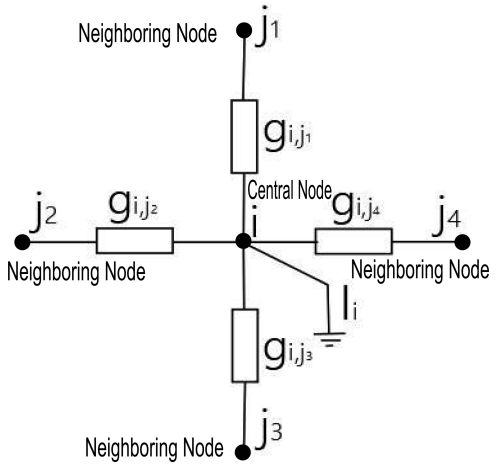


Figure 1: A representative node in the 2-D conductance Network with four neighboring nodes.

So the voltage of node i is calculated according to the leakage current of node i and the voltages of all neighboring nodes. Equation (1) is applied to all the nodes in the 2-D conductance network. A typical structure of power straps is shown in Figure 2. This model considers the network composed of power straps as the 2-D conductance network. And the intersection of power straps is regarded as a node in the calculation. The total number of nodes in the model network is N^2. The conductance between the neighboring nodes *(i,j)* and *(k,l)* is defined as $G_{(i,j)} = \sum_{(k,l) \in N_{(i,j)}} |g(i,j)(k,l)|$. And $V_{(i,j)}$ represents the gather of neighboring nodes of node *(i,j)*. The voltage of nodes in the 2-D conductance network is calculated as:

$$V_{(i,j)} = \sum_{(k,l) \in N_{(i,j)}} \frac{g_{(i,j)(k,l)}}{G_{(i,j)}} V_{(k,l)} - \frac{I_{(i,j)}}{G_{(i,j)}} \tag{2}$$

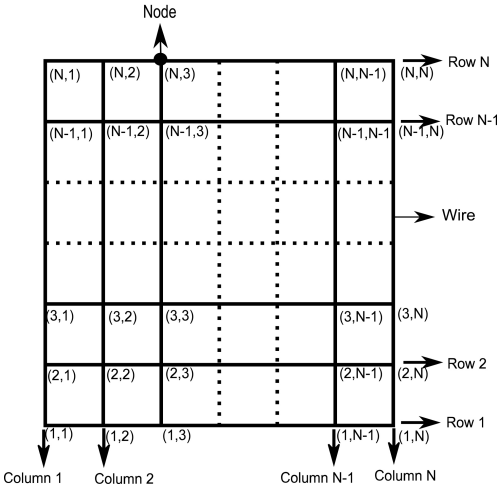


Figure 2: A typical structure of the 2-D conductance network.

The voltage distribution of the conductance network depends on the node leakage current, the voltages of neighboring nodes, and the neighboring conductance. The voltage of each node in the 2-D conductance network is calculated via Equation (2). Algorithm 1 details the Gauss-Seidel method, which uses recursive computation to calculate the voltage distribution. The leakage current of each node and the conductance between neighboring nodes should be provided before calculation. The initial voltages of nodes connected to power pad is VDD, which is constant, and the initial voltages of other nodes is 0. According to the initial voltage distribution and the conductance network, the voltage distribution of the first time recursive computation is calculated via Equation (2). In this way, the new voltage of every node is presented. The second time recursive computation is based on the voltage distribution calculated by the first time recursive computation. So this method provides a accurate voltage distribution with multiple recursive computation.

Algorithm 1 The GAUSS-SEIDEL Method in the 2-D conductance network

1: Calculate related conductance value and leakage current;
2: $V^{(0)}$=the initial value;
3: **for** each $n \in [1, P]$ **do**
4: **for** each $(i,j) \in Q$ **do**
5: $V_{(i,j)}^{(n+1)} = \sum_{(k,l) \in N_{(i,j)}} \frac{g_{(i,j)(k,l)}}{G_{(i,j)}} V_{(k,l)}^{(n)} - \frac{I_{(i,j)}}{G_{(i,j)}}$
6: **end for**
7: **if** max(abs($V^{(n)} - V^{(n-1)}$))< e ; **then**
8: break
9: **end if**
10: **end for**

Combining Algorithm 1, the first time recursive computation is finished in Line 4, and the new voltage distribution is applied to the second time recursive computation. In the process of recursive computations, the dynamic range of some parameters like the total leakage current, the total area, and the total conductance of the chip should be determined. In Algorithm 1, the parameter P on the Line 3 is the maximum times of recursive computation, which is set by users. The parameter Q represents all nodes in the power straps. The parameter e is the judgment factor to set the end of recursive computation. If the difference between adjacent computation is less than e, the recursive computation stops. The result calculated by the last time recursive computation is the output of Algorithm 1. Also, the parameter e can be used to adjust the accuracy and the time consumption of the recursive computation. With smaller judgment factor, the calculated voltage distribution is closer to the real distribution, i.e., the result is more accurate. And the results shows that the maximum value of the difference between the results calculated by the IGS-based optimization flow and traditional EDA optimization routine. Also, the smaller judgment factor leads to the increase of the iterations, which causes more time consumption of the recursive computation. However, the Gauss-Seidel Method ignores the conductance of power rails and TSVs, which influences the accuracy of voltage distribution.

2.2 The Improved Gauss-Seidel Method for the 3-D Conductance Network

The basic structure of the PGN is shown in Figure 3, which consists of three parts including power straps, TSVs, and power rails. In the PGN, the power straps connected to TSVs supplies power to TSVs. And some power rails are connected to the logic gates to feed the gates. In the actual PGN structure, the power straps can be placed in different metal layers while power rails only can be placed in one metal layer, which does not influence the analysis and calculation. This paper chooses a double layers PGN to analyse as shown in Figure 3. The leakage current only exists in power rails. The power straps and power rails with different width and density, as well as the TSVs, lead to complex conductance distribution. So the improved Gauss-Seidel method is proposed to calculate the IR-drop distribution of the 3-D PGN. The Gauss-Seidel method calculates the voltage distribution of the PGN, which is regarded as a 2-D conductance network. But the actual PGN has a 3-D structure consisting power straps, TSVs and power rails. Compared with the Gauss-Seidel method, the improved Gauss-Seidel method adopts a 3-D conductance network as the computation model to calculate the IR-drop distribution of the PGN, which is closer to the real structure of the PGN.

The first step of Algorithm 1 is calculating the conductance between neighboring nodes and the leakage current of nodes in the 2-D PGN. It is also suitable for the algorithm of the 3-D PGN. The initial voltage of power straps directly connected to the power is VDD, which is a constant in recursive computation. The initial voltage of other power straps is 0, which updates in each recursive computation. The total leakage current is measures by EDA tools, which exists in the power rails directly connected to logic gates. It is an important content to accurately calculate the equivalent conductance of metal lines in the related PGN, which depends on the volume and resistivity of metal lines. The resistivity of metal lines is provided in the standard library of chips. Figure 4 shows the equivalent transformation between the 3-D PGN and the conductance network. Each metal line is equivalent to a conductance and the metal lines network is equivalent to a conductance network. And the intersection area is regarded as a node in the equivalent transformation. The equivalent conductance networks are added to the basic conductance network to constitute the complete conductance network. And a 3-D conductance network is built in this way. There are three kind of nodes in the real PGN. The nodes formed by the horizontal and vertical power straps are defined as $node_a$. The nodes formed by the power straps and TSVs are defined as $node_b$, The nodes formed by power rails and TSVs are defined as $node_c$. And the conductance between $node_a$ is defined as g_a. The conductance between $node_b$ is defined as g_b and the conductance between $node_c$ is defined as g_c. The conductance of TSVs between $node_b$ and $node_c$ is already provided as a constant.

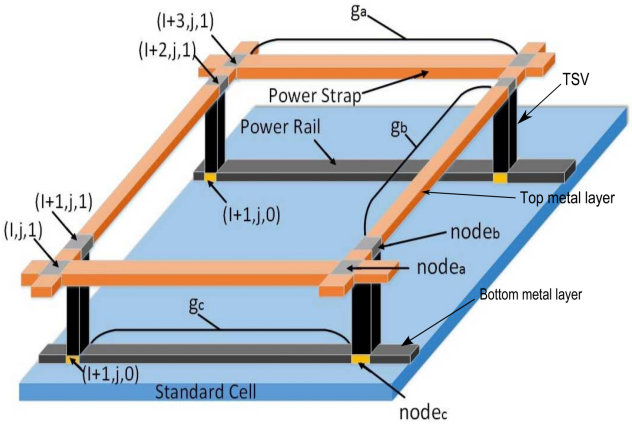
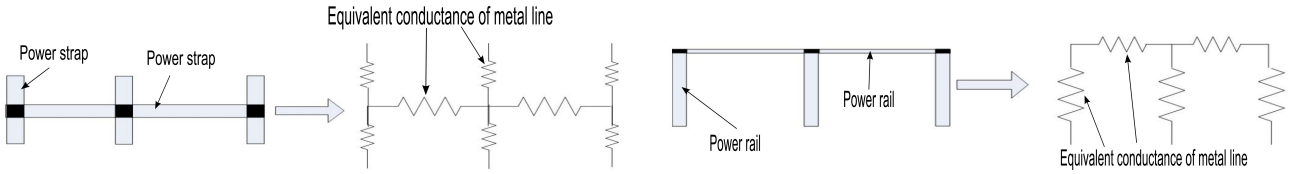


Figure 3: A two metal layers construction of the PGN consisting of power straps, TSVs and power rails.

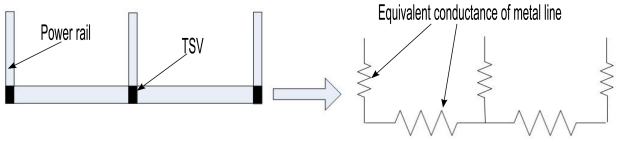
To calculate the conductance distribution of the PGN, the total length and density of power straps should be provided. The total length of power straps is set to L and the density is set to N. The distance of neighboring $node_b$ is defined as l_{rail}. The width of power straps is defined as W. And the width of power rails is defined as w. The electrical conductivity of the metal lines is defined as k. Then the conductance is calculated as:

$$g_a = k \cdot \frac{N-1}{L} \cdot W \tag{3}$$



(a) The equivalent conductance of metal lines at the power straps intersection.

(b) The equivalent conductance of metal lines at the intersection of power straps and TSVs.



(c) The equivalent conductance of metal lines at the intersection of power rails and TSVs.

Figure 4: The equivalent transformation between the 3-D PGN and the conductance network.

$$g_b = k \cdot l_{rail} \cdot W \tag{4}$$

$$g_c = k \cdot \frac{N-1}{L} \cdot w \tag{5}$$

The total length of chips and total leakage current, which are measured by EDA tools, are required to calculate the voltage distribution of chips. The parameters of power straps are constant except the width and density. In the PGN, which is placed in the 3-D system of coordinate as shown in Figure 3, the power straps and power rails are placed in different layers, which are connected by TSVs. The nodes in the power straps are defined as *(i,j,1)* and the nodes in the power rails are defined as *(i,j,0)*. The conductance between neighboring nodes *(i,j,k)* and *(l,m,n)* is defined as $G_{(i,j,k)} = \sum_{(l,m,n)\in N_{(i,j,k)}} |g(i,j,k)(l,m,n)|(k,n = 0,1)$. And $N_{i,j,k}$ represents all the neighboring nodes of node *(i,j,k)*. The leakage current only exists in power rails, so $I_{(i,j,1)} = 0$ and $I_{(i,j,0)} \neq 0$. The voltage of nodes in the power straps is calculated as:

$$V_{(i,j,1)} = \sum_{(l,m,n)\in N_{(i,j,1)}} \frac{g_{(i,j,1)(l,m,n)}}{G_{(i,j,1)}} V_{(l,m,n)} \tag{6}$$

And the voltage of nodes in the power rails is calculated as:

$$V_{(i,j,0)} = \sum_{(l,m,n)\in N_{(i,j,0)}} \frac{g_{(i,j,0)(l,m,n)}}{G_{(i,j,0)}} V_{(l,m,n)} - \frac{I_{(i,j,0)}}{G_{(i,j,0)}} \tag{7}$$

The voltage distribution is calculated via Equation (6) and (7). Algorithm 2 details the steps of the improved Gauss-Seidel method, which uses recursive computation to calculate the voltage distribution of the actual PGN. The conductance and leakage current of chips are provided before the algorithm.

The initial voltage of nodes connected to the power pad is VDD, which is a constant in the recursive compute process. And the initial voltage of other nodes is 0. According to the initial voltage distribution and the conductance distribution, the result of first time recursive computation is calculated via Equation (6) and (7). In this way, the new voltage distribution of the PGN is presented. Then the voltage distribution calculated by the first time recursive computation is applied to the second time recursive computation. By such analogy, the improved Gauss-Seidel method provides a accurate voltage distribution with multiple recursive computation. As well, the parameter P on the Line 3 is the maximum times of recursive computation, which is set by users. The parameter Q_1 represents all nodes in the power straps, and the parameter Q_2 represents all nodes in the power rails. The parameter e is a judgment factor to set the end of recursive computation in Algorithm 2. If the difference between adjacent computation is less than e, the recursive computation stops. The result of last time recursive compute is the output of Algorithm 2.Similar to the Algorithm 1, the judgement factor of Algorithm 2 determines the computational accuracy and the time consumption of the recursive computation. And the efficiency of the IGS method is inversely proportional to the value of the judgment factor. And the efficiency is proportional to the node number of the PGN as shown in Figure 4.

Algorithm 2 The Improved GAUSS-SEIDEL Method in the actual PGN

1: Calculate related conductance value and leakage current;
2: $V^{(0)}$=the initial value;
3: **for** each $n \in [1, P]$ **do**
4: **for** each $(i, j, 1) \in Q_1$ **do**
5: $V_{(i,j,1)} = \sum_{(l,m,n)\in N_{(i,j,1)}} \frac{g_{(i,j,1)(l,m,n)}}{G_{(i,j,1)}} V_{(l,m,n)}$
6: **end for**
7: **for** each $(i, j, 0) \in Q_2$ **do**
8: $V_{(i,j,0)} = \sum_{(l,m,n)\in N_{(i,j,0)}} \frac{g_{(i,j,0)(l,m,n)}}{G_{(i,j,0)}} V_{(l,m,n)} - \frac{I_{(i,j,0)}}{G_{(i,j,0)}}$
9: **end for**
10: **if** max(abs($V^{(n)} - V^{(n-1)}$))< e ; **then**
11: break
12: **end if**
13: **end for**

3 The Novel Fast Optimization Flow Based on the Improved Gauss-Seisel Method

To optimize the PGN efficiently, a novel fast optimization flow based on the improved Gauss-Seidel method is proposed, which calculates the voltage distribution for the PGN automatically without repeating full chip simulation. This proposed optimization flow optimizes the PGN according to IR-drop constraints, which considers the conductance of power rails and TSVs. The flow diagram of this optimization flow is shown in Figure 5 and the specific steps are detailed in Algorithm 3. The proposed IGS-based optimization flow involves a three-part process of measurement, calculation and verification. The optimization process is divided into three steps. Firstly, EDA tools are used to measure the total length and leakage current of chips. Secondly, the dynamic range of power strap parameters is limited, and the IR-drop distribution of power straps is calculated by the improved Gauss-Seidel method. Finally, this proposed optimization flow understands the 2-D IR-drop distribution constraints, and compare the IR-drop distribution calculated by the IGS method and the IR-drop constraints to verify whether the performance of the PGN meets the constraints. The detailed steps of this optimization flow are described below.

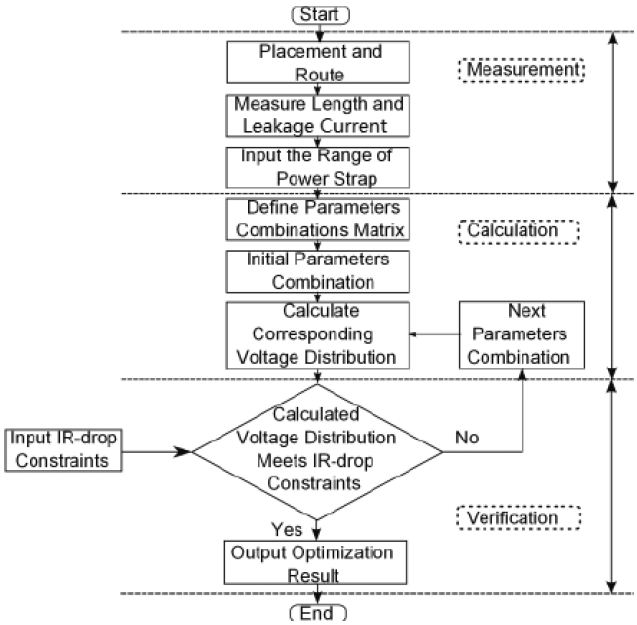


Figure 5: The flow diagram of the novel fast optimization flow diagram based on the improved Gauss-Seidel method.

3.1 Measuring the Total Length and Total Leakage Current of Chips

At the beginning of the proposed optimization flow, the total length and total leakage current of chips are provided to calculate the voltage distribution and the conductance of the PGN, which is measured by EDA tools directly. As the total leakage current is a constant when chips work at a unique frequency, EDA tools are used to place and route only once in the proposed optimization flow.

3.2 Voltage Distribution Calculation

The PGN of chips consists of three parts including power straps, power rails and TSVs as shown in Figure 3. The width and density of power rails are invariable while the width and density of power straps are adjustable. And the IR-drop can be reduced by adjusting the width and density of power straps. Before the optimization, the maximum density and minimum density of power straps are defined as $N_{\max}$ and $N_{\min}$. And the maximum width and minimum width of power straps are defined as $W_{\max}$ and $W_{\min}$. So the parameter group of power straps (W_α, N_β) should satisfy the condition that $W_\alpha \in (W_{\min}, W_{\max})$ and $N_\beta \in (N_{\min}, N_{\max})$. Each parameter group represents a kind of power strap distribution of the PGN.

To traverse all possible parameter groups, the traverse gradients of the power strap width and density are defined as Δw ($\Delta w \ll |W_{\min} - W_{\max}|$) and Δn ($\Delta n \ll |N_{\min} - N_{\max}|$). The traverse gradients are determined by the actual requirements of the optimization process, which determine the optimization precision. The object of the optimization flow is to detect the PGN with satisfactory power straps, whose performance meets the requirements of the IR-drop constraints. All parameter groups are listed in a matrix as follows:

$$\begin{bmatrix} (W_{\min}, N_{\min}) & (W_{\min}+\Delta w, N_{\min}) & \cdots & (W_{\max}, N_{\min}) \\ (W_{\min}, N_{\min}+\Delta n) & (W_{\min}+\Delta w, N_{\min}+\Delta n) & \cdots & (W_{\max}, N_{\min}+\Delta n) \\ (W_{\min}, N_{\min}+2\Delta n) & (W_{\min}+\Delta w, N_{\min}+2\Delta n) & \cdots & (W_{\max}, N_{\min}+2\Delta n) \\ \vdots & \vdots & \ddots & \vdots \\ (W_{\min}, N_{\max}) & (W_{\min}+\Delta w, N_{\max}) & \cdots & (W_{\max}, N_{\max}) \end{bmatrix}$$

The first element $(W_{\min}, N_{\min})$ in matrix is the initial parameter group. Combining the parameter group with the total width and total leakage of chips measured by EDA tools, the voltage distribution is calculated using the improved Gauss-Seidel method.

3.3 Verifying Whether the Voltage Distribution Satisfies IR-drop Constraints

The IR-drop constraints provide the minimum allowable voltage of IR-drop distribution in the PGN, which is presented in Algorithm 3 as $V^{(MIN)}$. And the $V^{(IGS)}$ in Algorithm 3 indicates the minimum value of the IR-drop distribution calculated by the IGS method. If the result calculated by the IGS method meets the requirements of the IR-drop constraints, the corresponding parameter group is the result of the optimization flow, otherwise the next parameter group is applied to the IGS method to calculate the IR-drop distribution until the satisfactory IR-drop distribution is obtained.

Algorithm 3 The Novel Fast PGN Optimization Flow

****** Measurement ******
Place and route;
Measure total length and total leakage current of the chip;
Input the Range of Power Strap and IR-drop Constraints;
Define parameters combinations matrix M;
for each $(W_\alpha, N_\beta) \in M$ **do**

****** Calculation ******
Calculate related conductance value and current drain;
$V^{(0)}$=the initial value;
for each $n \in [1, P]$ **do**
for each $(i, j, 1) \in Q_1$ **do**
$$V_{(i,j,1)} = \sum_{(l,m,n)\in N_{(i,j,1)}} \frac{g_{(i,j,1)(l,m,n)}}{G_{(i,j,1)}} V_{(l,m,n)}$$
end for
for each $(i, j, 0) \in Q_2$ **do**
$$V_{(i,j,0)} = \sum_{(l,m,n)\in N_{(i,j,0)}} \frac{g_{(i,j,0)(l,m,n)}}{G_{(i,j,0)}} V_{(l,m,n)} - \frac{I_{(i,j,0)}}{G_{(i,j,0)}}$$
end for
if $\max(\mathrm{abs}(V^{(n)} - V^{(n-1)})) < e$; **then**
break
end if
end for

****** Verification ******
if $V^{(IGS)} - V^{(MIN)} > 0$; **then**
break
end if
end for
Output Optimization Result;

The way to compare the IR-drop distributions calculated by the IGS method and the IR-drop constraints depend on the contents of the IR-drop constraints. For example, if the IR-drop constraints only contain the maximum allowable value of IR-drop distribution, the maximum value of the IR-drop distribution calculated by the IGS method should be extracted. And if the maximum value of the calculated result is less than the maximum allowable value of IR-drop distribution, the proposed optimization flow stops and outputs the satisfactory power strap parameters of the PGN, otherwise the optimization flow continues. The proposed IGS-based optimization flow can understand the 2-D IR-drop distributions constraints such as the area ratio of different IR-drop region, and determine whether the calculated IR-drop distribution meets the requirement of the constraints. Because of the feature of the PGN, the increased width and density of power straps causes the decrease of the power strap equivalent resistances, which reduces the IR-drop finally. Therefore, when a satisfactory PGN is determined, the performance of PGNs with the wider and greater density power straps also meets the requirements of the IR-drop constraints. So the result of the proposed optimization flow is not unique. For the convenient comparison in the next section, the result of the proposed optimization flow is the PGN with the minimum parameter group, which satisfies the IR-drop constraints. But in actual conditions, all PGNs with wider and greater density power straps are acceptable.

Benchmark	Gate Number	Leakage Current
s9234	2027	152mA
s13207	2573	76.19mA
s35932	12204	114.29mA
b19	67619	152.1mA

Table 1: Logic Gate Number and Leakage Current of Benchmarks

4 Experimental Results

4.1 Benchmarks Under Test

The experiments are based on the 28nm Synopsys Cell Library. In the experiment, the recursive parameter e is set to $10^{-7}V$ to ensure the accuracy of the IR-drop distribution. The experiments are performed on ITC99 benchmark s9234, s13207, s35932 and b19, which are presented on ISCAS conference as the dedicated test circuits. The number of logic gates and leakage current of the selected benchmarks are shown in Table 1. And the selected benchmarks have different sizes from thousands to ten thousands to testify the wide suitability of the IGS-based optimization flow. The power panels are placed at the center, which are connected to the power straps directly. The VDD is set to 1.05V. The working frequency of the benchmarks is set to 125MHz. The IR-drop is calculated by multiplying the equivalent conductance and the leakage current of the node. And the wavelength of the 10GHz electromagnetic wave is 3cm. So the size of the chips is the small electrical size even the operation frequency reaches 10GHz. As a result, the equivalent conductance of the PGN is approximate in the 0.1 10GHz range. At different operating frequencies, the IGS-based optimization flow uses EDA tools to measure the leakage current at step one, which is determined by the operating frequency. And the measured leakage current is the input of the following steps. So the change of the operating frequency does not influence the application of the proposed optimization flow. The benchmarks are optimized according to 2-D IR-drop distribution constraints on the area ratio of different IR-drop region, and the details of the contributions are shown in Table 2. The rule of setting IR-drop distribution constraints is based on the requirements of the blocks on chips. There are a few IP cores in the SoC, which are sensitive to the slack. As a result, the IR-drop of the area with these IP cores needs to be limited strictly. So the IR-drop constraints sets a large IR-drop value of the area with sensitive

Benchmark	The 2-D IR-drop Constraints on Area Ratio				
	0-20mV	20mV-40mV	40mV-60mV	60mV-80mV	80mV-100mV
s9234 s13207 s35932 b19	>5%	>15%	<80%	0	
	>5%			>15%	<80%

Table 2: The 2-D IR-drop Constraints on the Area Ratio of Different IR-drop region for Benchmarks

Benchmark	The Optimized Power Strap Width and Density					Time Consumption			
	Optimization Result by IGS		Optimization Result by EDA		Width Difference in Percentage	IGS	EDA	simulation time	Time-saving Ratio
	Strap Width	Strap Density	Strap Width	Strap Density					
s9234	0.34um	10	0.36um	10	5.5%	82.1s	1254s	12	93.4%
s13207	0.5um	15	0.54um	15	7.4%	48.2s	1835s	14	97.3%
s35932	0.66um	19	0.62um	19	6%	93s	2631.1s	21	96.5%
b19	1.06um	20	1.08um	20	2%	165.2s	6932s	11	97.7%

Table 3: Optimization Result and Time Overhead

components and a small IR-drop value of other area.

The accuracy of improved Gauss-Seidel method is analyzed to validate the reliability of the proposed optimization flow. The PGNs are optimized by the IGS-based optimization flow and traditional EDA optimization routine according to the IR-drop constraints. The IR-drop distributions calculated by two methods are compared to verify the efficiency of the proposed optimization flow. The time overhead of two optimization methods is also compared.

4.2 Verifying the Accuracy of Improved Gauss-Seidel Method

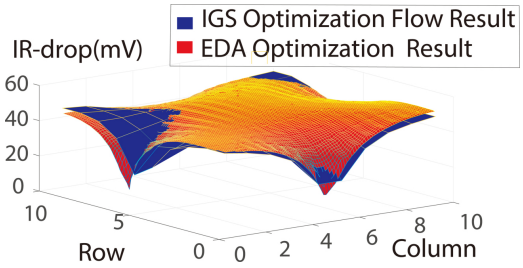
The proposed optimization flow is based on the improved Gauss-Seidel method. The IR-drop distribution, which is calculated by the improved Gauss-Seidel method, is a key factor to identify the performance of the PGN, so the accuracy of the improved Gauss-Seidel method determines the reliability of the proposed optimization flow. The IR-drop distributions of PGNs calculated by the improved Gauss-Seidel method and traditional EDA optimization routine are compared with the same IR-drop constraints to find the coherence of two methods. If the difference of two results is acceptable, the improved Gauss-Seidel method can be used to calculate the IR-drop distribution directly. Each benchmark is optimized by both optimization routines in the same condition. The IR-drop distributions of the PGN, which are calculated by the IGS method and traditional EDA optimization routine, are shown in Figure 6. Because of the different size of benchmarks, the IR-drop distributions are put into three dimensional coordinates to compare respectively.To highlight the gaps of the IR-drop distributions, the difference between two results is also presented in Figure 6. in Figure 6, X axis and Y axis denote the position of the nodes in power straps, and Z axis refers to the IR-drop value and the percentage of the difference as noted. The maximum value of the difference between the two results calculated by the IGS method and traditional EDA optimization routine is less than 4.7%, which proves that there is a high coherence between the two results. So the IGS method can calculate the IR-drop distribution of the PGN accurately.

Figure 7 shows the IR-drop distributions of the PGNs optimized by the IGS-based optimization flow and traditional EDA optimization routine according to the 2-D IR-drop constraints. The maximum value of the difference between two results is less than 5%. The result shows that the IGS-based optimization flow optimizes the PGN according to the IR-drop constraints on the area ratio of different IR-drop region successfully.

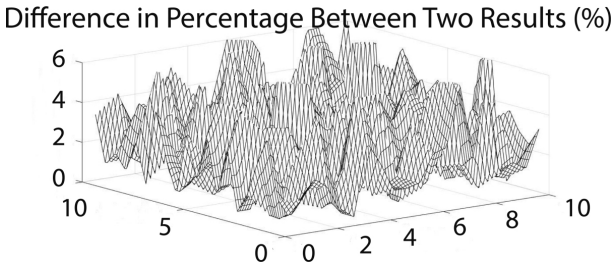
4.3 Analysis of Optimization Results and Time Overhead

In the optimization example, the width and density range of the power straps are set to $N_{\min} = 10$, $N_{\max} = 30$, $W_{\min} = 0.2um$, $W_{\max} = 2um$. And the variation gradient of the parameter groups is defined as $\Delta w = 0.02um$ and $\Delta n = 1$. The 2-D IR-drop distribution constraints on the area ratio of different IR-drop region are set to prove the effectiveness of the optimization flow based on the improved Gauss-Seidel method. The optimized PGNs of the proposed optimization flow are compared with traditional EDA optimization routine results to analyze the accuracy of optimization flow as shown in Figure 7. The optimization results and time overhead of two optimization methods are shown in Table 3. The maximum difference of width between two optimization methods is 7.4%, and the density of PGNs optimized by the proposed optimization flow is the same as the density of PGNs optimized by traditional EDA optimization routine. So the IGS-based optimization flow provides highly consistent results with traditional EDA optimization routine, and is reliable to optimize the PGN according to the IR-drop constraints.

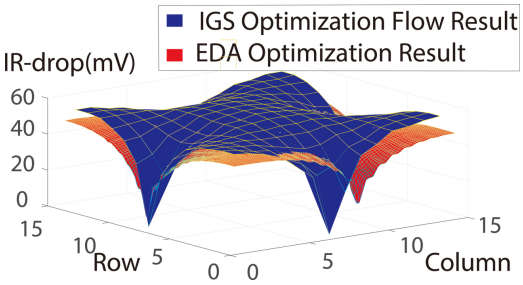
Time overhead is an important factor to judge the validity of the proposed optimization flow. The time overhead analysis of the optimization flow based on improved Gauss-Seidel method and traditional EDA optimization routine is shown in Table 3. In Table 3,



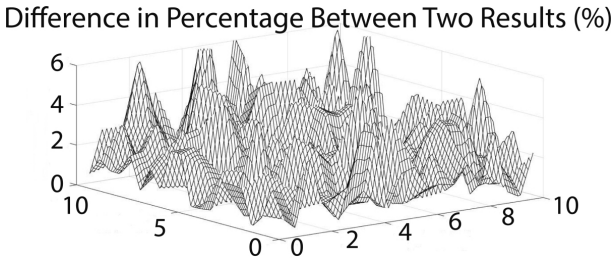
(a) The IR-drop distributions calculated by the IGS optimization flow and traditional EDA optimization routine for s9234.



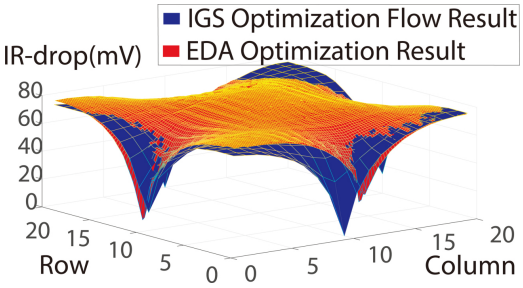
(b) The difference in percentage between the IR-drop distributions calculated by the IGS optimization flow and traditional EDA optimization routine for s9234.



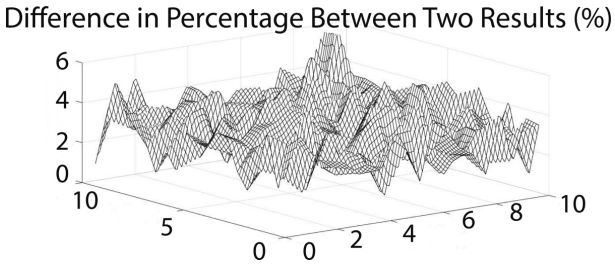
(c) The IR-drop distributions calculated by the IGS optimization flow and traditional EDA optimization routine for s13207.



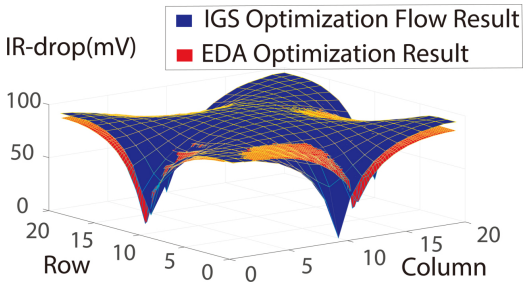
(d) The difference in percentage between the IR-drop distributions calculated by the IGS optimization flow and traditional EDA optimization routine for s13207.



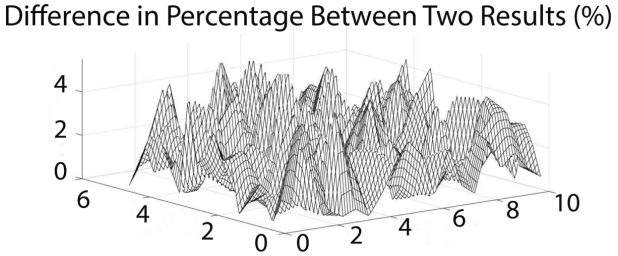
(e) The IR-drop distributions calculated by the IGS optimization flow and traditional EDA optimization routine for s35932.



(f) The difference in percentage between the IR-drop distributions calculated by the IGS optimization flow and traditional EDA optimization routine for s35932.



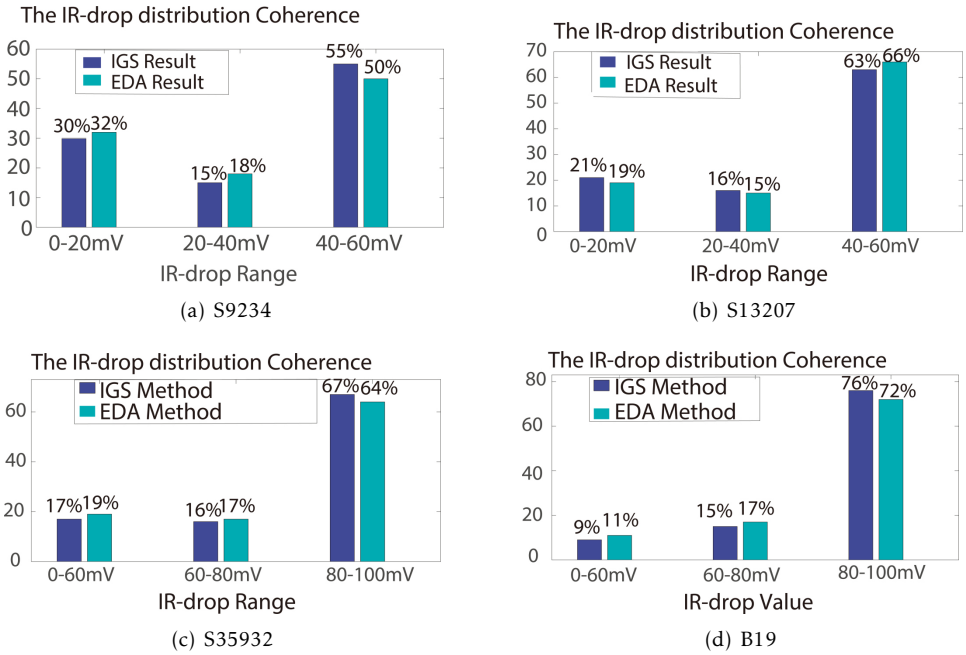
(g) The IR-drop distributions calculated by the IGS optimization flow and traditional EDA optimization routine for b19.



(h) The difference in percentage between the IR-drop distributions calculated by the IGS optimization flow and traditional EDA optimization routine for b19.

Figure 6: The IR-drop distributions calculated by IGS and traditional EDA optimization routine as well as the difference between two results for benchmark s9234 s13207 s35932 b19. It shows that the IGS result is close to the standard result provided by traditional EDA optimization routine. And the maximum value of difference is less than 4.7%, so the IGS method is reliable to calculate the IR-drop distribution.

the time consumptions of the IGS-based optimization and traditional EDA optimization routine are listed in Column 7 and 8. And the number of steps of the traditional EDA optimization routine is listed in Column 9, while the IGS-based optimization flow needs only one step to get the optimization result. To get an appropriate optimization result, traditional EDA optimization routine repeats the steps such as generating layout, extracting parameter and simulation when the parameter group of power straps is changed, which greatly increases time overhead. The optimization flow based on the improved Gauss-Seidel method only



(a) S9234 (b) S13207 (c) S35932 (d) B19

Figure 7: The IR-drop distribution coherence between the IGS calculation and traditional EDA optimization routine results for benchmark (a) s9234 (b) s13207 (c) s35932 and (d) b19. The maximum IR-drop difference is within the range of 5%, which shows the proposed IGS method successfully optimizes the PGN to satisfy the IR-drop constraints.

uses EDA tools once for place and route in the whole process. So when benchmark has a large size, traditional EDA optimization routine optimization spends a lot of time. For example, the time overhead of the proposed optimization flow to optimize benchmark b19 is 165.2s. Considering the universal judgment factor, the computation time does not increase significantly. But traditional EDA optimization routine spends 630s to simulate the full chip once, and spends 6930s to finish the optimization. The PGN optimization time has been reduced by 96.2% on average for ITC99 benchmark s9234, s13207, s35932 and b19 by the IGS-based optimization flow. The improved Gauss-Seidel method adopts a 3-D conductance network as the computation model to calculate the IR-drop distribution of the PGN, which is closer to the real structure of the PGN. In the application of the proposed optimization flow, the IGS-based optimization flow works together with the EDA tools, and the IR-drop constraints can be set flexibly and multiply. The IGS-based optimization flow can understand the complex IR-drop constraints, and optimize the particular area with the sensitive components emphatically to avoid the problems such as timing failure, abnormal reset and SRAM ?ipping. On the premise of increasing the complexity of the IR-drop distribution constraints, the time consumption is reduced by 15 to 40 times compared to the traditional EDA optimization routine to optimize the benchmark s9234, s13207, s35932 and b19. So the proposed optimization flow can improve the accuracy of the IR-drop distribution calculation compared with the past work, and avoid the functional failure by setting complex block-wise IR-drop constraints.

5 Conclusion

In this paper, a novel fast optimization flow of PGNs based on the improved Gauss-Seidel method is presented. In this optimization flow, the improved Gauss-Seidel method is provided to calculate the IR-drop distribution of the PGN according to the initial voltage distribution and the parasitic resistance distribution of the PGN. The IR-drop distributions calculated by IGS method and traditional EDA optimization routine are compared to verify the accuracy of IGS method. The difference between two results is less than 4.7%, which proves that the IGS method is reliable to calculate the IR-drop distribution of the PGN. The proposed IGS optimization flow can understand 2-D IR-drop distribution constraints and provide highly consistent results with traditional EDA simulation routine. And the IGS optimization flow only uses EDA tools once for place and route once in the whole optimization process while traditional EDA simulation routine repeats the steps such as generating layout, extracting parameter and simulation when the parameters of power straps is changed. So the IGS optimization flow saves 96.2% of the traditional EDA simulation routine time overhead on average for benchmark s9234, s13207, s35932 and b19.

References

[1] M. Chew, A. Aslyan, J.-h. Choy, X. Huang, "Accurate full-chip estimation of power map, current densities and temperature for em assessment," in Proceedings of the 2014 IEEE/ACM International Conference on Computer-Aided Design. IEEE Press, 440–445, 2014.

[2] B. H. Calhoun, K. Craig, "Flexible on-chip power delivery for energy efficient heterogeneous systems," in Proceedings of

the 50th Annual Design Automation Conference. ACM, 160, 2013.

[3] Q. Zhou, J. Shi, B. Liu, Y. Cai, "Floorplanning considering ir drop in multiple supply voltages island designs," Very Large Scale Integration (VLSI) Systems, IEEE Transactions on, **19**(4), 638–646, 2011.

[4] M. Han, A. Amirkhany, W. Xiong, "An enhanced power integrity analysis flow based on the interdependence between simultaneous switching output noise and static ir drop," in Electronic Components and Technology Conference (ECTC), 2014 IEEE 64th. IEEE, 560–565, 2014.

[5] K. Wang, B. H. Meyer, R. Zhang, M. Stan, K. Skadron, "Walking pads: Managing c4 placement for transient voltage noise minimization," in Design Automation Conference (DAC), 2014 51st ACM/EDAC/IEEE. IEEE, 1–6, 2014.

[6] K. Wang, B. H. Meyer, R. Zhang, K. Skadron, M. Stan, "Walking pads: Fast power-supply pad-placement optimization," in Design Automation Conference (ASP-DAC), 2014 19th Asia and South Pacific. IEEE, 537–543, 2014.

[7] Z. Li, Y. Ma, Q. Zhou, Y. Cai, Y. Wang, T. Huang, Y. Xie, "Thermal-aware power network design for ir drop reduction in 3d ics," in Design Automation Conference (ASP-DAC), 2012 17th Asia and South Pacific. IEEE, 47–52, 2012.

[8] N. H. Khan, S. M. Alam, S. Hassoun, "System-level comparison of power delivery design for 2d and 3d ics," in 3D System Integration, 2009. 3DIC 2009. IEEE International Conference on. IEEE, 1–7, 2009.

[9] Y. Shinozuka, H. Fuketa, K. Ishida, F. Furuta, K. Osada, K. Takeda, M. Takamiya, T. Sakurai, "Reducing ir drop in 3d integration to less than 1/4 using buck converter on top die (bct) scheme," in Quality Electronic Design (ISQED), 2013 14th International Symposium on. IEEE, 210–215, 2013.

[10] Y.-H. Huang, M.-S. Zhang, H.-Z. Tan, "A novel method for ir-drop reduction in high-performance printed circuit boards," in Electronic Packaging Technology (ICEPT), 2014 15th International Conference on. IEEE, 583–586, 2014.

[11] S. Wang, F. Firouzi, F. Oboril, M. B. Tahoori, "P/g tsv planning for ir-drop reduction in 3d-ics," in Proceedings of the conference on Design, Automation & Test in Europe. European Design and Automation Association, 44, 2014.

[12] T. Raja, V. D. Agrawal, M. L. Bushnell, "Variable input delay cmos logic for low power design," in VLSI Design, 2005. 18th International Conference on. IEEE, 598–605, 2005.

[13] H. Qian, S. R. Nassif, S. S. Sapatnekar, "Random walks in a supply network," in Proceedings of the 40th annual Design Automation Conference. ACM, 93–98, 2003.

[14] M.-H. Tsai, W.-S. Ding, H.-Y. Hsieh, J. C.-M. Li, "Transient ir-drop analysis for at-speed testing using representative random walk," IEEE Transactions on Very Large Scale Integration (VLSI) Systems, **22**(9) 1980–1989, 2014.

[15] G. Katti, M. Stucchi, K. De Meyer, and W. Dehaene, "Electrical modeling and characterization of through silicon via for three-dimensional ics," Electron Devices, IEEE Transactions on, **57**(1), 256–262, 2010.

[16] Y. Zhong and M. D. Wong, "Fast algorithms for ir drop analysis in large power grid," in Proceedings of the 2005 IEEE/ACM International conference on Computer-aided design. IEEE Computer Society, 351–357, 2005.

[17] S. Dietel, S. Hoppner, T. Brauninger, U. Fiedler, H. Eisenreich, G. Ellguth, S. Hanzsche, S. Henker, and R. Schu?ffny, "A compact on-chip ir-drop measurement system in 28 nm cmos technology," in Circuits and Systems (ISCAS), 2014 IEEE International Symposium on. IEEE, 1219–1222, 2014.

[18] M. Sadi and M. Tehranipoor, "Design of a network of digital sensor macros for extracting power supply noise profile in socs," IEEE Transactions on Very Large Scale Integration (VLSI) Systems, **24**(5), 1702–1714, 2016.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 2, No. 3, 722-731 (2017)
www.astesj.com
Special Issue on Recent Advances in Engineering Systems

ASTES Journal
ISSN: 2415-6698

GNSS Positioning in Non-line-of-Sight Context—a Survey for Technological Innovation

Julia Breßler[*,1], Marcus Obst[2]

[1]*Chair for Innovation Research and Technology Management, Technische Universität, 09107 Chemnitz, Germany*

[2]*BASELABS GmbH, 09126, Chemnitz, Germany*

ARTICLEINFO

Article history:
Received: 10 April, 2017
Accepted: 12 May, 2017
Online: 16 June, 2017

Keywords:
Non-Light-of-Sight
GNSS / Glonass Positioning
Urban Areas

ABSTRACT

Reliable, available and accurate localization of people and vehicles is one crucial requirement for intelligent transportation systems (ITSs). This has led to a variety of technical approaches in recent years. Towards an exact positioning via Global Navigation Satellite Systems (GNSSs) the visibility of satellites is an important factor. In this paper, we review the literature regarding the positioning under non-line-of-sight conditions and analyze a range of topics from explicit modeling to empirical data collection. The variety of research indicates a growing interest in robust localization in urban areas. Furthermore, we develop a taxonomy of the technical design for GNSS applications dealing with non-line-of-sight and a typology of the measuring instruments. Both is based on an extensive analysis and review of the state-of-the-art. In doing so, we characterize the current research road map. Thus, the goal of this work is to provide a starting point in terms of the state-of-the-art for further research activities in robust GNSS positioning and the opportunity for technological innovations to ITS.

1 Introduction

This paper is an extension of work originally presented in the international conference on Intelligent Transportation Systems 2016 [1].

Localization and navigation are two key features for our industrial and globally connected world. Reliable knowing the ego position at any time around the globe – in real-time – is a necessary prerequisite for many applications, which include movement from one point to another. One particular field is the area of autonomous driving (AD) which opens up a lot of new and interesting applications. AD as a system innovation involves technological, social and infrastructural changes. Right now, several technical systems have been presented that operate on level two (partial automation) of the SAE[1] scale. In order to reach the next level of automation, that also involves higher awareness of the automated vehicle, further technological innovations are needed. Here, robust and reliable GNSS is considered a key-technology. Moreover, the ongoing progress of urbanization causes a steady extension of satellite based navigation to urban canyon environments [2]. Within inner cities, users would like to navigate seamlessly without any obstacles in both outdoor and indoor [3].

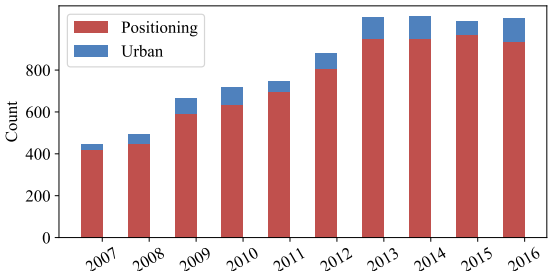


Figure 1: Number of relevant GNSS publications from 2007 to 2016.

However, the dense urban area causes several errors to satellite navigation-based positioning and thus decreases the potential for various positioning applications [4]. Due to the fact that the urban environment provides perfect surfaces for reflections, non-line-of-sight (NLOS) situations are a source for the worst errors in cities. This effect is also known outside of cities such as airports [5] or in harsh natural environments such as mountains or forests with dense foliage.

[*]Julia Breßler, Technische Universität Chemnitz, 09107 Chemnitz, Germany & julia.bressler@wirtschaft.tu-chemnitz.de
[1] https://www.sae.org/misc/pdfs/automated_driving.pdf

https://dx.doi.org/10.25046/aj020392

Traditional approaches are not able to identify and eliminate several stochastic and deterministic errors in satellite based navigation under both, urban and indoor, conditions [6], [7]. In [8] various approaches to address the urban canyon problem by minimizing the negative influences to the navigation solution are described. Moreover, [4] presents four different methods for improving positioning performance and categorizes in

- usage of several common radio signals such as radio signals, Wi-Fi or new technologies such as Long Term Evolution (LTE) or 5G,
- integration of different sensors such as dead reckoning (DR), mainly low-cost inertial sensors and odometry data from vehicles [9],
- techniques for the detection or modeling of the surrounding environment of the localization object [10], [11], [12], [13].

However, no cost-efficient system which provides a reliable sub-meter-level positioning in urban areas based on one single method is known, today. Consequently, we expect that future applications will combine multiple solutions to increase the accuracy and reliability. That means the most sophisticated localization system requires the combination of the single advantages of heterogeneous techniques within one system.

According to [14], the main aspects of cost-efficient GNSS-based localization system are *availability* (e. g. by using more satellite systems [15]), *accuracy* [16] and *reliability*. Some approaches improve one particular aspect without evaluating the potential negative influence to the other criteria. We believe, than an innovative localization system should strive to increase all three aspects at once while keeping costs at a reasonable level. For this reason, navigation and real-time positioning in challenging environments still requires a lot of efforts in research [17] and industry.

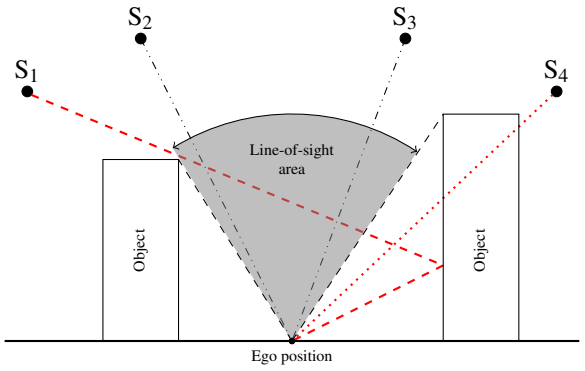


Figure 2: A typical situation in an urban area. The reception of each satellite signal depends on the position of the receiver and the satellite with respect to each other. Satellites S_2 and S_3 are in line-of-sight (LOS), satellite S_4 is blocked and satellite S_1 has non-line-of-sight (NLOS). This is a schematic representation of the electronic waves emitted by the satellites.

Based on [18], we focus on intelligent urban positioning (IUP) which includes new techniques for the detection of non-line-of-sight (NLOS) signal propagation. Thus, we will outline the current roadmap of scientists who deal with GNSS positioning under NLOS conditions. Within this survey, we intend to present a systematic overview of relevant data for NLOS handling. An explicitly in-depth assessment by benchmarking of the entire techniques is not aimed. However, we discuss topics such as measurement data generation. This survey is motivated, by the development of conference proceedings and journal publications of GNSS positioning as depicted in Figure 1, which has been stagnating since 2013. In fact, we present a subproblem of GNSS localization: the techniques for handling NLOS as well as the evaluation methods of the scientists. The paper is structure as as follows: in section 2 we first describe the NLOS phenomena by itself in order to prepare the introduction of an appropriate taxonomy. In section 3, we present a taxonomy of different NLOS handling techniques and evaluation methods. In the subsequent section an outlook to future trends and new approaches is given. The paper closes with a comprehensive conclusion in section 5.

2 Non-Line-of-sight

When using GNSS localization there are various influencing factors, which might decrease the quality of a position estimate in urban scenarios. [8] claims two important challenges: Multipath and NLOS situations. Both phenomena appear in several application domains where electromagnetic wave propagation between a sender and a receiver is used. This includes the signal reflection from surrounding buildings, a poor satellite visibility or an unfavorable satellite constellation. For the sake of clarity, we deliberately distinguish between *multipath* and *NLOS*. If a satellite signal is received via multiple paths at the antenna, that is, via the direct path (also known as line-of-sight or LOS) and the NLOS path, it is characterized as multipath. If it is solely received via the reflection path it is simply characterized as NLOS. In inner cities with high buildings and narrow streets the receiver performance is usually affected by both, multipath and NLOS. Thus, both effects are the dominant sources for unwanted disturbances [19]. Figure 2 depicts a typical situation within an urban canyon. It is worthwhile to mention, that advanced antenna concepts (e.g. based on dual polarization [20], antenna arrays [21] or choke ring designs) are able to exclude the NLOS part from a multipath signal. However, this no longer works if only the NLOS signal is received. Moreover, these effects cannot be removed by differential techniques [7]. Thus, an efficient and explicit estimation/identification step is required.

In [14] the NLOS problem is divided into eight individual sectors. Based on the aforementioned work, we abstract three different levels: the technical design for the detection of NLOS; the objective of the approach itself; and the application domain (see Figure 3).

- The *technical design* level consists of four different sub-levels—*error analysis*, *hardware sensors*, *environmental modeling*, *fault detection/outlier classifi-*

cation. This level summarizes all technical implementations and designs.

- *Objective* includes the sub-levels *handling* and *scope*. It describes how the contained information within the NLOS signal is used. For example, if it is just ignored or exploited to a certain extend. Moreover, we separate between *autonomous* (stand-alone) and *cooperative* architectures.
- The level *application domain* comprises the main areas where GNSS-based localization techniques are used for the positioning process.

In the subsequent text, we use the combination of technical design and the level of objectives (especially handling) as categories for our survey. This includes the analysis of current NLOS methods under the aspects of availability, accuracy and reliability.

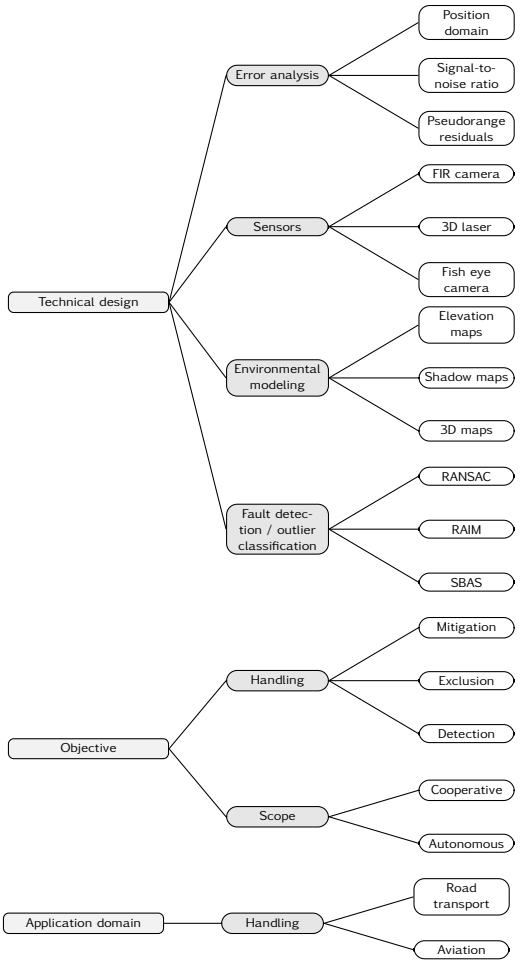


Figure 3: Overview of created taxonomy in reference to [14].

3 Classification & Taxonomy

In this section we will create a taxonomy, in which the levels of handling, objectives and techniques are linked (see Figure 3). Based on these levels, we will assign different research activities. In addition, we provide two further categorizations based on an inductive approach. We summarize current research approaches, which lead us to the taxonomy stages. For the sake of clarity, we present a table which categorizes the evaluation methods found in literature. Finally, we present an overview that summarizes current approaches which address NLOS.

3.1 Taxonomy of NLOS Handling Approaches including Techniques

After analyzing the literature, we came to the conclusion that there are four different high-level solutions to the NLOS problem. This includes: *ignorance* of the NLOS effect, *mitigation* of NLOS signals, *identification* of NLOS signals as well as *avoiding* NLOS. The latter one is equivalent to avoiding to go to cities or denied environments. Ignorance means, the disregard of multipath effects and NLOS signals at all. Although, this approach is not constructive in the sense of a modern localization system as highlighted in the introduction section, we included it for the sake of completeness. By mitigation we subsume approaches which basically assume that the effect of multipath can be minimized through the usage of improved features of hardware. In contrast, there is the field of identification. Here, we include approaches which aim to identify NLOS signals in order to use this information for further processing. In a second step, the identified NLOS signals can be eliminated or exploited.

It is worth mentioning that some approaches use different methods but generate similar results. For example, the sub-level *receiver-based correlator design* might mitigate the influence of NLOS without exact identification of the defect signal, while a similar work, which belongs to the sub-level *modified receiver processing strategy*, performs an identification followed by a subsequent elimination of that signal from the positioning process.

Mitigation

The work of [22] proposes two categories of mitigation techniques: antenna design and receiver-based correlator design. We augment this classification with two additional categories in our taxonomy, that is, *weighting* and *using quality parameters*. Due to the nature of these techniques, the algorithms are only able to decrease the influence of the NLOS error to the positioning solution without explicitly modeling which signals are received under NLOS or LOS conditions. Hence, they are for example not suitable for explicit modeling of pseudorange signal characteristics.

- Receiver-based correlator - different designs or implementations of the correlator at the HF-level are a key feature of work within this category and can be found in [23] and [24].
- Antenna design - based on a special design of the antenna, the NLOS part within a multipath signal can be extracted before the signal arrives the receiver. Representatives of this approach are: [8], [25], [26], [27] and [28]. Recurring on [8], six antenna designs can be distinguished, where the angle of arrival measurement (AOA)is rated the most performant technology.

Although the AOA principle provides a high performance, it is worth mentioning that this antenna design is not widely adopted yet.

- Weighting - refers to the so-called elevation cut-off angle for different satellites. In general, low elevated satellites are more affected by NLOS than satellites with a high elevation. For this reason measurements from these satellites will be excluded before a position solution is computed. Implementations of that technique can be found in [29] and [30].
- Using quality parameters - the approaches in this category indirectly use different indicators such as signal-to-noise ratio (SNR), Doppler frequency, etc. to assess the probability of a measurement being a NLOS signal. Based on a user-defined threshold in this domain, the potential NLOS signals are either excluded or tagged in order to apply a weighting. Implementations of this can be found in [31], [32], [33], [34], [35], [36] and [37].

Identification

We split the top-level category identification (as proposed by [8] and [38]) into the sub-levels *elimination* and *exploitation*. For both sub-levels, the identification of NLOS signals is a necessary condition. Regarding the elimination of NLOS signals, the concept initially presented in [4] is extended with a further category named *environmental modeling*. Thus, the final four sub-levels of the category elimination are:

- Environmental modeling - describes the explicit modeling of potential external influences such as buildings or infrastructure of cities: [12], [39], [18], [40], [41], [42], [43], [44], [45], [46], [47], [48], [49], [50], [51]. We also propose to include approaches such as a sky-positioning camera within this category.
- Integration with inertial navigation systems (INSs) - combines the benefits of two heterogeneous systems into one localization system: [14], [52], [53], [54], [55], [56], [57], [58].
- Modified receiver processing strategies - describes approaches were the receiver architectures is modified in order to achieve better signal processing aspects: [59], [60], [61], [62], [63], [64], [65], [66], [67], [68].
- Using multi-constellation GNSS - for increasing the overall amount of available satellite signals to choose from and thus improving the dilution of precision (DOP) metric: [13], [69], [70].

The level exploitation includes a large amount of approaches and consequently comprises several categories. The handling-technique is used to obtain further important information about accuracy, availability and reliability. Accordingly, the technique is appropriately utilized. This changes the view of NLOS: from a negative effect to an important information provider. Also, we have gathered techniques that allow to provide necessary information. In [43], the two categories *signal-to-noise ratio* and *shadow matching* are proposed. The category *signal delay matching* (SDM) was introduced in [4]. The other categories are taken from [14] and augmented with different collections of papers as shown in the following listing. It should be highlited, that recent research approaches usually propose a combination of the individual techniques. The categories of the sub-level exploitation are divided in:

- Signal-to-noise ratio - approaches sorted to this category aim to use the information about different SNR levels in several environments: [27], [47], [71], [72], [73], [74].
- Ray tracing - this technique uses several models (e.g. 3D city models) to predict an intersection of the LOS signal path with a building between a satellite and the receiver. In case of an intersection, the probability for NLOS is considered high: [4], [22], [71].
- SDM or signal delay matching - describes a technique to predict and observe parameters such as delta code phase delay. The observed parameter is then compared to a known model as shown in [4].
- Shadow matching - shadow matching is a pattern matching approach which utilizes a 3D city model in order to predict whether a satellite will be visible or not. Based on the decision whether a signal was directly received, the most likely position candidate from the 3D map can be selected: [11], [75], [76], [77], [78], [79] and [80].
- 3D building model - this level encompasses all approaches based on 3D building models which do not rely on ray tracing or shadow matching: [4], [22], [71], [81], [82], [83].
- Consistency checking - The rationale behind this approach is, that LOS measurements usually produce a more consistent navigation solution (e.g. in data fusion systems) than NLOS measurements [8]. Therefore, different combinations of satellite signals are used to compute individual position candidates which are afterwards compared to each other or another reference: [29], [84], [85], [86] and [87]. The technique of innovation filtering is also included in this category [8].
- Pseudorange residual - the least squares method aims to minimize the residuals produced by all considered measurements. The result is a more consistent (i.e. balanced) positioning solution as shown in [71].
- Carrier phase - this category is similar to the SNR level. However, this time the observed parameter is the carrier phase [88].
- Time delay and Doppler frequency - This category is similar to the SNR level. However, this time the observed parameters are the time delay and Doppler frequency: [89], [90] and [91].

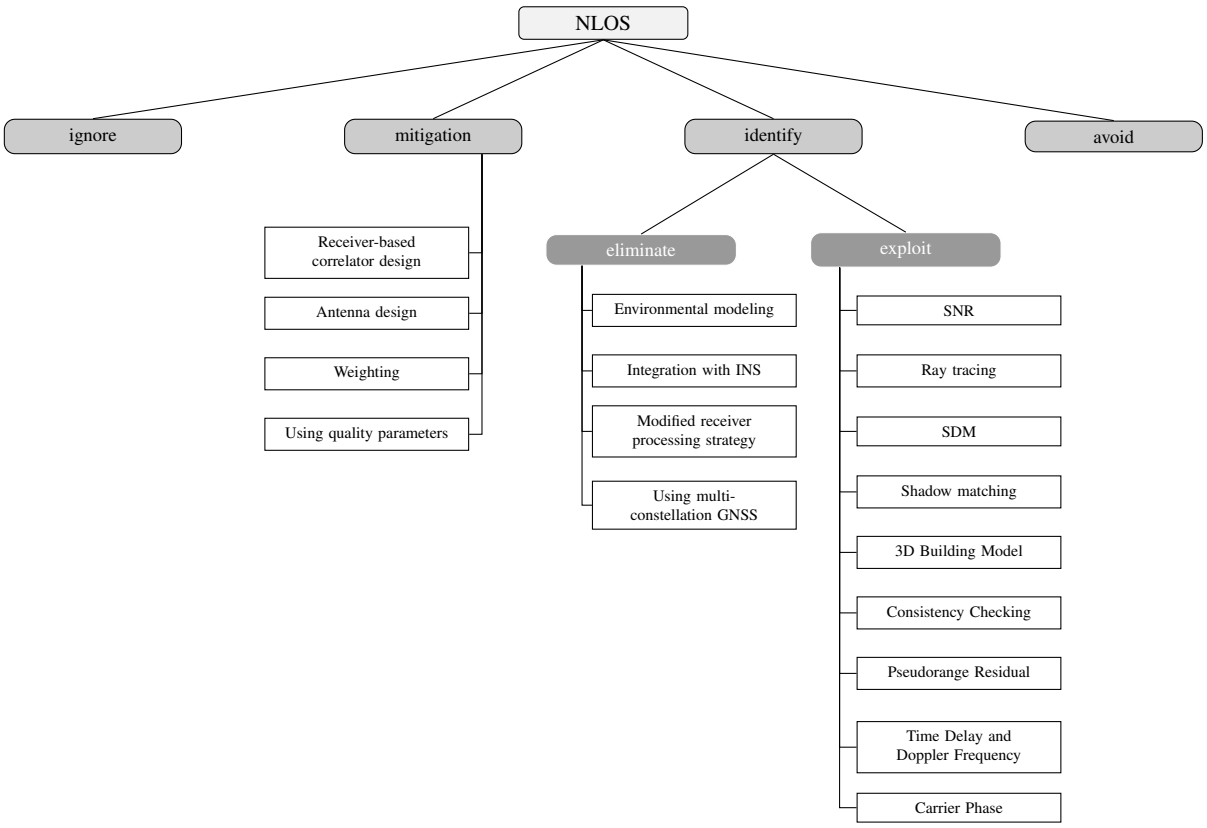


Figure 4: Classification of the NLOS problem into four different levels. Additionally, the level identify is divided into the two sub-levels eliminate and exploit. The different sub-levels are inspired by [4], [38], [8], [14] and extended by own categories.

Some approaches found in literature cannot be clearly assigned to one of the aforementioned categories. Moreover, papers which cover more than one technique are also missing at this stage. For the sake of completeness, we would like to add these papers to this survey. For example, techniques which utilize cooperative methods to exchange GNSS raw data in order to mitigate the influence of degraded satellite signals can be found in [92] and [93]. Other approaches try to detect jumps of the mean value within the measurement domain over time [94]. Moreover, various forms of receiver autonomous integrity monitoring (RAIM) have been investigated [95]. A further class of complex algorithms utilizes Dirichlet process mixtures (DPM) for multipath mitigation [96], [97].

3.2 Evaluation Methods

In the introduction, we mentioned the characterization of the scientific road map in the context of NLOS. Similarly, to the handling methods, we now want to present the analysis of the evaluation methods. At this point, we ask ourselves how the technical approaches for NLOS mitigation, identification and modeling are assessed. Therefore, we provide an overview of the used experimental setups. The focus is clearly on papers, which address NLOS or multipath handling. We proceed in an inductive manner. This means, we try to derive relevant categories from the reviewed papers. Within 53 papers we found static and dynamic measurement campaigns, which were destined for either vehicles (trucks, aircrafts, drones, cars, etc.) or persons. Moreover, these measurement campaigns can be categorized into simulation and real-world data collections. Table 1 illustrates our results and allows a classification. The row *no information* summarizes items that do not report the evaluation method. The classification is focused on real-world data collection for vehicles. We observed that real-world data collection campaigns for persons is constantly increasing since 2013. In contrast, simulations are underrepresented in connection with the exploration of identification and mitigation strategies of NLOS. We assume different obstacles in the process of transferring real world phenomena into a simulation model. The influences and dependencies between constantly changing environments during a dynamic maneuver are highly complex. Hence, the majority of the investigated papers use real-world data to cover most relevant side effects. During the characterization of the evaluation methods we also realized that a comparison of the results can hardly be produced. This is mainly due to a lack of a common dataset, which allows consistent benchmarking of the used methods and techniques.

3.3 Recent solutions for positioning under NLOS conditions

In this section, we will look at the different categories of our Survey for current research projects. Similarly, we test the statement by [8] that the combination of techniques is essential. The study of the reviewed work showed 7 research projects in the period from 2013 to 2015, which dealt with the NLOS problem. The methodology of the programs can

be classified into the field of identification. Based on this current research, we analyzed solutions in form of technical design, scope and data collection. Due to this sampling we found:

- one approach with a cooperative scope [56]. This approach was evaluated by simulation.
- six approaches, which are specialized in environmental modeling with 3D ([4], [18], [22], [43], [47], [71]). Here, real-world data was collected. Thus, building boundary data, digital terrain models and building models were used.

Table 1: Types of Measurement Data

Real Data collection	**Static measurement**	**Dynamic measurement**
Via vehicle	[8], [14], [33], [50], [80], [98], [99]	[4], [14], [30], [33], [35], [92], [47], [48], [50], [51], [58], [67], [68], [70], [74], [81], [82], [83], [85]
Via person	[18], [22]	[43], [49], [66], [71]
(No information)	[28], [29], [31], [32], [34], [37], [73], [84], [86], [87], [88]	
Simulation	**Static measurement**	**Dynamic measurement**
Via vehicle	[90]	[14], [27], [56], [68], [91], [94]
Via person (No information)	[25], [41], [89], [95]	

Furthermore, this analysis reveals that these projects used a combination of technical designs in order to identify the NLOS signal. For example, [18] used building models, consistency checking, signal geometry and signal strength information within one algorithm. In addition, we were able to determine which criteria were used to evaluate the research projects. In all studies, accuracy was the main objective, while some studies focused on the availability aspect as well. Only four studies addressed the performance indicator reliability.

Furthermore, an assumption from our introduction was confirmed. Thus, the costs represent a significant influencing factor on the techniques, which were used. Four out of the seven research programs address the low-cost aspect: [4], [18], [43] and [47]. For this reason, we conducted further meta-analysis on the criteria for NLOS handling techniques by investigating the following sources: conference proceedings as well as journal publication from the years 2013 to 2016. It became clear that accuracy was given to 59%, availability to 27% and reliability to 14% as observation and influencing factors for the assessment of NLOS approaches. Figure 5 illustrates this evolution. In addition, we observed an increase in low-cost conditions for NLOS techniques since 2014. Basically, these criteria can be considered evaluation standards for a common benchmarking of NLOS handling techniques.

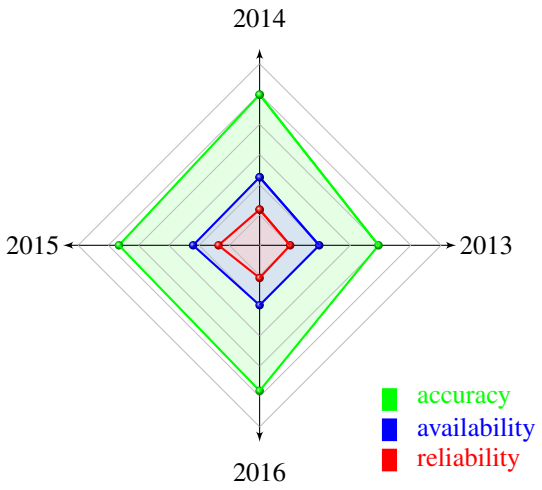


Figure 5: Evolution of evaluation criteria from 2013 to 2016.

4 Current Development

Despite the steadily growing field of satellite navigation for pedestrians, in combination with other techniques a seamless navigation independent from the environments, the majority of usage is for various vehicles on roads. This fact is shown in the rows for vehicle of Table 1. On the other hand, there are some gaps of research and experience with satellite-based navigation for pedestrians. Another fact is the ongoing development of autonomous vehicles in the near future. A reliable and accurate navigation system is one of the enabling-technologies for improved mobility. For this purpose the vehicle have to trust the navigation solution, independently of harsh environments or good visibility of satellites. In theory, there are several techniques to fulfill this task. However, we expect that the combination of multiple approaches will result in a safe navigation system that can be used in case of outage or partial degradation of a sub-system.

Although, this survey paper focuses on an analysis of the state-of-the-art of GNSS positioning under NLOS conditions, we would like to mention that there are a couple of national and European initiatives which aim to improve satellite navigation in challenging areas. For example, the goal of the German research project *smartLoc*[2] is to develop a robust positioning algorithm based on the factor graphs framework as used in the robotics community. The main motivation is to enable low-cost receivers to be used in urban areas by deliberately modeling the LOS and NLOS statistics inside of the navigation algorithm. By that ap-

[2]http://www.smartLoc.eu

proach, an appropriate integrity as well as an improve accuracy is expected. The algorithm is designed to explicitly handle temporal mulit-modalities of the GPS pseudoranges as caused under NLOS conditions. Similarly, the European research project *InDrive*[3] proposes to use a software GNSS receiver in combination with a Bayesian data fusion framework in order to estimate and identify NLOS signals already at the correlator level. In contrast to a standard GNSS receiver which typically solely emits one PVT solution per epoch, the *InDrive* concept allows to provide multiple solutions in case of temporal ambiguities. Due to the probabilistic interface on the output side, later stages (e.g. inside of a complex system architecture as in autonomous vehicles) can be used to finally select the most appropriate solution candidate based on further information (e.g. from digital maps or other exteroceptive sensors).

5 Conclusion

In this paper we addressed a specific segment or phenomenon of robust localization: NLOS. With our survey approach, we intend to present a systematic search for specific data relevant to the NLOS handling. Starting from the work presented in [14] and an analysis of several papers we have developed a taxonomy of handling NLOS in GNSS positioning. Thereby we integrated the fields of handling, the technology used and the intended objective. Afterwards, we identified types of evaluation methods and analyzed which are primarily used. Finally, we looked at current research programs and their technical design, data collection, handling, scope and application domain.

The output of Table 1 and the classification in Figure 4 give a good impression of the usage of satellite-based navigation. The majority of the reviewed papers are related to outdoor environments for any kind of vehicles. Many of them used real-world data to compare their solutions. As the used reference data (ground truth) is not the same within the data sets, an absolute comparison regarding the algorithmic performance is difficult. Hence, a quantitative comparison of the various solutions should be addressed in future work. Then, we applied the taxonomy of handling and evaluation methods to current research projects. It became clear that cooperative handling approaches as well as real-data measurements belong to the current standard. In addition, we were able to characterize the evaluation criteria of the researchers for their chosen technique and to outline an essential condition for the technique. Thus, we could determine the current researcher roadmap for dealing with NLOS. The roadmap is defined by the choice of several cooperating NLOS handling techniques in coordination with the respective research objective (mitigation, exploitation or elimination). Then, the evaluation of this technique is carried out under the conditions of low-cost, due to the availability, accuracy and reliability of the GNSS localization system. Real data are usually used for this purpose. In the future, we would like to use a common dataset to benchmark the used methods and techniques. These steps are needed to create technological innovations for our industrial and globally connected world - especially for system innovations in the area of ITS.

[3] http://www.indrive-project.eu/

Conflict of Interest The authors declare no conflict of interest.

Acknowledgment We want to thank Pierre Reisdorf for his contributions to the initial ITSC conference paper.

References

[1] J. Breßler, P. Reisdorf, M. Obst, and G. Wanielik, "Gnss positioning in non-line-of-sight context a survey," in *Intelligent Transportation Systems (ITSC), 2016 IEEE 19th International Conference on*. IEEE, 2016, pp. 1147–1154.

[2] R. Ercek, P. D. Doncker, and F. Grenez, "Study of Pseudo-Range Error Due to Non-Line-of-Sight-Multipath in Urban Canyons," in *Proceedings of the 18th International Technical Meeting of the Satellite Division of The Institute of Navigation (ION GNSS 2005)*, 2005, pp. 1083 – 1094.

[3] J. R. A. Watson, "High-Sensitivity GPS L1 Signal Analysis for Indoor Channel Modeling," Master's thesis, The University of Calgary, 2005.

[4] R. Kumar and M. G. Petovello, "A Novel GNSS Positioning Technique for Improved Accuracy in Urban Canyon Scenarios Using 3D City Model," in *Proceedings of the 27th International Technical Meeting of The Satellite Division of the Institute of Navigation (ION GNSS+ 2014)*, 2014, pp. 2139 – 2148.

[5] A. Guilloton, A. C. Escher, and D. Koenig, "Multipath study on the airport surface," in *Position Location and Navigation Symposium (PLANS), 2012 IEEE/ION*, April 2012, pp. 355 – 365.

[6] P. Misra and P. Enge, *Global Positioning System: Signals, Measurements, and Performance*. Ganga-Jamuna Press, 2011.

[7] P. W. Ward, J. W. Betz, and C. J. Hegarty, *Understanding GPS: Principles and Applications, Second Edition*. Artech House, 2006, ch. Chapter 6: Interference, Multipath, and Scintillation.

[8] P. D. Groves, Z. Jiang, M. Rudi, and P. Strode, "A Portfolio Approach to NLOS and Multipath Mitigation in Dense Urban Areas," in *Proceedings of the 26th International Technical Meeting of The Satellite Division of the Institute of Navigation (ION GNSS 2013)*, 2013, pp. 3231 – 3247.

[9] P. Groves, *Principles of GNSS, Inertial, and Multisensor Integrated Navigation Systems, Second Edition*, 2nd ed., ser. GNSS/GPS. Artech House, 2013.

[10] P. Furgale and et al., "Toward automated driving in cities using close-to-market sensors: An overview of the V-Charge Project," in *Intelligent Vehicles Symposium (IV), 2013 IEEE*, June 2013, pp. 809 – 816.

[11] L. Wang, P. Groves, and M. Ziebart, "Urban Positioning on a Smartphone: Real-time Shadow Matching Using GNSS and 3D City Models," in *Proceedings of the 26th International Technical Meeting of The Satellite Division of the Institute of Navigation (ION GNSS+ 2013)*, 2013, pp. 1606 – 1619.

[12] A. Bourdeau, M. Sahmoudi, and J.-Y. Tourneret, "Tight Integration of GNSS and a 3D City Model for Robust Positioning in Urban Canyons," in *Proceedings of the 25th International Technical Meeting of The Satellite Division of the Institute of Navigation (ION GNSS 2012)*, 2012, pp. 1263 – 1269.

[13] P. D. Groves, "Shadow Matching: A New GNSS Positioning Technique for Urban Canyons," *Journal of Navigation*, vol. 64, no. 03, pp. 417–430, 2011.

[14] M. Obst, "Bayesian Approach for Reliable GNSS-based Vehicle Localization in Urban Areas," Ph.D. dissertation, Technische Universität Chemnitz, 2015.

[15] C. J. Hegarty, "GNSS Signals—An Overview," in *2012 IEEE International Frequency Control Symposium Proceedings*, May 2012, pp. 1 – 7.

[16] F. van Diggelen, "GNSS Accuracy—Lies, Damn Lies and Statistics," *GPS World*, vol. 18, no. 1, Jan. 2007. [Online]. Available: http://www.frankvandiggelen.com/wp-content/uploads/2009/03/2007-gps-world-accuracy-article-0107-van-diggelen-1.pdf

[17] P. D. Groves, L. Wang, D. Walter, H. Martin, K. Voutsis, and Z. Jiang, "The Four Key Challenges of Advanced Multisensor Navigation and Positioning," in *2014 IEEE/ION Position, Location and Navigation Symposium - PLANS 2014*, May 2014, pp. 773 – 792.

[18] M. Adjrad and P. D. Groves, "Enhancing Conventional GNSS Positioning with 3D Mapping without Accurate Prior Knowledge," in *Proceedings of the 28th International Technical Meeting of The Satellite Division of the Institute of Navigation (ION GNSS+ 2015) September 14 - 18, 2015 Tampa Convention Center Tampa, Florida*, 2015.

[19] M. S. Braasch, "Performance comparison of multipath mitigating receiver architectures," in *Aerospace Conference, 2001, IEEE Proceedings.*, vol. 3, 2001, pp. 3/1309 – 3/1315 vol.3.

[20] P. D. Groves, Z. Jiang, B. Skelton, P. A. Cross, L. Lau, Y. Adane, and I. Kale, "Novel Multipath Mitigation Methods Using a Dual-Polarization Antenna," 2010.

[21] J. Ray, M. Cannon, and P. Fenton, "GPS Code and Carrier Multipath Mitigation Using a Multiantenna System," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 37, no. 1, pp. 183–195, 2001.

[22] L.-T. Hsu, Y. Gu, and S. Kamijo, "3D building model-based pedestrian positioning method using GPS/GLONASS/QZSS and its reliability calculation," *GPS Solutions*, pp. 1 – 16, 2015.

[23] L. Garin, F. van Diggelen, and J.-M. Rousseau, "Strobe & Edge Correlator Multipath Mitigation for Code," in *Proceedings of the 9th International Technical Meeting of the Satellite Division of The Institute of Navigation (ION GPS 1996)*, 1996, pp. 657 – 664.

[24] A. J. V. Dierendonck, P. Fenton, and T. Ford, "Theory and Performance of Narrow Correlator Spacing in a GPS Receiver," *NAVIGATION, Journal of the Institute of Navigation*, vol. 39, no. 3, pp. 115 – 124, 1992.

[25] G. A. McGraw, R. S. Young, K. Reichenauer, J. Stevens, and F. Ventrone, "GPS Multipath Mitigation Assessment of Digital Beam Forming Antenna Technology in a JPALS Dual Frequency Smoothing Architecture," in *Proceedings of the 2004 National Technical Meeting of The Institute of Navigation*, 2004, pp. 561 – 572.

[26] M. S. Braasch, "Multipath effects," in *Global Positioning System: Theory and Applications*, 1996.

[27] O. Osechas, K. J. Kim, K. Parsons, and Z. Sahinoglu, "Detecting Multipath Errors in Terrestrial GNSS Applications," in *Proceedings of the 2015 International Technical Meeting of The Institute of Navigation*, 2015.

[28] M. H. Keshvadi, A. Broumandan, and G. Lachapelle, "Analysis of GNSS Beamforming and Angle of Arrival Estimation in Multipath Environments," *Proceedings of the Institute of Navigation 2011 International Technical Meeting (ION ITM), San Diego, CA, USA*, pp. 24 – 26, 2011.

[29] P. D. Groves and Z. Jiang, "Height Aiding, C/N_0 Weighting and Consistency Checking for GNSS NLOS and Multipath Mitigation in Urban Areas," *Journal of Navigation*, vol. 66, no. 05, pp. 653 – 669, 2013.

[30] O. Le Marchand, P. Bonnifait, J. Bañez-Guzmán, F. Peyret, and D. Betaille, "Performance Evaluation of Fault Detection Algorithms as Applied to Automotive Localisation," in *European Navigation Conference - GNSS 2008*, Toulouse, France, Apr. 2008.

[31] C. Rost and L. Wanninger, "Carrier Phase Multipath Mitigation Based on GNSS Signal Quality Measurements," *Journal of Applied Geodesy*, vol. 3, no. 2, pp. 81 – 87, 2009.

[32] A. Bilich, K. M. Larson, and P. Axelrad, "Modeling GPS phase multipath with SNR: Case study from the Salar de Uyuni, Boliva," *Journal of Geophysical Research (Solid Earth)*, vol. 113, p. B04401, Apr. 2008.

[33] C. J. Comp and P. Axelrad, "Adaptive SNR-based Carrier Phase Multipath Mitigation Technique," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 34, no. 1, pp. 264 – 276, Jan 1998.

[34] A. Hirrle, H.-J. Jentschel, and O. Michler, "Estimation of Multipath Parameters using Hilbert Huang Transform," in *Proceedings of the 25th International Technical Meeting of The Satellite Division of the Institute of Navigation (ION GNSS 2012)*, September 2012, pp. 67 – 78.

[35] M. Kim, K. Kim, J. Lee, and S. Pullen, "High Integrity GNSS Navigation and Safe Separation Distance to Support Local-Area UAV Networks," in *Proceedings of the 27th International Technical Meeting of The Satellite Division of the Institute of Navigation (ION GNSS+ 2014)*, 2014, pp. 869 – 878.

[36] O. Le Marchand, P. Bonnifait, J. Ibanez-Guzmán, D. Betaille, and F. Peyret, "Characterization of GPS multipath for passenger vehicles across urban environments," *ATTI dell'Istituto Italiano di Navigazione*, no. 189, pp. 77 – 88, 2009.

[37] L. Li, J. Zhong, and M. Zhao, "Doppler-Aided GNSS Position Estimation With Weighted Least Squares," *IEEE Transactions on Vehicular Technology*, vol. 60, no. 8, pp. 3615 – 3624, Oct 2011.

[38] Z. Jiang and P. D. Groves, "NLOS GPS signal detection using a dual-polarisation antenna," *GPS solutions*, vol. 18, no. 1, pp. 15 – 26, 2014.

[39] L. Wang, P. D. Groves, and M. K. Ziebart, "Multi-Constellation GNSS Performance Evaluation for Urban Canyons using Large Virtual Reality City Models," *Journal of Navigation*, vol. 65, no. 03, pp. 459 – 476, 2012.

[40] Y. Suh and R. Shibasaki, "Evaluation of Satellite-Based Navigation Services in Complex Urban Environments Using a Three-Dimensional GIS," *Ieice Transactions*, vol. 90-B, pp. 1816 – 1825, 2007.

[41] J. Bradbury, M. Ziebart, and P. Cross, "Code Multipath Modelling in the Urban Environment Using Large Virtual Reality City Models: Determining the Local Environment," *The Journal of Navigation*, vol. 60, pp. 95 – 105, 2007.

[42] T. Suzuki and N. Kubo, "Correcting GNSS Multipath Errors Using a 3D Surface Model and Particle Filter," in *Proceedings of the 26th International Technical Meeting of The Satellite Division of the Institute of Navigation (ION GNSS 2013)*, 2013, pp. 1583 – 1595.

[43] A. T. Irish, D. Iland, J. T. Isaacs, J. a. P. Hespanha, E. M. Belding, and U. Madhow, "Using Crowdsourced Satellite SNR Measurements for 3D Mapping and Real-time GNSS Positioning Improvement," in *Proceedings of the 6th Annual Workshop on Wireless of the Students, by the Students, for the Students*, ser. S3 '14. New York, NY, USA: ACM, 2014, pp. 5 – 8.

[44] M. Obst, S. Bauer, and G. Wanielik, "Urban Multipath Detection and Mitigation with Dynamic 3D Maps for Reliable Land Vehicle Localization," in *Position Location and Navigation Symposium (PLANS), 2012 IEEE/ION*, April 2012, pp. 685 – 691.

[45] P. D. Groves, Z. Jiang, L. Wang, and M. K. Ziebart, "Intelligent Urban Positioning Using Multi-constellation GNSS with 3D Mapping and NLOS Signal Detection," in *ION GNSS 2012*, 2012, pp. 458 – 472.

[46] B. Ben-Moshe, E. Elkin, H. Levi, and A. Weissman, "Improving Accuracy of GNSS Devices in Urban Canyons," in *CCCG*, 2011.

[47] S. Peyraud, D. Bétaille, S. Renault, M. Ortiz, F. Mougel, D. Meizel, and F. Peyret, "About Non-Line-Of-Sight Satellite Detection and Exclusion in a 3D Map-Aided Localization Algorithm," *Sensors*, vol. 13, no. 1, p. 829, 2013.

[48] F. Peyret, D. , Bétaille, and F. Mougel, "Non-Line-Of-Sight GNSS Signal Detection Using an On-Board 3D Model of Buildings," in *ITS Telecommunications (ITST), 2011 11th International Conference on*. IEEE, 2011, pp. 280 – 286.

[49] L.-T. Hsu, Y. Gu, and S. Kamijo, "NLOS Correction/Exclusion for GNSS Measurement Using RAIM and City Building Models," *Sensors*, vol. 15, no. 7, p. 17329, 2015.

[50] J. i. Meguro, T. Murata, J. i. Takiguchi, Y. Amano, and T. Hashizume, "GPS Multipath Mitigation for Urban Area Using Omnidirectional Infrared Camera," *IEEE Transactions on Intelligent Transportation Systems*, vol. 10, no. 1, pp. 22 – 30, March 2009.

[51] F. Peyret, D. Betaille, M. Ortiz, S. Miquel, and L. Fontenay, "How to Improve GNSS Positioning Quality of Service for Demanding ITS in City Environments by Using 3D Digital Maps," in *19th ITS World Congress*, 2012.

[52] M. Petovello, C. O'Driscoll, and G. Lachapelle, "Weak Signal Carrier Tracking using Extended Coherent Integration with an Ultra-Tight GNSS/IMU Receiver," *Proc. ENC-GNSS 2008*, 2008.

[53] M. Petovello and G. Lachapelle, "Comparison of Vector-Based Software Receiver Implementations with Application to Ultra-Tight GPS/INS Integration," *Proceedings of Institute of Navigation GPS/GNSS Conference, Fort Worth, TX. Institute of Navigation*, 2006.

[54] A. Soloviev and F. V. Graas, "Use of Deeply Integrated GPS/INS Architecture and Laser Scanners for the Identification of Multipath Reflections in Urban Environments," *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 5, pp. 786 – 797, Oct 2009.

[55] A. Soloviev, C. Toth, and D. Grejner-Brzezinska, "Performance of Deeply Integrated GPS/INS in Dense Forestry Areas," in *Proceedings of the 24th International Technical Meeting of The Satellite Division of the Institute of Navigation (ION GNSS 2011)*, 2011, pp. 2427 – 2435.

[56] J. W. Cheong, E. Glennon, A. G. Dempster, D. Serant, and T. Calmettes, "Modelling and Mitigating Multipath and NLOS for Cooperative Positioning in Urban Canyons," 2015.

[57] P. D. Groves, C. J. Mather, and A. A. Macaulay, "Demonstration of Non-coherent Deep INS/GPS Integration for Optimised Signal-to-Noise Performance," 2007.

[58] E. López Casariego, E. Pérez, G. Falco, M. Campo-Cossío Gutiérrez, F. Zacchello, and S. Bories, "Low-cost Real-time Tightly-coupled GNSS/INS Navigation System Based on Carrier Phase Double Differences for UAV Applications," in *Proceedings of the 27th International Technical Meeting of The Satellite Division of the Institute of Navigation (ION GNSS+ 2014)*, Tampa, Florida, USA, 09/2014 2014, pp. 841 – 857.

[59] P. Xie and M. G. Petovello, "Improving High Sensitivity Receiver Performance in Multipath Environment for Vehicular Applications," in *Proceedings of the 25th International Technical Meeting of the Satellite Division of the Institute of Navigation (ION GNSS 2012), Nashville, TN, USA*, vol. 1721, 2012, p. 448458.

[60] P. Xie, M. G. Petovello, and C. Basnayake, "Multipath Signal Assessment in the High Sensitivity Receivers for Vehicular Applications," in *Proceedings of the ION GNSS*, 2011.

[61] L. R. Weill, "A High Performance Code and Carrier Tracking Architecture for Ground-Based Mobile GNSS Receivers," in *Proceedings of the ION GNSS Conference*, 2010.

[62] F. van Graas, A. Soloviev, M. U. de Haag, and S. Gunawardena, "Closed-Loop Sequential Signal Processing and Open-Loop Batch Processing Approaches for GNSS Receiver Design," *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 4, pp. 571 – 586, Aug 2009.

[63] M. Petovello, C. O'Driscoll, and G. Lachapelle, "Carrier Phase Tracking of Weak Signals Using Different Receiver Architectures," in *Proceedings of the ION NTM Conference, San Diego, CA, USA*, vol. 2830, 2008, p. 781791.

[64] M. Uijt de Haag, "An Investigation into the Application of Block Processing Techniques for the Global Positioning System," Ph.D. dissertation, Ohio University, 1999.

[65] D. Flament, D. Brocard, W. Ochieng, and C. Milner, "RAIM in dual frequency / multi constellation APV/LPV operations in aeronautics," in *2010 5th ESA Workshop on Satellite Navigation Technologies and European Workshop on GNSS Signals and Signal Processing (NAVITEC)*, Dec 2010, pp. 1 – 7.

[66] L.-T. Hsu, P. D. Groves, and S.-S. Jan, "Assessment of the Multipath Mitigation Effect of Vector Tracking in an Urban Environment," 2013.

[67] M. Kaddour, N. A. Tmazirte, M. E. E. Najjar, Z. Naja, and N. Moubayed, "Fault detection and exclusion of cycle slips for carrier-phase in GNSS positioning," in *Information Fusion (Fusion), 2015 18th International Conference on*, July 2015, pp. 393 – 398.

[68] J. Marais, E. Duflos, N. Viandier, D. F. Nahimana, and A. Rabaoui, "Advanced signal processing techniques for multipath mitigation in land transportation environment," in *Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on*, Sept 2010, pp. 1480 – 1485.

[69] C. O'Driscoll, G. Lachapelle, and M. Tamazin, "Investigation of the Benefits of Combined GPS/Glonass Receivers in Urban Environments," in *Proceeding on RIN NAV10 Conference on Position, Location, Timing: Everyone, Everything, Everywhere*, 2010.

[70] S. Feng, S. Wang, and W. Ochieng, "A Core Constellation Based Multiple-GNSS Positioning and Integrity Monitoring Algorithm," in *Proceedings of the 27th International Technical Meeting of The Satellite Division of the Institute of Navigation (ION GNSS+ 2014)*, 2014, pp. 307 – 314.

[71] S. Miura and S. Kamijo, "GPS Error Correction by Multipath Adaptation," *International Journal of Intelligent Transportation Systems Research*, vol. 13, no. 1, pp. 1–8, 2014.

[72] M. Obst and G. Wanielik, "Probabilistic Non-Line-of-Sight Detection in Reliable Urban GNSS Vehicle Localization Based on an Empirical Sensor Model," in *Intelligent Vehicles Symposium (IV), 2013 IEEE*, June 2013, pp. 363 – 368.

[73] A. L. Bilich, "Improving the precision and accuracy of geodetic GPS: applications to multipath and seismology," Ph.D. dissertation, University of Colorado, 2006.

[74] S. Bauer, R. Streiter, and G. Wanielik, "Non-line-of-sight mitigation for reliable urban gnss vehicle localization using a particle filter," in *Information Fusion (Fusion), 2015 18th International Conference on*, July 2015, pp. 1664–1671.

[75] L. Wang, P. D. Groves, and M. K. Ziebart, "GNSS Shadow Matching: Improving Urban Positioning Accuracy Using a 3D City Model with Optimized Visibility Prediction Scoring," in *Proceedings of the 25th International Technical Meeting of The Satellite Division of the Institute of Navigation (ION GNSS 2012)*, 2012, pp. 423 – 437.

[76] R. Yozevitch, B. Ben-Moshe, and A. Dvir, "GNSS Accuracy Improvement Using Rapid Shadow Transitions," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 3, pp. 1113 – 1122, June 2014.

[77] J. T. Isaacs, A. T. Irish, F. Quitin, U. Madhow, and J. P. Hespanha, "Bayesian Localization and Mapping Using GNSS SNR Measurements," in *2014 IEEE/ION Position, Location and Navigation Symposium - PLANS 2014*, May 2014, pp. 445 – 451.

[78] L. Wang, P. D. Groves, and M. K. Ziebart, "Smartphone Shadow Matching for Better Cross-street GNSS Positioning in Urban Environments," *Journal of Navigation*, vol. 68, pp. 411 – 433, 5 2015.

[79] R. Yozevitch and B. B. Moshe, "A Robust Shadow Matching Algorithm for GNSS Positioning," in *NAVIGATION, Journal of the Institute of Navigation*, vol. 62, no. 2, 2015, pp. 95 – 109.

[80] P. D. Groves, Z. Jiang, L. Wang, and M. Ziebart, "Intelligent Urban Positioning, Shadow Matching and Non-Line-of-Sight signal detection," in *Satellite Navigation Technologies and European Workshop on GNSS Signals and Signal Processing, (NAVITEC), 2012 6th ESA Workshop on*, Dec 2012, pp. 1 – 8.

[81] D. Betaille, F. Peyret, M. Ortiz, S. Miquel, and L. Fontenay, "A New Modeling Based on Urban Trenches to Improve GNSS Positioning Quality of Service in Cities," *Intelligent Transportation Systems Magazine, IEEE*, vol. 5, no. 3, pp. 59 – 70, 2013.

[82] V. Drevelle and P. Bonnifait, "Global positioning in urban areas with 3-D maps," in *Intelligent Vehicles Symposium (IV), 2011 IEEE*, June 2011, pp. 764 – 769.

[83] V. Drevelle and P. Bonnifait, "iGPS: Global Positioning in Urban Canyons with Road Surface Maps," *IEEE Intelligent Transportation Systems Magazine*, vol. 4, no. 3, pp. 6 – 18, 2012.

[84] Z. Jiang, P. D. Groves, W. Y. Ochieng, S. Feng, C. D. Milner, and P. G. Mattos, "Multi-Constellation GNSS Multipath Mitigation Using Consistency Checking," *Proceedings of the 24th International Technical Meeting of The Satellite Division of the Institute of Navigation (ION GNSS 2011)*, pp. 3889 – 3902, 2011.

[85] N. Alam, A. T. Balaei, and A. G. Dempster, "Relative Positioning Enhancement in VANETs: A Tight Integration Approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 1, pp. 47 – 55, March 2013.

[86] H. Isshiki, "A New Method for Detection, Identification and Mitigation of Outliers in Receiver Autonomous Integrity Monitoring (RAIM)," in *Proceedings of the 26th International Technical Meeting of The Satellite Division of the Institute of Navigation (ION GNSS 2013)*, 2013, pp. 151 – 158.

[87] Z. Jiang and P. D. Groves, "GNSS NLOS and Multipath Error Mitigation using Advanced Multi-Constellation Consistency Checking with Height Aiding," 2012.

[88] R. Moradi, W. Schuster, S. Feng, A. Jokinen, and W. Ochieng, "The carrier-multipath observable: a new carrier-phase multipath mitigation technique," *GPS Solutions*, vol. 19, no. 1, pp. 73 – 82, 2014.

[89] J. Li, R. Wu, D. Lu, and W. Wang, "GPS Multipath Mitigation Algorithm Based on Signal Correlation Character," *International Journal of Digital Content Technology and its Applications(JDCTA)*, vol. 7, no. 3, pp. 837 – 846, February 2013.

[90] J. Li, Y. Li, and Y. Zhou, *The Proceedings of the Second International Conference on Communications, Signal Processing, and Systems*. Cham: Springer International Publishing, 2014, ch. GPS Multipath Mitigation Algorithm Using C/A Code Correlation Character, pp. 1047 – 1058.

[91] P. Closas, C. Fernandez-Prades, and J. A. Fernandez-Rubio, "A Bayesian Approach to Multipath Mitigation in GNSS Receivers," *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 4, pp. 695 – 706, Aug 2009.

[92] N. Mattern, M. Obst, R. Schubert, and G. Wanielik, "Simulative Analysis of Accuracy Demands of Co-Operative Localization in the COVEL Project," in *Intelligent Vehicles Symposium (IV), 2011 IEEE*, June 2011, pp. 516 – 521.

[93] F. de Ponte Müller, E. M. Diaz, B. Kloiber, and T. Strang, "Bayesian cooperative relative vehicle positioning using pseudorange differences," in *2014 IEEE/ION Position, Location and Navigation Symposium - PLANS 2014*, May 2014, pp. 434 – 444.

[94] M. Spangenberg, J. Y. Tourneret, V. Calmettes, and G. Duchateau, "Detection of Variance Changes and mean Value Jumps in Measurement Noise for Multipath Mitigation in Urban Navigation," in *2008 42nd Asilomar Conference on Signals, Systems and Computers*, Oct 2008, pp. 1193 – 1197.

[95] J. Wang and P. B. Ober, "On the Availability of Fault Detection and Exclusion in GNSS Receiver Autonomous Integrity Monitoring," *Journal of Navigation*, vol. 62, no. 02, pp. 251 – 261, 2009.

[96] A. Rabaoui, N. Viandier, J. Marais, and E. Duflos, "Using Dirichlet Process Mixtures for the Modelling of GNSS Pseudorange Errors in Urban Canyon," in *Proceedings of the 22nd International Technical Meeting of The Satellite Division of the Institute of Navigation (ION GNSS 2009)*, 2009, pp. 2391 – 2399.

[97] A. Rabaoui, N. Viandier, J. Marais and E. Duflos, "On the use of Dirichlet process mixtures for the modelling of pseudorange errors in multi-constellation based localisation," in *Intelligent Transport Systems Telecommunications,(ITST), 2009 9th International Conference on*, Oct 2009, pp. 465 – 470.

[98] P. Axelrad, C. J. Comp, and P. F. Macdoran, "SNR-Based Multipath Error Correction for GPS Differential Phase," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 32, no. 2, pp. 650 – 660, April 1996.

[99] L. Serrano, "Carrier-Phase Multipath Mitigation In RTK-Based GNSS Dual-Antenna Systems," Ph.D. dissertation, Ph.D. dissertation, Department of Geodesy and Geomatics Engineering, Technical Report No. 287, University of New Brunswick, Fredericton, New Brunswick, Canada, 227 pp., 2013.

ASTES

Advances in Science, Technology and Engineering Systems Journal

Vol. 2, No. 3, 732-740 (2017)
www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTES Journal
ISSN: 2415-6698

Collision avoidance between a wheelchair front wheels and a step wall during step climbing using a care robot

Hidetoshi Ikeda[*,1], Kazuki Hashimoto[2], Daisuke Murayama[2], Rikuto Yamazaki[2], Eiji Nakano[3]

[1]*Department of Mechanical Engineering, National Institute of Technology, Toyama College, Toyama, 939-8630, Japan*
[2]*Department of Electrical and Control Systems Engineering, National Institute of Technology, Toyama College, Toyama, 939-8630, Japan*
[3]*Robofesta org., Tsukuba, 300-2636, Japan*

ARTICLEINFO

Article history:
Received: 27 April, 2017
Accepted: 06 June, 2017
Online: 16 June, 2017

Keywords:
Robot
Wheelchair
Step climbing collision
Avoidance step wall

ABSTRACT

This paper discusses a step climbing strategy using a wheelchair and a teleoperated care robot that avoids collisions between the wheelchair's front wheels and the step wall. In our method, the two vehicles are connected and use the velocity difference between them to climb steps. Since the wheelchair's front wheels, or another part of the wheelchair, tend to collide with the step being mounted while the wheelchair front wheels are in the lifted position, we investigated a system that permits switching between teleoperation and autonomous control to avoid this problem. In our proposed system, the wheelchair step climbing process is performed under autonomous control, while the robot step climbing process is performed under teleoperation control by a remote caregiver. Theoretical analysis and accompanying calculations were performed to clarify the most suitable lifting position for wheelchair front wheels when climbing, and our experimental results show the overall effectiveness of this system.

1 Introduction

This paper is an extension of work originally presented in *The 2016 IEEE International Conference on Simulation, Modeling, and Programming for Autonomous Robots* [1].

For disabled people, using wheelchairs is an effective means for locomotion. However, since it is difficult for wheelchairs to negotiate steps, user movements are often restricted. Some reasons for the wheelchair's difficulty include driving wheel slippage and/or vehicle imbalance. Numerous research programs aimed at improving wheelchair mobility on steps have been conducted, and the following potential solutions have been investigated: additional legs [2], multiple wheels and variable wheelbases [3], a combination of an adjustable center of gravity (COG) and multiple wheels [4], tracked vehicles [5], an additional driving wheel system [6], and changing the wheelchair posture to facilitate climbing [7]. In addition, Yoshikazu Mori studied a new step climbing and descending strategy using a manual wheelchair equipped with linear actuator mechanisms and a light portable ramp [8].

This study's research group has previously investigated a cooperative step climbing method that can be used by a robot and a manual wheelchair connected via a passive link [9], and by using a teleoperated robot controlled by a caregiver through an intranet and a conventional or electric drive wheelchair [10][11]. Our step climbing system, which is capable of reducing the burden imposed on the caregiver, takes advantage of the wheelchair user's ability to judge external information (e.g., obstacles and distances between vehicles and steps), and the caregiver's ability to teleoperate the system through an intranet.

To operate this step climbing system, both the wheelchair user and the caregiver will need sufficient training, particularly for the period when the wheelchair's front wheels are close to the step, which is where the risk of a collision between the wheels and step is high. Furthermore, since this system uses autonomous control during the wheelchair step climbing process, it was necessary to implement a support system that can switch between teleoperation and autonomous control.

While the primary purpose of this paper is to describe our proposed support system for avoiding collisions between the wheelchair front wheels and step during stair ascents [1], we will also describe the concept behind our step climbing method, and provide a detailed theoretical analy-

[*]Corresponding Author: Hidetoshi Ikeda, 13, Hongou-chou, Toyama, Japan, 939-8630, & ikedah@nc-toyama.ac.jp

sis, similation and the results of an experiment describing its use.

Other reports of multiple vehicles cooperating to cross irregular terrain include those of Asama *et al.* [12], who considered a forklift system, and Gaston *et al.* [13], who studied stair climbing robots using a mechanism similar to that of Asama *et al.*

Based on our preliminary measurements of the friction coefficients for the robot under wet and dry conditions on asphalt, concrete, wood, and interior flooring, the ground surface considered in this research was assumed to have a friction coefficient within the range of 0.6 to 0.9. The heights of steps located at the entrances of typical buildings and other structures were also measured, and a target step height of 120 mm, which accounts for more than 80% of the observed heights, was set.

This remainder of this paper is organized as follows: Section 2 describes the concept of cooperative step climbing, and Section 3 describes the robot and wheelchair used. Section 4 describes the step climbing process, and Section 5 provides a theoretical analysis of our proposed method. Section 6 describes an experiment using our proposed system and the results, and our conclusion is given in Section 7.

2 Concept behind wheelchair care robot

As described above, since it is difficult for wheelchairs to climb steps, wheelchair user movements tend to be restricted. Most wheelchairs have two casters employed as front wheels and two driving rear wheels, and the center of gravity of a wheelchair with a seated user is high. Thus, when a wheelchair attempts to climb a step, high potentials for driving wheel slippage (Figure 1 (a)) and/or tipping over the vehicle (Figure 1 (b)) arise. In addition to step climbing situations, wheelchair users face difficulties at numerous other times. For example, it is difficult for a wheelchair user to open a door (Figure 2 (a)) or pick up an object that is on the floor or placed on a high shelf (Figure 2 (b)). Thus, under most current circumstances, most wheelchair users need a human assistant. However, taking care of a wheelchair and its user is hard work.

(a) (b)

Figure 1: Reasons why wheelchairs are weak against steps: (a) driving wheel slippage, and (b) possibility of tipping over when the rear wheels transverse the step

(a) (b)

Figure 2: Obstacles for wheelchair users other than step climbing: (a) door opening, and (b) picking up objects

The aim of this research is to create a teleoperated robot system that can be used by caregivers to provide assistance to vulnerable but esteemed persons (such as their elderly parents) from a distant location. In the next section, we will begin by evaluating a method for avoiding collisions between the front wheels of a wheelchair and a step wall.

3 Wheelchair and robot

The robot used in this research is “Tateyama”, which is a wheeled robot developed in this laboratory (Figure 3).

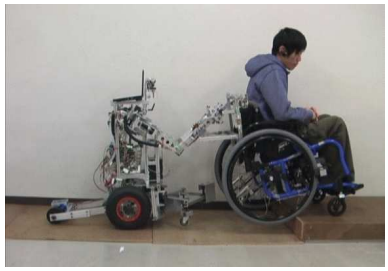


Figure 3: The wheelchair and robot

Table 1: Robot specifications

Overall length	230 to 800 mm
Overall height	747 mm
Radius of front wheels (r_{Bf})	25 mm
Radius of middle wheels (R_B)	145 mm
Radius of rear wheels (r_{Br})	19 mm
Wheelbase (WB_f)	190 to 440 mm
Wheelbase (WB_r)	270 mm
Distance of center of gravity from the middle wheel axes (l_{rB})	93 mm
Height of the mass above the middle wheel axes (h_{mB})	286 mm
Distance of Joint 2 from the rear axes (l_{LB})	90 mm
Height of Joint 2 above the rear axes (h_{LB})	532 mm
Mass of the robot’s body	55 kg
Mass of upper links (from Joint 2 to 4)	2.55 ×2 kg
Mass of forearm links (form Joint 4 to hand)	0.8 ×2 kg
Length of Link 2 (l_2)	330 mm
Length of Link 4 (l_4)	300 mm
Length of the hand (l_6)	105 mm
Length from Joint 4 to the connecting position (l_{4C})	370 mm

Table 2: Wheelchair specifications

Overall length	1060 mm
Overall height	985 mm
Radius of front wheels (r_A)	63 mm
Radius of rear wheels (R_A)	300 mm
Wheelbase (l_A)	430 mm
Hand rim position (l_{LA})	250 mm
Distance of center of gravity from the rear wheel axles (l_{rA})	149 mm
Height of mass above the rear wheel axles (h_{mA})	371 mm
Height of push-handle (h_{LA})	435 mm
Mass (wheelchair + driver) (M_A)	92.7 kg

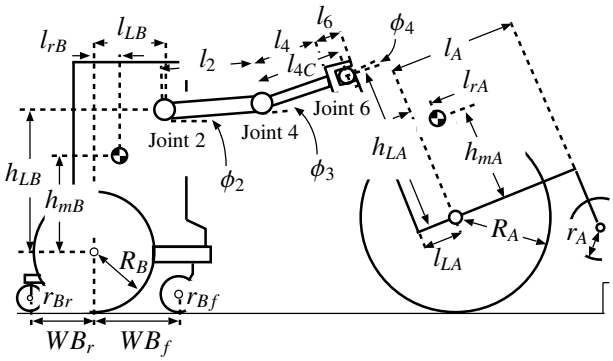


Figure 4: Model of wheelchair and robot

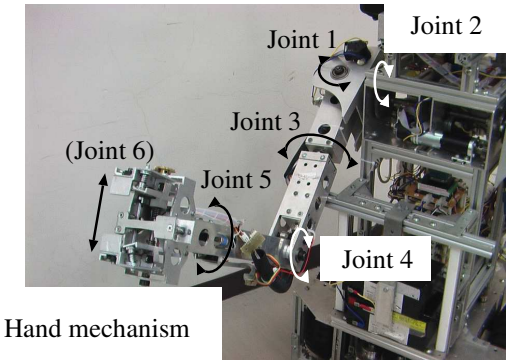
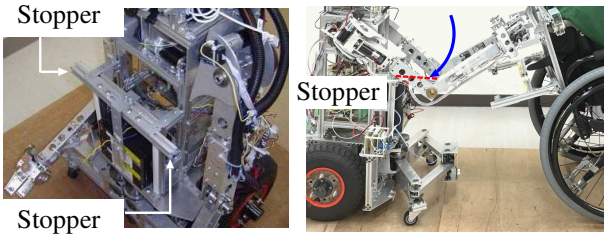


Figure 5: Robot manipulator



(a) (b)

Figure 6: Rotary motion control of the robot shoulder using the robot body: (a) Front body of the robot, and (b) side view of stopper on the robot chest

The process begins with the wheelchair and robot deployed in a forward-and-aft configuration to traverse the step (Figure 4). When the wheelchair and the robot encounter a step, the robotic hands grasp the rotary shaft of the wheelchair push handle [14]. Table 1 and Table 2 list the specifications of the robot and wheelchair, respectively. The wheelchair and robot move at a constant speed (0.76 km/h). This robot is equipped with three pairs of wheels, and each pair consists of left and right wheels. The front pair are casters, whereas the middle and rear pairs are driving wheels. The front and rear pairs can be folded.

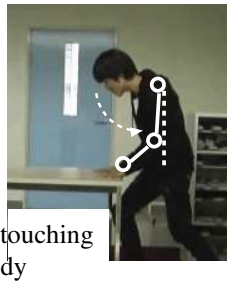
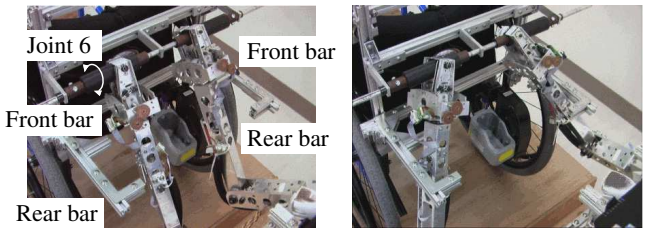


Figure 7: Human pushing an object



(a) (b)

Figure 8: Sides of the robot: (a) in wheelchair climbing position, and (b) in robot climbing position

(a) (b)

Figure 9: Stopper actions: (a) lifting the robot wheels, and (b) preventing the robot from falling down

The robot has manipulators attached to the left and right joints of its upper half. Since each arm has five degrees of freedom (5-DOF) and the hand has 1-DOF, the arm has a total of 6-DOF (Figure 5). In this paper, the length of the upper arm link (from Joint 2 (shoulder) to Joint 4 (elbow)) is l_2; the length of the forearm link (from Joint 4 to Joint 6) is l_{4C}. The manipulator joint angles are -90 deg $\leq \phi_2 \leq +90$ deg and $0 \leq \phi_3 \leq +100$ deg. The right and left hands each have two fingers (Joint 6) in order to grasp and hold items (Figure 5).

The robot is equipped with a stopper mounted on the front part of its body (Figure 6 (a)), as described below, and is capable of limiting the passive rotational travel of the manipulators when it is pushing the wheelchair (Figure 6

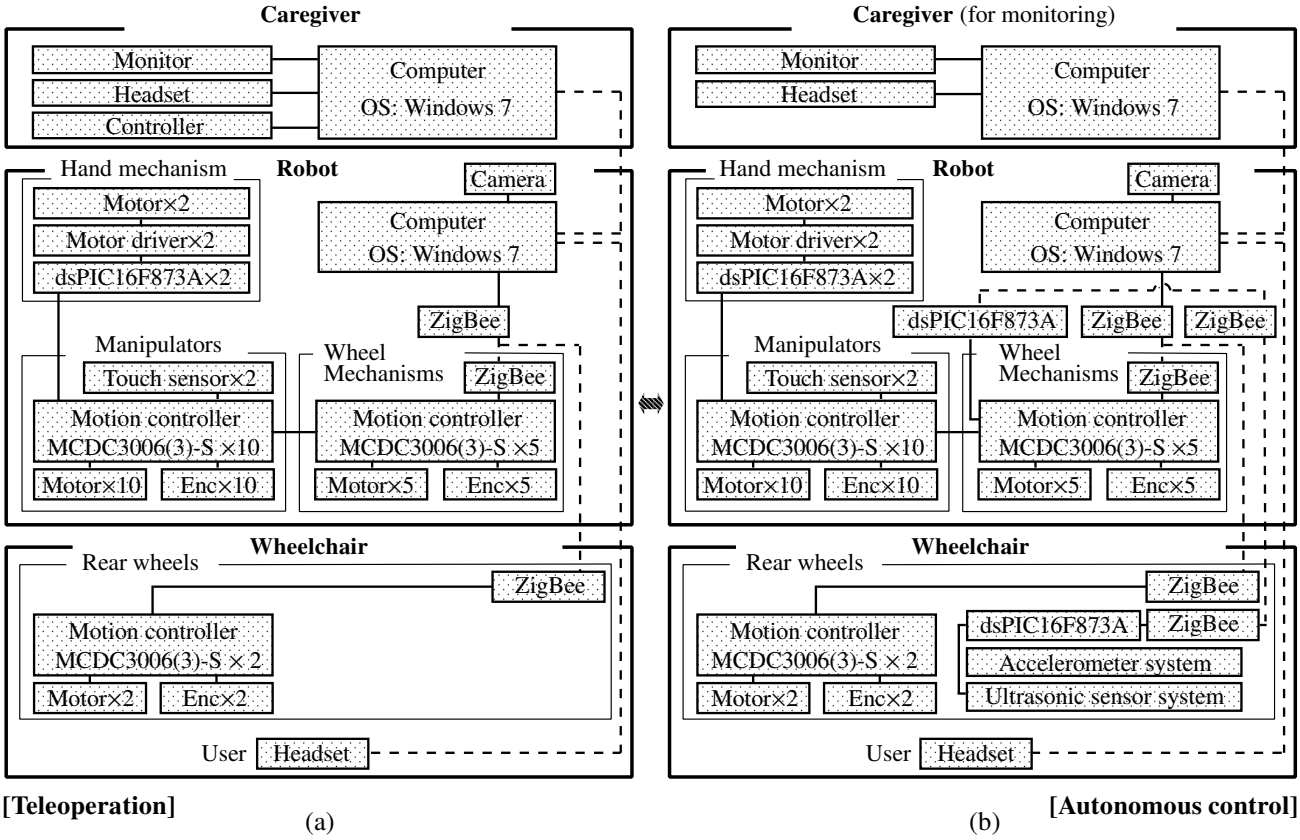


Figure 10: Working devices of the entire system: (a) teleoperation mode, and (b) autonomous control mode

(b)). This enables the robot to imitate the operation of a human pushing an object (Figure 7). The robot does not need to exert force around the shoulder axes because its upper arms push against the chest (i.e., the stopper) when pushing the wheelchair [11]. The motors mounted on the robot are connected to Faulhaber MCDC3003-S and MCDC3006-S motion controllers (Faulhaber-Global, Schnaich, Germany). These are connected to a notebook personal computer (PC) mounted on the robot. The shape of the Integral-ME wheelchair (NOVA, Carson, Ca) used in this study is typical among wheelchairs available on the market (Figure 3). The rear-wheel drive wheelchair is equipped with a TRD-1 electric drive unit (Acritech Co., Ltd.). Two motion controllers are added for the electric drive unit. The motion controllers on the wheelchair are connected to the notebook PC installed on the robot via a ZigBee wireless communications module. The wheelchair has a push handle mechanism on its back (Figure 8 (a)) that is equipped with a rotary shaft in order to permit passive rotation of the handle. When the robot hands grasp this shaft, the two vehicles are connected. The stopper of the wheelchair is composed of front and rear bars. During the robot climbing process, the sides of the robot are opened, and the right and left manipulators are inserted into the stopper (Figure 8 (b)). The dual manipulators of the robot are equipped with touch sensors that allow them to detect contact with the stopper. The robot pushes the front bars by using the forearm links of the manipulators to lift its front wheels (Figure 9 (a)). The rear bars are used to prevent the robot from tipping over backward when its center of gravity shifts behind the contact point between the middle wheels and the ground (Figure 9 (b)).

The robot control mode can be switched between network-based teleoperation and autonomous control by the caregiver. In Figure 10, (a) is a system configuration diagram that shows the working device in the teleoperation mode, and Figure 10 (b) shows the system diagram of the autonomous control mode. These system configuration diagrams show only the working devices in each mode. When engaged in wheelchair step climbing, this system uses the autonomous control. For robot step climbing, the system uses the teleoperation mode.

3.1 Teleoperation

The robot is equipped with a camera, and the video imagery transmitted from it and the Motion Manager 5 operating window are displayed on the robot's notebook PC. The screen on this notebook PC uses RealVNC (RealVCN Ltd., Cambridge, UK) software to transmit imagery as-is over the intranet to the display of the caretaker's PC (robot operator). The caregiver can control the robot by operating Motion Manager 5 (Faulhaber-Global) from his PC. The caregiver and the wheelchair user both wear headsets. The caregiver's headset is connected to his/her PC, and the chair user's headset is connected to the robot PC via Bluetooth. The caregiver and the wheelchair user verbally telecommunicate via the Skype (Microsoft Corp., Redmond, Wa) communication application. Keyboard commands, which are activated by pushing the buttons of a joypad, correspond to the controller manipulations necessary to teleoperate the robot. Switching between teleoperation and autonomous control is carried out by the controller.

3.2 Autonomous control

To lift the front wheels of the wheelchair, the robot stops while the wheelchair drive continues moving (see Section 4). However, network teleoperation is difficult for this process because the wheelchair's raised front wheels tend to collide with the step when they are in near proximity with each other (Figure 11).

An accelerometer system and an ultrasonic sensor system were added onto the chair, and a support system for the caregiver was created. The accelerometer system detects the incline of the chair, and the ultrasonic sensor system measures the distance between the chair and the step.

Both sensor systems are connected to the robot's PC by a ZigBee wireless communication module. Since the support system for the caregiver is capable of performing the wheelchair step climbing process automatically, the caregiver (robot operator) does not need to control the vehicles in the autonomous control mode. However, the caregiver can watch the live video from the robot's camera, and the wheelchair user and the caregiver are in direct verbal communication, so if the vehicles encounter a problem, the caregiver switches from autonomous to teleoperation control to handle it.

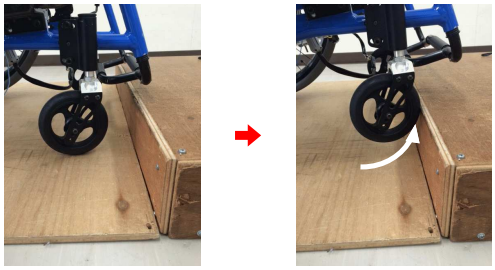


Figure 11: Collision between the front wheels and step

4 Step climbing process

When the vehicles encounter a step, the robot hands grasp the rotary shaft of the wheelchair's push handle mechanism, thus linking the two vehicles [14]. Joints 2, 4, and 6 are allowed to rotate passively (Figs. 5 and 8). In this study, Stages 1 and 2 signify the processes in which the wheelchair front and back wheels, respectively, ascend the step. Similarly, Stages 3 and 4 signify the processes in which the robot front and other wheels, respectively, climb the step.

Both vehicles move automatically during the wheelchair step climbing process (Stages 1 and 2) by using the caregiver's support system. After that, the caregiver switches from autonomous control to teleoperation, and the robot step climbing process (Stage 3 and 4) is carried out. The ascent process is described below (the states shown in Figure 12 correspond to (1) through (16) below.

[Stage 1]

(1) From Stage 1 to Stage 2, the vehicles are under autonomous control. The wheelchair and the robot move forward until the ultrasonic sensor system detects a suitable position where the wheelchair front wheels do not collide with the step. (2) The robot stops and the wheelchair continues forward until the accelerometer system detects the wheelchair inclination level that would place the front wheels on the step. (3) At this point, the wheelchair center of mass is in front of the contact point between the rear wheels and the ground. If it shifts to behind the contact point as the wheelchair tilt increases, the stopper of the robot limits the passive rotation around Joint 2, which prevents the wheelchair from tipping over backward (Figure 6 (b)). (4) Both vehicles move forward, and the front wheels of the wheelchair are placed on the step.

[Stage 2]

(5) Both vehicles continue to move forward. (6) The back wheels of the wheelchair come into contact with the step. (7) The robot continues to push the wheelchair so that the rear wheels of the wheelchair climb up onto the step. The robot supports the wheelchair during this process to prevent the wheelchair from tipping over backward. (8) Both vehicles move forward, and the wheelchair's rear wheels climb the step.

After the wheelchair's rear wheels have reached the upper level of the step, the robot control mode is switched from autonomous to teleoperation control by the caregiver. The caregiver directs the vehicles to stop.

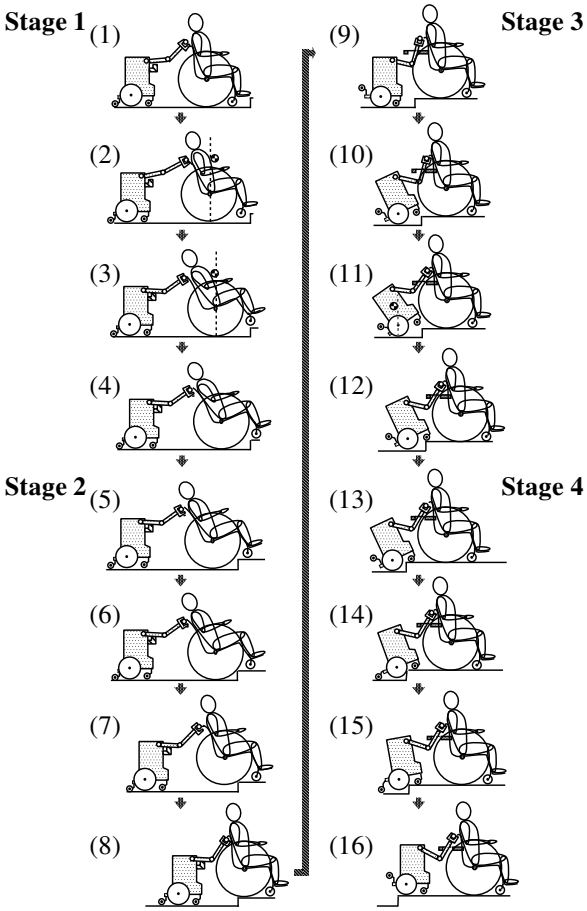


Figure 12: Step climbing process (Stages 1 through 4)

[Stage 3]

(9) The robot's front and rear wheels are folded, its

sides are opened, and the manipulator forearm links are inserted into the wheelchair stopper (Figs. 8 (a) and (b)). The wheelchair stops while the robot moves forward, and the manipulator forearm links comes into contact with the front bars of the wheelchair stopper (Figure 9 (a)). (10) The robot continues to push the wheelchair using the forearm links, and the front wheels of the robot are lifted. (11) When the robot center of mass shifts behind the contact point between its middle and the ground, the robot's tilt increases and it begins to tip over backward. However, at that point, part of the manipulator forearm links come into contact with the rear bars of wheelchair stopper, which limits the extent of rotation (Figure 9 (b)). Thus, the wheelchair prevents the robot from tipping over backward. (12) Both vehicles move forward, and the front wheels of the robot are placed on the upper level of the step.

[Stage 4]

(13) Both vehicles continue to move forward. (14) When the robot middle wheels come into contact with the step, the wheelchair pulls it forward. The robot rear wheels are lowered and provide climbing help. Then, the robot middle wheels begin to climb the step. (15) Both vehicles continue to move forward. (16) The robot middle wheels can now climb the step. Once the robot middle wheels have reached the upper level of the step, both vehicles are stopped, and the robot's rear wheels are folded upward.

5 Theoretical Analysis

When the distance between the step and the wheelchair is too short to lift the front wheels, the wheels will collide with the step (Figure 11). In this section, we clarify the requirement for avoiding a collision between the chair front wheels and the step.

In Figure 13, Σ_B is the basic coordinate system for the robot and wheelchair, where p_0 is the origin as well as the contact position between the robot middle (driving) wheels and the ground. The position of the step edge is p_{se}, where p_i (i = 1–7) are the joints (p_1: axes of the robot middle wheels, p_2: shoulder joints, p_3: elbow joints, p_4: robot hands (location where the hands hold the wheelchair push handle), p_5: axes of the wheelchair rear wheels, p_6: axes of the wheelchair front wheels and, p_7: the tread of the wheelchair front wheels).

The position vectors for the joints in system Σ_B are expressed as ${}^B\boldsymbol{p}_i = [x_i \;\; z_i]^T$ ($i = 1-7$). The position vectors of the step edge are ${}^B\boldsymbol{p}_{se} = [L \;\; h]^T$.

In the local coordinate system, in the case that Σ_i parallels Σ_0, ${}^0\boldsymbol{p}_1 = [0 \;\; R_B]^T$, ${}^1\boldsymbol{p}_2 = [l_{LB} \;\; h_{LB}]^T$, ${}^2\boldsymbol{p}_3 = [l_2 \;\; 0]^T$, ${}^3\boldsymbol{p}_4 = [l_{4C} \;\; 0]^T$, ${}^4\boldsymbol{p}_5 = [l_{LA} \;\; -h_{LA}]^T$, ${}^5\boldsymbol{p}_6 = [l_A \;\; -R_A + r_A]^T$, and ${}^6\boldsymbol{p}_7 = [r_A \;\; 0]^T$ (Figure 13).

Moreover, ϕ_i is the angle of Σ_i formed by Σ_{i-1}. In this system, ϕ_1 is zero in *Stage 1* because Σ_1 parallels Σ_0, and ϕ_3 is zero when the wheelchair is pulled by the robot because the elbow joints are controlled passively. In addition, $\phi_5 = 0$ is always zero because Σ_5 parallels Σ_4. Hence, the wheelchair inclination is $\Sigma_{k=1}^{5}\phi_i = \phi_2 + \phi_4$.

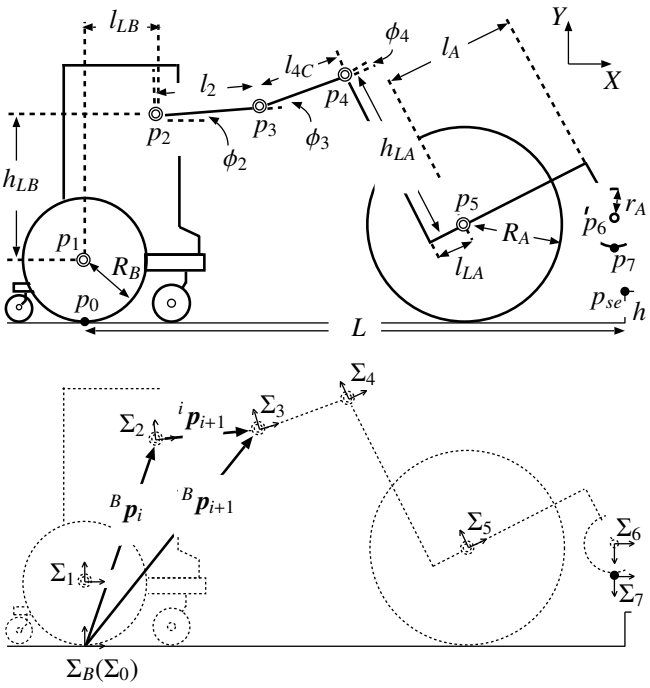


Figure 13: Wheelchair and robot model

In the basic coordinate system Σ_B, the homogeneous transformation matrices ${}^B\boldsymbol{T}_5$, ${}^B\boldsymbol{T}_6$ are given below.

$$ {}^B\boldsymbol{T}_5 : \begin{bmatrix} \cos\phi_{24} & -\sin\phi_{24} & x_5 \\ \sin\phi_{24} & \cos\phi_{24} & y_5 \\ 0 & 0 & 1 \end{bmatrix} \tag{1} $$

$$ {}^B\boldsymbol{T}_6 : \begin{bmatrix} \cos\phi_{246} & -\sin\phi_{246} & x_6 \\ \sin\phi_{246} & \cos\phi_{246} & y_6 \\ 0 & 0 & 1 \end{bmatrix} \tag{2} $$

Here, $\phi_{24} = \phi_2 + \phi_4$, $\phi_{246} = \phi_2 + \phi_4 + \phi_6$. Equations (3) and (4) show the positions of the wheelchair rear and front wheels axes, ${}^B\boldsymbol{p}_5 = [x_5 \;\; y_5]^T$, ${}^B\boldsymbol{p}_6 = [x_6 \;\; y_6]^T$, respectively.

$$ {}^B\boldsymbol{p}_5 : \begin{bmatrix} x_5 \\ y_5 \end{bmatrix} = \begin{bmatrix} \cos\phi_{24} & -\sin\phi_{24} \\ \sin\phi_{24} & \cos\phi_{24} \end{bmatrix} \begin{bmatrix} l_A \\ -h_{LA} \end{bmatrix} + (l_2 + l_{4C}) \begin{bmatrix} \cos\phi_2 \\ \sin\phi_2 \end{bmatrix} + \begin{bmatrix} l_{LB} \\ h_{LB} + R_B \end{bmatrix} \tag{3} $$

$$ {}^B\boldsymbol{p}_6 : \begin{bmatrix} x_6 \\ y_6 \end{bmatrix} = \begin{bmatrix} \cos\phi_{24} & -\sin\phi_{24} \\ \sin\phi_{24} & \cos\phi_{24} \end{bmatrix} \begin{bmatrix} l_A + l_{LA} \\ r_A - R_A - h_{LA} \end{bmatrix} + (l_2 + l_{4C}) \begin{bmatrix} \cos\phi_2 \\ \sin\phi_2 \end{bmatrix} + \begin{bmatrix} l_{LB} \\ h_{LB} + R_B \end{bmatrix} \tag{4} $$

Similarly, from a calculation of the transformation matrix ${}^B\boldsymbol{T}_7$, ${}^B\boldsymbol{p}_7 = [x_7 \;\; y_7]^T$ is given below.

$$ {}^B\boldsymbol{p}_7 : \begin{bmatrix} x_7 \\ y_7 \end{bmatrix} = r_A \begin{bmatrix} \cos\phi_{246} \\ \sin\phi_{246} \end{bmatrix} + \begin{bmatrix} \cos\phi_{24} & -\sin\phi_{24} \\ \sin\phi_{24} & \cos\phi_{24} \end{bmatrix} \begin{bmatrix} l_A + l_{LA} \\ r_A - R_A - h_{LA} \end{bmatrix} + (l_2 + l_{4C}) \begin{bmatrix} \cos\phi_2 \\ \sin\phi_2 \end{bmatrix} + \begin{bmatrix} l_{LB} \\ h_{LB} + R_B \end{bmatrix} \tag{5} $$

Here, when ${}^B\boldsymbol{p}_7$ is the bottom position of the wheelchair front wheels, $\Sigma_{k=1}^{6}\phi_i = -90$ deg. Hence, $\cos\phi_{246} = 0$ and $\sin\phi_{246} = -1$.

Meanwhile, $y_5 = R_A$ in Stage 1 because y_5 is the height of the wheelchair rear wheel axes (Figure 13). Hence, Equation (6) is obtained from (3).

$$l_{LA}\sin\phi_{24} - h_{LA}\cos\phi_{24} + (l_2 + l_{4C})\sin\phi_2 + h_{LB} + R_B = R_A \quad (6)$$

Thus,

$$\sin\phi_2 = \frac{h_{LA}\cos\phi_{24} - l_{LA}\sin\phi_{24} + R_A - R_B - h_{LB}}{l_2 + l_{4C}} \quad (7)$$

and

$$\cos\phi_2 = \sqrt{1 - \sin^2\phi_2} \quad (8)$$

Equation (6) is substituted for y_7 of (5) when ${}^B\boldsymbol{p}_7$ is the bottom of the front wheels.

$$y_7 = -r_A + l_A\sin\phi_{24} + (-R_A + r_A)\cos\phi_{24} + R_A \quad (9)$$

where

$$\sin\phi_{24} = \sqrt{1 - \cos^2\phi_{24}} \quad (10)$$

From Equations (9) and (10), we obtain Equation (11).

$$\cos\phi_{24} = \frac{(R_A - r_A)(R_A - r_A - y_7) + l_A\sqrt{2(R_A - r_A)y_7 + {l_A}^2 - {y_7}^2}}{(R_A - r_A)^2 + {l_A}^2} \quad (11)$$

We substitute $y_7 = h$ for Equation (11) if the system detects the step height h, and can then ascertain the necessary minimum inclination of the wheelchair in order to put the front wheels on the step.

When the height of the front wheel axes is lower than the step height, the requirement needed to avoid a collision between the wheelchair front wheels and the step is Equation (12) (Figure 14 (a)).

$$\|{}^B\boldsymbol{p}_s - {}^B\boldsymbol{p}_6\| > r_A \quad (0 < x_6 < L,\ 0 < y_6 < h) \quad (12)$$

Here,

$${}^B\boldsymbol{p}_s = [L \ \ y_6]^T \quad (13)$$

is a vector whose horizontal position is on the step wall (L), and the height is the same as the wheelchair front wheel axes (y_6).

Similarly, when the front wheel axes are above the step, Equation (14) shows the requirement to avoid a collision (Figure 14 (b)).

$$\|{}^B\boldsymbol{p}_{se} - {}^B\boldsymbol{p}_6\| > r_A \quad (0 < x_6 < L,\ h \le y_6) \quad (14)$$

Then, using Equations (12) and (14), the system avoids a collision between the wheelchair front wheels and the step by detecting the chair inclination ($\phi_2 + \phi_4$), the distance from the step (L), and the step height (h).

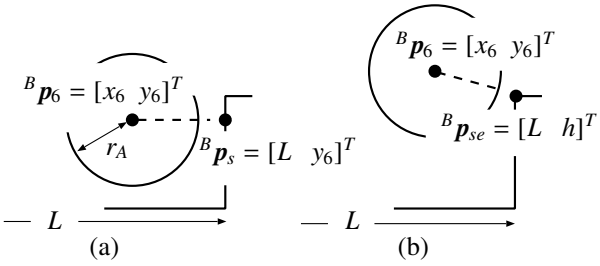


Figure 14: Collision between the front wheels and a step. (a) $0 \le y_6 \le h$, (b) $h < y_6$

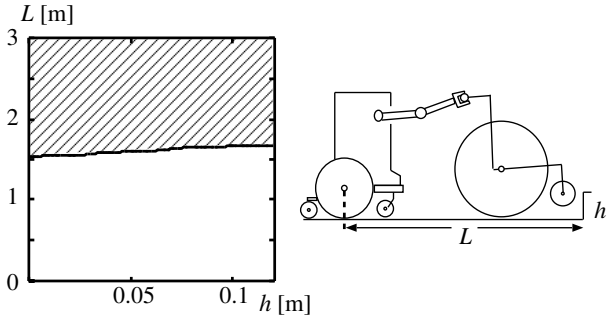


Figure 15: Step height and distance capable of avoiding collision

The requirement for avoiding a collision between the wheelchair front wheels and the step (Equation (12), Figure 14 (a)) is more difficult than the case where the front wheel axes are above the step (Equation (14), Figure 14 (b)). Thus, we performed a numerical calculation to clarify the lifting position of the front wheels using Equation (12). Equation (12) is solved using (4), (7), (8), (10), (11), and (13). Here, y_7 is the minimum bottom height of the front wheels (Figure 13) needed for the system to lift the wheels up to the step height (h).

Substituting $y_7 = h$ for Equation (12), the shading in Figure 15 shows a combination of the step height (h) and the distance (L) that enables it to avoid the collision, where the horizontal axis indicates h (step height) and the vertical axis indicates L (distance from the robot's rear wheel axes to the step wall). By selecting the distance (L) in the shadings when the step height is (h), the wheelchair front wheels can avoid collision during climbing.

6 Experiment

An experiment was carried out in a situation involving a step 120 mm high and a friction coefficient $\mu = 0.72$. The wheelchair user was on one floor of the National Institute of Technology, Toyama College, and the caregiver (robot operator) was on another floor. The movement speeds of the vehicles were constant (0.76 km/h). The wheelchair user and the robot operator were both able-bodied adult males. We assumed that the system knew the step height was below 120 mm.

In Stages 1 and 2, when using autonomous control, the wheelchair front wheels did not bump into the vertical wall of the step and the front wheels and rear wheels were able to traverse the step successfully (Figure 16). Here, $L = 2.07$ m. Operation changed from autonomous to teleoperation control following Stage 2, and the caregiver was able to move the robot forward into a suitable position for climbing the step. Next, the robot successfully climbed the step (Stages 3 and 4). In Stages 3 and 4, the system was operated by the caregiver who was watching video imagery from the robot's camera. However, the operation was difficult because the robot inclined during its step climbing process and the mounted camera lost sight of the wheelchair. This and other experiments have shown that we need to construct

a fully autonomous control system that can handle robot climbing as well as wheelchair climbing.

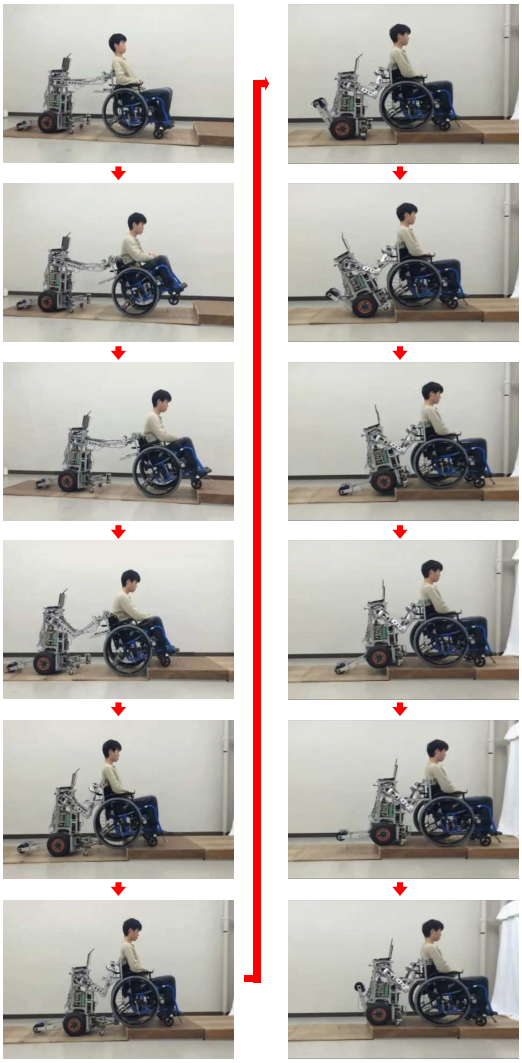


Figure 16: Step climbing experiment

7 Conclusion

This report describes a support system for a robot operator that can be used to perform wheelchair step climbing. The support system was designed and constructed in a way that permitted switching between teleoperation and autonomous control. We then conducted a theoretical analysis to show the requirements for preventing the wheelchair's front wheels from colliding with the step, which involved a numerical calculation to clarify the position for lifting the wheelchair's front wheels.

In experiments carried out with the system, autonomous control for wheelchair step climbing and teleoperation control for robot step climbing were both demonstrated to be effective. However, teleoperation control by a caregiver was complicated by the need to watch video imagery from the robot camera. This indicates that the system needs autonomous control system for the robot as well.

Nevertheless, it is notable that our theory for collision avoidance between the step wall and wheelchair's front wheels during the step climbing was confirmed, and that both vehicles were able to climb a step using the theory and system.

In the future, we will construct a fully autonomous control system will be able to perform step climbing for both the wheelchair and the robot, and we will analyze a suitable step climbing trajectory for both vehicles.

Acknowledgment This research has been supported by the Kayamori Foundation of Informational Science Advancement (K25 *XVII* 430).

References

[1] H. Ikeda, K. Hashimoto, D. Murayama, R. Yamazaki and E. Nakano, "Robot teleoperation support system for collision avoidance between heelchair front wheels and a step", Proceedings of The 2016 IEEE International conference on Simulation, Modeling, and Programming for Autonomous Robots, 203-209, 2016

[2] V. Kumar and V. Krovi, "Optimal Traction Control In A Wheelchair With legs And Wheels", Proceedings of the 4th National Applied Mechanisms and Robotics Conference, 95-030-01 - 95-030-07, 1995

[3] A.Gonzalez, E.Ottaviano and M.Ceccarelli "On the Kinematic Functionality of s Four-bar Based Mechanism for Guiding Wheels in Climbing Steps and Obstacles," Mechanism and Machine Theory, **44**, 1507-1523, 2009.

[4] Independence Technology, L.L.C., iBOT, http://www.ibotnow.com/, 2008.

[5] K.Sugiyama, T.Ishimatsu, T.Shigechi and M.Kurihara "Development of Staeir-climbing Machines at Nagasaki", Proceedings of the 3rd International workshop of Advanced Mechatoronics, 214-217, 1999.

[6] Y.Munakata and M.Wada, "Modeling and analysis of static wheelie of a five-wheeled wheelchair for climbing over a step", Proceedings of IEEE/ASME International Conference on Advanced intelligent Mechatoronics, 1006-1011, 2014.

[7] M.Lawn and T.Ishimatsu, "Modeling of a Stair-Climbing Wheelchair Mechanism with High Single Step Capability. IEEE Transactions on Neural Systems and Rehabilitation Engineering", **11**(3), 323-332, 2003.

[8] Y. Mori, K. Katsumura, and K. Nagase. "A pair of step-climbing units for a manual wheelchair user: Passing over several steps using a pair of portable slopes", Advances in Mechanical Engineering, **9**(3), 1-11, 2017

[9] H.Ikeda, Y.Katsumata, M.Shoji, T.Takahashi and E.Nakano "Cooperative strategy for a wheelchair and a robot to climb and descend a step," ADVANCED ROBOTICS, **22**, 1439-1460, 2008

[10] H.Ikeda, H.Kanda, N.Yamashima and E.Nakano, "Step climbing and Descending for a Manual wheelchair with a Network care robot,"Proceeding of The Second International Conference on Intelligent System and Applications, 95-102, 2013.

[11] H. Ikeda, N. Hatakeyama, A. Kinoshita and E.Nakano, "Step Moving for an Electric Wheelchair Using a Robot Programmable over the Intranet", Proceedings of 2nd IEEE/SAE/IFAC International Conference on Connected Vehicles and Expo, 791-797, 2013.

[12] H.Asama, M.Sato, N.Goto, H.Kaetsu, A.Matsumoto and I.Endo, "Mutual Transportation of Cooperative Mobile Robots Using Forklift Mechanisms," Proceedings of the 1996 IEEE International Conference on Robotics and Automation, 1754-1759, 1996.

[13] J. Gaston, K.Raahemifar, P.Hiscocks, "A cooperative Network of Reconfigurable Stair-Climbing Robots", Proceeding of 2006 IEEE International Sysposium on circuits and Systems, 4-7, 2006

[14] H.Ikeda, N.Yamashima, S.Haruta and E.Nakano, "Docking method for a wheelchair and a care robot using both teleoperation and autonomous control", International Journal of Machanical Engineering and Robotics Research, **5**(2), 82-89, 2016

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 741-748 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Adapting Model Predictive Control for a PV Station and Evaluating two different MPPT Algorithms P&O and FLC

Imad Elzein[*1], Yury Petrenko[2]

[1]*Dep. of Electrical Engineering and Control systems, Belarusian National Technical University, Minsk, Belarus, imad.zein@liu.edu.lb*

[2]*Dep. of Electrical Engineering and Control systems, Belarusian National Technical University, Minsk, Belarus, ypetrenko@bntu.by*

ARTICLE INFO

Article history:
Received: 20 March, 2017
Accepted: 03 June, 2017
Online: 16 June, 2017

Keywords:
Fuzzy logic controller
MPPT
Photovoltaic

ABSTRACT

In this paper first discussion approach will stress out the integration of model predictive control in maximum power point tracking MPPT and as progressing a second approach identified as fuzzy logic controller FLC and perturb & Observe P&O algorithms are analyzed. All are interrelated to MPPT model for a photovoltaic module, PVM, to search for and generate the maximum power; in this case what's called P_{max}. As per the first technique the focus is on the optimal duty ratio, D, for a series of multi diverse types of converters and load matching. The design of the MPPT for a stand-alone photovoltaic power generation system is applied where the system will consist of a solar array with nonlinear time varying characteristics, and a converter with appropriate filters. The integration of model predictive control will be addressed first in this paper. The second fold will implement an MPPT system that use the FLC and compare it with a P&O) algorithm through the utilization of Simulink.

1. Introduction

It has been obvious that maximum power point tracking algorithm is currently playing a vital role to enhance the efficiency of the solar panel as less than 40% of energy incident is being converted into electrical energy. Due to the growing dependency and the increasing need in acquiring electricity, and due to the limitations in supply and the uprising prices of conventional sources (such as the continuous increase in electrical bills, generation, distribution, and maintenance of the electrical plants, fluctuating petroleum prices, etc.), photovoltaic (PV) energy vitality turns into a promising option as it is inescapable, openly accessible, environmentally promising, and has less operational and upkeep costs. Along these lines, the interest of PV era systems is by all accounts expanded for both standalone and grid-connected modes of PV systems. As a result, an efficient maximum power point tracking (MPPT) technique is vital for tracking the MPP at all environmental conditions and then push forward the PV system to functionally be operable at that MPP point. Undesirably, photovoltaic generation systems have two note-worthy issues: the conversion efficiency in electric power generation is somehow low (normally below 17 percent particularly under low irradiation conditions), and the amount of electric power generated by solar arrays changes persistently with climate conditions. Numerous MPPT methodologies have been recommended in the literature; the Perturb and Observe (P&O), the Incremental Conductance (IC), the Artificial Neural Network, and the Fuzzy Logic methods, etc. It has been noticed that the efficiency of the PV is influenced by the following two parameters: the panel's irradiance and temperature which are stochastic and unpredictable. In any PV module a DC/DC converter is accountable for transferring maximum power to the load.

Unfortunately and since the MPP point must be sought, this can be achieved in either computation models or search algorithms. Figure 1 illustrates an MPPT module diagram.

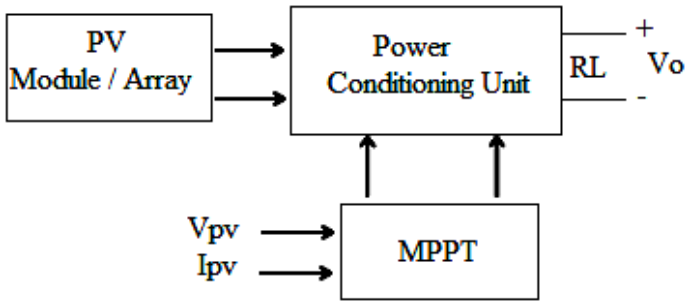


Figure1.MPPT module diagram

[*]Corresponding Author: Imad Elzein, Belarusian National Technical University, Minsk, Belarus | Email: imad.zein@liu.edu.lb

We have to stress out that the voltage across the power conditioning unit (DC-DC converter) is fed to get an isolated load. The input-output (I/P-O/P) voltage relationship for converter conduction mode is given by duty cycle. As a result, the forthcoming of this paper is an extension of a previous research outcome that was originally presented in the International Conference on Smart Systems and Technologies (SST 2016) to recompense for the parametric variations that took place due to the effect of temperature and irradiance variations, where a controller was utilized to be adaptive to environmental changes. The extension of the already published paper will be expanded towards proposing an innovative fuzzy logic controller (FLC) for DC–DC converters that yield to an effective element of MPPT system so that it integrate itself into enhancing the photovoltaic modules to work under changeable operating conditions as well as dealing with the nonlinear properties of DC-DC power converters [1].

2. Overview of model predictive control

MPC is a model in view of line control approach with the following accompanying modules: a prediction horizon, a receding horizon procedure, and a regular update of the model and re-computation of the optimal control input [2,3,4,5]. A block diagram of MPC system is shown in Figure. 2. A process model is used to predict the current values of the output variables. The residuals, the differences between the actual and predicted outputs, serve as the feedback signal to a Prediction block. The predictions are utilized as a part of two types of MPC calculations that are performed at each sampling instant: the first is the set-point calculations and the second is the control calculations.

The set points for the control calculations, which are called targets, calculated from an economic optimization based on a steady-state model of the process, conventionally, a linear model. In MPC the set points are customarily computed each time the control calculations are conducted.

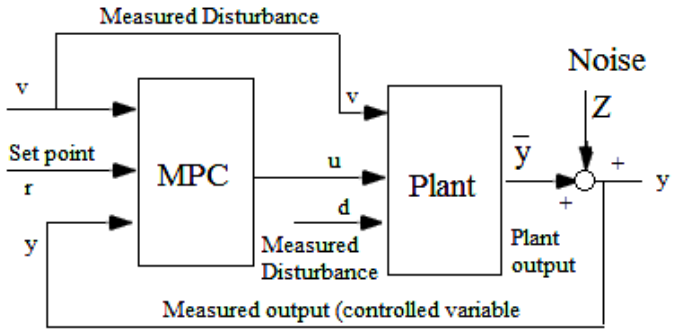


Figure 2.Block diagram for MPC

3. Nonlinear predictive control

The basic principle of model predictive control is shown in Figure 3. At a denoted time (t) certain measurements are provided, which will trigger the controller to predicts the future dynamic behavior of the system over a prediction horizon *Tp* and furthermore determine (over a control horizon *Tc*≤ *Tp*) the input under the condition a predefined open-loop performance objective is optimized.

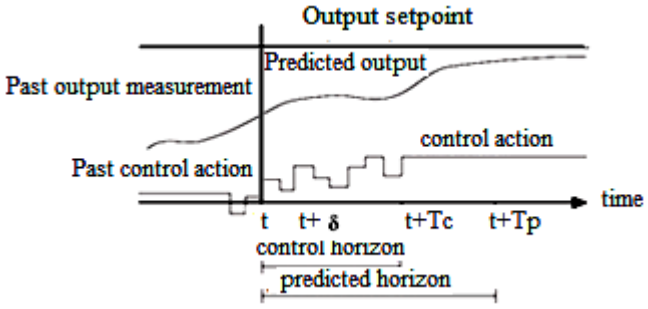


Figure 3.Principle of model predictive control

When neither disturbances exist nor mismatching model plant presence is evident, and if the optimization problem could be solved for infinite horizons, then we can apply the input function found at time t=0 to the system for all times $t \geq 0$.As a matter of fact, this is not feasible generally. Because of the factors of disturbances and the stressing model-plant mismatch, the real system behavior is going to differ from the predicted behavior. To integrate a feedback scheme, the open-loop manipulated input function acquired will be ready for implementation as the next measurement becomes available. The time difference between the recalculation/measurements can vary, however often it is assumed to be fixed, i.e. the measurement will take place every δ sampling time-unit. Using the new measurement at time (t + δ), then the whole process for the prediction and optimization is going to be repeated to find a new input function with the control and prediction horizons moving forward. Knowing that in Figure 3 the input is depicted as arbitrary function of time. For numerical solutions of the open-loop optimal control problem it is often necessary to parameterize the input in an appropriate way. This is normally achieved by employing a finite number of basic functions; as an example the input could be approximated as piecewise constant over the sampling time δ. The computation of the applied input (based on the predicted system behavior) permits the inclusion of constraints on states and inputs as well as the optimization of a given cost function. In general the predicted system behavior will differ from the closed-loop one; and thus further cautionary should be taken into account to achieve closed-loop stability.

We consider the stabilization problem for a class of systems described by the following nonlinear set of differential equations:

$$\dot{X}(t) = f\big(x(t), u(t)\big), \quad X(0) = x_0, \qquad (1)$$

subject to input and state constraints of the form:

$$U(t) \in \mu \ \forall t \geq 0 \ \ x(t) \in X, \mu \ \forall t \geq 0 \qquad (2)$$

Where x (t) $\in X \subseteq \mathbb{R}^n$ *and* $u(t) \in \mu \subseteq \mathbb{R}^m$ denotes the vector of states and inputs, respectively. The set of feasible input

values is denoted by X. In the simplest form $\mu$and X are given by box constraints of the form:

$$\mu: = \{u \in \mathbb{R}^m | Umin \leq u \leq Umax\} \quad (3)$$

$$X: = \{x \in \mathbb{R}^n | Xmin \leq x \leq Xmax\},$$

Here u_{min}, u_{max} and x_{min}, x_{max} are given constant vectors.

We assume that the purpose of the control object (1) is to ensure that the equalities

$$\lim_{t\to\infty} \|x(t) - r_x(t)\| = 0;$$

$$\lim_{t\to\infty} \|u(t) - r_u(t)\| = 0 \quad (4)$$

Where a given vector functions; $r_x(t)$ and $r_u(t)$ define a desired motion of the object.

We introduce the concept of quality control, setting a functional to control the movement of the object (1).

$$J_{0=} J_0\,(x(t), u(t)) \quad (5)$$

Any optimal control problem is to find such a control action of a given class (if the task takes into account the feasible set U), which ensures the achievement of goals (4) taking into account the constraints $x(t) \in X$ and $\forall t \in [0, \infty)$ and minimizes the functional (5).

To explain the basic tenets of the theory of predictive control, along with a mathematical model (1) of the control object, a system of differential equations of the form:

$$\dot{\bar{X}}_i(\tau) = \bar{f}\big(\tau, \bar{x}(\tau), \bar{u}(\tau)\big)\bar{x}\big|_{\tau \dot{=} t} = x(t) \quad (6)$$

Is considered, where$\bar{X} \in E^n$ - the state vector, $\bar{u} \in E^m$- the vector control, $\tau \in [t, \infty)$. We assume that the function $\bar{f}$has the same properties as the function f and vectors,$\bar{X}$and$\bar{u}$taking the value of the admissible sets,$X$and $u$respectively.

In addition, let us assume that the function $\bar{f}$ is set in such a way that for any admissible control$\bar{u}(\tau) \equiv u(\tau)$ vector functions $x(\tau)$ and $\bar{x}(\tau)$, satisfy the system (1) and (6) respectively; and are close to each other at a rate for each $\tau \in [t, \infty)$.

The system of differential equations (6) is called predictive model in relation to the mathematical model (1) of the control object.

A practical view of photovoltaic energy conversion station is presented in Figure 4, where the source of energy, represented by the solar array PHVS with filter C delivers power to DC-DC power converter, such as buck, boost or buck-boost converters; which depends on specific architecture of the entire system to select one of the above-mentioned types. For instance, if PHVS is considered to supply power to a DC distribution network (i.e. a space station power network) than power interface will provide conditioned DC power. In case the main utilization was for smart-home application with AC power distribution network then independently controlled DC-AC inverter with built in PWM control may be used. The interface links between PHVS and the consumer grid may include other elements such as energy-storage devices (i.e. batteries or super capacitors), which require specific controllers to run "charge-discharge" mode of operation. We have to bear in mind that control designs for interfaces devices are ignored.

The objective of DC-DC converter on PHV station MPPT system is to regulate the output voltage V_0 (Figure 4) under wide range of operations and PHV nonlinearity characteristics, caused by existing atmospheric conditions and load variation. The controller task is to deliver output signal D (Figure 4) to follow MPPT [6] and constraints of the circuit.

DC-DC boost converter schematic is in a view of lumped parameter circuits used for control design and represented in Figure 5.

The coil non-linearity's is neglected, and the switch is also considered ideal. At each switching instant, the stray inductor and parasitic capacitor are also neglected. After this assumptions the lumped parameter switched model X_l represent the linear inductance value associated to the coil L, which losses are accounted for r_l, x_c , $and r_c$ represent capacitance and equivalent series resistor of C.

Output load presented by resistor R_0. The switching stages of the converter are formalized through the switch S, representing the dual operated semiconductor component [7]. The converter operation characterizes by switching period T_s (corresponding frequencyf_s).

The DC component of the output voltage regulated through the duty cycle D[k], (Figure. 4), which is defined by:

$$\mathrm{D[k]} = \frac{t_1[k]}{T_s} \quad (7)$$

Where $t_1[k\,]$ represents the time interval during the kth switching period;T_s during which S is in the position 1 for the boost converter, connecting to the supply U_s.

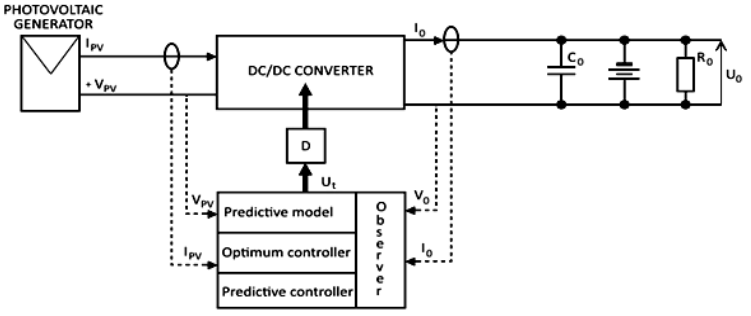


Figure 4.Photovoltaic station model predictive control block diagram

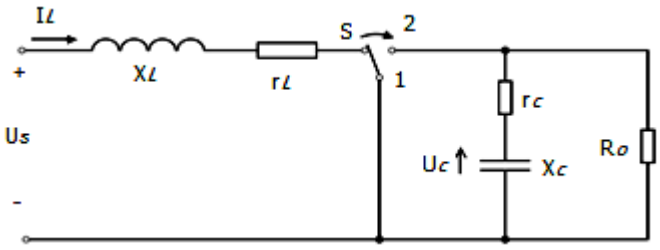


Figure 5 Lumped parameter model of the boost converter

$$x(t) = [i_l(t) \quad u_c(t)]^T \tag{8}$$

as the state vector, where $i_l(t)$ is the inductor current and $u_c(t)$ is the capacitor voltage, and with a given duty cycle D[k] for the kth period, the systems can be described by the following continuous-time state-space equations:

$$\dot{x}(t)=G_1 x(t)+g_1 u_s(t), \quad kT_s \leq t < (k + D[k]) \text{ when } S=1 \tag{9}$$

$$\dot{x}(t) =G_2 x(t)+g_2 u_s(t) (k + D[k]) \; T_s \leq t < (k+1)T_s \text{ when } S=0$$

Where G_1, G_2, g_1, g_2 are given by:

$$G_1 = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \text{ For the boost converter.} \tag{10}$$

Output of the converter u_o is equal to

$$u_o(t) =B_1^T x(t), \quad kT_s \leq t < (k + D[k])T_s,$$

$$u_o(t) =B_2^T x(t), \quad (k + D[k])T_s \leq t < (k +1)T_s \tag{11}$$

$$B_1=\begin{bmatrix} 0 & \frac{r_0 \; r_2}{r_0+r_c} \end{bmatrix}^T \tag{12}$$

$$B_2 =\begin{bmatrix} \frac{r_0 \; r_2}{r_0+r_2} & \frac{r_0}{r_0+r_c} \end{bmatrix}^T \tag{13}$$

4. Fuzzy logic controller approach for photovoltaic Station

A standout amongst the most noteworthy problems in PV module and MPPT effectiveness is DC-DC converter. As of late, there has been expanding eagerness for the advancement of proficient control procedures to enhance dynamic conduct of DC–DC converters by utilizing conventional PID based controllers and fuzzy logic controller (FLC), and other well dealt and known controllers which have been utilized to control buck, boost, and buck–boost converters.

Carrying out, many leaders in this domain have additionally proposed to utilize the FLC in circumstances where it could be valuable in (1) there is no exact mathematical model for the plant and (2) there are experienced personnel whom they can control a plant and give subjective control rules in regards to vague and fuzzy instances. The tracking of the maximum power point is of twofold: (a) through rough search, with a major move to improve the response of the MPPT controller, (b) the fine phase where the step is small, thus ensuring the system stability and decrease the maximum oscillations around the MPP [8].

As we move towards the components of fuzzy logic controller we need to address the three major components of FLC which are Fuzzification (input variables) that encompass transforming crisp values into grades of membership for linguistic terms of fuzzy sets; Fuzzy Inference which is involved in the transformation from a given input to an output using fuzzy logic inference rules; and Defuzzification which converts the degrees of membership of output linguistic variables into numerical values [9]. A general view of fuzzy inference system is shown below in figure 6.

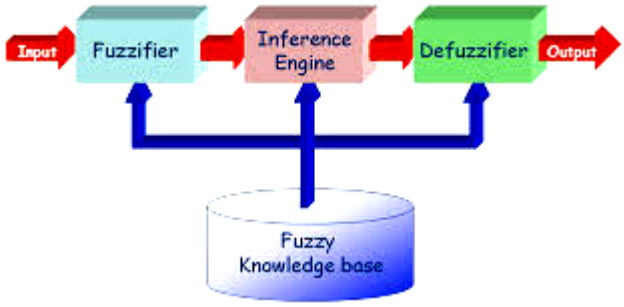


Figure 7 demonstrates the block diagram of fuzzy logic controller that is intended to control the DC-DC power converter with respect to MPPT of a photovoltaic system. Referring to Figure. 7 the inputs (error (e) and the difference in error (de) of the system are defined as follows [10] :

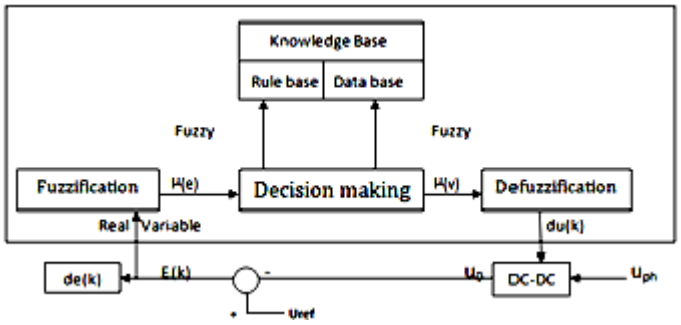


Figure 7.Fuzzy logic controller diagram

$$de(k) = e(k) - e(k-1) \tag{15}$$

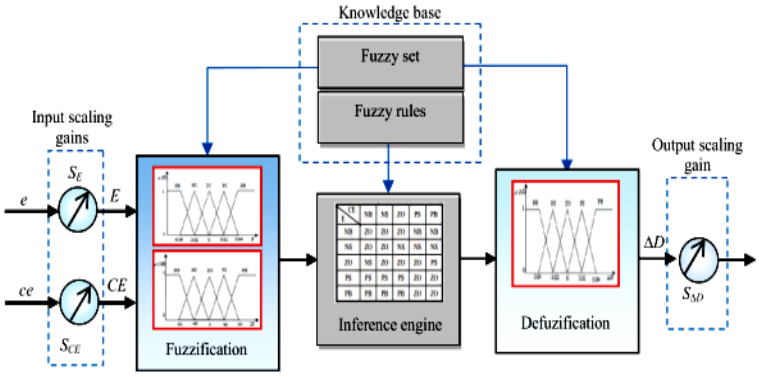
Knowing that U_{ref} = reference output voltage, U_0= *DC*–DC converter actual output voltage (at k[th] sampling time).

The fuzzy logic controller (FLC) output is considered as a change in duty ratio; *du (k).*

Where, duty ratio *d (k)*, is expressed at the kth sampling time as follows:

$$d(k) = d(k-1) + du(k) \tag{16}$$

A more detailed structuring of a fuzzy logic controller is shown below in figure 8.



5. Simulations and results for FLC and P&O MPPT techniques

A new adaptation and simulation of a PV system using fuzzy logic controller with an MPPT is addressed. As shown in figure.8, the input variables (*E*, *CE*) of the fuzzy controller are resulting from the actual signals (*e*, *ce*) through applying a product term in consequent scale gains (S_E, S_{CE}), which get converted in the form of linguistic variables utilizing fuzzy subset.

A sample of fuzzy rules are shown in table 1 and used to control the buck converter in a form to reach the maximum power point MPP of the photovoltaic PV generator, where the entries are a sort of fuzzy sets in the form of error (E), change of error (*CE*) and change of duty ratio (ΔD) to the converter.

Table 1. Fuzzy rules

E	CE				
	NB	**NS**	**ZO**	**PS**	**PB**
NB	ZO	ZO	NB	NB	NB
NS	ZO	ZO	NS	NS	NS
ZO	NS	ZO	ZO	ZO	PS
PS	PS	PS	PS	ZO	ZO
PB	PB	PB	PB	ZO	ZO

Figure 8 shows an approach for a main PV model structure composed of PV array, a DC-DC converter in the form of buck or step down converter along with MPPT controller which is connected to a load. This setup is manipulated inside the Matlab/simulink setting. Sunntech STb134.12/Tp represented the PV and its technical data is showing in table 2. Six PV modules connected in series formed the PV array which its total capacity is 810 W. The 105v DC input voltage was stepped down to 48v to fit the battery voltage level using a buck converter [11-14].

In addition, and to implement the fuzzy logic controller FLC, FL toolbox in Matlab/Simulink will be selected and the simulation of fuzzy logic controller for MPPT to be conducted and as we progress an evaluation comparison with a Perturb & Observe (P&O) for MPPT controller was performed.

Table 2.PV module specs

Electrical specifications	**Values**
Open-circuit voltage	22.3 V
Short-circuit voltage	8.20A
Optimum operating voltage (V_{mpp})	17.5 V
Optimum operating current (I_{mpp})	7.71A
Maximum power at STC (P_{max})	135W
Current temperature coefficient of Isc	(0.055 ±0.01) %K
Voltage temperature coefficient	-(75±10) mV/K

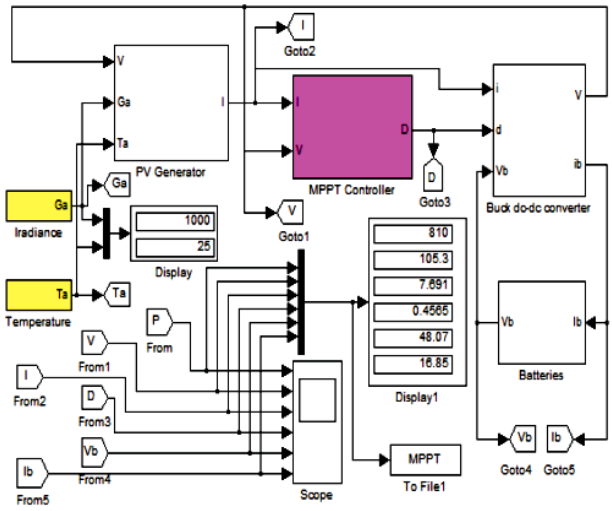


Figure 8a. Simulink implementation of the stand-alone PV system

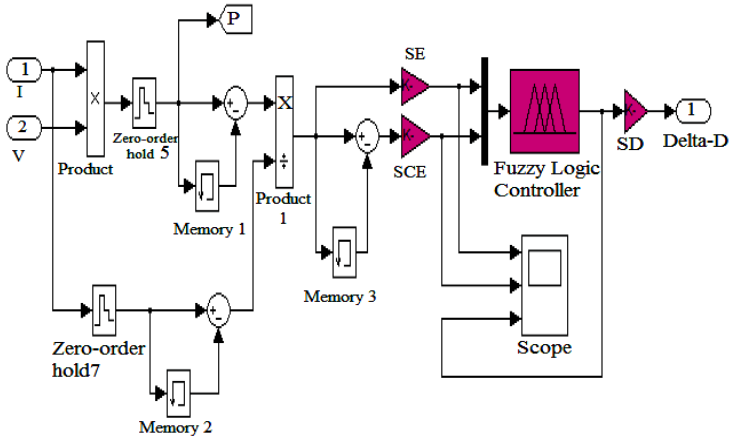


Figure 8b. FLC MPPT algorithm

Figure 9 demonstrates the outcome of the simulation result of PV generator O/P power, operating current & voltage, and eventually the duty ratio "D" through the use of a buck converter and these results were obtained at a standard test conditions (STC)

among the fuzzy logic controller "FLC" and the P&O based MPPT. To take the discussions a step forward obviously the outcome of the FLC MPPT had reduced clearly the response time of the PV system. On the other hand as we compare the above to P&O MPPT system P&O showed an impact of energy losses. Figure.9b is derived from comparing the MPPT signals at a frequency of 10HZ "STC". As we used the P&O technique there was an evidence of a continuous oscillation at the operation point and this was due to the incessant perturbation that took place at the operating voltage to reach the maximum power point "MPP". As compared to the FLC technique such oscillation wasn't existing in FL based MPPT technique, whereas the signals of the other parameters which were namely the power "P", voltage "V", current "I", and duty ratio "D" continued to stay constant, which would cause an impact on the reduction of power loss.

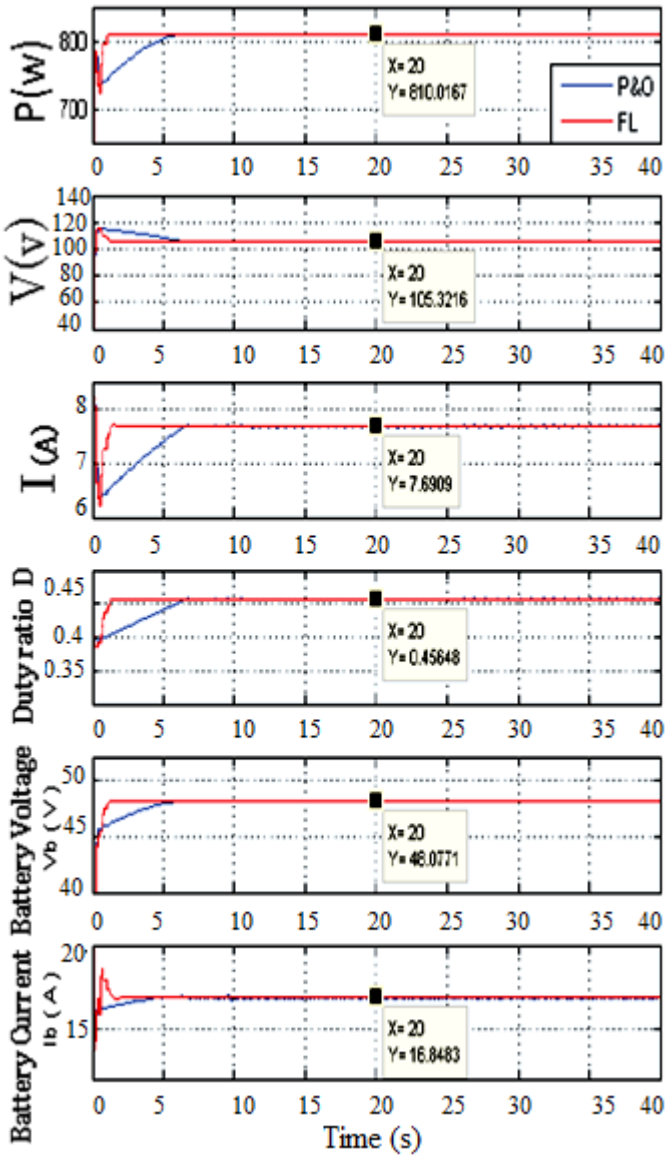


Figure 9a. Comparing P&O and FL signals

Figure 11 displays the simulation of the increase in cell temperature from "25°C" to "50°C" which was taking place at a "2 sec" time period.

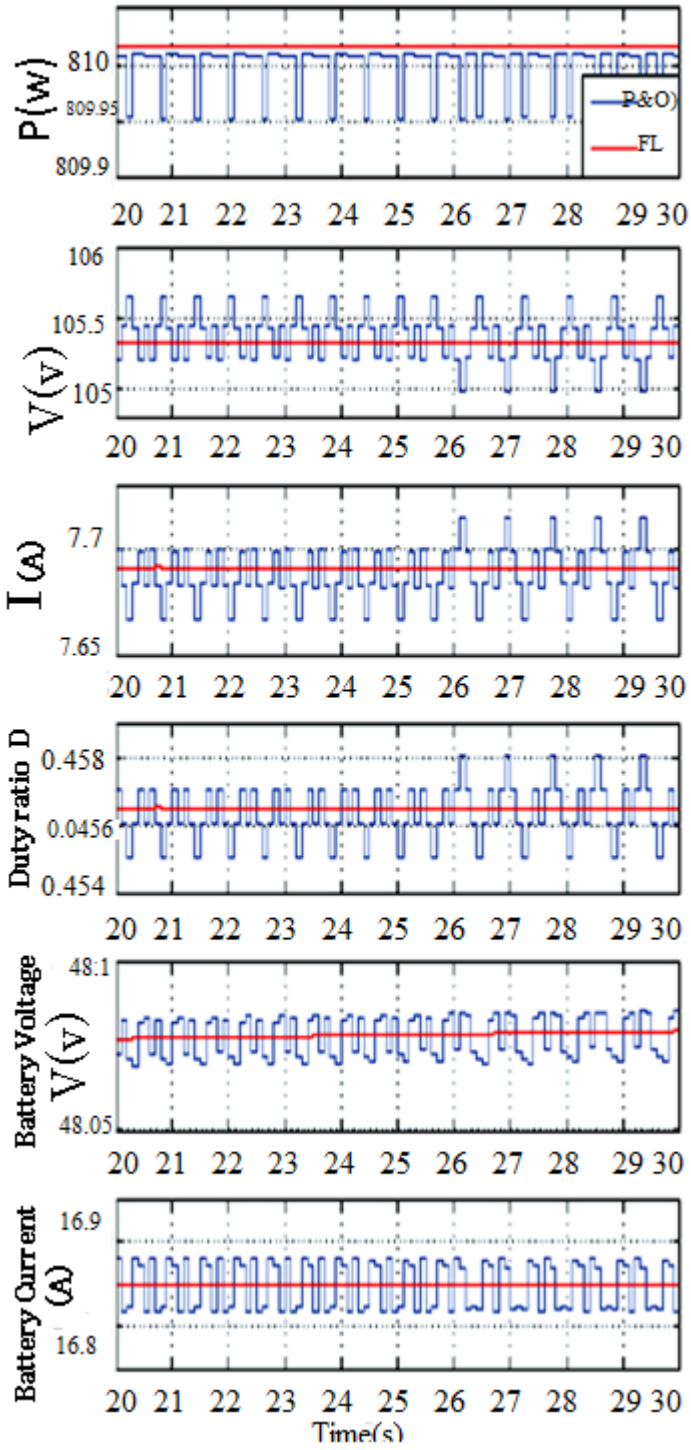


Figure 9b.Comparing MPPT methods at 10Hz frequency at the (STC) condition

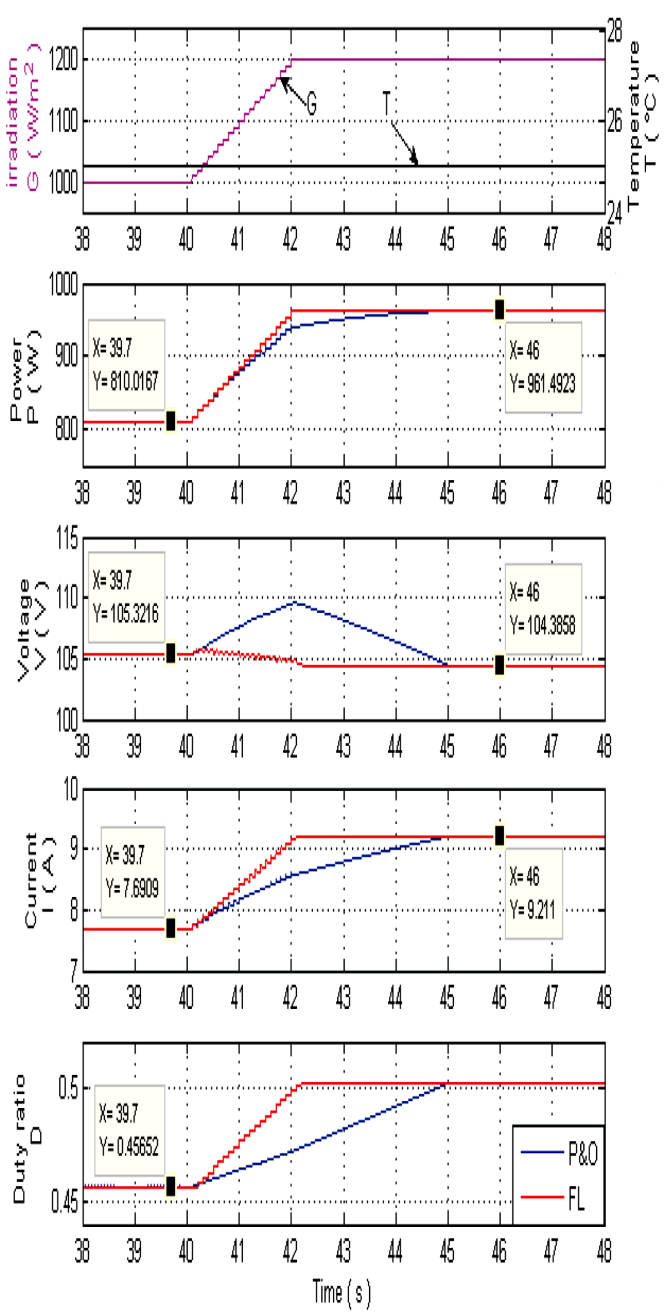


Figure 10. FL and P&O methods at fast increasing irradiation

From the simulation results we can clearly see that irradiation stayed constant on a 1000w/m2 value. In nutshell, Figure 11 shows that the output power with a decreased linearity as we applied both MPPT algorithms namely the fuzzy logic and Perturb and Observe. At the same time P&O MPPT method resulted in low deviation from the maximum power point MPP.

6. Conclusion

The employment of MPPT for PV stations and their applications were addressed in the form of two schemes; a first scheme was an adaptation of a model predictive control (MPC) for a PV station that is considered to be a promising technique to maximize the efficiency of the power utilization.

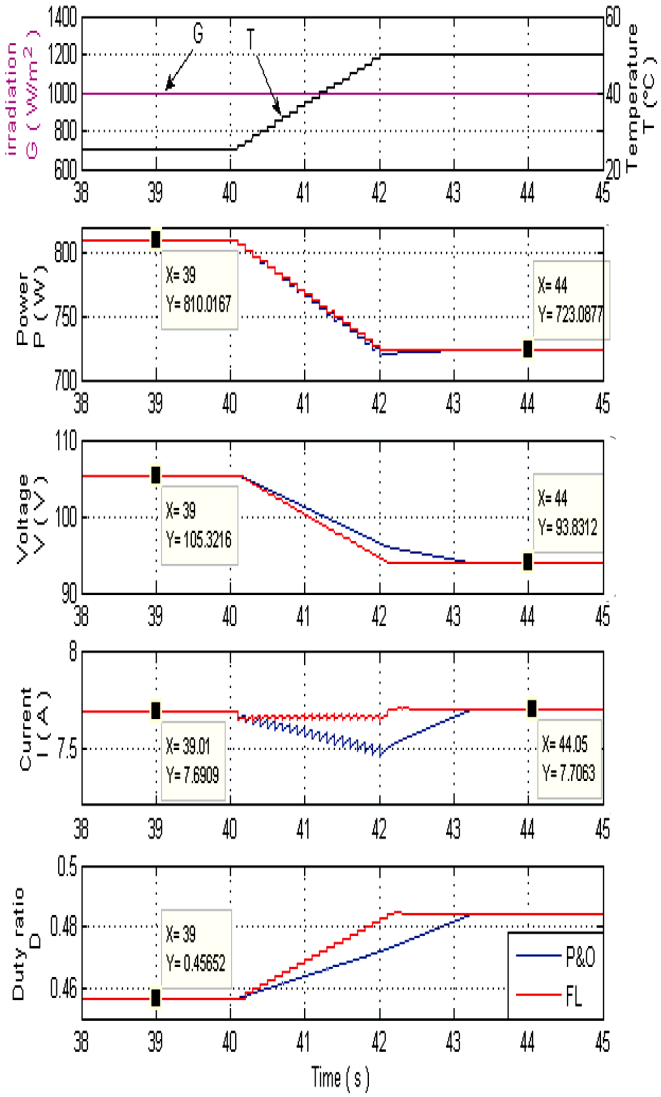


Figure 11. FL and P&O methods at fast increasing temperature

The second scheme of the paper was focusing on the implementation and simulation of fuzzy logic controller for MPPT to seek the maximum power point and comparing it to P&O. From the simulation results, FLC provided a reliable response as compared to the P&O controller in regards to main performance of MPPT.

References

[1] Elzein Imad, Petranko Yury. An Integration of a Predictive Control Model and MPPT for PV Station.IEEE Region 8, International Conference on Smart Systems and Technologies (SST 2016), proc. Of IEEE trans., Croatia, IEEE xplorer, region 8, pp. 275-280, ISBN: 978-1-5090-3718-6, IEEE catalog number: CFP16G03-PRT

[2] Sotnikova MV. Development toolkit for solving control problems with non-linear predictive models in MATLAB Tr. V Int. Conf. "Design of engineering and scientific applications in the environment MATLAB", 2011, p. 299-319.

[3] Veremey EI, MV Sotnikova, Plasma stabilization on the basis of the forecast of a steady linear approximation, St. Petersburg. Univ. Ser. 10. 2011, Vol. 1, p. 116-133.

[4] Sotnikova MV, Questions of stability of motion in control systems with predictive models, Bulletin of Voronezh State Technical University, 2012, p. 72-79.

[5] MargaritaV. Sotnikova, Evgeny I. Veremey. Dynamic Positioning Based on Nonlinear MPC, IFAC Proceedings Volumes (IFAC-Papers Online). Proceedings of the 9th IFAC Conf. on Control Appl. in Marine Systems (CAMS 2013). Osaka, Japan, September 17-20, 2013, Volume 9, Issue PART 1, 2013, Pages 37-42.

[6] Nevzat Onat. Recent Developments in Maximum Power Point Tracking Technologies for Photovoltaic Systems. International Journal of Photo energy, Vol. 2010, Article ID 245316, 11 pp., doi:10.1155/2010/245316.

[7] Elzein, I. Maximum Power Point Tracking System for Photovoltaic Station: a Review. System Analysis and Applied Information Science, No. 3, 2015, pp.15-20.

[8] D. P. Hohm and M. E. Ropp, "Comparative study of maximum power point tracking algorithms using an ex- perimental, programmable, maximum power point track- ing test bed," in Proc. Photovoltaic Specialist Conference, pp. 1699–1702, 2000.

[9] K. H. Hussein, I. Muta, T. Hoshino, and M. Osakada, "Maximum power point tracking: an algorithm for rapidly chancing atmospheric conditions," IEE Proc.-Gener. Transm. Distrib., Vol. 142, pp. 59–64, 1995.

[10] Mattavelli, P., Rossetta, L., Spiazzi, G., &Tenti, P. General purpose fuzzy logic controller for DC-DC converters. IEEE Transactions on Power Electronics, 12(1), 1997, pp. 79-85.

[11] E. Koutroulis, F.Blaabjerg, " A New Technique for Tracking Global Maximum Power Point of PV Arrays Operating Under Partial-Sheding Conditions" IEEE JORNAL OF PHOTOVOLTAICS, vol. 2, no.2, April 2012.

[12] Petrenko, Y.N. Fuzzy logic and genetic algorithm technique for non-liner system of overhead crane / Y.N. Petrenko, S.E. Alavi. Computational Technologies in Electrical and Electronics Engineering (SIBIRCON), 2010 IEEE Region 8 International Conference, 11-15 July 2010. pp. 848 - 851.

[13] Hassan, S.Z.; Li, H.; Kamal, T.; Arifoğlu, U.; Mumtaz, S.; Khan, L. Neuro-Fuzzy Wavelet Based Adaptive MPPT Algorithm for Photovoltaic Systems. *Energies* **2017**, *10*, 394.

[14] S. Z. Hassan, H. Li, T. Kamal, M. Nadarajah and F. Mehmood, "Fuzzy embedded MPPT modeling and control of PV system in a hybrid power system," *2016 International Conference on Emerging Technologies (ICET)*, Islamabad, 2016, pp. 1-6. doi: 10.1109/ICET.2016.7813236

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 749-754 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Human Sit Down Position Detection Using Data Classification and Dimensionality Reduction

Paul Rosero-Montalvo[1,2,*], Daniel Jaramillo[1], Stefany Flores[1], Diego Peluffo[1,3], Vanessa Alvear[1], Milton Lopez[2]

[1]*Universidad Técnica del Norte, Ibarra- Ecuador*

[2]*AInstituto Tecnológico Superior 17 de Julio-Yachay, Yachay-Ecuador*

[3]*Corporación Universitaria Autónoma de Nariño, Nariño-Colombia*

A R T I C L E I N F O

Article history:
Received: 24 March, 2017
Accepted: 19 May, 2017
Online: 16 June, 2017

Keywords:
Chair position
Knn
Pca
Embedded system

A B S T R A C T

The analysis of human sit down position is a research area allows for preventing health physical problems in the back. Many works have proposed systems that detect the sitting position, some open issues are still to be dealt with, such as: Cost, computational load, accuracy, portability, and among others. In this work, we present an alternative approach based on an embedded system to acquire the position-related variables and machine learning techniques, namely dimensionality reduction (DR) and classification. Since the information acquired by sensors is high-dimensional and therefore it might not be saved into embedded system memory, for this reason the system has a DR stage based on principal component analysis (PCA) is performed. Subsequently, the posed detection is carried out by the k-nearest neighbors (KNN) classifier between the matrix stored in the system and new data acquired by pressure and distance sensors. Thus, regarding using the whole data set, the computational cost is decreased by 33 % as well as the data reading is reduced by 10 ms. Then, sitting-pose detection task takes 26 ms, and reaches 75% of accuracy in a 4-trial experiment.

1. Introduction

The area of analysis and detection of human sit down during workdays and in classrooms has been explored by several studies that uses electronic devices and a wide range of sensors. Mostly, such studies aim at detecting of sitting static positions and determining their influence on the normal development of daily activities [1] [2]. In [3], authors explain that keeping bad/incorrect postures habits may cause attention problems, and loss of motor and brain skills for problem solution and decision making, as well as physical health problems such as muscle rigidity, fatigue and muscle pain [4]. The recognition/detection of sitting poses can be conducted by develop mainly two approaches: Image-processing-based and sensor-based approach [5]. The latter requires presure sensors able to be located on different weight-supporting surfaces by taking into account the pressure distribution over the whole surface.

The most common human positions in chair have been categorized in different manners. For instance, in [6], after performing several experiments, it is determined that there are 10 typic poses. Another study [5] considers 5 poses to develop the detection system. Despite that different systems have been proposed for sitting-pose detection, some open issues are still to be dealt with, such as: Cost, computational load, accuracy, portability, and among others.

In this work, we present an alternative approach based on a embedded electronic system to acquire the position-related variables and machine learning techniques, namely dimensionality reduction (DR) and classification. To design our detection system, we start with a comfortable chair that helps a user to readily reach a good posture so that his/her body keeps the natural state wherein the spin position and the space between the vertebrae do not change and the muscles of the waist and back relax. Otherwise,

*Paul Rosero-Montalvo, Ibarra-Ecuador, roseropal@utn.edu.ec

when changing to an incorrect position, intervertebral discs are narrowed and ligaments are stressed causing pain and discomfort [7]. The correct chair selection si necessary to be taken int account the human height, for this reaseon it must be able to adjust in order to provide a support for the user's lumbar region to be able to execute his task with comfort [7].

The selection is done of set of poses is highly dependent on the detection system design. Particularly, we set 4 poses detected through direct visualize method.

Once data are acquired by installed sensors inside the chair seat, we perform a data analysis stage using machine learning techniques. Such a stage involves a classification step using the well-known K-Nearest Neighbors (K-NN) multilabel classifier that depends of a metric used for the calculus of the Euclidian distance between two points for select the k neighbors nearest in terms of p attributes [11]. As well, due to the high dimension of acquired data (four dimensions) , we perform a DR step based on Principal Component Analysis (PCA) [8]. To implement our system, we opt by an Arduino platform because it is not only a powerful embedded system but is by construction highly versatile. In addition, there is a great amount of free code available at internet [9] [10] [12].

The electronic system proposed is aimed at detecting different poses that student may make during studying time. It works as follows: we use strength and ultrasonic sensors, which send the information to an embedded system. Such sensors are appropriately located and distributed on the chair.

The microprocessor inside the arduino does not have sufficient resources to storage big size of data and limited RAM memory for reading program, also the system is embedded to the chair and does not have communication for a computer. For this reason, we need to development a stages of machine learning.

To process the acquired data, we first apply PCA to represent the initial data –arranged into a data matrix - in lower-dimensional space [13]. Subsequently, we perform the KKN [14] [15] algorithm on the reduced data matrix. Doing so, we determine the poses that users (students) have made during a pre-established time period. To assess the accuracy of our system, we perform quantitative comparisons with a standard system running the same algorithms over the original higher-dimensional data matrix. As a result, regarding using the whole data set, the computational cost is decreased by 33 % as well as the data reading is reduced by 10 ms. Then, sitting-pose detection task takes 26 ms, and reaches 75% of accuracy in a 4-trial experiment.

The remaining of this paper is structured as follows: Section II, on one hand, presents the design of the electronic system and explains the data acquisition process. On the other hand, it outlines the data analysis made to detect sitting poses. Section III gathers the obtained results. Finally, Section IV draws the concluding remarks and future work.

2. Materials and Methods

The design of the Human Position detection in a chair system proposed in this work can be divided into two main stages: Acquisition and data analysis. An explaining diagram of the proposed system is shown in Figure 1.

2.1. Considered sitting poses and data settings

To establish the poses to be considered for our study, we use direct observation for setting the types of sitting made by a study group. The analysis is done over a University student group, who are monitored during their studying activities at library, as shown in Figure 2. As a conclusion, we find that there is a sitting pattern evidenced in 4 poses, similarly as proved in another study [16].

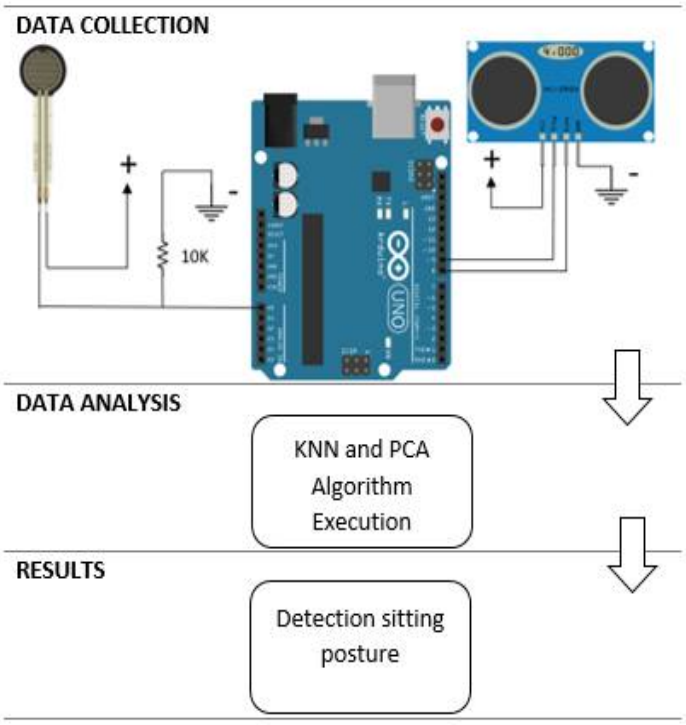


Figure 1. Explaining diagram of the proposed system. It mainly involves stages for data collection and analysis.

Figure 2. Sample picture of the student's type of sitting. Four poses are determined: Upright (red), leaning forward without back support (yellow), bowed forward (green), and tilted backwards (blue)

Table I gathers the name and description of all the considered poses. Figure 1. Comparatively, Poses **a)** (in red) and **c)** (in green) are better than the remaining ones since they have lumbar support, preventing from extra loads on the intervertebral discs of the spine, while relaxing the hip and back muscles.

TABLE I. DESCRIPTION FOR THE POSITIONS

#	Pose	Brief description
a)	Upright	The back and thighs form an angle of 90◦.
b)	Leaning forward without back support	The lower back have no support. Slope of up to 40º.
c)	Bowed forward	It has lumbar support. Slope of up 40º.
d)	Tilted backwards	Upper back against the chair backrest. Relaxed.

Data are acquired from observing 7 students (4 men and 3 women). Each of them is monitored for a while and making the 4 considered poses depicted in Figure 3. Our system linearly picks samples over a period of 1 minute. The processor works at 16MHz with own clock.

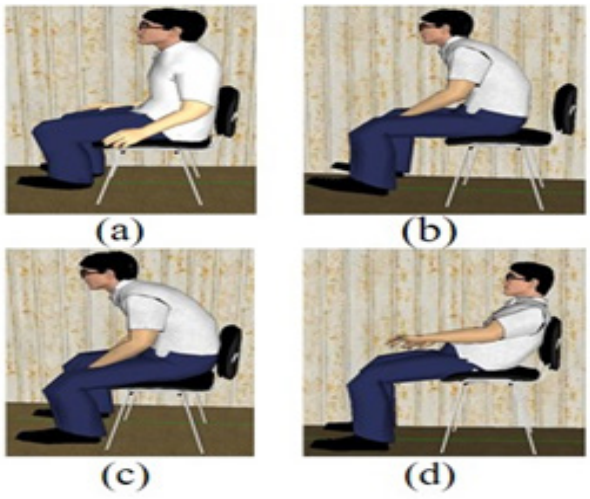


Figure 3. Sitting poses to be performed by the students from study group.

2.2. Electronic System Design

We design the circuit using as a main element the Arduino 1 R3 as embedded system. Three strength sensors (FSR-402) and one ultrasonic sensor (HCSR-05) are connected to Arduino through a transmission line that carries data while the user is sitting on strength sensors surface. Sensors are utilized as follows: Strength sensors are placed together on the chair being adapted inside the seat, while the ultrasonic sensor is placed on top of backrest as shown in Figure 4

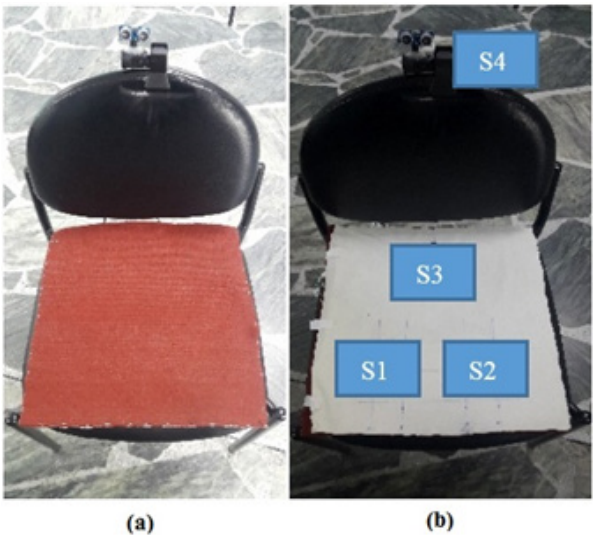


Figure 4. Sensor distribution on the chair. The figure shows the front (a) and backhand (b) view of fabric.

2.3. Data Analysis

The acquired data are arranged into a data matrix **M** (training set) which is **m** x **n**, where m is the number of samples or data points and n is the number of attributes or variables representing each data point. Initially, since 4 sensors are used, we get **n** = 400 and **m** = 4 for the higher-dimensional matrix **M**. Then, such a matrix is reduced by using PCA [10]. Particularly, we extract a new, 2-dimensional matrix. Over the reduced matrix, we run a K-NN classification algorithm. It depends on a metric used as a distance to measure how similar are the data points. Here we consider the Euclidean distance [11]. The data matrix **N** (test set) which is **j** x **l**, where j=100 and l=4.

Since strength sensors tend to be very sensitive, during data collection, the outliers must be clearly identified, especially those generated in the descent to the chair and sitting movements before establishing a fix sitting pose. We perform a manual outlier removal procedure. A total of 100 data points per position are considered for training the system. PCA is used to find the most relevant variables following a variance criterion. We obtain that sensors 3 and 4 represents the 86.66% of the data representation (see Figure 5). To perform the DR procedure, we set the resultant dimension as 2. So, we can draw 2D scatter plots as shown in Figure 6.

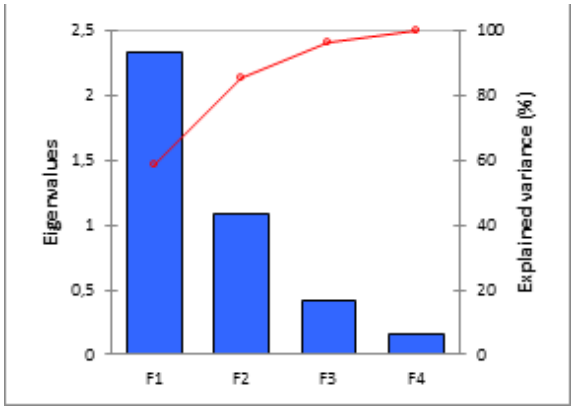


Figure 5. Eigenvalues bar diagram and explained variance plotting for the original higher-dimensional data matrix

The considered variables are those for the strength sensor S3 and proximity sensor S4. Another matrix size reduction should still be performed to prevent from saturating the RAM of the microcontroller [17] [18]. Following a K-NN scheme the number of rows (samples) can be reduced by excluding the data points being close from the average data distribution. Doing so, we get that the new number of rows becomes 80. Here we name the new, reduced matrix is named S that shows at the figure 6

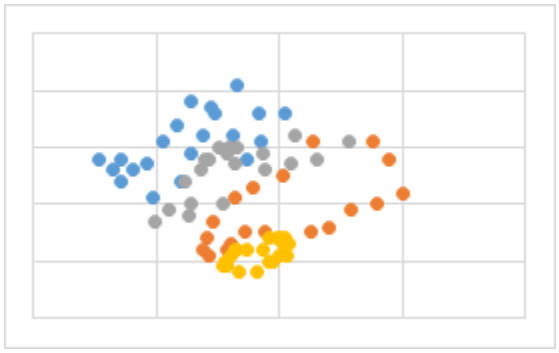


Figure 6. Scatter plot of reduced matrix. Each color represents one of the considered classes (poses) as follows: Pose 1 in blue, pose 2 in orange, pose 3 in gray and pose 4 in yellow.

The **S** matrix is stored in the embedded system RAM, in this way is faster the access to registers in relation to data EEPROM memory, but the CPU had overload when integrated all sensors and KNN algorithm. For this reason, we implemented a restart system after the electronic system has classified each five same times the user postures on the chair.

The flow diagram of the implemented approach is graphically described in Figure 7.

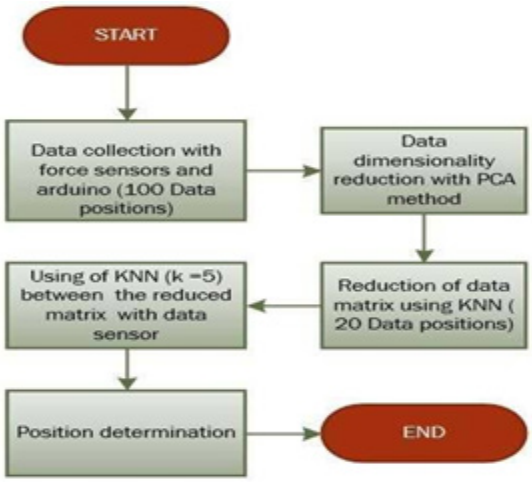


Figure 7. Flow diagram of the proposed data classification approach used for sitting-pose detection

The most important part in this work, is the embedded KNN algorithm inside an 8-bit microcontroller. After obtain the reduced matrix in 82 points with a position label, it necessary create a hexadecimal type file which must compilate in RAM memory. In the principal program, we call the file to read and compare each matrix's point with new data acquired for the accelerometer sensor and determinate the shorter distance and predict the position with the matrix label. The picture 8 shows a pseudocode that describes the K-NN function implemented [19].

```
KNN ALGORITHM
Include position's database //Insert database in hexadecimal format with four positions and twenty data
//points each one. Database sin mxn points where m=82 and n=4.
Integer column, row // variables for the matrix lecture
Integer cont_vector;// vector count variable
Double data_x, data_y, data_z // data acquisition variables of the matrix
Double lecture_x, lecture_y, lecture_z // data acquisition of the accelerometer
Double knn_dist // distance variable
Vector [5] five_less_distance; // knn vector
  Begin
        Read for accelerometer x, y, z in lecture_x, lecture_y, lecture_z
        Repeat
              Repeat
                    Case column
                          Case 0
                                Store in data_x
                          Case 1
                                Store in data_y
                          Case 2
                                Store in data_z
                    End cases
                    knn_dist=((data_x-lecture_x)^2+(data_x-lecture_x)^2+(data_x-lecture_x)^2)^1/2
                    Repeat
                          If vector less distance at position 0 is greater than position n + 1
                                Change positions.
                                Remember label by position
                          Repeat cont_vector <5
              Column < position
        Row < position database row length
        Repeat
              Print label saved in the top 3 positions
        Result <3
        Wax variables
        end
```

Figure 8. K-NN pseudocode write in C language

3. Results

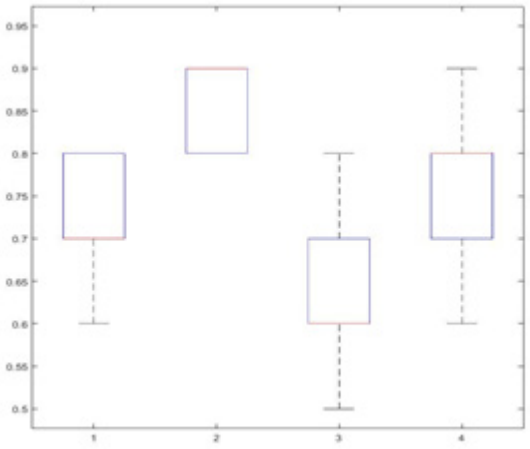
The tests were conducted on both a standard computer (PC) and the proposed system. Ten data points per pose taken at random were tested considering the matrices ***N*** and ***S***. The former matrix is only tested on the PC.

Results were obtained for all the analyzed poses a. Using matrix ***N***, for pose 1, an average accuracy rate (AAC) of 72% success rate was reached. Likewise, for poses 2, 3 and 4 the AAC was 86%, 63% and 77%, respectively. Figure 9 shows the ACC boxplots. Given its simplicity, the system outcomes are comparable with those reported in literature. As a result, the rules for x and y to reduce the dimension were:

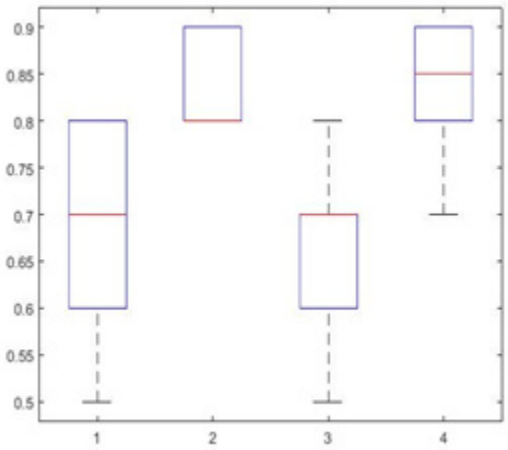
$$x = \sum_{i=0}^{t} (-0.0945 * i) + (0.7541 * i) + (0.6487 * i) + (-0.0384 * i),$$

and

$$y = \sum_{i=0}^{t} (-0.0232 * i) + (0.0274 * i) + (0.0239 * i) + (0.9991 * i) .$$



Using ***S***, AAC is of 68% for pose 1, 83% for pose 1, 67% for pose 3 and 83% for pose 4. These results can be graphically appreciated in Figure 10.



Results regarding memory usage of the embedded system processor are significant good. When inputting the higher-dimensional matrix ***M*** 99% of the flash memory of 8Kbytes is used, and therefore there would not be enough space for typing more script lines. On the contrary, using the lower-dimensional matrix ***S***, only the 30% of the memory is used, leaving enough to add lines of programming space. As a result, regarding using the whole data set, the computational cost is decreased by 33 % as well as the data reading time is reduced by 10 ms. Then, sitting-pose detection task takes 26 ms, and reaches 75% of accuracy in a 4-trial experiment.

4. Conclusions and Future Work

This work presents an alternative approach to deal with the problem of sitting-pose detection, which is based on a sensor network to acquire the position-related variables and machine learning techniques. Particularly, dimensionality reduction (DR) and classification techniques are considered. In this regard, both K-NN and PCA show to be suitable.

For future work, other issues are to studied such as real-time applications, computational load optimization and the use of different classification techniques to reach better performance.

References

[1] A. G. Costa y R. Fiorito, «Evaluación de posturas estáticas sedentes: Método del TEPmetro,» Santa Fe, Argentina, 2014.

[2] B. Mutlu, A. Krause, J. Forlizzi, C. Guestrin y J. Hodgins, «Robust, Low-Cost, Non-Intrusive Sensing and Recognition of Seated Postures,» de *Research Showcase at CMU*, 2007.

[3] D. Cohen, «An objective measure of seat comfort,» de *Journal of Aviation, Space, and Environmental Medicine*, 1998.

[4] M. Halender y L. Zhang, «Field studies of comfort and discomfort in sitting.,» *Ergonomics,* pp. 895-915, 1997.

[5] Y.-R. Huang y X.-F. Ouyang, «Sitting Posture Detecting And Recognition Using Force Sensor,» de *International Conference on BioMedical Engineering and Informatics* , Chongqing, 2012.

[6] M. Zhu, A. Martínez y H. Tan, «Template-based Recognition of Static Sitting Postures,» de *Computer Vision and Pattern Recognition Workshop*, Madison, Wisconsin, USA, 2003.

[7] Y. Mingjiu y Q. L. Changde, «Ergonomics Analysis for Sitting Posture and Chair,» *International Conference on Computer Aided Industrial Desing and Conceptual Desing ,* vol. 7, pp. 1-4, 2007.

[8] L. Vilhena Martins, M. Santos, C. Quaresma y P. Vieira, «Intelligent Chair Sensor, Classification of Sitting Posture,» de *Communications in Computer and Information Science*, 2013.

[9] Q. Jiang y J. Shi, «Sparse kernel entropy component analysis for dimensionality reduction of neuroimaging data,» de *36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Chicago, 2014.

[10] P. Gonzáles Martín, A. Diaz de Pascual, E. Torres lezama y E. Garnica Olmos, «Una Aplicación del Análisis de Componentes Principales en el Área Educativa,» *Revista Económica Universidad de los Ándes Venezuela,* vol. 26, pp. 1-18, 2012.

[11] R. Min, D. Stanley, Z. Yuan, A. Bonner y Z. Zhang, «A Deep Non-linear Feature Mapping for Large-Margin kNN Classification,» de *Ninth IEEE International Conference on Data Mining*, Miami, FL, 2009.

[12] P. Jamieson y J. Herdtner, «More missing the Boat — Arduino, Raspberry Pi, and small prototyping boards and engineering education needs them,» de *Frontiers in Education Conference (FIE)*, El Paso, TX, 2015.

[13] M. P. y V. Stopjakova, «PCA Data Preprocessing for Neural Network-Based Detection of Parametric Detecs in Analog IC,» *2006 IEEE Desing and Diagnostics of Electronic Circuits and systems,* nº 9007211, pp. 129 - 133, 2006.

[14] X. Hao, X. Tao, C. Zhang y Y. Hu, «An Effective Method To Improve kNN Text Classifier,» de *Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, 2007. SNPD 2007. Eighth ACIS International Conference*, Qingdao, 2007.

[15] Y. Komai , Y. Sasaki, T. Hara y S. Nishio, «A kNN Query Processing Method in Mobile Ad Hoc Networks,» de *2011 IEEE 12th International Conference on Mobile Data Management*, Lulea, 2011.

[16] Y. Zheng y J. Morrell, «Comparison of Visual and Vibrotactile,» *IEEE Transactions on Haptics,* vol. 6, pp. 13-23, 2012.

[17] B. Qui, V. Prinet, E. Perrier y O. Monga, «Multi-block PCA method for image change detection,» de *Image Analysis and Processing, 2003.Proceedings. 12th International Conference on*, 2003.

[18] P. Dhillon, J. Gee, L. Ungar y B. Avants, «Anatomically-Constrained PCA for Image Parcellation,» de *Pattern Recognition in Neuroimaging (PRNI), 2013 International Workshop on*, Philadelphia, 2013.

[19] S. Nuñez-Godoy *et al*., "Human-sitting-pose detection using data classification and dimensionality reduction," *2016 IEEE Ecuador Technical Chapters Meeting (ETCM)*, Guayaquil, 2016, pp. 1-5.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 755-764 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Designing a Smart Car Parking System (PoC) Prototype Utilizing CCTV Nodes: A vision of an IoT parking system via UCD process

Muftah Fraifer*, Mikael Fernström

Interaction Design Center, dept. Computer Science and Information Systems, University of Limerick, V94 T9PX, Ireland

ARTICLE INFO

Article history:
Received: 04 April, 2017
Accepted: 02 June, 2017
Online: 16 June, 2017

Keywords:
Smart parking
Video processing
Interaction systems
UCD
HCI
IoT

ABSTRACT

Using smart parking systems has become very important, and particularly so for metropolitan areas, because of the benefits for drivers in many aspects, such as time, frustration, stress, and anger, in addition to the increased consumption of fuel while searching for a vacant parking space. This paper proposes a review of recent advances in sensing and communication technology concerning smart parking systems. It includes a brief study of the selected topics and provides an implementation process of those selected systems. Moreover, this work proposes a design approach for a smart car parking system prototype based on utilizing CCTVs (nodes), it is also illustrates the algorithms used for computer vision detection through simulation and real environments, as the system has been deployed in both these environments. Furthermore, the system has been tested and evaluated by stakeholders via a user-centred design process by applying a qualitative research; the promising results demonstrate the effectiveness of our prototype. Finally, this paper discusses the benefits of engaging the stakeholders to develop the prototype.

1. Introduction

Installing a parking system at any car park can be expensive, but recently, many researchers have been exploring techniques, such as vehicular ad hoc networks (VANET), or integrated RFID with Wireless Sensor Networks (WSN). As the Internet of Things (IoT) has evolved, further possibilities to provide services have emerged. Here, IoT can connect objects, such as smart-phones, sensors, and actuators, as nodes (in our case, CCTV cameras). The development of IoT may transform a number of domains dramatically, for example, automation, transportation, healthcare, and financial services. IoT can also be applied to new concepts and developments for smart homes to provide improved comfort, security, and quality of life. Some parking systems need extensive infrastructure investment, for example, digging holes in the ground at each parking spot to install a sensor. However, we are proposing an approach that can utilise an existing system instead, i.e., CCTV (node) coverage. The proposed system has many advantages. Firstly, it is cost efficient, as it is using the existing CCTV cameras (nodes). Secondly, no infrastructure changes are needed for the parking lots. Finally, and according to the interviews with stakeholders that were done during the evaluation process using the User-centred Design (UCD) method, the prototype can be useful for small and large parking areas whether the parking area is active or not where the user can easily identify and locate the vacant parking spots on the fly. This paper presents a prototype of a low-cost and flexible smart parking monitoring system using an embedded micro-web server, with IP connectivity for accessing and controlling devices and appliances remotely [1].

2. Contributions

We introduce a novel IoT parking system porotype based on computer vision and build a model of the system operation.

The general approach to design a usable Interactive Technology (IoT) smart parking system for stakeholders (end-users). The methodology for accomplishing this goal was to apply User Centred Design (UCD), which included observation, surveys with stakeholders, and in-depth interviews with users. This article and research contribution covers different aspects:

- It review and presents a related works of different parking systems.
- Design, implement, verify, validate, test and evaluate a smart parking system utilizing CCTVs (camera nodes) facilitating provision of mobile services in indoor and outdoor environments within a university campus as an example.

*Corresponding Author: Muftah Fraifer, Interaction Design Centre -Dep of CSIS, V94 T9PX, Ireland | Email: muftah.fraifer@ul.ie

- In the field of IoT, facilitating the parking systems process by enabling a camera (thing) to be a smarter thing using a Galileo board (microprocessor) by providing a solution by links smart objects to the Internet.
- In the field of HCI, the knowledge and insights about UCD and how it can be put into practice to design a usable and effective IoT Parking System.

3. Related Work

To provide an overview of the different technologies used by others, the following is a summary of the systems we have reviewed. Foshini et al. discussed the design and implementation of an M2M application in the field of road traffic management study which used an integrated IoT retractable bollard barrier management system to allow vehicular access to restricted city areas, based on a standard infrastructure and software components [2]. Pala and Nihat used RFID for finding the nearest vacant parking space and reporting to the driver [3]. Tang et al. developed a system using crossbow motes that have a low unit cost; the system enabled a car to detect entry to the car park and guide the driver to an empty parking space by signs displayed to the driver [4]. Benson et al. proposed that an RF transceiver and micro controller system could monitor the availability of car-parking spaces and send this information to drivers and car park administrators [5]. Funck et al. proposed a system using CCTV cameras fitted in car parks to automatically detect car parking spaces. However, all these methods have the problem that sometimes they are not accurate [6]. Panayappan et al. described a parking system using VANET, which located available parking spaces. The system depended on roadside units to relay parking messages and GPS coordinates to locate parking positions [7]. Kianpisheh et al. presented a smart parking system using ultrasonic sensors. For each individual car park, a sensor was fixed in the ceiling above each parking space. Such ultrasonic sensors operate based on echolocation. The sensor emits a sound, which is reflected back to the sensor; the presence of a vehicle changes the timing of the reflection [8]. Mathur et al. discussed a centralized solution, with cars equipped with ultrasonic sensors driving past parking spaces to collect occupancy data and upload the data to a database; the cars that require parking query the database [9]. In 2010, Hazrin et al. proposed a smart parking using SMS services to communicate with drivers. A parking reservation system was developed whereby users could reserve a parking spot with an SMS message, including GPS information [10]. Bong et al. proposed a system with image processing algorithms, the motivation being to minimise the cost. The system was called Car-Park Occupancy Information System (COINS) [11]. Klappenecker et al. modelled a parking lot as a continuous-time Markov chain. The parking area was modelled as a grid, and schemes for information aggregation and dissemination over the grid were proposed [12]. Jain et al. proposed a system with a Gate-PC Controller and Embedded Gate Hardware, RFID System, Modular Parking Management Platform [13]. Lee et al. proposed the use of a combination of magnetic and ultrasonic sensors for the accurate and reliable detection of vehicles in a parking lot, also describing a modified version of the Min-max algorithm for the detection of vehicles using magnetometers [14].

4. Research Methodology and Background

Our general approach is based on UCD. This includes observations of user behaviour, surveys with stakeholders, and in-depth interviews with users (ranging from the parking users to system operators and facilities management staff).

This is carried out in parallel with technical probes, i.e., incremental development and integration of system components. At present, we have a low-fidelity prototype where parking and CCTV nodes are emulated in an office environment (using toy cars) and then actual parking spaces. We explore the desired services informed and shaped by our user groups. We strive towards using open source software as far as possible, and in our preliminary sketches of the system, we may also be using Google APIs, including both maps and calendar functions that can be used for reserving parking spaces based on the user's planned activity on campus. It is thought that a minimally intrusive system that users' experience as augmenting their activity may be more likely to succeed. To address this, we are using part of the University of Limerick campus as a living lab to engage with users and evaluate a range of services. The campus has full coverage with open Wi-Fi, 3G, CCTV and a range of architecturally complex spaces with buildings interspersed with car parking [1].

5. Hardware and Software Requirements

- Micro-controller Intel Galileo board or, as in our case, desktop computer with Linux.
- CCTV cameras (nodes) A,B,C, etc.
- Parking Server Computer (PS).
- Software components with Java and Intel OpenCV library, Parking Server PS (LAMP stack)

Figure 1 below shows a Galileo board, which is used to run the OpenCV code directly. Also, the code can be compiled and run using the live-video streaming link for the live scene of the parking areas to monitor the parking status in real time.

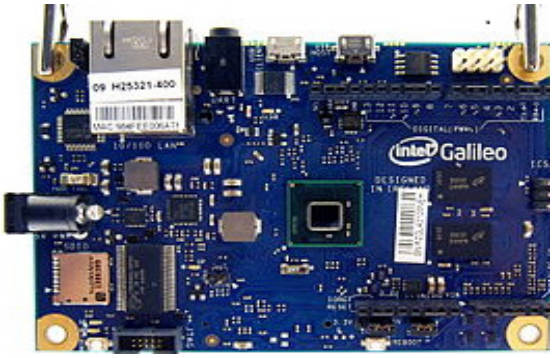


Figure 1: Galileo board

6. Application Architecture Layers

From the high level of the IoT based smart parking system; there are three main layers ; As shown in Figure 3.

- Application layer: the main purpose of this layer is data processing and service providing.
- Transmission layer: which is about transferring data over long distances and for wide areas.
- Perception layer: the system aims to collect and process the data from the physical world (in our case is using computer vision technology)

Figure 2 : Prototype CCTV camera node with the cover shield to protect the Galileo board

7. Design Goals

Based on our literature review and having access to the existing infrastructure on campus, we set out to explore a smart parking system based on CCTVs/Web-cams (nodes) that can provide drivers with convenient and simple parking service information. The prototype has been evaluated and tested using stakeholders, and the UCD process has been used to gather the end-users' needs and requirements. The process is also used to improve the prototype iteratively.

The system must fulfil a range of criteria. Firstly, it has to provide real-time navigation support to the driver, for example, by utilizing Google maps. Secondly, it has to provide a friendly parking information service to both the parking user and the parking facility's operator/owner. Finally, using CCTV cameras (nodes) for this purpose as well as for security purposes, it is an economical option. The main function the system will perform is to gather information about the parking status in real-time, and help drivers to navigate directly to vacant parking spaces [1]. Figure 3 and Figure 4 shows (admin GUI) when run the system in outdoor and indoor environments

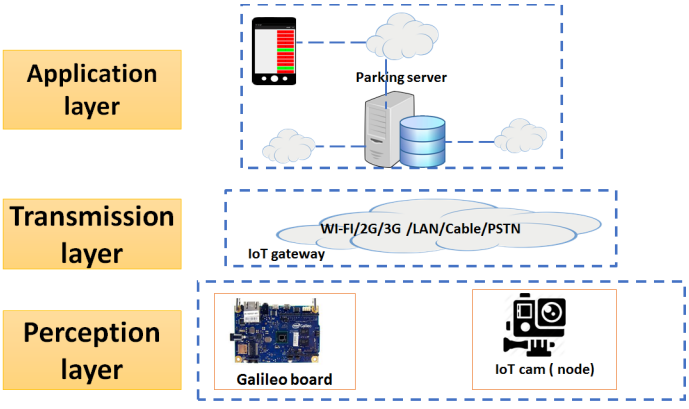


Figure 3: smart parking IoT application architecture

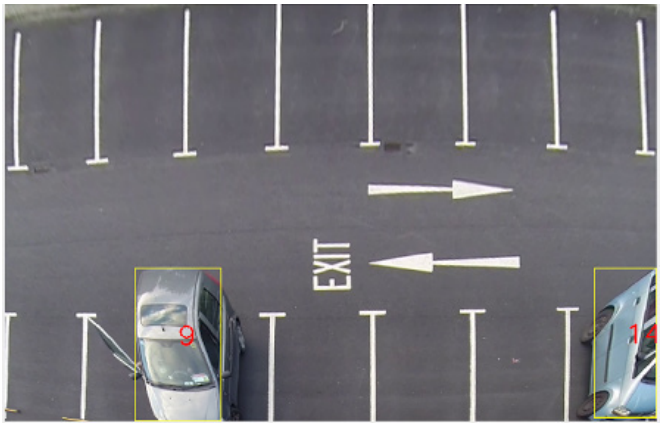


Figure 4: snapshot of testing scenarios in out-door environment covered by one node from admin GUI (monitoring (parking status on real-time)

8. User-Centred Design and its Influence on our Porotype

The UCD methodology in this process is used as a framework of processes in which the needs of stakeholders of a product or service are given extensive attention at each stage of the design process. It begins by defining the problem, which in our case is congestion during peak times and the effect on drivers, as it appears difficult for the driver to find a parking spot [1].

Another aspect to this research-study is carried out to explore the approach of UCD with a focus on stakeholders' involvement in the development of a smart parking technology prototype. We tried to alleviate this everyday issue of finding a parking space through the proposed prototype. Although some of the stakeholders' suggestions may involve some costs, the stakeholders have been identified and a thorough investigation of their needs has been conducted. By performing tasks and needs analyses, designers can develop alternative design solutions to be evaluated by users. The data analysis was based on a qualitative content analysis from a interaction design perspective.

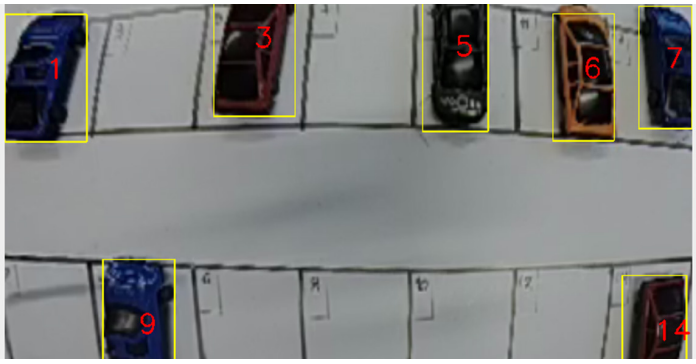


Figure 5: snapshot of Testing scenario in in-door environments from admin GUL (monitoring (parking status on real-time)

9. Prototype of the Proposed System

There are three main components:

- firstly, an app installed on the user's smartphone that connects the user with a parking server (PS);

- secondly, an IoT cams (which in our prototype is a Web Camera) nodes that is connected to a microcontroller (Intel Galileo);
- and lastly, a PS (Parking Server) hosting a database.

When the user installs the app on their Smartphone, they are given unique ID. The app can also hold a user's profile and preferences including all the information entered by the user during the registration with the system and the server. This allows the app to provide services that match a user's parking preferences with actual available spaces in the actual parking scenario at UL [1].

Our main scenario is the aforementioned openly accessible areas. Each parking lot is divided into areas that are covered by nodes, as shown in Figure 4 below.

Our prototype uses IoT camera nodes. An IoT cam can be set to report changes to the scene as a grid of polygons. At the same time, the user can interact with the system through the app. The IoT cam uses a DHCP connection to connect with the PS, which is on a fixed IP.

A simple blob-detection algorithm (using Intel's OpenCV library) is used for computer vision, which monitors and reports changes in a defined grid of polygons (i.e. if parking spaces are vacant or occupied). IoT cams connect to the PS, which aggregates the data and updates the database of available car parking spaces in close to real-time. It should be possible to define and redefine parking areas by system operators. The system will operate in close to real-time, so we can aim to improve the parking experience rather than forcing and regimenting it. Status information regarding the parking area within the coverage of any IoT cam (nodes) is continuous, and changes are sent to the PS. We are using both scenarios (indoor and outdoor environments); the first is only to test the feasibility of the system, which depends on toy cars in a model parking lot, as shown in Figure 3, while the second prototype was tested on real cars with an actual car parking area, as shown in Figure 4 (both snapshots were taken from admin GUI). The camera, in a simulated environment (CCTV), is fixed in position with a side view of the parking space. The height of the camera must be sufficient to obtain a clear plan view of the parking

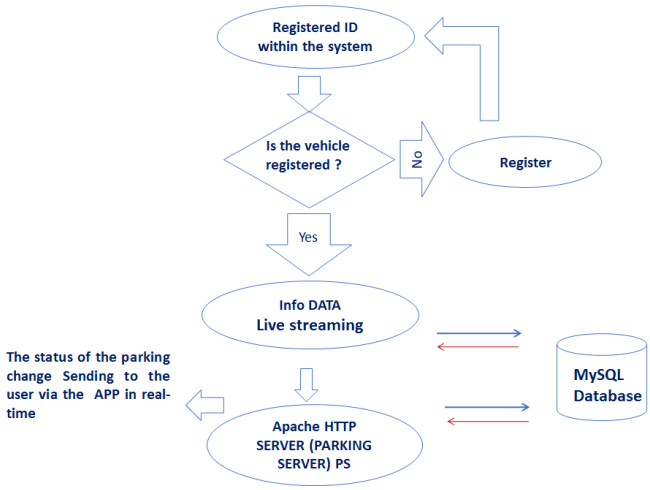


Figure 6: The operations triggered by the car-in and car-out events.

space, as in Figure 1. The IoT cam is positioned with an angled view of the parking space acquiring a fixed scene [1].

10. Interaction with MySQL Database

We are using a LAMP server as the open-source platform, which uses Linux as the operating system, Apache as the Web server, MySQL as the relational database management system, and PHP for scripting.

The system has real-time communication between the mobile app and the server, using a RESTful API. Each car park can be identified with labels, for instance, (Node A, Space 1, 2, 3, 4 etc.), so when a driver enters the parking area, they can easily identify each parking space. The main technique of the system is to monitor the system through IoT cams, while the video is processed by analysing the live streaming of data and changes are stored in a MySQL database. Figure 7 illustrates the UML use-case diagramor the interaction of the prototype Open CV Algorithms (1)(2)(3)(4).

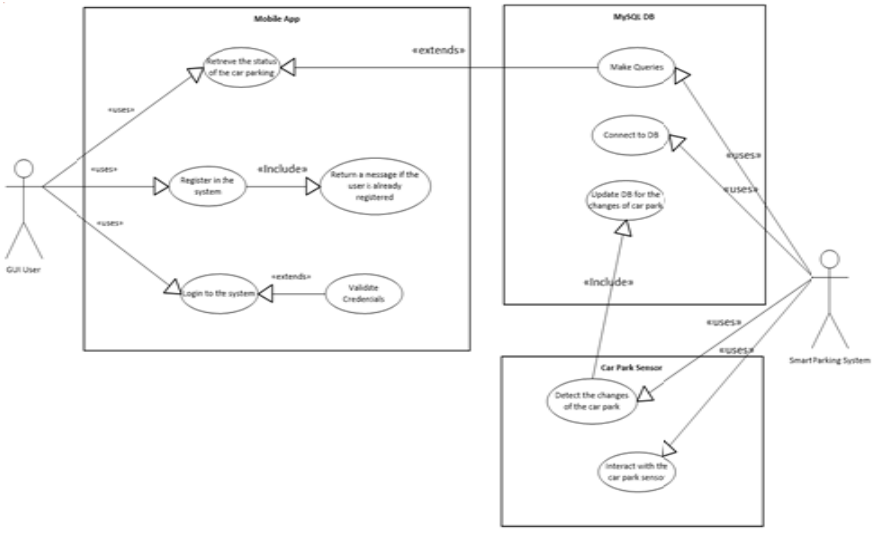


Figure 7: UML -USE CASE Diagram for the protpype

11. OpenCV Algorithms and Senescing Techniques

- The code depends on the OpenCV library, which has been used for video processing, to detect vacant parking spaces.
- The code retrieves frames from streaming video by applying a prospective transform Matrix on source images.
- For each pixel (x,y) of the destination image, the functions compute the coordinates of the corresponding "donor" pixel in the source image and copy the pixel value. A perspective transformation is then applied to the image. The function ***warpPerspective*** transforms the source image using the specified [3x3] matrix:

$$dst(x,y)=src\left(\frac{M_{11x}+M_{12y}+M_{13x}}{M_{31x}+M_{32y}+M_{33x}},\frac{M_{21x}+M_{22y}+M_{23}}{M_{31x}+M_{32y}+M_{33}}\right) \quad (1)$$

Where:

src: input image
dst: output image
M : 3x3 transformation matrix.

- Image loading, including erosion, smoothing, noise reduction, blurring, contrast fix, and brightness are applied to the uploaded images to get a uniform output from the camera input image.

$$dst(x,y)=\min_{(x',y'):\mathbf{element}\,(x,y)\neq 0} src(x+x'+y+y') \quad (2)$$

- We are using the ISODATA method (Iterative Self Organizing Data Analysis Technique) for thresholding of image segmentation to get a high degree of accuracy and to consistently find the contours for each polygon (car) [2].
- A fixed-level threshold is applied to each array element. (THRESH BINARY)

As below

$$dst(x,y)=\begin{cases} maxval & if\ src(x,y) > thresh \\ 0 & otherwise \end{cases} \quad (3)$$

src : input array (single-channel, 8-bit or 32-bit floating point).
dst : output array of the same size and type as src.
thresh: threshold value.

maxval: maximum value to use with the THRESH BINARY and THRESH BINARY INV thresholding types.

- One layer of the algorithm has been applied to detect vacant spaces in the parking lot using a contour method (rectangular detection) to identify the number of cars in the parking lot. The findRect method matches the detected shapes to the shape of a parked car, detecting the cars as rectangles.
- Rectangle X and Y coordinates as in (4), where variables are points, namely, first point and second point, and then a rectangle is drawn around the image.

$$Point\ (x,y) \rightarrow Point2f\ (x* 2^{-shift}, y*2^{-shift}) \quad (4)$$

- This data is sent to the MySQL database through a MySQL connector, with a list of all parking areas, number of possible spaces, etc., so the system can count how many parking spots are available. Using the marked parking spaces as bearings, the code can identify the central position of a car as illustrated in the pseudocode below.

12. Illustration the stages of Implementation of the Code

The algorithm passes through different stages; starting by getting the prospective transform matrix of the image. Then the code retrieves frames from streaming video using the prospective transform Matrix as clarified in the pseudo code below.

Pseudo code for the detection algorithm

```
// If a contour is detected in the parking spot, the contour
represents a parked car (occupied parking spot)
If (contours.size() != 0 )
{

 // checking each contour
 For (int idx = 0; idx < contours.size(); idx++)
  {

   //To get the boundaries of the rectangular contour.
   Rect contourRect=Imgproc.boundingRect(contours.get(idx));

   // If bounding rect is between the two aspect ratio x,y.
  If
  ((contourRect.width < min_x)||(contourRect.height <
min_y)||(contourRect.width > max_x)||(contourRect.height >
max_y))
  continue;

   // To get the aspect ratio.
  double ratio = [(double)contourRect.height
/(double)contourRect.width];
  if ((ratio < x)||(ratio > y))
  continue;
  If (the detected contour is a car)
  End If
   {
  // Draw a rectangle of the car.
  R = rectangle of car,
  // Get the position of each car. Pi = central position of car(i),

  // Give each occupied parking spot an ID.
  ParkID(i) = occupied_parking_spot(i),
  End If
  End for

 }
}
```

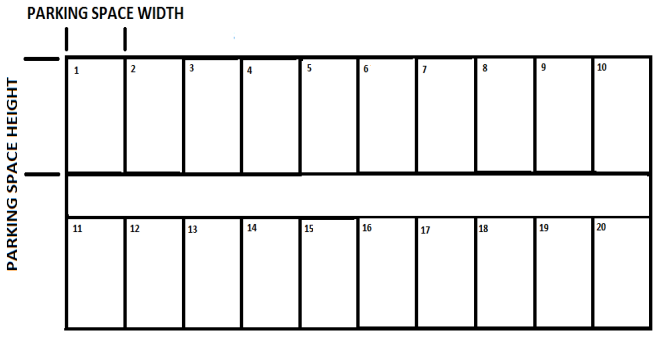


Figure 8: OpenCV algorithm design

13. System Implementation

A smart parking system based on CCTV (computer vision) has been proposed and tested in this paper. There are two Java applications; one is called ***SPSystemApp*** (OpenCV) and the other is called ***SPMobileApp*** (GUI). The interaction between them is via the MySQL database. The main purpose of the first application is to detect the status or changes of the car park and send this information periodically to the database in real time. At the same time, the GUI retrieves these changes for the user.

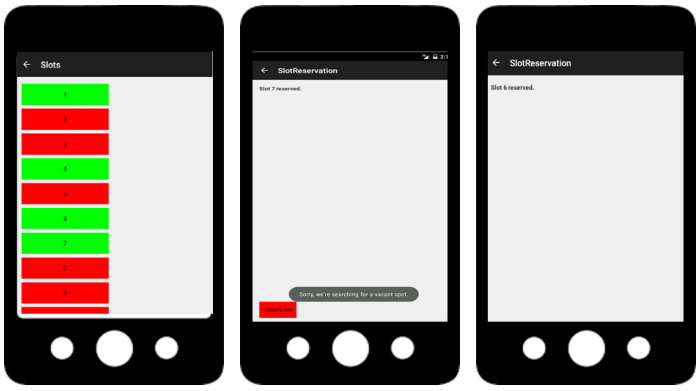


Figure 9: a,b,c Shows (a snapshots)AVD with a reservation scenario (it shows the interactive process by giving another parking spot

Figure 3 shows the implementation of the system, which shows red and green buttons where the user can reserve the parking space. The test cases used to evaluate the functionality of the system are listed in Table 1.

Table 1. Testing the system with different functions

Test scenario	Result
MySQL connector can make a database update record according to status change	Pass
Can the user reserve a specific parking space	Pass
View the number of available parking spots in real-time	Pass
Parking area status will update info to PS when a car leaves the car park	Pass
Able the user to register in the system using GUI	Pass
Ability to do several tasks (Registration/reserve)	Pass

14. The System and its Maximum Effectiveness

Regarding the implementation of the work of the system on the Galileo board (microcomputer); Which distinguishes its use during its work in real time in addition to being a small board next to each camera (node) and do not need to connect video over the Internet to a central computer, where the delay here in the live broadcast to nearly half a minute or more. What also distinguishes our use of the motherboard in this system is the possibility of operating it in a real car par. We used a special cover to protect the processor from atmospheric conditions as shown in the picture. Again, we want to clarify that this article came as part of a research with several purposes to demonstrate the proof of concept and the prototype of the system of a complete illustration of the general perception to take advantage of the work of cameras not only in the parking lot, but also anywhere to include temporary parking on both sides of roads and others Through our experience of the reality of the work of the system and found that its work is in the best conditions and give good results identical to the reality of the case in terms of the number of cars and their location and the possibility of detection by the system if the speed of the Internet is not less than 120 Mbps ; to be able to stream data with a smooth and fast way to emulate reality and thus the quality of images through video broadcast in real time of the parking lot , which is dedicated to a particular camera (node) or more. Also, The image resolution required is no less (1400X960) pixels .To enable the system to operate with acceptable efficiency. The system handles (capturing, analyzing and grabbing frames from the connected nodes) up to 22 frames per second ; This is what distinguishes using OpenCV in this research.

15. System Requirements and Implantation Techniques

The system can detect vacant spaces using two different methods. The first method is using the IoT-Cam, where it pushes any changes in vacancy to the server (the microcontroller/server (in this scenario) while the PS can get the parking area status. The second way is using smart cameras that just capture an image, and all the processing happens on the server where the server can get an image from the camera providing the server is able to get smart camera status.

The IoT cameras (web cameras nodes) are connected via Wi-Fi to the Internet, so we can get the IP-number for live streaming of the video, which covers a certain area of the parking space; When the code is executed online, it assigns the link to the IoT cam to get the video stream analysed on the Galileo board or via the Linux machine..

16. System Reliability in Different Scenarios

Below is a collection of screenshots for application during use; The captured images were placed within the Android framework for clarity. Several scenarios have been considered; For example, the first user booked a certain parking place number 7 as in Fig x.a ; Upon arrival he found the place vacant, In this scenario, the user has only to click on the button called a complaint. Although such

behavior is unlikely to occur. Because it will usually apply fines to prevent such violations as any other system which provides services to the public as it shows in Fig x.b Which shows a second option to the user automatically.

14.a System Efficiency

With regard in the reliability of our prototype. The system has been tested for several different times and for several different scenarios. The first tests were done in the laboratory (indoor environments) .While other tests were conducted in a real atmosphere. The tests carried out in the selected paper for the conference were carried out with a group of stakeholders and the number of them was 12 completed in the laboratory. All tests were ethically proven, and approved by ethical committee in the university. In terms of tests and reliability of experimental system. The system worked well and especially when the cars in both cases (toys and real) was properly parked. We noticed that those cars that are not parked along the lines of the stand cannot be detected by the system. The dark atmosphere makes the system fail to perform its functions, so we mentioned the need to light the parking lot in order to increase its efficiency. It is worth mentioning here. We connected the system to one of the cameras in a nearby parking lot .by Enabling VLC media player for Real Time Streaming Protocol (UDP); latency was quite high somethings, So there is some delay in the video displayed directly.

Based on the speed of the Internet and the quality of the images transmitted from the parking lot be the quality of the system and its efficiency and reliability.

We noticed that the system is delayed in this scenario to approximately 35 to 45 seconds. Sometimes the image is frozen in a fixed format as a result of the speed of the Internet or low performance of the processor. As the main purpose of this research is to clarify the potential of the work of the system in ideal circumstances.

14.b Various Activities for Application and Features Available

Regarding the real implementation process of the system, the stakeholders themselves, who are students, staff, and visitors contributed voluntarily from the experience of the system in practice. This is what distinguishes this work. As the method used in the design process is User-centred Design (UCD).The images below illustrate the practical tests of the system [10a] [10b] [10c] Low-fidelity prototype (paper prototyping) technique was used; as it is a method in the UCD process to ensure that stakeholders meet with satisfaction and acceptance. Then; The application is designed to run on the Android system. In a simple way, the application contains several activities. The most important is the possibility of registering in the system.

Where it is easy for the subscriber to enter his personal data including the number plate of his/her car. The car number has been added to its registration rules through suggestions for some stakeholders to follow the violators of the parking laws that want to apply such a system to facilitate the service to find a vacant position. The second activity relates to the possibility of seeing vacant positions in real time with the advantage of the possibility of reserving a particular position. The application is linked to a database as described above.

And each interaction with the system appears in the picture short message as shown in the figure below.

As previously reported, all trials were completed as shown in the table below. All the tests were in the appropriate conditions in terms of speed of the Internet and the quality of cameras and the weather and lighting. In addition to the emphasis on who was driving cars for the purpose of testing the need to be in a straight position parallel to the place allocated and not be tilted even a little and these are one of the biggest disadvantages of the system now.

17. Graphical user interface GUI for User and Admin

The system has a car parking mobile app as a key component, with a user interface for everyday use. Through the app, car parking information and services can be delivered to the user in a functional and user-friendly way and that are relevant to the user's present goals and location and according to the user's dynamic profile. Below is a collection of screenshots of the application during use.

Figure 10: GUI shows the status of parking area

When a user registers with the system, they are given a unique ID. The emulator used (an Android Virtual Device (AVD)) as in Fig[9] and Fig[10, 11].also the figure below shows real the app on a real device .

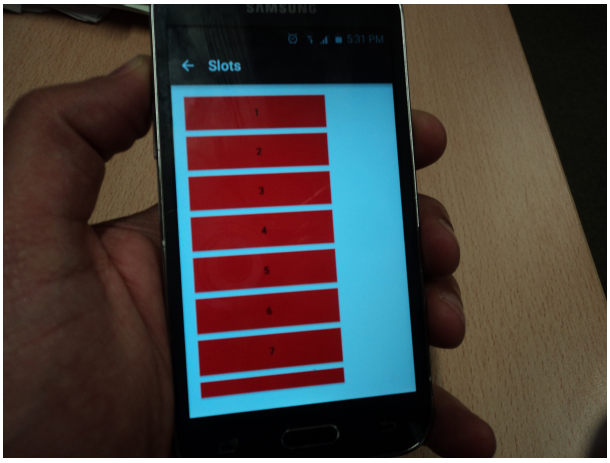


Figure 10 run the app on real device

Also, should someone else have deliberately or by mistake taken a pre-booked space, the system is currently able to allocate the user another parking space by the user clicking on the complain button as illustrated below in Figure 8(a,b,c) where red indicates an occupied and green a vacant spot.

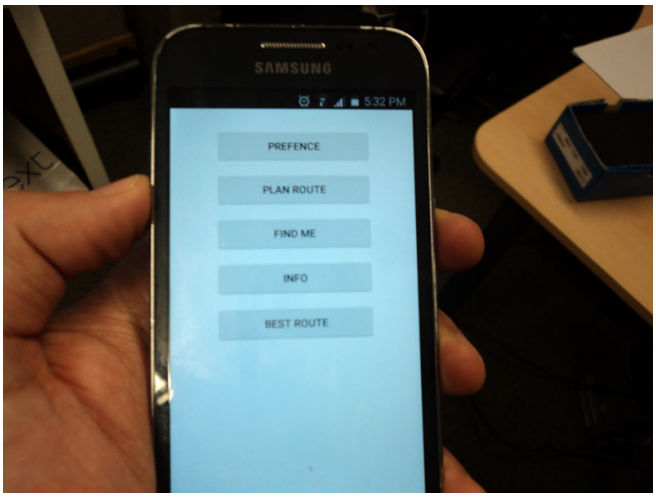


Figure 11 : run the app on real device for registration process

The figure below shows different snapshots of the admin GUI which is designed to help the system administrator to monitor the work and performance of the system.

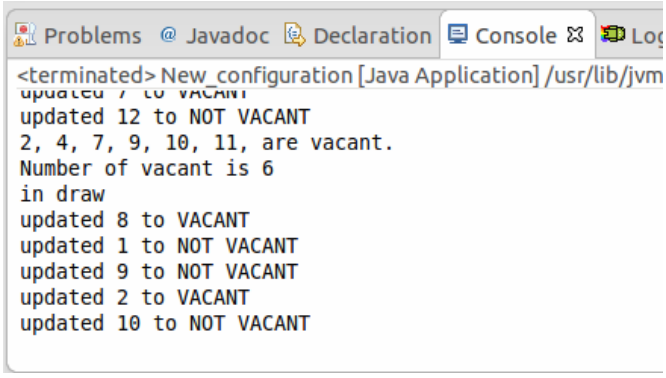


Figure 12 : Snapshot 1 of the admin GUI (monitoring parking status on real-time)

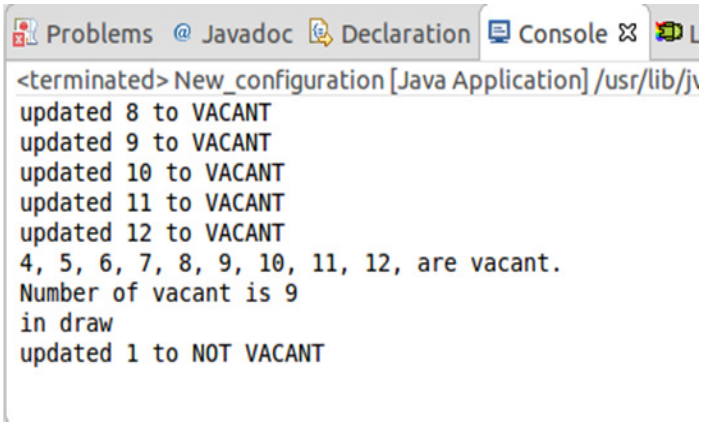


Figure 13: snapshot 2 of admin GUI (monitoring parking status on real-time)

18. Discussion

This smart parking system requires cooperation from multiple users searching for parking spaces at the same time, but there is no barrier to the parking spots. In addition, there are some situations where the system may not work. So far, the system is not designed for use at night, as the car park in our scenario is a scarce resource only during office-hours. If night-time operation was required, night vision cameras or artificial illumination would have to be used. [1]

To create a potentially successful parking locator system, we first considered user needs and requirements [15]. Users are not a homogenous group, and an individual user's goals can change rapidly. The heterogeneity of users may range from residents/workers, short/long-term visitors, deliveries, services and emergency access. As a whole, users may engage in a range of disparate and dynamic activities and roles. In addition, accidents and temporary disruptions that may render dedicated parking spaces temporarily unusable, which presents the problem of getting the system quickly updated to reflect any changes.

In addition, if a car is parked across two spaces, then the system may detect this as two separate shapes, neither of which is the correct shape to be a car and may be ignored [1].

19. Evaluation and Testing Our Prototype upon Empirical Study via Qualitative Research

In particular, intended smart parking solutions for urban environments necessarily and primarily need to involve all the stakeholders in an early phase of the solution development. The stakeholders involved in the UCD process were in general enthusiastic, and they were happy to be able to contribute to the system development. That is in line with research that found that users who are involved in the process indicated that the system was designed to meet their needs. The prototype was evaluated and tested in two iterations using the UCD methodology. The first iteration on the UCD process was informed by 15 participant interviews. This qualitative study builds on a case study to propose a methodological framework for the implementation of a smart parking IoT system.

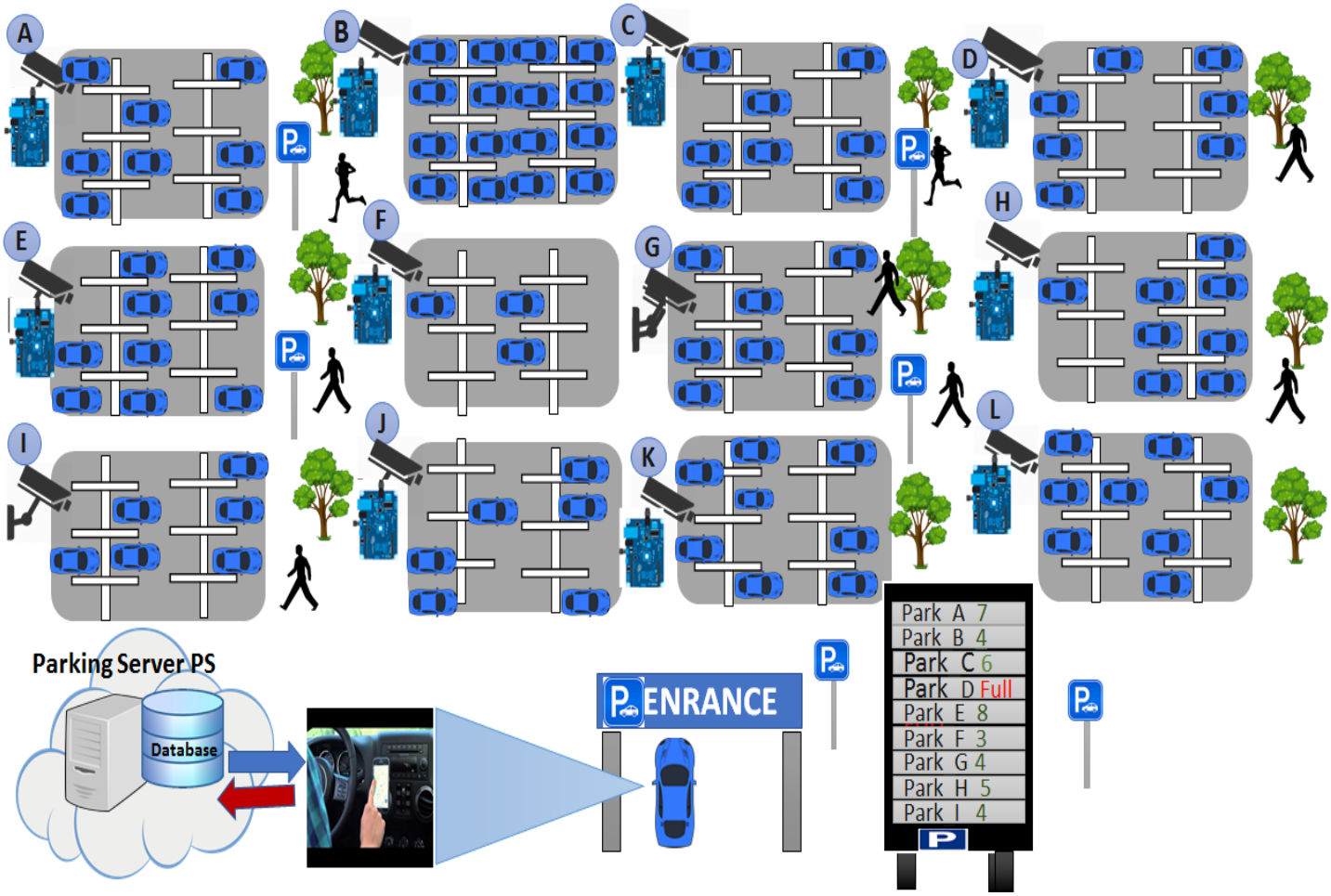


Figure 14: General outline of the IoT prototype using different nodes

The majority of participant stakeholders were impressed by the prototype and believed that the system would be useful, at least to alleviate the issue of finding parking spaces. They believed that even if the system is able to tell the users that there are no vacant spaces available, which would be useful information instead of the driver wasting time and fuel searching for a space when none is available. However, they had a variety of ideas regarding how to improve the prototype. We took this into consideration in the figure (9) above as a general prospective for our IoT parking system prototype.

20. Empirical Study and Ideas from Stakeholders Used to Improve the Prototype:

Our findings from the empirical data indicate that there is a real need for a smart parking system to alleviate the issues when finding a parking spot. Some stakeholders suggested letting all drivers know about the fully occupied parking areas via large information screens while others suggested using small screens connected to our prototype instead of the apps at each parking space/area to help the drivers avoid crowding and congestion. In addition, some recommended utilising large screens at each entrance to the university to provide information to the drivers. Also, interviewees recommended enabling the user to click on a button to confirm that he/she has arrived. Furthermore, stakeholders suggested applying rules about carpooling to convince the students and staff to set up a car-sharing system (a website would be created to support this). Moreover, some suggested using the system to cover all the parking spaces to let the driver know about the parking status. Besides, some stakeholders advised strongly against using a reservation option in the app. (It should be used to get information about the status of parking spaces, they believed, so they need it only to allow them to go directly to the less occupied ones). And to make the system more accurate, there was a suggestion from stakeholders regarding security staff - many stakeholders suggested that there was a need for crew safety, which are employed by the Buildings and Estates department to intervene if a driver finds someone has taken his/her place.

Finally, the ultimate problem may emerge from users, whose mistakes or deliberate ignorance may challenge the system. This raises the question of whether the system serves the users or if the users are serving the system. While it is technically possible to design a system that leaves almost no room for a user's ad-hoc behaviour, such a system may be too complex to be deployed (e.g. physical barriers preventing entry or exit from unauthorized individual locations) [1].

21. The Final Visualization of our Model System

According to the qualitative research with stakeholders, we can conclude that the system would be beneficial for two cases. On the one hand, it would be useful when the parking area is large and by guiding a driver to vacant spaces. Figure 4 represents the model, and a general perception of the system in the final image was that different size lots are desired so the user can easily can recognize the vacant parking space. On the other hand, it is also beneficial in a case when a parking area is small and active.

Conclusion and Future Work

To conclude, the objectives of this research have been achieved. The aggravation involved in searching for vacant parking spaces can been minimised by informing the driver about the status of parking areas. Although much work has been done on smart parking, the work reported in this paper is based upon IoT cams (nodes). We propose a novel architecture for low cost and flexible smart parking using CCTVs (nodes). The system is intended to use previously installed CCTV cameras at any parking space instead of using dedicated systems, which would be more expensive.

Upgrades are currently being done to ensure better reliability and scalability for the proposed system. Some adjustments are being done which focus on making the algorithms effective to cover more parking spots. More iterations of UCD will be applied to the process regarding evaluation of our prototype in order to improve the quality of service. Also, In our future work, we will consider the security aspects of our prototype system as well as implement the system in large scales in real environments.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The authors would like to thank our colleagues in Interaction Design Centre- IDC in UL for their support in pursing this work. We also would like to acknowledge the support of the Ministry of Higher Education and Scientific Research-Tripoli for their funding of the project. Also, part of the article is taken from our previous work as referenced in [1].

References

[1] This paper is an extension of work originally presented in WF-IoT Conference By M. Fraifer and M. Fernström, "Smart car parking system prototype utilizing CCTV nodes: A proof of concept prototype of a novel approach towards IoT-concept based smart parking," in Internet of Things (WF-IoT), 2016, IEEE 3rd World Forum on, 2016, pp. 649-654.

[2] L. Foschini, T. Taleb, A. Corradi, and D. Bottazzi, "M2M-based metropolitan platform for IMS-enabled road traffic management in IoT," Communications Magazine, IEEE, vol. 49, pp. 50-57, 2011.

[3] Z. Pala and N. Inanc, "Smart parking applications using RFID technology," in RFID Eurasia, 2007 1st Annual, 2007, pp. 1-3.

[4] V. W. S. Tang, Y. Zheng, and J. Cao, "An intelligent car park management system based on wireless sensor networks," in Pervasive Computing and Applications, 2006 1st International Symposium on, 2006, pp. 65-70

[5] J. P. Benson, T. O'Donovan, P. O'Sullivan, U. Roedig, C. Sreenan, J. Barton, et al., "Car-park management using wireless sensor networks," in Local Computer Networks, Proceedings 2006 31st IEEE Conference on, 2006, pp. 588-595.

[6] S. Funck, N. Mohler, and W. Oertel, "Determining car-park occupancy from single images," in Intelligent Vehicles Symposium, 2004 IEEE, 2004, pp. 325-328.

[7] R. Panayappan, J. M. Trivedi, A. Studer, and A. Perrig, "VANET-based approach for parking space availability," in Proceedings of the fourth ACM international workshop on Vehicular ad hoc networks, 2007, pp. 75-76.

[8] A. Kianpisheh, N. Mustaffa, P. Limtrairut, and P. Keikhosrokiani, "Smart Parking System (SPS) Architecture Using Ultrasonic Detector," 2012.

[9] S. Mathur, S. Kaul, M. Gruteser, and W. Trappe, "ParkNet: a mobile sensor network for harvesting real time vehicular parking information," Proceedings of the 2009 MobiHoc S, vol. 3, p. 3, 2009.

[10] N. Hanif, M. Badiozaman, and H. Daud, "Smart parking reservation system using short message services (SMS)," in Intelligent and Advanced Systems (ICIAS), 2010 International Conference on, 2010, pp. 1-5

[11] K. C. T. a. K. C. L. D.B.L. Bong, "Integrated Approach in the Design of Car Park Occupancy Information System (COINS)."

[12] H. L. Andreas Klappenecker , , Jennifer L. Welch, "Finding available parking spaces made easy," journal homepage: www.elsevier.com/locate/adhoc, 2012.

[13] S. Jian, K. S. Yang, and C. L. Lee, "Modular RFID parking management system based on existed gate system integration," WSEAS Transactions on Systems, vol. 7, pp. 706-716, 2008.

[14] S. Lee, D. Yoon, and A. Ghosh, "Intelligent parking lot application using wireless sensor networks," in Collaborative Technologies and Systems, 2008. CTS 2008. International Symposium on, 2008, pp. 48-57.

[15] T. Lowdermilk, User-centered design: a developer's guide to building user-friendly applications: " O'Reilly Media, Inc.", 2013.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 765-772 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Multilingual evaluation of voice disability index using pitch rate

Shuji Shinohara[*, 1], Yasuhiro Omiya[2], Mitsuteru Nakamura[1], Naoki Hagiwara[2], Masakazu Higuchi[1], Shunji Mitsuyoshi[1], Shinichi Tokuno[1]

[1]*Verbal Analysis of Pathophysiology Graduate School of Medicine, The University of Tokyo, 113-0033, Japan*

[2]*Research and Product Development, PST Inc., 231-0023, Japan*

A R T I C L E I N F O

Article history:
Received: 05 April, 2017
Accepted: 24 May, 2017
Online: 16 June, 2017

Keywords:
Pitch detection ratio
Voice disability
Read speech
Jitter
Shimmer
Harmonic to noise ratio
Multilingual evaluation

A B S T R A C T

We propose the use of the pitch rate of free-form speech recorded by smartphones as an index of voice disability. This research compares the effectiveness of pitch rate, jitter, shimmer, and harmonic-to-noise ratio (HNR) as indices of voice disability in English, German, and Japanese. Normally, the evaluation of these indices is performed using long-vowel sounds; however, this study included the recitation of a set passage, which is more similar to free-form speech. The results showed that for English, the jitter, shimmer, and HNR were very effective indices for long-vowel sounds, but the shimmer and HNR for read speech were considerably worse. Although the effectiveness of jitter as an index was maintained for read speech, the pitch rate was better in distinguishing between healthy individuals and patients with illnesses affecting their voice. The read speech results in German, Japanese, and English were similar, and the pitch rate showed the greatest efficiency for identification. Nevertheless, compared to English, the identification efficiency for the other two languages was lower.

1. Introduction

This paper is an extension of the work originally presented in IEEE Engineering in Medicine and Biology Society(EMBS) 2016 [1], with the main addition being the evaluation of German and Japanese samples, in addition to the previously evaluated English data.

In recent years, the widespread use of smartphones has created interest in pathological analysis that utilizes voice data [2-4]. Voice analysis by smartphones has the advantages of being non-invasive and requiring no specialized equipment; therefore, it can be performed easily and even remotely.

It is known that changes in voice characteristics due to various illnesses can be observed. For example, patients with Parkinson's disease often also suffer from dysarthria and stuttering. Additionally, patients with illnesses affecting the vocal cords or larynx may have hoarse voices.

To date, jitter, shimmer, and the harmonic-to-noise ratio (HNR) have been proposed as indices to describe the degree of voice disability [5–8]. Previous research on the use of these indices in the analysis of illness has shown that patients with depression pronounce vowels with higher shimmer and jitter values. Additionally, it has been shown that the frequencies of the first and second formants of depression patient speech are lower than those of healthy individuals [9]. In other research, the Lyapunov exponents and Kolmogrov entropy of the voices of depression patients were measured using chaos analysis methods [10]. Additionally, Zhou et al. have proposed new feature values derived from the Teager energy operator for the classification of voices under stress [11].

Research in the relationship between mental illness and voice has been conducted by analyzing the speaking rate [12-14], as well as the switching pause and percent pause [13, 15] of depression patients.

The objective of this study is to identify and evaluate characteristics that can be used to discriminate between voices of healthy individuals and those of disorder patients based on voice recordings of free-form speech.

[*]Shuji Shinohara, The University of Tokyo, 7-3-1 Hongo Bunkyuo-ku Tokyo 113-8655 Japan, +81-3-5800-9168 & shinokan99@m.u-tokyo.ac.jp

https://dx.doi.org/10.25046/aj020397

For this purpose, we focused on pitch degradation as an index representing sonic perturbations in patients' voices. As a measure of this feature, we devised the concept of the pitch rate to express the ease of pitch detection in voices.

2. Materials and Methods

2.1. Voice acquisition

For the voice evaluations, we used the Disordered Voice Database and Program, Model 4337, from KayPENTAX, a division of Pentax Medical Inc. This database contains recordings of the voices of approximately 700 persons (healthy individuals and patients), with two types of sounds from each individual: the long-vowel sound "ah" (lasting approximately 3 s for healthy individuals and approximately 1 s for patients) and a read speech from the "Rainbow passage" (approximately the first 12 s).

The voice recordings of read speech from 715 individuals contain samples from 53 healthy individuals and 662 patients. The voice recordings of long vowels from 711 persons include samples from 53 healthy individuals and 658 patients. The various disorders in the patients are (duplication observed): hyperfunction (288), paralysis (77), Anterior-Posterior squeezing (182), gastric reflux (54), vocal fold edema (46), and ventricular compression (109). The format used for recording had a sampling frequency of either 25 kHz or 50 kHz with 16-bit quantization.

Conventionally, long-vowel sounds are used in the evaluation of jitter, shimmer, and HNR.

Among the data of 658 patients, no jitter, shimmer, or HNR could be detected in 13 patients regarding the long-vowel sound "ah". Therefore, we used the data of 698 persons, including 53 healthy individuals and 645 patients.

The long-vowel sound "ah" does not change fundamentally between languages. In contrast, reciting a passage involves language differences based on the actual words used and pronunciation. In our research, we used the read speech of passages in Japanese and German, in addition to English, to investigate the language dependence of the proposed indices. Details regarding German and Japanese speech are provided below.

German speech was evaluated using the "Saarbruecken Voice Database" (http://stimmdb.coli.uni-saarland.de/) from the Institute of Phonetics of Saarland University. This voice database is available for free download, and it includes samples from over 2000 people. In our research, we used the samples of healthy individuals and patients with dysphonia saying "Guten morgen, wie geht esIhnen?" ("Good morning, how are you?") from this database. The number of voice files used was 632 from healthy individuals and 101 from patients with dysphonia. These samples had a sampling frequency of 50 kHz with 16-bit quantization.

Japanese speech was evaluated using voice samples from "Assessment of Motor Speech for Dysarthria" [16]. This CD contains voice samples from patients with dysarthria caused by various illnesses and from healthy individuals. All samples included recordings of readings from Aesop's fable "The North Wind and the Sun"; however, the samples from dysarthria patients also contained free-form speech. For this reason, the patient files were classified into a read-speech group and a free-form speech group. For consistency, only the read-speech samples of "The North Wind and the Sun" were used in the comparisons between patients and healthy individuals. Meanwhile, the free-form speech samples were compared to the read-speech samples of the same patients and evaluated to determine the presence of any differences between them. In this collection of samples, speech samples of the same patient were categorized and saved under multiple illnesses, so the duplicate voice files were erased. As a result, the total number of voice files from healthy individuals and patients was 6 and 35, respectively. These voice files had a sampling rate of 44.1 kHz with 16-bit quantization.

2.2. Evaluation method

Using the aforementioned data, we calculated the pitch rate, jitter, shimmer, and HNR, and compared the results of healthy individuals and patients. To evaluate the performance of each index in discriminating between patients and healthy individuals, we used the area under the curve (AUC) in the receiver-operating characteristic (ROC) plot, the sensitivity, and the specificity.

2.3. Evaluation index

We used the Praat software version 5.4.10 (http://www.fon.hum.uva.nl/praat/) to analyze the jitter, shimmer, and HNR. While shimmer measures the fluctuations in voice loudness, jitter measures the fluctuations in voice pitch and is affected by the tension and hardness of the vocal cords. The HNR is a measure of stability of the harmonic structure and indicates voice disorders, in particular, the degree of hoarseness. In this study, we used ppq5 for jitter, apq5 for shimmer, and harmonicity for HNR. In addition to these evaluation indices, we have introduced the pitch rate, which was devised by us, as an index for voice disability.

2.4. Pitch rate

Conventionally, jitter, shimmer, and HNR have been used as the measure representing the disturbance of the periodicity of speech waves. However, these cannot be measured with respect to the portion of the data in which the fundamental frequency cannot be detected. Furthermore, we focused on the ease of detection of the pitch rather than its value. The pitch detection ratio (pitch rate) is defined as the percentage of frames in which the fundamental frequency can be detected.

In normal voice recordings, there are silent intervals between utterances. In these intervals, the pitch cannot be detected. This means that the pitch rate is affected by the length of such intervals.

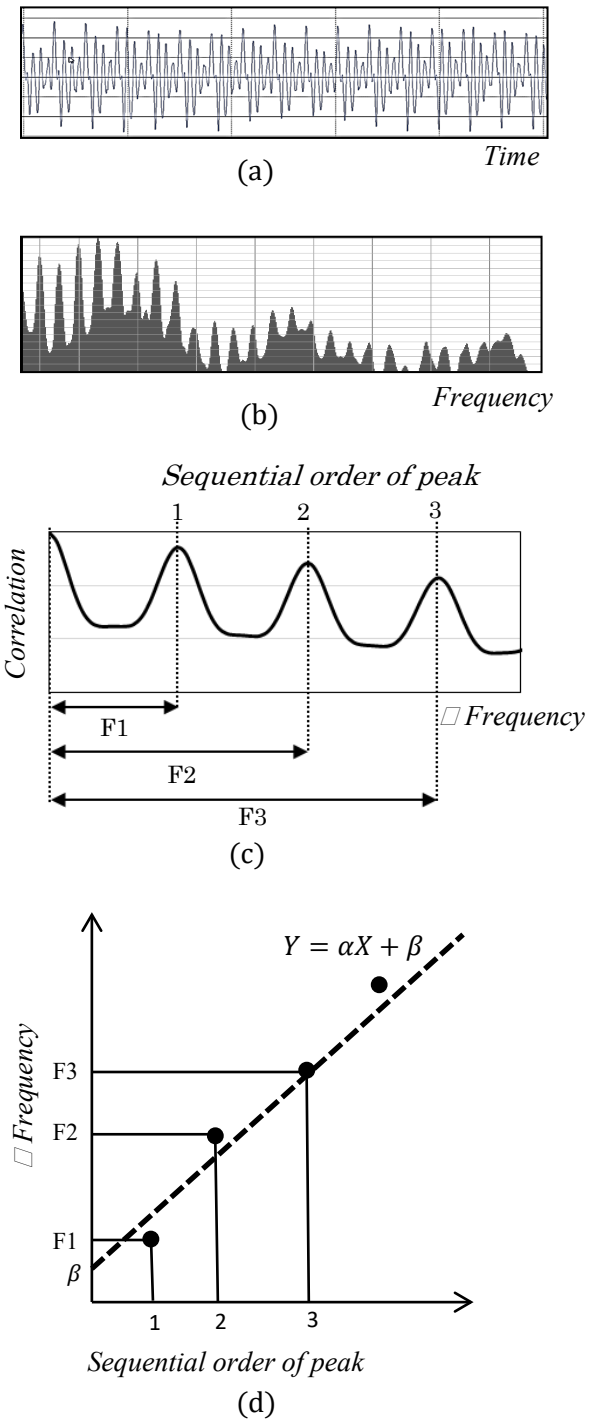


Figure 1. The pitch detection process. Details are listed in the text. (a) Voice signal, (b) frequency spectrum, (c) autocorrelation of the frequency spectrum, (d) plot of peak number and frequency shift exhibiting linear regression.

Therefore, to calculate the pitch rate, the data of these intervals must be excluded from the voice data. For this purpose, we used the Sensibility Technology software, version 3.0 (AGI Inc., Tokyo, Japan) [17-19].

After the speech interval data were extracted, the data were separated into frames of length L. However, each frame was offset from the previous frame by L/4 (i.e., 3/4 of each frame overlapped with the previous frame). We set L to equal 46.44 ms, which corresponds to 512 data points at a sampling rate of 11025 Hz.

The pitch was calculated for each frame using the methods described below [20, 21]. First, the voice signal waveform (Figure 1a) was processed by a fast Fourier transform (Figure 1b). Next, the Fourier spectrum autocorrelation was calculated, and the corresponding peaks were determined. Each peak was assigned a sequential number, and the change in frequency, ΔFrequency (i.e., F1, F1, F3…) was determined (Figure 1c). Lastly, the peak number was plotted on the X axis, the change in frequency was plotted on the Y axis, and a linear regression of the form $Y=\alpha X+\beta$ was determined (Figure 1d). If the linear regression had a coefficient of determination R larger than the threshold θ_1, and the absolute value of the y-intercept β was smaller than the threshold θ_2, then the pitch was considered as detected. Furthermore, if the pitch was detected, the slope of the linear regression α represented the pitch value.

$$pitch\ detection = \begin{cases} 1(detected)\ if\ R > \theta_1\ and\ |\beta| < \theta_2 \\ 0(not\ detected) \qquad otherwise \end{cases} \quad (1)$$

The pitch rate of every speech interval was determined by dividing the total number of frames by the number of frames in which the pitch was detected (which were subsequently analyzed). The frames that were analyzed were selected by the following methods. First, the value of the average power of the entire speech interval was determined. Next, the average power of each frame was calculated; if it was at least θ_3% higher than the average power in the entire speech interval, the frame was selected for analysis. In the research presented in IEEE EMBS 2016, we established a threshold for the absolute power was to select frames for analysis: if the average power of a frame exceeded this threshold, that frame was selected for analysis. However, in this study, we defined a relative value for this threshold. It must be noted that because of this change, the number of values provided in the Results section is different from that stated in IEEE EMBS 2016.

After the pitch rate was calculated for each speech interval, the average pitch rate of all speech intervals included in the voice files was determined; this average value was considered as the pitch rate for the corresponding voice files.

While jitter and shimmer are measures representing small disturbances of the periodicity within the range in which pitch can be detected, the pitch rate can be described as a measure representing a large disturbance of periodicity, such that the pitch cannot be detected.

If the intensity or the pitch varies continuously, the jitter and shimmer are affected. On the other hand, even if the intensity or the pitch fluctuates, the pitch rate is not affected, as long as the fluctuations are sufficiently small to detect the pitch. That is, the pitch rate is robust against fluctuations. Even if the amplitude, period, or the harmonic structure is greatly disturbed, the pitch rate diminishes. That is, the pitch rate can be described as a compositive index that included jitter, shimmer, and HNR. Therefore, it can be considered to be a rough indicator in comparison with conventional indicators.

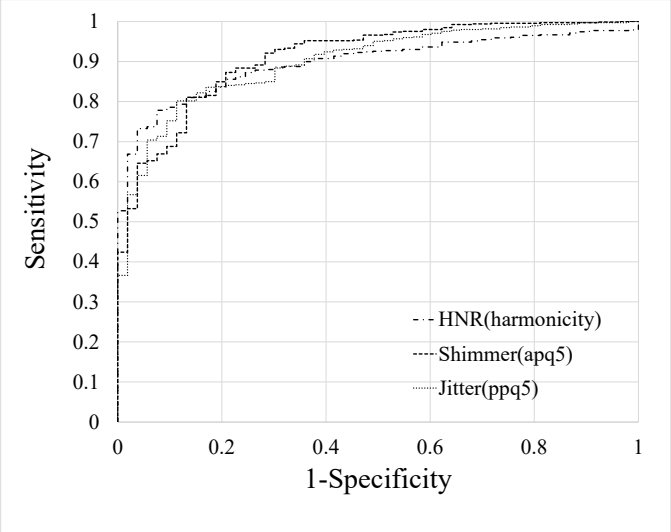


Figure 2 ROC curves used for discriminating between healthy individuals and patients using the jitter, shimmer, and HNR in the data of the long-vowel sound "ah".

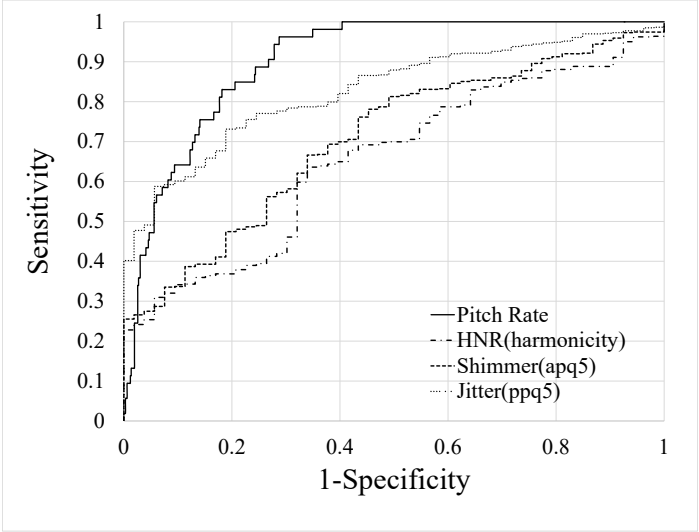


Figure 3 ROC curves used for discriminating between healthy individuals and patients using the pitch rate, jitter, shimmer, and HNR in the "Rainbow passage" read speech data.

The purpose of this study is to distinguish patients based on free-form speech rather than long-vowel sounds. Because more uncertainties are included in free-form speech compared with long vowels, we consider that a rough indicator is more adequate for free-form speech.

3. Results

3.1. Ability to discriminate based on long-vowel sound

TABLE 1. The mean values for healthy individuals and patient in the data of the long-vowel sound "ah"

Index	*Healthy Individuals (n=53)*	*Patients (n=645)*	*P-value*
Jitter (ppq5)	0.0016±0.00066	0.0061±0.0078	$1.23\cdot10^{-39}$
Shimmer (apq5)	0.012±0.0062	0.039±0.0293	$7.04\cdot10^{-55}$
HNR (harmonicity)	23.94±2.86	15.18±6.19	$4.76\cdot10^{-34}$

Table 1 shows the mean values for healthy individuals and patients in the data of the long vowel sound “ah”.

With respect to jitter, the mean values for healthy individuals and patients were 0.0016 (Number of Subjects (n) = 53, Standard Deviation (SD) = 0.00066) and 0.0061(n = 645, SD= 0.0078), respectively. Based on the t-test, significant differences were observed between the two groups ($t(695) = -14.05$, $p = 1.23\cdot10^{-39}$).

Regarding shimmer, the mean values for healthy individuals and patients were 0.012 (n = 53, SD = 0.0062) and 0.039 (n = 645, SD = 0.0293), respectively. Based on the t-test, significant differences were observed between the two groups ($t(329) = -19.01$, $p = 7.04\cdot10^{-55}$).

With respect to the HNR, the mean values for healthy individuals and patients were 23.94 (n = 53, SD = 2.86) and 15.18 (n = 645, SD = 6.19), respectively. Based on the t-test, significant differences were observed between the two groups ($t(97) = 18.77$, $p = 4.76\cdot10^{-34}$).

Figure 2 shows the ROC curves, which demonstrate the discrimination between healthy individuals and patients using these indices. The horizontal axis represents 1-specificity (false positive rate) and the vertical axis represents sensitivity (true positive rate).

Table 2 shows the AUC for the ROC curve, sensitivity, and specificity for these indices. The best discriminability was achieved by the shimmer in terms of AUC and sensitivity and by the HNR regarding specificity. For all the indices, the AUC was approximately 0.9, showing a strong discriminability between healthy individuals and patients.

TABLE 2. The discriminability of the indices (for the long-vowel sound "ah")

Index	*AUC*	*Sensitivity*	*Specificity*
Jitter (ppq5)	0.900	0.802	0.887
Shimmer (apq5)	0.911	0.811	0.868
HNR (harmonicity)	0.897	0.778	0.925

3.2. Ability to discriminate based on read speech in English

TABLE 3. The mean values for healthy individuals and patients in the data of read speech of "rainbow passage"

Index	*Healthy Individuals (n=53)*	*Patients (n=662)*	*P-value*
Pitch rate	0.903±0.047	0.643±0.25	$8.98\cdot10^{-39}$
Jitter (ppq5)	0.0084±0.0016	0.014±0.0076	$2.84\cdot10^{-39}$
Shimmer (apq5)	0.042±0.010	0.057±0.026	$4.55\cdot10^{-14}$
HNR (harmonicity)	13.12±2.18	11.41±3.48	$1.91\cdot10^{-6}$

Regarding the total speech time, the average values of healthy individuals and patients were 10089.23 ms (n = 53, SD = 734.33), and 9270.70 ms (n = 662, SD = 1394.42), respectively. As a result of the t-test, there was a significant difference between them (t(86) = 7.15, p = $2.71 \cdot 10^{-10}$).

Table 3 shows the mean values for healthy individuals and patients in the data of read speech of “Rainbow Passage”.

With respect to pitch rate, the mean values for healthy individuals and patients were 0.903 (n = 53, SD = 0.047) and 0.643 (n = 662, SD = 0.25), respectively. Based on the t-test, significant differences were observed between the two groups (t(394) = 22.285, p = $8.98 \cdot 10^{-39}$).

With respect to jitter, the mean values for healthy individuals and patients were 0.0084 (n = 53, SD = 0.0016) and 0.014 (n = 662, SD = 0.0076), respectively. Based on the t-test, significant differences were observed between the two groups (t(321) = −15.09, p = $2.84 \cdot 10^{-39}$).

With respect to shimmer, the mean values for healthy individuals and patients were 0.042 (n = 53, SD = 0.010) and 0.057 (n = 662, SD = 0.026), respectively. Based on the t-test, significant differences were observed between the two groups (t(117) = −8.58, p = $4.55 \cdot 10^{-14}$).

Regarding the HNR, the mean values for healthy individuals and patients were 13.12 (n = 53, SD=2.18) and 11.41(n=662, SD=3.48), respectively. Based on the t-test, significant differences were observed between the two groups (t(75) = 5.166, p = $1.91 \cdot 10^{-6}$).

TABLE 4. Discriminability of the indices (for read speech of "rainbow passage")

Index	*AUC*	*Sensitivity*	*Specificity*
Pitch rate	0.902	0.962	0.713
Jitter (ppq5)	0.827	0.588	0.943
Shimmer (apq5)	0.706	0.761	0.566
HNR (harmonicity)	0.655	0.636	0.660

Figure 3 shows the ROC curves, which demonstrate the discrimination between healthy individuals and patients using these indices.

Table 4 shows the AUC for the ROC curve, sensitivity, and specificity for these indices. The pitch rate had the best values for the AUC and sensitivity. However, the jitter had the best value for specificity. Compared to Table 1, the AUC for the shimmer and HNR was considerably reduced.

3. 3. Ability to discriminate based on read speech in German

Regarding the total speech time, the average values of healthy individuals and patients were 1561.81 ms (n = 632, SD = 331.24) and 1920.62 ms (n = 101, SD = 488.37), respectively. As a result of the t-test, there was a significant difference between them (t (115) = −7.13, p = $9.67 \cdot 10^{-11}$).

TABLE 5. The mean values for healthy individuals and patients in the data of read speech of "Guten Morgen, wie geht es Ihnen?"

Index	*Healthy Individuals (n=632)*	*Patients (n=101)*	*P-value*
Pitch rate	0.826±0.094	0.701±0.180	$6.55 \cdot 10^{-10}$
Jitter (ppq5)	0.0101±0.0028	0.0122±0.0056	0.00039
Shimmer (apq5)	0.036±0.012	0.045±0.019	$4.48 \cdot 10^{-5}$
HNR (harmonicity)	14.89±2.29	14.47±3.27	0.22

Table 5 shows the mean values for healthy individuals and patients in the data of read speech of “Guten Morgen, wie geht es Ihnen?”.

With respect to the pitch rate, the mean values for healthy individuals and patients were 0.826 (n = 632, SD = 0.094) and 0.701 (n = 101, SD = 0.180), respectively. Based on the t-test, significant differences were observed between the two groups

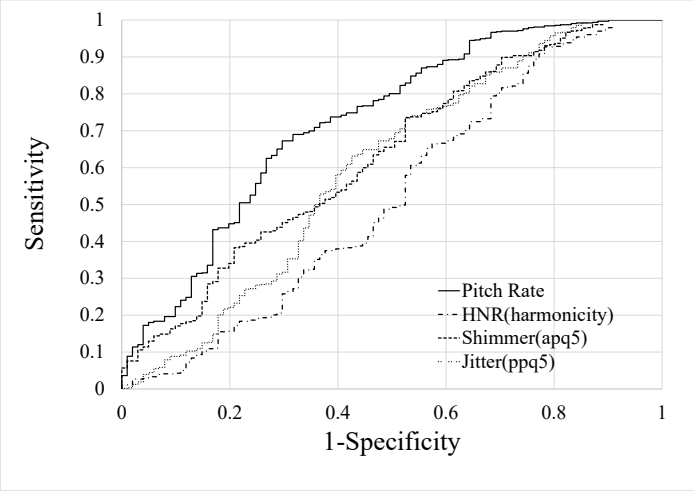


Figure 4 ROC curves used for discriminating between healthy individuals and patients using the pitch rate, jitter, shimmer, and HNR in data containing read speech of "Guten Morgen, wie geht es Ihnen?"(German).

(t(109) = 6.78, p = $6.55 \cdot 10^{-10}$).

With respect to jitter, the mean values for healthy individuals and patients were 0.0101 (n = 632, SD = 0.0028) and 0.0122 (n = 101, SD = 0.0056), respectively. Based on the t-test, significant differences were observed between the two groups (t(632) =−3.66, p = 0.00039).

Regarding shimmer, the mean values for healthy individuals and patients were 0.036 (n = 632, SD = 0.012) and 0.045 (n = 101, SD = 0.019), respectively. Based on the t-test, significant differences were observed between the two groups (t(112) = −4.25, p = $4.48 \cdot 10^{-5}$).

With respect to the HNR, the mean values for healthy individuals and patients were 14.89 (n = 632, SD = 2.29) and 14.47 (n = 101, SD = 3.27), respectively. Based on the t-test, no significant differences were observed between the two groups (t(116) = 1.230, p = 0.22).

As shown above, in all indexes other than the HNR, there were significant differences between healthy individuals and patients.

Figure 4 shows the ROC curves, which demonstrate the discrimination between healthy individuals and patients using these indices.

TABLE 6. Discriminability of the indices (for read speech of "'Guten Morgen, wie geht es Ihnen?'")

Index	*AUC*	*Sensitivity*	*Specificity*
Pitch rate	0.725	0.672	0.703
Jitter (ppq5)	0.598	0.734	0.475
Shimmer (apq5)	0.628	0.736	0.475
HNR (harmonicity)	0.519	0.927	0.218

Table 6 shows the AUC for the ROC curve, sensitivity, and specificity for these indices. For the AUC and specificity, the pitch rate was the most suitable. In terms of sensitivity, the HNR showed the best value. However, the value of the AUC tended to be lower than in the case of English.

3.4. Ability to discriminate based on read speech in Japanese

Regarding the total speech time, the average values of healthy individuals and patients were 26144.17 ms (n = 6, SD = 2995.73) and 18676.11 ms (n = 35, SD = 5145.64), respectively. As a result of the t-test, there was a significant difference between them (t(39) = 3.43, p = 0.0014).

TABLE 7. The mean values for healthy individuals and patients in the data of read speech of " Kitakaze to Taiyo "

Index	*Healthy Individuals (n=6)*	*Patients (n=35)*	*P-value*
Pitch rate	0.831±0.051	0.668±0.232	0.00078
Jitter (ppq5)	0.011±0.0026	0.013±0.0084	0.12
Shimmer (apq5)	0.040±0.014	0.0625±0.035	0.013
HNR (harmonicity)	13.64±2.46	11.85±3.59	0.25

Table 7 shows the mean values for healthy individuals and patients in the data of read speech of "Guten Morgen, wie geht es Ihnen?".

Regarding pitch rate, the mean values for healthy individuals and patients were 0.831 (n = 6, SD = 0.051) and 0.668 (n = 35, SD = 0.232), respectively. Based on the t-test, significant differences were observed between the two groups (t(36) = 3.67, p = 0.00078).

With respect to the jitter, the mean values for healthy individuals and patients were 0.011 (n = 6, SD = 0.0026) and 0.013 (n = 35, SD = 0.0084), respectively. Based on the t-test, no significant differences were observed between the two groups (t(26) = −1.61, p = 0.12).

Regarding shimmer, the mean values for healthy individuals and patients were 0.040 (n = 6, SD = 0.014) and 0.062 (n = 35, SD = 0.035), respectively. Based on the t-test, no significant differences were observed between the two groups (t(19) = −2.76, p = 0.013).

With respect to the HNR, the mean values for healthy individuals and patients were 13.64 (n = 6, SD = 2.46) and 11.85 (n = 35, SD = 3.59), respectively. Based on the t-test, no significant differences were observed between the two groups (t(39) = 1.17, p = 0.25).

As shown above, there was a significant difference between healthy individuals and patients only in the pitch rate.

Figure 5 shows the ROC curves, which display the discrimination between healthy individuals and patients using these indices.

TABLE 8. Discriminability of the indices (for read speech of "Kitakaze to Taiyo")

Index	*AUC*	*Sensitivity*	*Specificity*
Pitch rate	0.752	0.833	0.686
Jitter (ppq5)	0.562	1	0.343
Shimmer (apq5)	0.733	1	0.486
HNR (harmonicity)	0.624	1	0.314

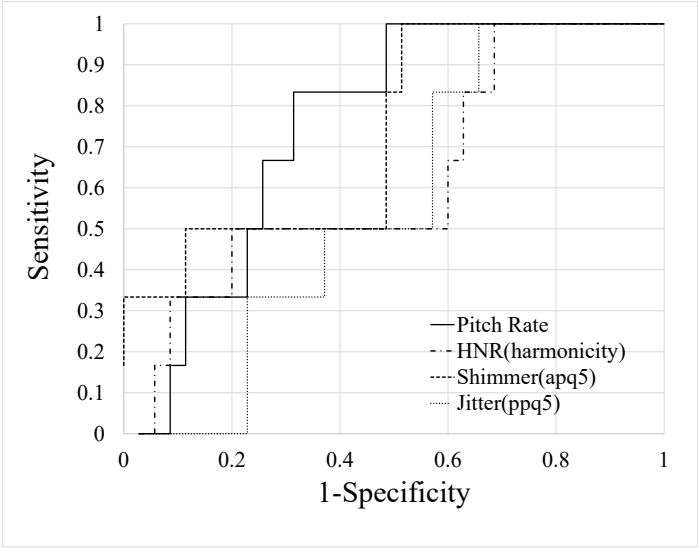


Figure 5 ROC curves used for discriminating between healthy individuals and patients using the pitch rate, jitter, shimmer, and HNR in the data containing read speech of "Kitakaze to taiyo" (Japanese).

Table 8 shows the AUC for the ROC curve, sensitivity, and specificity for these indices. For the AUC and specificity, the pitch rate was the most suitable. The sensitivity was 1 for all indices other than the pitch detection rate. As with German, the value of the AUC tended to be lower than in English.

Patient speech in Japanese included both read speech and free-form speech; these were separated manually. The differences between free-form speech and the recitation of a passage were investigated using these voice samples.

With respect to jitter, the mean values for reading and free-form speech were 0.013 (n = 35, SD = 0.007) and 0.013 (n = 35, SD = 0.008), respectively. Based on the t-test, no significant differences were observed between the two groups (t(68) = −0.304, p = 0.762). The correlation coefficient between them was 0.88.

With respect to shimmer, the mean values for reading and free-form speech were 0.059 (n = 35, SD = 0.030) and 0.067 (n = 35, SD = 0.034), respectively. Based on the t-test, no significant differences were observed between the two groups (t(68) = −0.95, p = 0.348). The correlation coefficient was 0.88.

With respect to the HNR, the mean values for reading and free-form speech were 12.30 (n = 35, SD = 3.23) and 11.65 (n = 35, SD = 3.56), respectively. Based on the t-test, no significant differences were observed between the two groups ($t(68) = -0.806$, $p = 0.423$). The correlation coefficient was 0.85.

With respect to pitch rate the mean values for reading and free-form speech were 0.668 (n = 35, SD = 0.232) and 0.582 (n = 35, SD = 0.277), respectively. Based on the t-test, no significant differences were observed between the two groups ($t(68) = 1.41$, $p = 0.162$). The correlation coefficient was 0.79.

4. Discussion and Conclusion

In this article, we propose a method for predicting whether a voice sample belongs to a healthy individual or a patient based on the pitch rate of a speech interval in English, German, and Japanese speech. Similar studies conducting comparative analyses of healthy individuals and patients have been performed using the same samples [22, 23]. However, those studies utilized voice samples of the long-vowel sound "ah." Instead, our research focused on the detection of illness via a smartphone using free-form speech in the form of reciting a set passage.

As shown in Table 2, in the case of the long-vowel sound, the jitter, shimmer, and HNR all resulted in an AUC of approximately 0.9, showing that these indices have good discriminability. However, as shown in Table 4, in the case of read speech, the discriminability diminished. In particular, for the shimmer and HNR, the AUC decreased considerably, to approximately 0.7 or lower.

In contrast, in the case of read speech, the pitch rate resulted in an AUC exceeding 0.9, demonstrating better discriminability between healthy individuals and patients than the conventional indices. However, in the case of the long-vowel sound, the pitch rate for both healthy individuals and patients was approximately 1.0, making it unsuitable for discrimination between healthy individuals and patients.

As stated above, the AUC of the pitch rate for the English passage was 0.902, which was a favorable outcome. In German and Japanese, the pitch rate showed the highest identification efficiency, as was the case in English. However, the AUC values were 0.725 and 0.752 for German and Japanese, respectively, which is more than 0.15 lower that the AUC value for English. In English, the average pitch rate values for healthy individuals and for patients were 0.903 and 0.643, respectively. The corresponding values in German were 0.826 and 0.701 and in Japanese 0.831 and 0.668, respectively. These results demonstrate that the healthy individual pitch rates in German and Japanese are lower than the healthy individual pitch rates in English and the patient pitch rates in German and Japanese are higher than the patient pitch rates in English. Differences in patient pitch rates may be due to differences in the severity of illnesses among the patients in the databases.

However, the question of why healthy individuals reciting in English have a higher pitch rate than in the other two languages remains. This could be owing to the characteristics of the languages, and the differences in the contents of the recited passages could also be a factor. Furthermore, the different sampling frequency used for each database and differences in the noise from the recording environment could be factors contributing to the pitch rate discrepancy.

Regarding the difference in the speech contents, in German, for example, the passage used was a greeting used in daily life, which is close to free-form speech. The influence of these differences was examined in Japanese patient speech through a comparison of free-form speech and read speech. The results showed that all indices—jitter, shimmer, HNR, and pitch rate—had a strong correlation of about 0.75 to 0.85 between free-form speech and read speech. In addition, there was a significant difference between free-form speech and read speech only in the pitch detection rate. The average value of the pitch detection rate was 0.582 in the case of free-form speech and 0.668 in case of read speech, and the free-form speech tends to be lower. Perhaps the German pitch detection rate is lower than that of English because the content of the speech is close to free conversation.

Additionally, the average total speech time for the German passage used was 1561.81 ms for healthy individuals, which is extremely short compared to the English passage (10089.23 ms) and the Japanese passage (26144.17 ms). However, because the average pitch rate value was calculated for each utterance, the difference in the total length of the voice samples is not expected to have an effect on the results.

The thresholds θ_1, θ_2, and θ_3, used for the pitch rate calculation algorithm described in section 2.4 had the same value for every language. In a separate pilot study, it was determined that if the thresholds are set at an appropriate value for each separate language, the AUC for German and Japanese exceeds 0.8. Therefore, these thresholds may need to be tuned separately for each language. These points will be further evaluated in future studies.

Although we have demonstrated that for read speech, the pitch rate has high discriminability, it must be noted that read speech and everyday conversation in telephone calls are not equivalent. In the read speech data used here, both the healthy individuals and the patients read the same sentences. However, in everyday conversation, the content is diverse. For example, the percentage of consonants varies in each case and the pitch rate decreases when there are more consonants. It has been reported [24] that compared with read speech, spontaneous speech can better classify patients with major depressive disorders. We believe that further studies are required on the effects of the speech content on the pitch rate or the differences between read speech and spontaneous speech.

Although long vowels can be recorded by using a smartphone, it is necessary that day-to-day recording is performed regularly. It has been reported [25] that the frequency of voice recording in such conditions gradually decreases compared with telephone calls.

In this study, we used the algorithm proposed in [20, 21]. The pitch rate itself is naturally affected by the performance of the algorithm. However, if the algorithms are used to process the same pitch, although there may be variations in the absolute pitch rate values, we believe that the relative magnitude relationship between healthy individuals and patients will not change. This means that it is likely that differences in the algorithm do not adversely affect the discriminability between healthy individuals and patients. This is also a topic for future study.

In this study, each analysis was performed for well-recorded speech signals using a PC processor. However, when analyzing real speech recorded in real conditions using smartphones, the analysis might be influenced by the audio transducer or processor of the smartphone. These points are under verification.

The voice database used in the study contained data from patients with diverse disorders. It can therefore be argued that although we have suggested an index to broadly discriminate between healthy individuals and patients, this index cannot be applied to diagnose a specific disorder. In the future, we aim to develop algorithms for diagnosing specific disorders.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This research is (partially) supported by the Center of Innovation Program from the Japan Science and Technology Agency, JST. This work was supported by JSPS KAKENHI Grant Numbers JP16K01408 and JP15H03002.

References

[1] S. Shinohara, M. Nakamura, S. Mitsuyoshi, S. Tokuno, Y. Omiya and N. Hagiwara. Voice disability index using pitch rate." In Proceedings of IEEE EMBS, 557–560, 2016.

[2] S. Arora, V. Venkataraman, A. Zhan, S. Donohue, K. M. Biglan, E. R. Dorsey, and M. A. Little, "Detecting and monitoring the symptoms of Parkinson's disease using smartphones: A pilot study" Parkinsonism Relat. D. **21**(6), 650–653, 2015.

[3] K. K. Rachuri, M. Musolesi, C. Mascolo, P. J. Rentfrow, C. Longworth, and A. Aucinas, "EmotionSense: a mobile phones based adaptive platform for experimental social psychology research," In Proceedings of the 2012 ACM Conference on Ubiquitous Computing, 650–653, 2010.

[4] H. Lu, M. Rabbi, G. T. Chittaranjan, D. Frauendorfer, M. S. Mast, A. T. Campbell, D. Gatica-Perez, and T. Choudhury, "Stresssense: Detecting stress in unconstrained acoustic environments using smartphones," In Proceedings of the 2012 ACM Conference on Ubiquitous Computing, 351–360, 2012.

[5] O. Amir, N. Amir, and O. Michaeli, "Evaluating the influence of warmup on singing voice quality using acoustic measures" J. Voice, **19**(2), 252–260, 2005.

[6] J. P. Teixeira, C. Oliveira, and C. Lopes, "Vocal Acoustic Analysis - Jitter, Shimmer and HNR Parameters" Procedia Technology, **9**. 1112–1122, 2013.

[7] E. Yumoto, W. J. Gould, and T. Baer, "Harmonics-to-noise ratio as an index of the degree of hoarseness" J. Acoust. Soc. Am., **71**(6), 1544–1549, 1982.

[8] M. Farrús, J. Hernando, and P. Ejarque, "Jitter and Shimmer Measurements for Speaker Recognition," In Proceedings of the International Conference Interspeech 2007, 778–781, 2007.

[9] K. Vicsi, and D. Sztaho, "Examination of the sensitivity of acoustic-phonetic parameters of speech to depression," In IEEE 3rd International Conference on Cognitive Infocommunications, 511–515, 2012.

[10] T. Shimizu, N. Furuse, T. Yamazaki, Y. Ueta, T. Sato, and S. Nagata, "Chaos of vowel /a/ in Japanese patients with depression: A preliminary study" J. Occup. Health, **47**(3), 267–269, 2005.

[11] G. Zhou, J. H. L. Hansen, and J. F. Kaiser, "Nonlinear feature based classification of speech under stress," IEEE Trans. on Speech and Audio Processing, **9**(3), 201–216, 2001.

[12] M. Cannizzaro, B. Harel, N. Reilly, P. Chappell, P.J. Snyder, "Voice acoustical measurement of the severity of major depression" Brain and Cognition. **56**, 30–35, 2004.

[13] J.C. Mundt, P.J. Snyder, M.S. Cannizzaro, K. Chappie, D.S. Geralts, "Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology." J. Neurolinguistics. **20**(1), 50–64, 2007.

[14] E. II. Moore, M. Clements, J. Peifert, L. Weisser, "Analysis of prosodic variation in speech for clinical depression" In Proceedings of IEEE EMBS, 3, 2925–2928, 2003.

[15] Y. Yang, C. Fairbairn, J.F Cohn, "Detecting Depression Severity from Vocal Prosody" IEEE Transactions on Affective Computing. **4**(2), 142–150, 2013.

[16] M. Nishio, "Assessment of Motor Speech for Dysarthria" (ISBN 978-4-900637-38-2), Interuna Publishers, Inc., 2009. (in Japanese)

[17] S. Mitsuyoshi, F. Ren, Y. Tanaka and S. Kuroiwa, "Non-verbal voice emotion analysis system" Int. J. Innovative Computing, Information and Control, **2**(4), 819–830, 2006.

[18] S. Mitsuyoshi, Y. Tanaka, R. Ren, K. Shibasaki, M. Kato, T. Murata, T. Minami, and H. Yagura. "Emotion voice analysis system connected to the human brain." In IEEE NLP-KE 2007, 479–484, 2007.

[19] S. Mitsuyoshi, "Emotion recognizing method, sensibility creating method, device, and software" WO0223524, Mar. 21, 2002.

[20] S. Mitsuyoshi, "Research on the phonetic recognition of feelings and a system for emotional physiological brain signal analysis," PhD thesis, Tokushima University, Tokushima, Japan, 2006.

[21] S. Mitsuyoshi, K. Ogata and F. Monma, "Speech analyzer detecting pitch frequency, speech analyzing method, and speech analyzing program" WO2006132159, Dec. 14, 2002.

[22] A. Al-nasheri, G. Muhammad, M. Alsulaiman, Z. Ali, T. A. Mesallam, M. Farahat, K. H. Malki, and M. A. Bencherif. "An Investigation of Multidimensional Voice Program Parameters in Three Different Databases for Voice Pathology Detection and Classification." J. Voice, **31**(1), pp. 113.e9–113.e18, 2017.

[23] A. Al-nasheri, G. Muhammad, M. Alsulaiman, and Z Ali. "Investigation of Voice Pathology Detection and Classification on Different Frequency Regions Using Correlation Functions." J. Voice, **31**(1), 3–15, 2017.

[24] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Breakspear, and G. Parker, "Detecting depression: A Comparison between spontaneous and read speech," In Proceedings of 2013 IEEE International Conference on ICASSP, 7547–7551, 2013.

[25] N. Hagiwara, Y. Omiya, S. Shinohara, M. Nakamura, Y. Kogure, S. Mitsuyoshi, and S. Tokuno, "Verification of effects on the mental health monitoring system by voice using smartphone caused by the difference of the recording methods," In proceedings of symposium of Japanese Society for Medical and Biological Engineering, 147, 2016. (in japanese)

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 773-787 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Efficacy of multiple anticancer therapies may depend on host immune response

Kritika Karri[1, 2*], Dhundy R Bastola[1]

[1]*University of Nebraska at Omaha, College of Information Science and Technology, 68182, USA*

[2]*Boston University, Bioinformatics Graduate Program, 02215, USA*

ARTICLE INFO

Article history:
Received: 05 April, 2017
Accepted: 16 May, 2017
Online: 16 June, 2017

Keywords:
Biomarker
Immune response
Progressive Disease
Complete Remission
Lung adenocarcinoma
Colon adenocarcinoma
Alkylating agents
Antimetabolite
Mitotic inhibitors
Anti-angiogenesis
GSEA

ABSTRACT

The host immune system is a key player in anticancer therapy response and resistance. Although the impact of host immune response in the 'war against cancer' has been studied and it has been the basis for immunotherapy, understanding of its role in attenuating the action of conventional anticancer therapies is an area that has not been fully explored. In spite of advances in systemic therapy, the 5-year survival rate for adenocarcinoma is still a mere 13% and the primary reason for treatment failure is believed to be due to acquired resistance to therapy. Hence, there is a need for identifying reliable biomarkers for guided treatment of lung and colon adenocarcinoma and to better predict the outcomes of specific anticancer therapies. In this work, gene expression data were analyzed using public resources and this study shows how host immune competence underscores the efficacy of various anticancer therapies. Additionally, the result provides insight on the regulation of certain biochemical pathways relating to the immune system, and suggests that smart chemotherapeutic intervention strategies could be based on a patient's immune profile.

1. Introduction

Cancer immunology has been studied and researched over many years to unravel the key mechanisms that determine therapeutic synergy or antagonism for cancer patients based on individual host immune environments. Traditional chemotherapy and other targeted therapies have been effective in combating many types of cancer. However, the efficacy of these interventions is determined by off-target effects within the host immune system [1]. Review of the literature shows that host immune response is key to the success of anticancer therapies through a process known as immune-surveillance [2].

Studies have shown that type-1 interferon can convert tumor associated macrophages (TAMs) into tumor-antagonizing macrophages by up-regulating the expression of dendritic cells in breast cancer. This results in increased CD8+ T cell population with reduced tumor growth and metastasis. As such, the combination of chemotherapy together with drugs that repolarize TAMs may be exploited to achieve greater patient responses and prevent resistance mechanisms within the immune system [3]. Some chemotherapeutics skew the polarization of macrophages directly, or indirectly via regulating cancer cell secreted factors [4]. FOXP3 expressing regulatory T cells (Tregs) are well-known for their immunosuppressive function in tumors [5]. They are a subset of CD4+ T cell population, which are involved with adaptive immune response. Tregs are known to be sensitive to cyclophosphamide and their depletion improves the potency of platinum-containing chemotherapeutic drugs against lung adenocarcinoma [6].

Another study has shown that Interleukin-17 (IL-17) producing CD4+ T cells limit the ability of 5-Flouro Uracil (5-FU) to delay the growth of subcutaneous thymoma cells. In contrast, IL-17 elevates the therapeutic efficacy of doxorubicin in sarcoma. Hence, the mode of action of these CD4+ T cell is paradoxical and depends widely on the cancer type [6]. However, their manipulation may be a useful strategy to determine the efficacy of chemotherapy.

*Kritika Karri, College of Information Science and Technology, University of Nebraska at Omaha, 68182, USA, Email: kkarri@unomaha.edu

Host immune response is known to significantly regulate the response of anti-angiogenic therapies [7]. Studies suggests that TAMs counteract the potency of anti-angiogenic agents; one such example is anti-angiopoietin 2 (ANGPT2), which decreases the blood vessel density of tumor and promote its remission. Interestingly, it also prevents the TIE2-expressing macrophages from associating with endothelial cells [8]. Together, these findings support the rationale for combining anti-angiogenic drugs with macrophage-targeting strategies to increase the efficacy of the former, particularly in tumors that develop resistance to anti-VEGF therapy. Despite the scientific progress in the field, a complete understanding of the role of the host immune system's response in improving the efficacy of anticancer therapies is greatly needed. There has been no systematic analysis carried out to evaluate the efficacy of different anticancer treatments based on host immune profile and subsequent application of this knowledge to develop reliable biomarkers to predict how well patients will respond to a treatment interventions.

Lung cancer continues to be the leading cause of death in industrialized countries and most commonly occurring disease among smokers. It can be broadly categorized into small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC). About 80-89% of patients suffer from the latter with most common being lung adenocarcinoma (LUAD), which has the average survival rate of 12-15% [9]. It develops on the outer parts of the lung or 'in-situ' and tends to grow slower than other types of cancer resulting in late-stage detection and a paucity of late stage treatments.

Colon adenocarcinoma (COAD) is the third most common cause of cancer mortality worldwide and fourth in the United States. Even though the 5-year survival rates have improved over the last 3 decades with ~95% in stage-I, it still remains low for stage III (~61%) and reaches only 8% for patients in stage IV. The treatment regimen involves mainly two classes of chemotherapy-alkylating agents and antimetabolites. However, the therapy is often insufficiently effective, with some patients exhibiting strong chemo-resistance. Therefore, host immune-competency plays a vital role to promote immunosurvelliance and disease regression [10].

Chemotherapy is one of the most common interventions for cancer care. It uses certain drugs to kill cancer cells or to stop them from growing and spreading to other parts of the body. This therapy works by damaging the DNA inside the nucleus of the transformed cells. While some drugs damage cells at the point of cell division, others damage the cells during DNA replication prior to division. Based on these differences in the mechanism of action, the chemo drugs can be classified into three main therapy types [11]:

(i) Alkylating agents: This class of drug directly targets DNA to repress a cell from reproducing. These drugs are not cell cycle specific and can be used for wide range of cancers.

(ii) Antimetabolites: This class of drug interferes with DNA and RNA synthesis by substituting for the normal building blocks of nucleotides. These agents damage cells during the S phase of cell cycle, when the cell's chromosomes are being copied.

(iii) Mitotic inhibitors: This class of drug is primarily derived from plant alkaloids and other natural products, and they operate by stopping mitosis in the M phase of the cell cycle by disrupting microtubule formation.

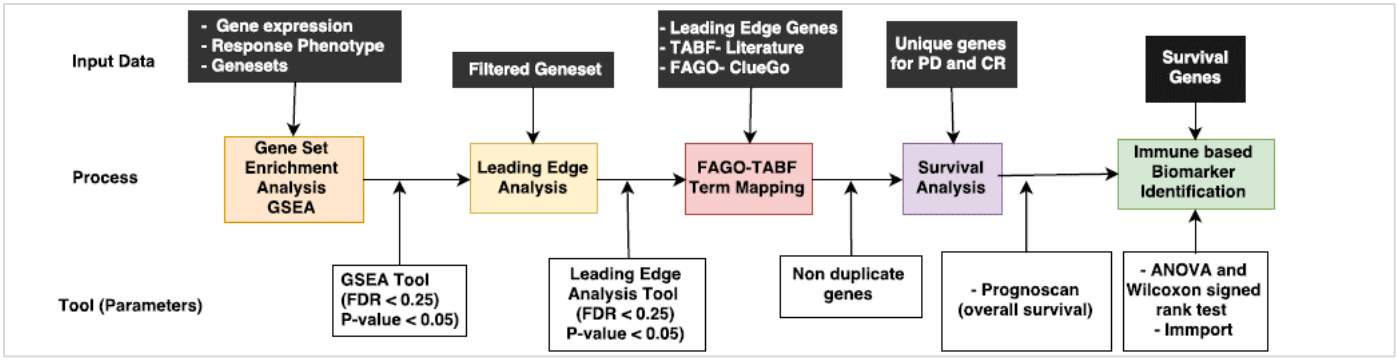
(iv) Anti-Angiogenesis: This class of drug is most widely used in the domain of vascular-mediated therapy [12], which is another class of treatment strategy that involves the impairment of tumor nutrition by targeting tumor blood vessel.

The single gene analysis method has been instrumental in our understanding of cell-biological processes, especially in elucidating the correlation between mutation and cancer. However, with increased understanding of the disease etiology, we have come to realize that it is not usually a single gene but a set that contribute to the clinical manifestation of cancer. Hence, it is more relevant to study the changes initiated by a set of genes, which can dramatically alter various cell biological and metabolic pathways. Along this vein the commonly used approaches to analyze a geneset are by over representation or aggregate score calculation. The latter is more reliable as it calculates the aggregate score for each geneset based on the gene-specific scores for that geneset and overcomes the limitations of over-representation analysis, which relies on the cutoff used in generating the geneset and can vary considerably depending on the gene list [13].

High throughput data like RNA-seq and microarray have fueled knowledge-based approaches like Gene Set Enrichment Analysis (GSEA), with access to publically available datasets for gene expression and clinical outcomes from TCGA. GSEA is widely used in studies that compare and contrast two conditions (for example disease vs. normal states) and usually focuses on a handful of genes that are either up or down regulated in the process. Therefore, in the present study we used GSEA to extract significant information at the level of functionally related geneset between the two clinical conditions (Fig 1.).

While GSEA provides a 'bird's eye view' of interactions between the drug treatment and the geneset significantly overrepresented in the phenotype being compared, not all the members in the geneset contribute equally to attain significant enrichment. As described by Subramanian et al., there are leading edge subsets of genes within the set that appear in the ranked-list before the point at which the running sum reaches its maximum deviation from zero [14]. These sets of genes are called 'leading edge genes' as they contribute more to the enrichment score of a geneset during geneset enrichment analysis. A gene that is in many of the leading edge subsets is more likely to be of higher significance or interest compared to other genes. called 'leading edge genes' as they contribute more to the enrichment score of a geneset during geneset enrichment analysis. A gene that is in many of the leading edge subsets is more likely to be of higher significance or interest compared to other genes.

In the current study, we used publically available datasets from TCGA for lung and colon adenocarcinoma samples along with the therapy regimen and response type - *Progressive Disease* (PD) (ineffective therapy with continued growth and spreading of cancer), and *Complete Remission* (CR) (state of disappearance of all signs of cancer in response to a treatment regimen) - to come up with putative biomarkers specific for both cancer types and different anticancer therapies using GSEA. Finally, to determine the prognostic ability of these biomarkers, we performed survival analysis.



2. Material and Methods

2.1. Data Access and Processing

The Cancer Genome Atlas (TCGA) public resource was used to download RNAseq gene expression datasets (level 3) for lung adenocarcinoma (LUAD) and colon adenocarcinoma (COAD). The expression set was quantified using RSEM method (RNA-seq by expectation maximization). Additionally, the genesets were obtained from MsigDB (http://software. broadinstitute.org/gsea/msigdb), which was explicitly integrated with the GSEA tool [15]. Five geneset categories, namely, *hallmark, canonical pathways (CP), GO genesets (C5), oncogenic signatures (C6) and immunologic signatures (C7)* were selected from the MsigDB collection, which included genes involved in cancer, immunology and biochemical pathways. The information relating to therapeutic interventions including drug name and response type for these patients was obtained from Nationwide Children's Hospital, which serves as the Biospecimen Core Resource (BCR) for TCGA.

The clinical data for lung adenocarcinoma consisted of four therapies discussed in the previous section, whereas alkylating agents and antimetabolites were the most preferred intervention for colon adenocarcinoma [16]. The drug information and patient response were then mapped with gene expression data using the TCGA ID as the primary identifier as shown in Table 1.

Table 1. Four therapy-classes were identified based on the information obtained from BCR for TCGA- LUAD tumor samples. Each named-drug used on the patients are listed under respective therapy-class.

Alkylating Therapy	Mitotic Inhibitor	Antimetabolite	Anti-angiogenesis
Cholrambucil, carmustine, dacarbazine, temozolomide, altretamine, cisplatin, carboplatin, oxalaplatin	Taxol, Ixabepilone, vinblastine, vinorelbine	5- Flourouracil, capecitabine, floxuridine, gemcitabine	Avastin, sunitnib

2.2 GSEA

2.3 Leading Edge Genes

We performed leading edge analysis on the enrichment results from GSEA, to identify genes that were significantly enriched across the PD vs. CR in LUAD and COAD. The leading edge analysis was carried out for genesets that were significantly enriched in the GSEA (FDR < 0.25, p-value < 0.05) across different geneset categories (C5, CP, hallmark, C7 and C6) and for all anticancer therapies. The leading edge genes obtained from the analysis were merged for each geneset category into the PD upregulated and CR upregulated groups.

2.4 Functional Association of Leading Edge Genes with GO terms (FAGO)

In order to extract functional annotation for the leading edge genes we used ClueGo, a Cytoscape plugin [17] that gives biological interpretation and systematic organization for a large list of genes. The list of leading edges genes for each anticancer therapy (for both PD and CR separately) was provided to ClueGo to obtain a list of gene-clusters that are functionally associated with a specific GO functional categories including biological processes (BP), cellular component (CC), molecular function (MF) and immune based terms. From a large list of GO terms, the terms that had significant p-value and high percentage term association were extracted and labeled as *'functionally associated GO terms'* (FAGO). There were several repetitive terms which performed similar function. Merging of all such FAGO terms helped us in downstream analysis.

2.5 Therapy Associated Biological Functions (TABF)

Each anticancer therapy has a different mechanism of action to stop the cancer cells from growing; each mechanism evokes perturbations in a complex system with several immunological processes and biological pathways involved. To evaluate the efficacy of each intervention, a comprehensive understanding of the mode of drug-action and the processes/pathways affected is necessary. Therefore, we searched the literature for concepts related to action mechanism in each drug category which were known to affect the efficacy of the therapy. These concepts/terms were labeled as '*therapy associated biological functions'* (TABF).

2.6 Mapping of TABF with FAGO terms

To identify relevant biological function affected by a therapy, we integrated TABF term for each therapy with related FAGO terms. While TABF term provided experimental evidence through the literature, FAGO defined concepts used to describe gene function. For instance, STAT3 signaling is known to be involved with tumorigenesis in all chemotherapies [18]. Here *'STAT3 Signaling'* is TABF and it is mapped with all the related GO terms (FAGO terms) present in PD.

2.7 Extraction of Unique Genes and Survival Analysis

The genes associated with mapped FAGO–TABF terms were retrieved for both disease outcome phenotype (PD and CR) as well as for all therapy groups. The duplicate gene entries were discarded and the list of genes from each therapy were compared to determine common and mutually exclusive genes between all therapies in both phenotypes (PD and CR). Survival analysis was performed for genes unique to each therapy using PrognoScan - a survival prediction tool with a collection of publicly available cancer survival data [19]. A list of genes specific to each treatment and phenotype was provided as input for the PrognoScan tool [20]. The survival performance was evaluated based on p-value, where a value < 0.05 indicated good overall survival while a greater value indicated poor survival. The genes obtained from survival analysis were further evaluated for immune related function using Immport which is a comprehensive list of immune related genes [21].

2.8 Study of Expression Variance for Biomarker Prediction

We performed statistical tests (ANOVA and Wilcoxon signed rank test) to detect putative biomarker genes whose expression significantly varied across the two phenotypes (PD vs CR) for each anticancer therapy in LUAD and COAD samples.

3. Results

3.1. Data

As shown in Table 1. 129 samples along with therapy information were obtained from BCR-TCGA for LUAD. Alkylating agents had the highest number of samples (60) due to its wide drug target spectrum, whereas 42 samples were divided amongst alkylating agents and antimetabolite therapy for COAD with clinical outcome (PD vs. CR) as shown in Table 1 with almost equal number of samples.

3.2. Gene Set Enrichment Analysis (GSEA)

Gene sets were filtered based on the p-value cutoff (<0.05) and FDR (<0.25), and we found significantly higher enrichment for complete remission in all four therapies for LUAD as reported in Table 2. High enrichment of oncogenic signatures (C6) in CR phenotype were reported for antimetabolite (PD: 0, CR: ~53%), mitotic inhibitor (PD: ~5%, CR: ~24%) and anti-angiogenesis (PD: 0, CR: ~58%). No significant enrichment was recorded for canonical pathway (CP) in both PD and CR for alkylating agents. However, in COAD oncogenic signature geneset (C6), PD phenotype was highly enriched in both therapies and a similar trend was observed for the hallmark geneset category as shown in Table 2. That said, the C7 category (immunogenic signatures) was equally enriched in PD and in CR for alkylating agents. Between lung and colon adenocarcinomas, we found higher overall enrichment for PD in COAD as opposed to LUAD.

Table 1. Distribution of LUAD and COAD samples based on therapy and outcome (PD vs CR) obtained from BCR-TCGA.

Therapy	Lung adenocarcinoma (LUAD)			Colon Adenocarcinoma (COAD)		
	Progressive Disease	Complete Remission	Total (Therapy wise)	Progressive Disease	Complete Remission	Total (Therapy wise)
	PD	CR		PD	CR	
Alkylating Agent	13	47	60	7	12	19
Antimetabolite	7	26	33	13	10	23
Mitotic Inhibitor	5	10	15	-----	-----	-----
Anti-angiogenesis	6	15	21	-----	-----	-----
Total(Response Type)	31	98	129	20	22	42

Table 2. Enrichment percentage for significantly enriched (p< 0.05, FDR < 0.25) genesets from GSEA across PD vs. CR in each anticancer therapy for LUAD and COAD samples.

Lung Adenocarcinoma (LUAD)	Progressive disease (PD)					Complete Remission (CR)				
	C5	C6	C7	CP	Hallmark	C5	C6	C7	CP	Hallmark
Alkylating agent	0	4.20%	0.30%	0	26%	1.4%	7.4%	0.4%	0.0%	6%
Antimetabolite	0.2%	0	0	0.9%	0	0.2%	51.3%	1.2%	9.9%	20%
Mitotic Inhibitor	1.9%	4.8%	2.4%	10.7%	38%	8.2%	24.3%	4.7%	8.2%	32%
Anti-angiogenesis	2.5%	0.0%	1.7%	1.4%	14%	10.9%	57.7%	3.8%	5.8%	36%
Colon Adenocarcinoma (COAD)	Progressive disease (PD)					Complete Remission (CR)				
	C5	C6	C7	CP	Hallmark	C5	C6	C7	CP	Hallmark
Alkylating agent	9.2%	20.6%	7.5%	4.0%	16.0%	0.0%	0.0%	8.2%	5.2%	0.0%
Antimetabolite	5.9%	23.2%	6.6%	2.7%	22.0%	1.2%	0.0%	0.1%	0.1%	2.0%

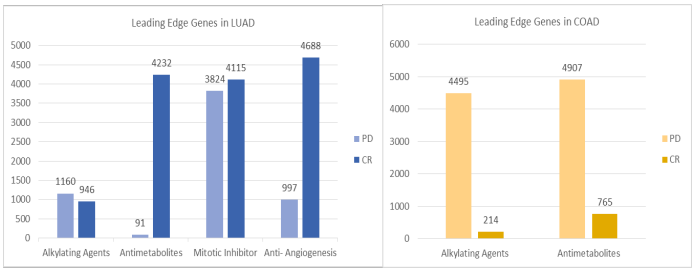
3.3. Leading Edge Genes

This trend was reversed in the case of COAD, where large number of leading edge genes enriched across PD for both the alkylating agents (PD:4495, CR:214) and antimetabolite therapies (PD:4907, CR:765) as shown in Figure 2. This result is consistent with the GSEA results shown in Table 2.

3.4. TABF-FAGO term mapping

We next sought to identify the functions associated with the genes within various enriched genesets. The *Functionally Associated GO-terms* (FAGO) for all the genesets were obtained, using ClueGO analysis. The number of genes associated with each of the GO terms is shown in Table 3. Majority of the FAGO terms belong to biological processes (BP) or immune-related processes

for both LUAD and COAD. The TABF terms and mapped FAGO terms are reported in Tables 1-12 in Appendix I section and have been explained in detail in the discussion section.

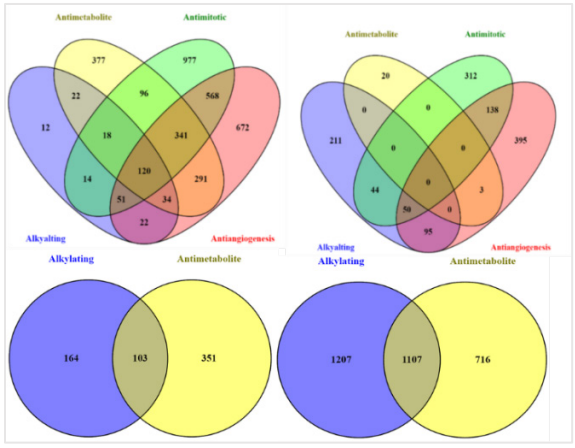


3.5. Survival analysis and Biomarker Discovery

The set of non-repetitive genes from each TABF-FAGO term mapping was extracted for survival analysis. However, due to the interest in seeking specific biomarkers for each therapy, only genes that were unique and non-overlapping between the anticancer therapies were extracted. The result showed 120 common leading genes across all four therapies for CR and 0 genes common across PD in LUAD. Similar to the analysis of genes unique to the therapy, analysis of these 120 common genes was also completed as shown in Figure 3. However with COAD data, both alkylating agents and antimetabolite therapies target the cell cycle and are involved in its disruption, which explains the large number of overlapping genes for both PD (103 genes) and CR (l107 genes) as shown in Figure 3.

Table 3. The number of FAGO terms mapped with TABF terms in each GO term category for all the anticancer therapies across PD and CR for LUAD and COAD samples.

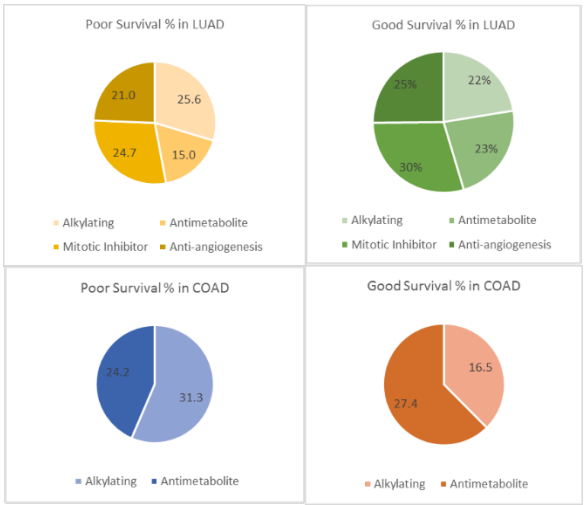
Lung Adenocarcinoma (LUAD)												
Therapy	Alkylating agents						Antimetabolite					
	PD			CR			PD			CR		
GO term Category	FAGO	TABF-FAGO Mapped Terms	% mapping	FAGO	TABF-FAGO Mapped Terms	% mapping	FAGO	TABF-FAGO Mapped Terms	% mapping	FAGO	TABF-FAGO Mapped Terms	% mapping
Biological Process (BP)	9	9	100	339	74	21.8	3	0	100	57	13	22.8
Cellular Component (CC)	5	0	0	28	0	0.0	0	0	NA	6	0	0.0
Molecular Function (MF)	6	0	0	38	3	7.9	0	0	NA	34	4	11.8
Immune system process	6	6	100	70	3	4.3	0	0	NA	27	27	100.0
KEGG	20	3	15	17	4	23.5	0	0	NA	40	8	20.0
Total	46	18	39.13	492	84	19	3	3	NA	164	52	33
Therapy	Anti-angiogenesis						Mitotic Inhibitor					
	PD			CR			PD			CR		
GO term Category	FAGO	TABF-FAGO Mapped Terms	% mapping	FAGO	TABF-FAGO Mapped Terms	% mapping	FAGO	TABF-FAGO Mapped Terms	% mapping	FAGO	TABF-FAGO Mapped Terms	% mapping
Biological Process (BP)	36	29	80.6	136	37	27.2	22	22	100	76	32	42.1
Cellular Component (CC)	7	0	0.0	26	0	0.0	0	0	NA	3	0	0.0
Molecular Function (MF)	7	0	0.0	13	4	30.8	0	0	NA	9	9	100.0
Immune system process	25	23	92.0	57	40	70.2	3	3	100	148	49	33.1
KEGG	3	1	33.3	27	6	22.2	1	0	0	30	9	30.0
Total	78	53	67.9	259	87	42	26	25	96.2	266	99	42.4
Colon Adenocarcinoma (COAD)												
Therapy	Alkylating agents						Antimetabolite					
	PD			CR			PD			CR		
GO term Category	FAGO	TABF-FAGO Mapped Terms	% mapping	FAGO	TABF-FAGO Mapped Terms	% mapping	FAGO	TABF-FAGO Mapped Terms	% mapping	FAGO	TABF-FAGO Mapped Terms	% mapping
Biological Process (BP)	10	4	40.0	10	5	50.0	25	7	28	100	6	6
Cellular Component (CC)	12	4	33.3	12	5	41.7	0	0	NA	13	4	30.8
Molecular Function (MF)	0	0	NA	0	0	NA	0	0	NA	0	0	NA
Immune system process	36	17	47.2	23	8	34.8	20	11	55	56	17	30.4
KEGG	1	1	100.0	0	0	NA	4	1	25	0	0	NA
Total	59	26	44.0	45	18	40.0	49	19	39.0	169	27	16



Simultaneously, the number of immune-related genes in each case using a comprehensive list of known immune related genes from Immport was evaluated. Results show a good percentage of immune related genes. Of all these total genes involved in good and poor survival, ~39% genes (in CR) ~48% genes (in PD) for LUAD, and ~42% genes (in PD) and ~14.6% (in CR) for COAD were related to immune response as shown in Table 4.

3.6. Expression Variance

The gene expression data was used in statistically analyzing expression variance across the PD vs. CR for all survival genes using ANOVA and Wilcoxon ranked test. The immune-related genes, which showed significant expression variance (p-value < 0.05), are presented in Table 5 for LUAD and Table 6 for COAD. In lung adenocarcinoma, the data resulted in 2 immune-related biomarkers for alkylating agents and 6 for antimetabolites, 17 gene biomarkers for mitotic inhibitors, and 33 gene biomarkers



for anti-angiogenesis, as shown in Table 5. For colon adenocarcinoma, 5 biomarkers for alkylating agents and 8 expression biomarkers for antimetabolites were seen in Table 6. The expression biomarkers are represented as boxplots in Figure 5 and Figure 6, where a clear distinction between PD upregulation compared to CR for DEF126 in alkylating genes are seen for LUAD, while JUNB has very high overall expression with high variance for PD as compared to CR. In antimetabolites, we found a good differentiation in expression for two groups: PD vs. CR response phenotype, for all the 6 genes including SIVA1, DNM1 and TNFRSF14, which shows a significant variance with high expression for CR as shown in Figure 5. The highest number of gene expression biomarkers were found in mitotic inhibitors, with a mixture of PD and CR upregulated genes with immune-based function. SRGN, MAPT, ABR, NPBL, STXPB1, NFE2L1, AATF, and VEZF1 were genes with striking difference in expression across PD and CR. The anti-angiogenesis therapy had 17 genes with striking expression variance, with 3 genes upregulated for PD and 14 for CR. A significant expression variance was seen across PRAKAA2, F13A1, NUPR1, RAPGEF4 and STK38L in PD compared to CR.

In COAD, we found 35 significant immune-related biomarkers - NPPB, S100A3, PPBP, LCNL1, PRDX2, TNFRSF11A, S100G, LEFTY1, IL13RA2, CXCR1, S100A12, FAM3C, CCK, PSMD14, PROK2, IL20RB, S100A9, WNT5A, TNFRSF11B, INHBC, APLN, MALT1, IFNE, RLN2, KLRC2, NCR1, KLRC3, PSPN, CRLF3, IL15, CDH1, COLEC10, GAST, SP1, and CRABP2 were upregulated for PD while two genes, IL18RAP and IL24, were significantly upregulated for CR in alkylating agents. In antimetabolite therapy, 8 immune-related biomarkers were indicated: EIF2AK2, FABP5, FCN2, HRAS, IL10, NAMPT, SEMA4G, and TAP2 were upregulated for PD, and three genes, IGF2R, NOV, and IL17C, were upregulated for CR as shown in Figure 6.

Table 4. The number of survival analysis genes with known immune function for PD and CR in all the therapies for LUAD and COAD.

Lung Adenocarcinoma						
Therapy	Progressive Disease (PD)			Complete Remission (CR)		
	Poor Survival	immune related	immune %	Good Survival	immune related	immune %
Alkylating	54	18	33.3	6	4	66.7
Antimetabolite	3	3	100.0	193	85	44.0
Mitotic Inhibitor	77	49	63.6	642	211	32.9
Anti-angiogenesis	83	35	42.2	377	179	47.5
Total	217	105	48.3	1218	479	39.3
Colon Adenocarcinoma						
Therapy	Progressive Disease (PD)			Complete Remission (CR)		
	Poor Survival	immune related	immune %	survival	immune related	immune %
Alkylating	378	221	58.5	27	3	11.1
Antimetabolite	173	13	7.5	96	15	15.6
Total	551	234	42.5	123	18	14.6

Therapy	PD_UP	CR_UP	Immune genes with high expression variance based on anova and	Total number of immune related genes after survival analysis	%
Alkylating Agents	DEFB126, JUNB		2	22	9
Antimetabolite		HPRT1, PFKP, SIVA1,TNFRSF14,RAB34,DNM1	6	88	6.8
Anti-angiogenesis	XRCC5,PPP2R1A,AIPL1	GNB1,TNFRSF25, ID4, RAPGEF4, NRP2, IL20RA, STK38L, IGF2R, PRKAA2, NUPR1, GLG1, SH3BP4, BNIP3L, F13A1	17	213	7.9
Mitotic Inhibitor	MAPK3,BAG1, SRGN, IL10RB, MAEA, NPY5R, CCL25, PYDC1	HLF, MAPT, ABR, HSPG2, CASP2, PPARD, STXBP1, TMEM173, HIVEP2, IRAK1, MLLT6, HDAC5, STIM1, ITPR3, NLK, SRC, PRKACA, NFE2L1, RIPK1,SLIT1, NOD1, NIPBL, DAB2IP, DGKZ,VEZF1 ,AATF	33	260	12.6

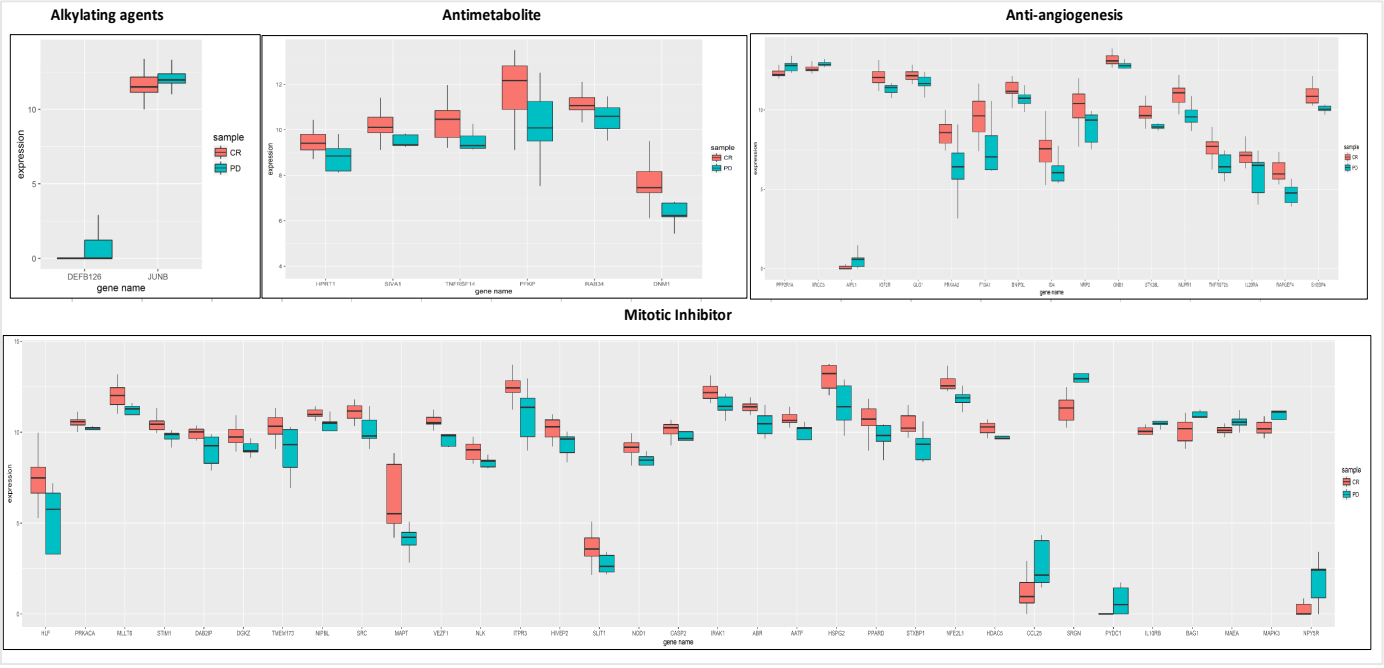


Figure 5. Boxplot showing expression variance between immune related genes for all the four anticancer therapies in LUAD.

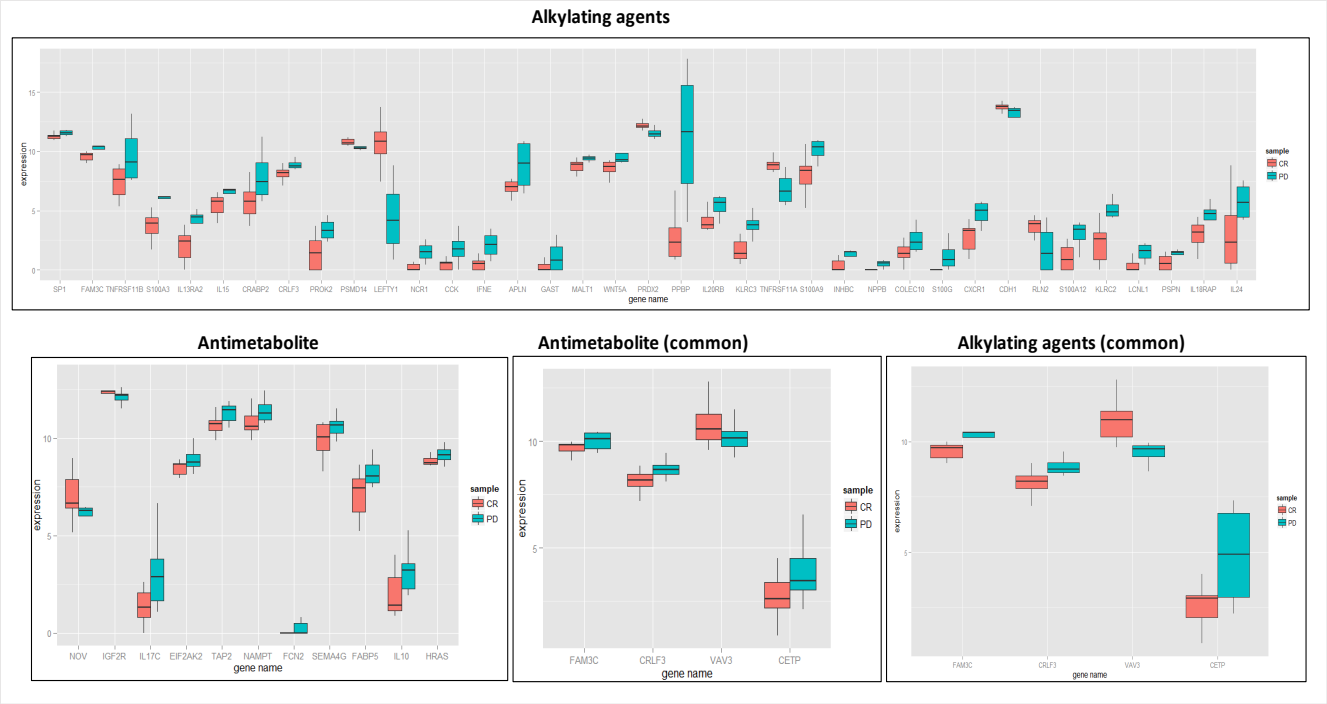


Figure 6. Boxplot showing expression variance between immune related genes for alkylating therapies and antimetabolite in COAD.

Therapy	PD_UP	CR_UP	Immune genes with high expresison varicne based on ANOVA	Total number of genes from survival analysis
Alkylating	SP1,TNFRSF11B,CABP1,S100A3,IL13RA2,IL15,CRABP2,PROK2,PSMD14,LEFTY1,CCK,IFNE,APLN,MALT1,WNT5A,PRDX2,PPBP,IL20RB,KLRC3,TNFRSF11A,S100A9,INHBC,NPPB,COLEC10,S100G,CXCR1,CDH1,RLN2,S100A12,KLRC2,PSPN,PLAUR,ANGPT2,IRAK4,FGFR1OP2,,TBX2,SLC12A6,ADAM10,MBD2,WASL,DGKA,MYL12A,ROCK1,ITPKB,SOCS6,SETX,RASAL2,MAPK6,RABGEF1,FGD2,C18orf54	RPL10,RPL18,RPS12,IL18RAP,IL24	5	61
Antimetabolite	EIF2AK2,FABP5,FCN2,HRAS,IL10,NAMPT,SEMA4G,TAP2,ASNS,ATP6V1A,BLM,CCNB2,FANCG,GLRX2,KIAA0101,LIN9,MAPKAPK5,MCM10,PA2G4,PIF1,PTGES3,RANBP1,SART3,SDS,TXNRD2,XPO1,ARPC5	CHD6,FZD8,HOXA11,KDM4B,KIDINS220,NCOA6,PHF8,PTCH1,SSBP3,IGF2R,NOV,IL17C	8	27

4. Discussion

4.1. GSEA and Leading Edge Analysis

In this work, we used the public dataset from TCGA for lung and colon adenocarcinoma patients who were treated with different anticancer drugs. Due to the lack of sufficient pre-cancerous samples, this study investigated the efficacy of the treatment outcomes, which were PD and CR. Since alkylating agents were the most commonly used drug therapy for non-small cell lung cancer (NSCLC) [22] compared to other therapeutic agents, we obtained a high number of study samples for patients treated with drugs in this therapeutic category. Anti-angiogenesis therapy is relatively new in contrast to the other three therapies [23], and consists of around 20 compounds being tested - only a few of them have been FDA-approved for commercial use, which explains the relatively small sample size for this therapy class. As for colon cancer, chemotherapy (alkylating agents and antimetabolite drugs) is the preferred choice of treatment.

A geneset enrichment analysis on two datasets to evaluate pathway level differences between the two clinical responses, i.e. PD vs. CR, is presented in this study. In case of LUAD, higher enrichment for CR samples was seen as opposed to PD in all five geneset categories. Since the disease under study is cancer, the result explains the large number of genesets enriched for C6, which is an oncogenic signature category based on MsigDB as shown in Table 2. Also this pattern was consistent across the leading edge analysis, where we selected the genes that significantly contributed towards the enrichment signal of a geneset. Overall enrichment for CR was higher than PD for LUAD as shown in Figure 2.

This pattern was reversed for COAD where PD samples were more enriched than CR, as presented in Table 2, and reflects upon the chemoresistant nature of this cancer type. This result was translated further to explain the higher leading edge genes enriched for PD as compared to CR as shown in Figure 2. Moreover, oncogenic signatures were highly enriched compared to other geneset categories. However, unlike LUAD, a significantly higher enrichment percentage for immunogenic signatures (C7) was also observed, which suggests a more immuno-active behavior of COAD. This is supported by the report from Tougeron et al., [24] which described the immune response as vital factor in the progression of colon cancer, and there exists a close association between the prognosis of the disease and the rate of tumor-infiltrating lymphocytes. These results are concordant with a number of TABF-FAGO terms related to immune response function, as shown in Table 3.

4.2 TABF- FAGO term mapping

The results from Table 3 show the number of FAGO terms obtained from ClueGO analysis and mapped with TABF terms for both complete remission (CR) and progressive disease (PD). In this section we will try to comprehend these associations and their biological interpretations.

4.2.1 Alkylating agents – Complete Remission

a. Lung adenocarcinoma

The TABF-FAGO term mapping results for alkylating agents can be comprehended from its mechanism of action. Cytotoxicity is one of the most highly correlated topics between cancer and immune system. It is a process involved in cell toxication which trigger processes like apoptosis, necrosis, and cell lysis. Alkylating agents, as described earlier, are cytotoxic agents that target the process of cell division. The host immune system has its own set of agents that cause cytotoxicity. The main ones are lymphocytes, cytotoxicity T-cells and natural killer cells. The recent review from Kulawik [25] explains the role of cytotoxic T-lymphocytes in antitumor response in lung adenocarcinoma. The host immune system also plays a crucial role in the antitumor response with a population of CD8+ and CD4 + lymphocytes [26], neutrophils [27], NK-T, B and T cells triggering the process of apoptosis [28]. The process is often amplified with the help of macrophages and dendritic cells which contributes towards antigen presentation for a successful cytolytic attack [29]. Therefore, the TABF terms found from the literature, were categorized as: *cell cytotoxicity, macrophage activation, immune response and myeloid cell activation, regulation of T cell and lymphocyte activation*, and *B cell proliferation and activation.* These terms are known to contribute synergistically towards the efficacy of alkylating therapy and are known to show tumor remission. This explains FAGO terms related to cytotoxicity in complete remission and TABF-FAGO term mappings involved with activation of T-cell, B-cell, macrophage activation, and immune response as shown in Table 1- appendix.

Targeting some signaling pathways can prevent DNA damage response and increase the efficacy of chemotherapy. While signaling pathways are essential for normal stem cells, their dysregulation can lead to tumorigenesis [30]. They form a complex interconnected network and can be crucial in regulating the cell structure, apoptosis and macrophage polarity [31]. This explains signaling pathway related term mapping, upregulated in complete remission for alkylating agents. The process of leukocyte activation and differentiation is connected with inflammatory response, which is beneficial in normal cells, and leukocytic adhesion to endothelial tissues marks the beginning of the inflammation process. But as much as these processes are important in the normal cell functioning, they can be the reason of poor survival in cancer and their regulation is crucial for improved efficacy of chemotherapy specifically alkylating agents [32]. The negative regulation of tumor promoting leukocytes is therefore one of the most important processes in improving the efficacy of this intervention. The role of leukocytes and lymphocytes is paradoxical and their balance mediates the process of tumor regression or promotion [33].

Cytokines are secretory proteins that are involved in the regulation and activation of immune cell. The role of host-derived cytokines is insightful in understanding the cytokine-tumor cell interactions in the inhibition of cancer development. According to the review by Dranoff [34], host-derived cytokines can suppress tumor formation by controlling infection, inflammation and immunity. Our results were consistent with cytokine-associated FAGO terms upregulated in CR for alkylating therapy. The review also talks about IL-12 cytokine production by macrophages and dendritic cells inhibiting chemical carcinogenesis. We found chemical carcinogenesis pathway enriched in progressive disease while cytokine-related FAGO terms were upregulated to the opposite response (CR), which further validates and justifies our results. The host immune response can be categorized into two broad categories: innate and adaptive, where the former is the first line of defense and involves the activation of macrophages, neutrophils, granulocytes etc. while the latter activates B and T cell response and production of immunoglobulins [35]. The activation of adaptive immune response and immunoglobulin production with the process involving somatic recombination have shown to sensitize tumors to alkylating drugs and increasing the drug efficacy [36] which further explains TABF-FAGO term mapping in CR for LUAD**.**

b. Colon adenocarcinoma

Bracci et al., [37] discussed in-depth the relevance of host immune-competency for improved efficacy of chemotherapy. There are evidences of enhanced T and B lymphocytes in the spleen of animal model with colon cancer showing synergistic association with chemotherapy drugs like oxaliplatin, gemcitabine and 5-FU which explains the upregulation of B and T lymphocyte activation, proliferation and other terms related to the activation of immune system in CR for both alkylating agents and antimetabolite therapy. Mitochondria organization is a key player in release of cytochrome-c with subsequent activation of caspase-9 and -3 which reduces cell cycle progression and carcinogen activation, as well as promoting immunosurvelliance [38]. Furthermore, non-coding RNA processing is often associated with improved response to chemotherapy in colon cancer and are directly involved in DNA methylation, histone modification and gene silencing. In turn, they are epigenetically targeted for repression of cancer and improved drug therapy [39]. All these biological processes explain the upregulation of FAGO terms in Table 2- appendix.

4.2.2 Alkylating Agents – Progressive Disease

a) Lung adenocarcinoma

We obtained 18 FAGO terms that were successfully mapped with the TABF terms as shown in Table 3-appendix. Holland-Feir Cancer Medicine (6th ed.) [40] presents a complete review on alkylating agents and the factors affecting its efficacy. Alkylating agents induce cell cytotoxicity but its potency is determined by the host system. One such finding is the host response to alkylation damage by arresting the cell in the G2 phase of the cell cycle which in turn activates the DNA repair process, where cells enter the mitotic phase despite alkylation damage, which leads to therapy resistance. A human tumor cell line has been shown to exhibit G2 arrest in response to alkylating damage, resulting in resistance to nitrogen mustard [41] which clearly explains the upregulation of FAGO term*: regulation of cell cycle G1/S phase transition* in PD for lung cancer. In some cases, transcription factors involved in DNA transcription initiation are released in response to genotoxic stress (agents involved in DNA damage), which stimulate RNA polymerase and promote RNA binding. This eventually leads to mRNA synthesis and the process is called gene activation, further stimulating DNA repair mechanisms by the host system and promoting defense against alkylating agents

[42]. We found related FAGO terms upregulated in PD as shown in Table 3- appendix, which strongly correlates with the findings discussed above.

The relationship between inflammatory response and cancer progression was first discovered in the 19th century by Rudolf Virchow and is one of the most widely studied phenomena. There exists an established connection between leukocyte chemotaxis and inflammation. The movement of leukocytes from blood to tissue in response to inflammatory stimuli was observed by Metchnikoff in 1891. During inflammation, leukocytes migrate to the site of inflammation to fulfill their role in defense response [43]. Pathways involved in inflammation and wound healing have shown to be associated with promoting tumorigenesis, with increased resistance to anticancer treatments. Interestingly, there also exists a strong link between complement activation and inflammation. Generally, the complement cascade system is activated to defend a host against microbes during an inflammatory response. The activation of complement and its role in tumor progression is evident in studies conducted for different cancer types, including lung adenocarcinoma [44], breast [45] and thyroid cancer [46]. Consequently, we observed similar processes from ClueGO results where genes involved in inflammation, complement activation and defense response were activated in PD alongside genes responsible for positive regulation of leukocyte chemotaxis.

b) Colon adenocarcinoma

Colon cancer is strongly linked with inflammatory microenvironment composed of intestinal epithelial cells, immune cells, etc. which determine the fate of the intervention. Moreover, cytokines acts as a mediator between the host immune response and malignant cells that activate oncogenic transcription factor STAT3 which is a major activator in cancer inflammation and promotes cell cycle progression [47]. Holohan et al., explains the role of JAK-STAT signaling pathway in cell proliferation, tumor progression, and chemo-resistance [48]. We found the upregulation of several interleukin factors, e.g. IL-1, 2, 4, 6, and 10, in progressive disease for colon cancer. IL-6 is a major player in inflammatory response in COAD that interacts with the membrane-associated gp130 subunit and activates Janus kinases (JAKs) with downstream effects on STAT3 [49], which explains their role in tumor progression as shown in Table 4-appendix.

IL-6 can increase resistance to various chemotherapeutic drugs [50, 51] by preventing apoptosis and downstream signaling of MAPK cascade [52]. Moreover, the release angiogenesis growth factors greatly influence the drug response in solid tumors and evidence suggests their involvement in cancer cell survival in the presence of chemo drugs [53]. Furthermore, IL-1b is a pro-inflammatory cytokine which can modulate the response of different drug transports [54]. Tumor necrosis factor is a major inflammatory cytokine and is well known for its role in tumor progression and drug resistance. Therefore, the upregulation of this term in the PD category for alkylating agents is well explained [55]. Mccubre et al., discusses the effect of the ERK-signaling cascade in promoting tumorigenesis and resistance to doxorubicin treatment [56]. All these points support our results from TABF-FAGO term mapping for PD in alkylating agents in Table 4-appendix [57].

4.2.3 Antimetabolite – Complete Remission

a) Lung adenocarcinoma

Antimetabolites are another class of chemotherapeutic drug that inhibit the action of metabolites and interfere with DNA and RNA growth [58]. They replace the normal compounds in the cell, resulting in disrupted cell division and specifically targeting the synthetic phase of cell cycle [59]. Since antimetabolites disrupt cell division and DNA damage, the TABF and FAGO terms upregulated for complete remission were very similar to alkylating agents. We found FAGO terms related to cell cytotoxicity upregulated in CR, and the majority of the FAGO terms were similar to alkylating agents.

Naïve CD4+ T cells can be categorized into four main lineage: T helper cells -1, 2, 17 and Treg cells [60]. Th2 promotes tumor growth and Th1 produce cytokines like interferon gamma, which causes cytotoxic T lymphocytes (CTL) proliferation, resulting in tumor elimination and potentiating the action of chemotherapy [61].The host immune milieu is rich in cytokine and chemokines which can produce immune-stimulatory effects. For instance, type-I IFNs can enhance the stimulation of dendritic cells which activate Th1 and inhibits Treg activity [62]. Therefore, we saw FAGO terms associated with Th cell activation and cytokines in Table 5-appendix.

Signaling pathways are popular targets and critical to the success of an anticancer therapy [63]. For instance, activated PI3K/Akt/mTOR pathways can regulate cell proliferation, growth, and survival, and are important targets for the development of potential antitumor agents [64]. According to Wang et al [65], the activation of signal transduction pathways enhances the cellular response to drug treatment. Hence, we found various signaling pathways terms associated with CR.

The link between skeletal system development and the immune system has been recognized recently with an interdisciplinary field studying and embracing the relationship under the name of "osteoimmunology". Studies have also shown that osteoimmunology can be a potential target for anticancer interventions as several systems that are related to host immune response are actually related to skeletal function. Such information has been useful to understand the improved efficacy of anticancer drugs, especially antimetabolites [66]. Receptor activator of NF-kB (RANK)/RANK Ligand (RANKL)/Osteoprotegerin (OPG) system is involved in the regulation of osteoclastogenesis, which influences immune, cardiovascular, endocrine, and nervous systems. They stimulate the antigen-specific immune response and are involved in dendritic cell maturation and survival [67, 68]. Hence, this clearly explains the upregulation of TABF-FAGO association in Table 5-appendix.

b) Colon Adenocarcinoma

We observed the upregulation of terms related to T cell lymphocyte activation, proliferation and positive regulation of immune response, a consistent pattern for CR category in all chemotherapies for both lung and colon cancer. Additionally, we

found histone modification terms enriched in CR. Histone modifications are crucial to gene regulation processes by compacting DNA. Histones are often involved in maintaining the balance between euchromatin and heterochromatin by undergoing post-translational modification at the amino terminal ends. These modifications, e.g. methylation, acetylation, ubiquitination, etc. [69] are correlated with the efficacy of chemo drugs in numerous types of cancer (prostate, breast, colorectal, lung) and hematological malignancies. All these TABF terms with their mapped FAGO terms are presented in Table 6-appendix.

4.2.4 Antimetabolites – Progressive Disease

a) Lung adenocarcinoma

Almost 90% of genomic DNA is transcribed into RNA, but only 2% of this RNA is protein-coding; the remaining "dark matter" genomic content is transcribed into non-coding RNA (ncRNA) and has been implicated with biological significance particularly in cancer. According to Liu et al. [70], ncRNAs play a crucial role in tumor progression and resistance to chemotherapy. Expression of metastasis-associated lung adenocarcinoma transcript 1 (MALAT1), a long non-coding RNA (lncRNA) is associated with resistance to chemotherapy and poor prognosis in patients in non-small cell lung cancer [71]. Another interesting study was conducted by Mader and colleagues [72] that showed increased expression of thymidylate synthase (TS) resulted in resistance to antimetabolite drug –5 fluorouracil (5-FU), to incorporate with RNA. Our results for progressive disease were in consensus with the discussion as shown in Table 7 -appendix.

b) Colon adenocarcinoma

The upregulated terms in the PD category for antimetabolite therapies were similar to alkylating agents - *IL1 production, tumor necrosis factor production, inflammation,* and *ERK cascade*. This is explained by their both being chemotherapy drugs with similar modes of action. Slow growth of tumors or G0-G1 cell cycle promotes resistance against antimetabolite drugs, which explains the connection between the upregulation of G1/s , G2/M phase transition and the progression of disease in Table 8 -appendix. Furthermore, literature based study shows TNF-alpha increase post-drug recovery and tumor survival in colon cancer from 5-FU (antimetabolite drug) by obstructing S phase entry [73].

4.2.5 Mitotic Inhibitors – Complete remission

Mitotic inhibitors are derived from natural plant alkaloids. They disrupt microtubules involved in mitotic cell division process [74]. Even though the mechanism of action for this therapy is very similar to alkylating agents, it was interesting to see that the immune response varied with some unique FAGO – TABF term mapping. The terms that were similar to alkylating agents were: *T cell and lymphocyte activation, B cell activation, positive regulation of immune response* and *cell signaling*. The terms similar to antimetabolite therapies were related to *system development and KEGG signaling pathways*. Interestingly, we found some striking differences. Firstly, innate immune response was more active as compared to adaptive when compared to alkylating agents and antimetabolites. Secondly, we found significant upregulation of FAGO terms related to cyclic adenosine monophosphate (cAMP) processes, which is attributed to the fact that cAMP plays a fundamental role in cellular response. Regulation of cAMP pathways is associated with ion metabolism and apoptosis [75]. They are also known to intermingle with other signaling pathways like calcium [76], JAK/STAT inhibitors [77], and RAK-mediated MAPK kinases to modulate cell processes [78]. The upregulation of cAMP-related pathways is favorable for the action of antimitotic drugs [79] which justifies the upregulated FAGO terms in CR as shown in Table 9 -appendix.

Another paradoxical phenomenon that was observed in antimitotic agents was the upregulation of an inflammatory response in CR unlike the other therapies. Even though inflammatory response is often associated with tumor regression, in some cases it can enhance the action of antimitotic agents. Host macrophages secreting inflammatory cytokines have been shown to induce the expression of inducible nitric oxide synthase (iNOS). This potentiated the effect of paclitaxel to regress cancer in solid tumors and promoted apoptosis [80]. Since inflammatory response is a critical stage in the wound healing process [81], we could see that wound healing was upregulated alongside inflammatory response in CR as shown in Table 9 –appendix.

Another contradictory observation was upregulation of interleukin-10 (IL-10) which is an anti-inflammatory cytokine that downregulates cytokine production. The effect of host IL-10 in tumor progression or regression has been debated over a long time. While studies have shown that regulatory T cells (Treg) produce IL-10 which downregulates immune response [82], this is inconsistent with Mumm et al's [83] claim that IL-10 reduces tumor development by inducing cytolytic molecules in CD8+ T cells and IFN- gamma dependent mechanisms. The review by Teng et al. [84] talks about the significance of stable IL-10, which can be beneficial in the regressing tumor. So far, there has not been any strong evidence showing a positive correlation between host IL-10 expression and survival of cancer patients. Our results in Table 9 -appendix highlights the above discussion.

4.2.6 Mitotic Inhibitors – Progressive Disease

The key highlight in the PD category was the upregulation of terms: *antigen presentation and transporter-associated with antigen processing (TAP)*, which are known for their multidrug resistance in cancer. The experiment conducted on a human cancer cell line by Izquierdol and colleagues [85] present a strong correlation between TAP expression and resistance (~ 2 fold) to the antimitotic drugs doxorubicin and vincristine. Mitotic checkpoints and response to DNA damage have shown resistance against these drugs, as they sense microtubule attachment to kinetochore. According to Wang et al. [65], the activation or dysregulation of these mitotic checkpoints and DNA damage response leads to continued cell division and increased resistance to taxol (antimitotic drug). Table 10 -appendix shows the FAGO terms upregulated in PD for mitotic inhibitors.

4.2.7 Anti-Angiogenesis

Angiogenesis is the formation of new blood vessels and is known to play an essential role in tumor growth, as blood vessels supply oxygen and nutrients for the cancerous cell's nourishment. Angiogenesis inhibitors bind to the vascular endothelial growth

factor (VEGF) and prevent them from binding to the receptor in order to facilitate the process of vasculogenesis [86].

a) Complete Remission

The mechanism of action for this therapy is different from the other three types of chemotherapies discussed previously. There is a strong link between host immune response and anti-angiogenesis therapy. T-cell activation and VEGF share an antagonistic relationship where the latter inhibits the production of T cells by interfering with thymus development [87]. Therefore, T-cell activation has a positive effect on anti-angiogenesis therapy. To back up this hypothesis, we have the case study of semaphorins, axon guidance molecules belonging to the class of transmembrane proteins which are known to play a role in immune response and organ morphogenesis. Class IV semophorin (Sem4A), expressed on dendritic and B-cells, enhance T-cell activation and differentiation [88]. Toyofoku and colleagues demonstrated in their experiments that mice lacking Sem4A exhibit enhanced angiogenesis in response to VEGF or inflammatory stimuli [89]. Therefore, deeming Sem4A as an inhibitor of VEGF and angiogenesis. Our results are concurrent with the above discussion and presented in Table 11- appendix.

Hypoxia is a process of oxygen depletion in cells which induce pro-angiogenic factors along with inflammatory response. According the study conducted by Brustugun [90], hypoxia caused treatment failure in lung carcinoma. Resistance to gefitinib, extensively used in EGFR-mutated lung cancer, is also induced by hypoxia via upregulation of growth factor. Blocking chronic inflammation can prevent pro-tumor polarization and contribute to anti-angiogenesis effect. Therefore, we see the process of negative regulation of hypoxia upregulated in CR. Even though the wound healing process is connected with an inflammatory response, in the case of anti-angiogenesis therapy response wounding is seen to contribute to increased efficacy. Activated host healing response is a crucial process after the extensive surgery of solid tumors. Interleukin-6 (IL-6) can act as an inflammatory cytokine and also as an anti-inflammatory cytokine [91]. There are other cytokines (IL-20) that possess anti-angiogenesis properties [92] and their regulation enhances the efficacy of the therapy as reflected in Table 11- appendix.

b) Progressive Disease

We observed several immune-related terms upregulated in progressive disease for anti-angiogenesis therapy. We discussed the role of VEGF and other growth factors in promoting angiogenesis, which results in tumor progression. The results in Table 12 from appendix section show the upregulation of all these pathways which induce therapy resistance. MHC and antigen presentation are associated with a multidrug resistance class of proteins especially in lung cancer; the overexpression of the transporter protein is known to induce high resistance to therapy [93]. But further investigation is needed to understand their involvement with cancer progression and resistance. The review conducted by Zaidi and Merlino [94] talks about the paradoxical role of interferon gamma in producing immune response against tumors while also producing tumorigenic effects under certain conditions. Hence, response to interferon gamma has a role in reducing drug efficacy and promoting tumor progression. Innate immunity can release angiogenic factors and promote tumor angiogenesis. Therefore, the release of inflammatory cytokines by the host innate immune system can further diminish the response to drug therapy [95]. The activated DNA repair mechanism has shown resistance to cetuximab/Erbitux (an anti-angiogenesis drug) used to treat EGFR-expressing tumors [96]. Table 12 in appendix shows upregulated FAGO terms mapped with all the biological functions discussed above.

4.3 Survival analysis and Biomarkers

The prognostic ability of genes specific to each therapy were tested using survival analysis. We found several genes with significant overall survival for both LUAD and COAD as shown in Table 4, which further validates the role of host immune response and therapy outcome. As for the common genes across different therapies, there was no complete overlap of genes for LUAD in PD; this result is obvious, as all four therapies have different modes of action. But, interestingly, there was an overlap of 120 genes for CR, shown in Figure 3, with one gene (IL4R-interleukin 4 receptor) that showed significant expression variance and survival for lung adenocarcinoma. The role of IL4R has been paradoxical in cancer immunology and has often been debated. In our results, IL4 was seen to be involved with complete remission when treated with any of the three therapies: mitotic inhibitors, anti-angiogenesis and antimetabolite. This was validated further with the experimental evidence of overexpressed IL4R contributing towards remission in lung cancer patients [97].

However, in colon cancer, the two treatments under study have similar cell cycle targets. Hence, we found more overlapping genes between alkylating agents and antimetabolites as shown in Figure 3. While we found only 4 genes (FAM3C, CRLF3, CETP, VAV3) that showed significant survival and expression variance as shown in Figure 6, the first three showed higher expression for disease progression, while VAV3 was seen as CR upregulated. We also found literature evidence of VAV3 as one of the molecular signatures for colon cancer [98]. IL-17 is often associated with tumor promotion and acts as a pro-angiogenesis factor, while in some cases it has also been regarded as tumor suppressing, as shown by experimental studies from Kryczek et al. [99]. Therefore, its role has remained undefined. However, in our results, we found IL-17 to contribute towards complete remission in antimetabolite therapy for colon cancer.

5. Conclusion

The era of translational medicine has opened new horizons for personalized therapeutic interventions and spurred the development of novel prognostic therapy-based biomarkers for each patient's individual immune profile. It is now evident that the tumor–host immune interaction dictates the magnitude, quality and efficacy of most anticancer strategies. Despite this, in our knowledge, no systematic analysis has been performed to study expression and pathway-based differences of multiple anticancer therapies depending on patient's immunocompetence.

In this study, we used expression based analysis to come up with biomarkers with prognostic validity for different anticancer therapies in two cancer datasets, LUAD and COAD. We found many gene biomarkers, unique to a cancer type or therapy, which were either associated with the progression or remission of the intervention. Additionally, we also found a number of

overlapping genes across different therapies, of which some had a controversial status towards cancer promotion or regression. Our results predicted the outcome of therapy resistance alongside its improved efficacy to some extent. The scope of this study went further to understand some of the pathways and their mode of mechanism in-depth for different therapies in both cancer types. However, one major limitation of this study is the absence of experimental validation for these biomarkers even though the majority of the indicated genes in our results have corresponding correlations within the literature. Therefore, such information would be highly valuable for translational medicine and in making an accurate intervention choice.

This work is an extension of our previous conference paper published at BIBM IEEE 2016 [100].

Conflict of Interest

The authors declare no conflict of interest.

References

[1] S. B. Coffelt, K. E. de Visser, "Immune-mediated mechanisms influencing the efficacy of anticancer therapies" Trends in immunology., **36**(4), 198–216, 2015.

[2] G. P. Dunn, L. J. Old, R. D. Schreiber, "The immunobiology of cancer immunosurveillance and immunoediting" Immunity., **21**(2), 137–148, 2004.

[3] K. N. Kodumudi, K. Woan, D. L. Gilvary, E. Sahakian, S. Wei, J. Y. Djeu, "A novel chemoimmunomodulating property of docetaxel: suppression of myeloid-derived suppressor cells in tumor bearers" Clinical Cancer Research., **16**(18), 4583–4594, 2010.

[4] E. M. Dijkgraaf, M. Heusinkveld, B. Tummers, L. T.C. Vogelpoel, R. Goedemans, V. Jha, J. W.R. Nortier, M. J.P. Welters, J. R. Kroep, S. H. van der Burg, "Chemotherapy alters monocyte differentiation to favor generation of cancer-supporting M2 macrophages in the tumor microenvironment" Cancer research., **73**(8), 2480–2492, 2013.

[5] P. A. Savage, S. Malchow, D. S. Leventhal, "Basic principles of tumor-associated regulatory T cell biology" Trends in immunology., **34**(1), 33–40, 2013.

[6] A.-P. Ganesan, M. Johansson, B. Ruffell, A. Yagui-Beltrán, J. Lau, D. M. Jablons, L. M. Coussens, "Correction: Tumor-infiltrating regulatory T cells inhibit endogenous cytotoxic T cell responses to lung adenocarcinoma" The Journal of Immunology., **191**(10), 5319, 2013.

[7] C. Murdoch, M. Muthana, S. B. Coffelt, C. E. Lewis, "The role of myeloid cells in the promotion of tumour angiogenesis" Nature Reviews Cancer., **8**(8), 618–631, 2008.

[8] R. Mazzieri, F. Pucci, D. Moi, E. Zonari, A. Ranghetti, A. Berti, L. S. Politi, B. Gentner, J. L. Brown, L. Naldini, "Targeting the ANG2/TIE2 axis inhibits tumor growth and metastasis by impairing angiogenesis and disabling rebounds of proangiogenic myeloid cells" Cancer cell., **19**(4), 512–526, 2011.

[9] D. M. Parkin, F. Bray, J. Ferlay, P. Pisani, "Global cancer statistics, 2002" CA: a cancer journal for clinicians., **55**(2), 74–108, 2005.

[10] S. Colak, J. P. Medema, "Cancer stem cells–important players in tumor therapy resistance" FEBS journal., **281**(21), 4779–4791, 2014.

[11] K. L. Manasa, "E7010: Investigational anticancer agents targeting the microtubules" International Journal of Pharmaceutical Sciences and Research., **6**(9), 3713, 2015.

[12] J. Folkman, "Clinical applications of research on angiogenesis" New England Journal of Medicine., **333**(26), 1757–1763, 1995.

[13] M. F. Misman, S. Deris, S. Z. M. Hashim, R. Jumali, M. S. Mohamad, "Pathway-based microarray analysis for defining statistical significant phenotype-related pathways: a review of common approaches" in 2009 International Conference on Information Management and Engineering, Kuala Lumpur, Malaysia, 2009.

[14] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles" Proceedings of the National Academy of Sciences., **102**(43), 15545–15550, 2005.

[15] A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdóttir, P. Tamayo, J. P. Mesirov, "Molecular signatures database (MSigDB) 3.0" Bioinformatics., **27**(12), 1739–1740, 2011.

[16] C. Kin, E. Kidess, G. A. Poultsides, B. C. Visser, S. S. Jeffrey, "Colorectal cancer diagnostics: biomarkers, cell-free DNA, circulating tumor cells and defining heterogeneous populations by single-cell analysis" Expert review of molecular diagnostics., **13**(6), 581–599, 2013.

[17] G. Bindea, B. Mlecnik, H. Hackl, P. Charoentong, M. Tosolini, A. Kirilovsky, W.-H. Fridman, F. Pagès, Z. Trajanoski, J. Galon, "ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks" Bioinformatics., **25**(8), 1091–1093, 2009.

[18] H. Xiong, Z.-G. Zhang, X.-Q. Tian, D.-F. Sun, Q.-C. Liang, Y.-J. Zhang, R. Lu, Y.-X. Chen, J.-Y. Fang, "Inhibition of JAK1, 2/STAT3 signaling induces apoptosis, cell cycle arrest, and reduces tumor cell invasion in colorectal cancer cells" Neoplasia., **10**(3), 287–297, 2008.

[19] H. Mizuno, K. Kitada, K. Nakai, A. Sarai, "PrognoScan" BMC medical genomics., **2**(1), 18, 2009.

[20] D. R. Cox, D. Oakes, "Analysis of survival data, CRC Press, 1984.

[21] S. Bhattacharya, S. Andorf, L. Gomes, P. Dunn, H. Schaefer, J. Pontius, P. Berger, V. Desborough, T. Smith, J. Campbell, "ImmPort: disseminating data to the public for the future of immunology" Immunologic research., **58**(2-3), 234–239, 2014.

[22] R. J. Scheff, B. J. Schneider, "Non–Small-Cell Lung Cancer: Treatment of Late Stage Disease: Chemotherapeutics and New Frontiers" in Seminar Interventional Radiology., **30**(02), 191–198, Thieme Medical Publishers, 2013.

[23] T. E. Stinchcombe, "Targeted therapy of advanced non-small cell lung cancer: the role of bevacizumab" Biologics: targets & therapy., **1**(3), 185, 2007.

[24] D. Tougeron, É. Fauquembergue, J.-B. Latouche, "Réponse immunitaire et cancers colorectaux" Bulletin du Cancer., **100**(3), 283–294, 2013.

[25] J. Domagala-Kulawik, "The role of the immune system in non-small cell lung carcinoma and potential for therapeutic intervention" Translational lung cancer research., **4**(2), 177–190, 2015.

[26] M. Salagianni, C. N. Baxevanis, M. Papamichail, S. A. Perez, "New insights into the role of NK cells in cancer immunotherapy" Oncoimmunology., **1**(2), 205–207, 2012.

[27] Z. Granot, J. Jablonska, "Distinct functions of neutrophil in cancer and its regulation" Mediators of inflammation., **2015**, 2015.

[28] J. G. Aerts, J. P. Hegmans, "Tumor-specific cytotoxic T cells are crucial for efficacy of immunomodulatory antibodies in patients with lung cancer" Cancer research., **73**(8), 2381–2388, 2013.

[29] K. Vermaelen, R. Pauwels, "Pulmonary dendritic cells" American journal of respiratory and critical care medicine., **172**(5), 530–551, 2005.

[30] G. Leon, L. MacDonagh, S. P. Finn, S. Cuffe, M. P. Barr, "Cancer stem cells in drug resistant lung cancer:Targeting cell surface markers and signaling pathways" Pharmacology & therapeutics., **158**), 71–90, 2016.

[31] G. S. Martin, "Cell signaling and cancer" Cancer cell., **4**(3), 167–174, 2003.

[32] D. G. DeNardo, D. J. Brennan, E. Rexhepaj, B. Ruffell, S. L. Shiao, S. F. Madden, W. M. Gallagher, N. Wadhwani, S. D. Keil, S. A. Junaid, "Leukocyte complexity predicts breast cancer survival and functionally regulates response to chemotherapy" Cancer discovery., **1**(1), 54–67, 2011.

[33] S. Ostrand-Rosenberg, "Immune surveillance: a balance between protumor and antitumor immunity" Current opinion in genetics & development., **18**(1), 11–18, 2008.

[34] G. Dranoff, "Cytokines in cancer pathogenesis and cancer therapy" Nature Reviews Cancer., **4**(1), 11–22, 2004.

[35] L. M. Weiner, R. Surana, S. Wang, "Monoclonal antibodies: versatile platforms for cancer immunotherapy" Nature Reviews Immunology., **10**(5), 317–327, 2010.

[36] G. P. Adams, L. M. Weiner, "Monoclonal antibody therapy of cancer" Nature biotechnology., **23**(9), 1147–1157, 2005.

[37] L. Bracci, G. Schiavoni, A. Sistigu, F. Belardelli, "Immune-based mechanisms of cytotoxic chemotherapy: implications for the design of novel and rationale-based combined treatments against cancer" Cell Death & Differentiation., **21**(1), 15–25, 2014.

[38] L. M. Coussens, Z. Werb, "Inflammation and cancer" Nature., **420**(6917), 860–867, 2002.

[39] C. Nervi, E. de Marinis, G. Codacci-Pisanelli, "Epigenetic treatment of solid tumours: a review of clinical trials" Clinical epigenetics., **7**(1), 127, 2015.

[40] D. W. Kufe, J. F. Holland, "Cancer medicine 6, Decker, 2003.

[41] P. M. O'Connor, D. K. Ferris, G. A. White, J. Pines, T. Hunter, D. L. Longo, K. W. Kohn, "Relationships between cdc2 kinase, DNA cross-linking, and cell cycle perturbations induced by nitrogen mustard" Cell Growth Differ., **3**(1), 43–52, 1992.

[42] M. Christmann, B. Kaina, "Transcriptional regulation of human DNA repair genes following genotoxic stress: trigger mechanisms, inducible responses and genotoxic adaptation" Nucleic acids research., **41**(18), 8403–8420, 2013.

[43] A. Rot, "The role of leukocyte chemotaxis in inflammation, Springer Netherlands, 1992.
[44] D. Ajona, C. Razquin, M. D. Pastor, M. J. Pajares, J. Garcia, F. Cardenal, M. Fleischhacker, M. D. Lozano, J. J. Zulueta, B. Schmidt, "Elevated levels of the complement activation product C4d in bronchial fluids for the diagnosis of lung cancer" PloS one., **10**(3), e0119878, 2015.
[45] F. Niculescu, H. G. Rus, M. Retegan, R. Vlaicu, "Persistent complement activation on tumor cells in breast cancer" The American journal of pathology., **140**(5), 1039, 1992.
[46] M. Yamakawa, K. Yamada, T. Tsuge, H. Ohrui, T. Ogata, M. Dobashi, Y. Imai, "Protection of thyroid cancer cells by complement-regulatory factors" Cancer., **73**(11), 2808–2817, 1994.
[47] J. Terzić, S. Grivennikov, E. Karin, M. Karin, "Inflammation and colon cancer" Gastroenterology., **138**(6), 2101–2114, 2010.
[48] C. Holohan, S. van Schaeybroeck, D. B. Longley, P. G. Johnston, "Cancer drug resistance: an evolving paradigm" Nature Reviews Cancer., **13**(10), 714–726, 2013.
[49] R. Catlett-Falcone, W. S. Dalton, R. Jove, "STAT proteins as novel targets for cancer therapy" Current opinion in oncology., **11**(6), 490, 1999.
[50] D. R. Hodge, E. M. Hurt, W. L. Farrar, "The role of IL-6 and STAT3 in inflammation and cancer" European journal of cancer., **41**(16), 2502–2512, 2005.
[51] A. Lichtenstein, Y. Tu, C. Fady, R. Vescio, J. Berenson, "Interleukin-6 inhibits apoptosis of malignant plasma cells" Cellular immunology., **162**(2), 248–255, 1995.
[52] A. Ogata, D. Chauhan, G. Teoh, S. P. Treon, M. Urashima, R. L. Schlossman, K. C. Anderson, "IL-6 triggers cell growth via the Ras-dependent mitogen-activated protein kinase cascade" The Journal of Immunology., **159**(5), 2212–2221, 1997.
[53] P. M. Navolanic, L. S. Steelman, J. A. McCubrey, "EGFR family signaling and its association with breast cancer development and resistance to chemotherapy (Review)" International journal of oncology., **22**(2), 237–252, 2003.
[54] F. Mosaffa, F. Kalalinia, H. Lage, J. T. Afshari, J. Behravan, "Pro-inflammatory cytokines interleukin-1 beta, interleukin 6, and tumor necrosis factor-alpha alter the expression and function of ABCG2 in cervix and gastric cancer cells" Molecular and cellular biochemistry., **363**(1-2), 385–393, 2012.
[55] F. Balkwill, "Tumour necrosis factor and cancer" Nature Reviews Cancer., **9**(5), 361–371, 2009.
[56] J. A. McCubrey, L. S. Steelman, W. H. Chappell, S. L. Abrams, E. W.T. Wong, F. Chang, B. Lehmann, D. M. Terrian, M. Milella, A. Tafuri, "Roles of the Raf/MEK/ERK pathway in cell growth, malignant transformation and drug resistance" Biochimica et Biophysica Acta (BBA)-Molecular Cell Research., **1773**(8), 1263–1284, 2007.
[57] P. J. Morin, "Drug resistance and the microenvironment" Drug Resistance Updates., **6**(4), 169–172, 2003.
[58] G. J. Peters, C. L. van der Wilt, C. J. van Moorsel, Kroep, JR, Am Bergman, S. P. Ackland, "Basis for effective combination cancer chemotherapy with antimetabolites" Pharmacology & therapeutics., **87**(2), 227–253, 2000.
[59] P. D. Cole, J. A. Zebala, B. A. Kamen, "Antimetabolites" Drug Discovery Today: Therapeutic Strategies., **2**(4), 337–342, 2006.
[60] J. Zhu, H. Yamane, W. E. Paul, "Differentiation of effector CD4 T cell populations" Annual review of immunology., **28**), 445–489, 2009.
[61] Z.-C. Ding, G. Zhou, "Cytotoxic chemotherapy and CD4" Clinical and Developmental Immunology., **2012**, 2012.
[62] A. Le Bon, D. F. Tough, "Links between innate and adaptive immunity via type I interferon" Current opinion in immunology., **14**(4), 432–436, 2002.
[63] G. Powis, "Signalling pathways as target for anticancer drug development" Pharmacology & therapeutics., **62**(1-2), 57–95, 1994.
[64] V. Asati, D. K. Mahapatra, S. K. Bharti, "PI3K/Akt/mTOR and Ras/Raf/MEK/ERK signaling pathways inhibitors as anticancer agents" European journal of medicinal chemistry., **109**), 314–341, 2016.
[65] X. Wang, H. W. Cheung, A. C. Chun, D.-Y. Jin, Y.-C. Wong, "Mitotic checkpoint defects in human cancers and their implications to chemotherapy" Frontiers in bioscience: a journal and virtual library., **13**), 2103–2114, 2007.
[66] C. Criscitiello, G. Viale, L. Gelao, A. Esposito, M. de Laurentiis, S. de Placido, M. Santangelo, A. Goldhirsch, G. Curigliano, "Crosstalk between bone niche and immune system: osteoimmunology signaling as a potential target for cancer treatment" Cancer treatment reviews., **41**(2), 61–68, 2015.
[67] M. L. Cheng, L. Fong, "Effects of RANKL-targeted therapy in immunity and cancer" Frontiers in oncology., **3**), 329, 2014.
[68] S. H. Lee, Y. Choi, "Communication between the skeletal and immune systems" Osteoporosis and Sarcopenia., **1**(2), 81–91, 2015.
[69] T. Kouzarides, "Chromatin modifications and their function" Cell., **128**(4), 693–705, 2007.
[70] Q. Liu, J. Huang, N. Zhou, Z. Zhang, A. Zhang, Z. Lu, F. Wu, Y.-Y. Mo, "LncRNA loc285194 is a p53-regulated tumor suppressor" Nucleic acids research., **41**(9), 4976–4987, 2013.
[71] Q. Ji, X. Liu, X. Fu, L. Zhang, H. Sui, L. Zhou, J. Sun, J. Cai, J. Qin, J. Ren, "Resveratrol inhibits invasion and metastasis of colorectal cancer cells via MALAT1 mediated Wnt/β-catenin signal pathway" PloS one., **8**(11), e78700, 2013.
[72] R. M. Mader, M. Müller, G. G. Steger, "Resistance to 5-fluorouracil" General Pharmacology: The Vascular System., **31**(5), 661–666, 1998.
[73] L.-C. Wang, C. Y. Okitsu, E. Zandi, "Tumor necrosis factor α-dependent drug resistance to purine and pyrimidine analogues in human colon tumor cells mediated through IKK" Journal of Biological Chemistry., **280**(9), 7634–7644, 2005.
[74] K. S. Chan, C. G. Koh, H. Y. Li, "Mitosis-targeted anti-cancer therapies" Cell death & disease., **3**(10), e411, 2012.
[75] B. Y. EARL, W. SUTHERLAND, T. W. RALL, "Fractionation and characterization of a cyclic adenine ribonucleotide formed by tissue particles, 1958.
[76] P. J. Rogue, J.-P. Humbert, A. Meyer, S. Freyermuth, M.-M. Krady, A. N. Malviya, "cAMP-dependent protein kinase phosphorylates and activates nuclear Ca2+-ATPase" Proceedings of the National Academy of Sciences., **95**(16), 9178–9183, 1998.
[77] M. David, E. Petricoin, A. C. Larner, "Activation of protein kinase A inhibits interferon induction of the Jak/Stat pathway in U266 cells" Journal of Biological Chemistry., **271**(9), 4585–4588, 1996.
[78] S. J. Cook, F. McCormick, "Inhibition by cAMP of Ras-dependent activation of Raf" SCIENCE-NEW YORK THEN WASHINGTON-., **262**), 1069, 1993.
[79] E. K. Rowinsky, "Signal events" The Oncologist., **8**(Supplement 3), 5–17, 2003.
[80] T. J. Mitchison, "The proliferation rate paradox in antimitotic chemotherapy" Molecular biology of the cell., **23**(1), 1–6, 2012.
[81] S. a. Guo, L. A. DiPietro, "Factors affecting wound healing" Journal of dental research., **89**(3), 219–229, 2010.
[82] O. Martinez-Maza, J. S. Berek, "Cancer Immunology" The Global Library of Women's Medicine, 2009.
[83] J. B. Mumm, J. Emmerich, X. Zhang, I. Chan, L. Wu, S. Mauze, S. Blaisdell, B. Basham, J. Dai, J. Grein, "IL-10 elicits IFNγ-dependent tumor immune surveillance" Cancer cell., **20**(6), 781–796, 2011.
[84] M. W.L. Teng, P. K. Darcy, M. J. Smyth, "Stable IL-10: a new therapeutic that promotes tumor immunity" Cancer cell., **20**(6), 691–693, 2011.
[85] M. A. Izquierdo, J. J. Neefjes, A. E. Mathari, M. J. Flens, G. L. Scheffer, R. J. Scheper, "Overexpression of the ABC transporter TAP in multidrug-resistant human cancer cell lines" British journal of cancer., **74**(12), 1961, 1996.
[86] S. Rafii, D. Lyden, R. Benezra, K. Hattori, B. Heissig, "Vascular and haematopoietic stem cells: novel targets for anti-angiogenesis therapy?" Nature Reviews Cancer., **2**(11), 826–835, 2002.
[87] J. E. Ohm, D. I. Gabrilovich, G. D. Sempowski, E. Kisseleva, K. S. Parman, S. Nadaf, D. P. Carbone, "VEGF inhibits T-cell development and may contribute to tumor-induced immune suppression" Blood., **101**(12), 4878–4886, 2003.
[88] A. Kumanogoh, S. Marukawa, K. Suzuki, N. Takegahara, C. Watanabe, E. Ch'ng, I. Ishida, H. Fujimura, S. Sakoda, K. Yoshida, "Class IV semaphorin Sema4A enhances T-cell activation and interacts with Tim-2" Nature., **419**(6907), 629–633, 2002.
[89] T. Toyofuku, M. Yabuki, J. Kamei, M. Kamei, N. Makino, A. Kumanogoh, M. Hori, "Semaphorin-4A, an activator for T-cell-mediated immunity, suppresses angiogenesis via Plexin-D1" The EMBO journal., **26**(5), 1373–1384, 2007.
[90] O. T. Brustugun, "Hypoxia as a cause of treatment failure in non-small cell carcinoma of the lung" Seminars in radiation oncology, 2014.
[91] S. A. Jones, J. Scheller, S. Rose-John, "Therapeutic strategies for the clinical blockade of IL-6/gp130 signaling" The Journal of clinical investigation., **121**(9), 3375–3383, 2011.
[92] N. Heuzé-Vourc'h, M. Liu, H. Dalwadi, F. E. Baratelli, L. Zhu, L. Goodglick, M. Põld, S. Sharma, R. D. Ramirez, J. W. Shay, "IL-20, an anti-angiogenic cytokine that inhibits COX-2 expression" Biochemical and biophysical research communications., **333**(2), 470–475, 2005.
[93] S. P. Cole, G. Bhardwaj, J. H. Gerlach, J. E. Mackie, C. E. Grant, K. C. Almquist, A. J. Stewart, E. U. Kurz, A. Duncan, R. G. Deeley, "Overexpression of a transporter gene in a multidrug-resistant human lung cancer cell line" SCIENCE-NEW YORK THEN WASHINGTON-., **258**), 1650, 1992.
[94] M. R. Zaidi, G. Merlino, "The two faces of interferon-γ in cancer" Clinical Cancer Research., **17**(19), 6118–6124, 2011.

[95] K. E. de Visser, A. Eichten, L. M. Coussens, "Paradoxical roles of the immune system during cancer development" Nature Reviews Cancer., **6**(1), 24–37, 2006.
[96] A. Weinandy, M. D. Piroth, A. Goswami, K. Nolte, B. Sellhaus, J. Gerardo-Nava, M. Eble, S. Weinandy, C. Cornelissen, H. Clusmann, "Cetuximab induces eme1-mediated DNA repair: a novel mechanism for cetuximab resistance" Neoplasia., **16**(3), 207–220, 2014.
[97] Z. Li, L. Chen, Z. Qin, "Paradoxical roles of IL-4 in tumor immunity" Cellular & molecular immunology., **6**(6), 415, 2009.
[98] J. F. Ferreiro, L. Rouhigharabaei, H. Urbankova, J.-A. van der Krogt, L. Michaux, S. Shetty, L. Krenacs, T. Tousseyn, P. de Paepe, A. Uyttebroeck, "Integrative genomic and transcriptomic analysis identified candidate genes implicated in the pathogenesis of hepatosplenic T-cell lymphoma" PloS one., **9**(7), e102977, 2014.
[99] I. Kryczek, S. Wei, W. Szeliga, L. Vatan, W. Zou, "Endogenous IL-17 contributes to reduced tumor growth and metastasis" Blood., **114**(2), 357–359, 2009.
[100] K. Karri, D. R. Bastola, "Immune based prognostic biomarkers for multiple anticancer therapies in lung adenocarcinoma" in Bioinformatics and Biomedicine (BIBM), Shenzen,China, 2016.

Appendix I

Table 1. TABF-FAGO terms upregulated in CR for alkylating agents (LUAD)

TABF Terms	FAGO terms upregulated in CR for Alkylating Agents
Cell cytotoxicity	positive regulation of leukocyte mediated cytotoxicity, regulation of leukocyte mediated cytotoxicity, natural killer cell mediated cytotoxicity, leukocyte activation involved in immune response, positive regulation of leukocyte activation, leukocyte differentiation, regulation of leukocyte differentiation, positive regulation of leukocyte mediated immunity, regulation of leukocyte mediated immunity, leukocyte proliferation, positive regulation of leukocyte differentiation, myeloid leukocyte differentiation, granulocyte migration, neutrophil chemotaxis
Marcophage Activation	macrophage differentiation
Immune Response and Myeloid Cell Activation	positive regulation of immune response, myeloid dendritic cell activation, activation of immune response, myeloid cell activation involved in immune response, mast cell activation, positive regulation of immune effector process, regulation of production of molecular mediator of immune response
Regulation of T cell and Lymphocyte Activation	positive regulation of lymphocyte differentiation, positive regulation of T cell differentiation, positive regulation of alpha-beta T cell activation, lymphocyte activation involved in immune response, alpha-beta T cell differentiation, CD4-positive, alpha-beta T cell differentiation involved in immune response, regulation of alpha-beta T cell differentiation, activated T cell proliferation, thymic T cell selection, regulation of T cell differentiation, regulation of lymphocyte activation, positive regulation of lymphocyte activation, T cell activation, lymphocyte differentiation, T cell differentiation, regulation of lymphocyte differentiation, regulation of T cell proliferation, positive regulation of T cell proliferation, T cell proliferation, regulation of alpha-beta T cell activation, T cell costimulation, positive regulation of lymphocyte mediated immunity, regulation of lymphocyte mediated immunity, lymphocyte mediated immunity
B cell Proliferation and Activation	regulation of B cell activation, B cell mediated immunity, B cell activation, B cell proliferation
Signaling pathways	Fc-epsilon receptor signaling pathway, Fc receptor signaling pathway, immune response-regulating cell surface receptor signaling pathway, immune response-regulating signaling pathway, immune response-activating signal transduction, antigen receptor-mediated signaling pathway, immune response-activating cell surface receptor signaling pathway, cell-cell signaling, cell surface receptor signaling pathway, cell-activation, cell migration, cell chemotaxis, Chemokine signaling pathway
Negative regulation of leukocyte and lymphocyte activation	negative regulation of leukocyte activation, negative regulation of lymphocyte activation, negative regulation of leukocyte differentiation
Cytokine activity	cytokine activity, cytokine receptor activity, cytokine receptor binding, Cytokine-cytokine receptor interaction, cellular response to cytokine stimulus, regulation of cytokine production
Adaptive Immune Response and Immunoglobulin Production	response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains, adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains positive regulation of adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains, regulation of immunoglobulin mediated immune response

Table 2. TABF-FAGO terms upregulated in CR for alkylating agents (COAD)

TABF Terms	FAGO terms Upregulated in CR for Alkylating Agent
ncRNA processing	ncRNA metabolic process, RNA processing, RNA splicing, mRNA processing, gene silencing by RNA
mitochondrion organization	mitochondrial transport, mitochondrial transmembrane transport, mitochondrial electron transport, NADH to ubiquinone, mitochondrial respiratory chain complex I assembly, mitochondrial ATP synthesis coupled proton transport
activation of immune response	myeloid cell activation involved in immune response, adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains, negative regulation of production of molecular mediator of immune response
T cell differentiation, proliferation, activation, lymphocyte aggregation	regulation of lymphocyte mediated immunity, lymphocyte costimulation, T cell mediated immunity, regulation of T cell proliferation, positive regulation of alpha-beta T cell activation, regulation of CD4-positive, alpha-beta T cell activation
B cell homeostasis	regulation of B cell proliferation, positive regulation of B cell proliferation, B cell proliferation

Table 3. TABF-FAGO terms upregulated in PD for alkylating agents (LUAD)

TABF Terms	FAGO Terms upregulated in PD for Alkylating agents
G1/S cell cycle regulation	regulation of cell cycle G1/S phase transition
Response to DNA damage	DNA-templated transcription, initiation intrinsic apoptotic signaling pathway in response to DNA damage, response to cycloheximide, transcription elongation from RNA polymerase I promoter
Leukocyte chemotaxis	positive regulation of leukocyte chemotaxis
Defense response	defense response, antibacterial humoral response, wound healing
Inflammation	inflammatory response
Complement activation	complement activation, alternative pathway
KEGG pathways in cancer	viral carcinogenesis, cancer carcinogenesis
Glutathione metabolism	Glutathione metabolism

Table 4. TABF-FAGO terms upregulated in PD for alkylating agents (COAD)

TABF Terms	FAGO terms Upregulated in PD for Alkylating Agents
negative regulation of immune response	negative regulation of production of molecular mediator of immune response, negative regulation of adaptive immune response, negative regulation of immune effector process
Interleukin production	interleukin-1 beta production, interleukin-1, 2,4 ,6 10 production
angiogenesis	positive regulation of angiogenesis, regulation of angiogenesis, blood vessel development
MAPK cascade	regulation of MAPK cascade, positive regulation of MAPK cascade
ERK1 and ERK2 cascade	positive regulation of ERK1 and ERK2 cascade,
cell proliferation/ differentiation	positive regulation of cell differentiation, positive regulation of mononuclear cell proliferation, cell differentiation, regulation of cell proliferation
cytokine production	regulation of cytokine biosynthetic process, cytokine-mediated signaling pathway, cellular response to cytokine stimulus, negative regulation of cytokine secretion
Inflammatory response	acute inflammatory response, chronic inflammatory response
negative regulation of wound healing	response to wounding, negative regulation of response to wounding, regulation of wound healing
interferon-gamma production	regulation of interferon-gamma production, positive regulation of interferon-gamma production
positive regulation of tumor necrosis factor production	chemokine production, positive regulation of tumor necrosis factor production, regulation of tumor necrosis factor superfamily cytokine production

Table 5. TABF-FAGO terms upregulated in CR for antimetabolites (LUAD)

TABF Terms	FAGO terms Upregulated in CR for Antimetabolite
Cell Cytotoxicity	leukocyte differentiation, negative regulation of leukocyte migration, leukocyte activation involved in immune response, monocyte chemotaxis, myeloid leukocyte differentiation, positive regulation of leukocyte migration, regulation of leukocyte chemotaxis, positive regulation of leukocyte chemotaxis, positive regulation of leukocyte activation, regulation of production of molecular mediator of immune response, regulation of leukocyte mediated immunity, positive regulation of
Cytokine and Chemokine Regulation	cytokine receptor activity, cytokine activity, cytokine receptor binding, chemokine receptor activity, Cytokine-cytokine receptor interaction, Chemokine signaling pathway
Adaptive Immune Response	regulation of adaptive immune response, regulation of adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains,
Regulation of T cell and Lymphocyte Activation	T cell activation involved in immune response, T cell differentiation involved in immune response, alpha-beta T cell differentiation involved in immune response, regulation of lymphocyte mediated immunity, regulation of T-helper 1 type immune response, stem cell differentiation
Cell Signaling	cell surface receptor signaling pathway, G-protein coupled receptor signaling pathway, cell-cell signaling, Phospholipase D signaling pathway, PI3K-Akt signaling pathway, immune response-
KEGG Signaling pathway	Rap1 signaling pathway, Calcium signaling pathway, G-protein coupled receptor signaling pathway, Phospholipase D signaling pathway, PI3K-Akt signaling pathway
System development	cardiovascular system development, skeletal system development

Table 6. TABF-FAGO terms upregulated in CR for antimetabolite (COAD)

TABF Terms	FAGO terms Upregulated in CR for Antimetabolite
system development	nervous system development, anatomical structure morphogenesis, embryonic morphogenesis, embryonic limb morphogenesis, regulation of growth, organ growth, aorta development
histone modification	histone lysine methylation, histone H3-K9 modification, histone H3-K9 methylation, histone H3-K36 demethylation
T cell proliferation activation, leukocyte differentiation, lymphocyte differentiation	leukocyte cell-cell adhesion, leukocyte aggregation, lymphocyte aggregation, positive regulation of leukocyte proliferation, T cell proliferation, lymphocyte activation involved in immune response, leukocyte activation involved in immune response, positive regulation of leukocyte migration, positive regulation of lymphocyte proliferation
positive regulation of immune response	myeloid cell activation involved in immune response, regulation of production of molecular mediator of immune response, positive regulation of production of molecular mediator of immune response, positive regulation of innate immune response, positive regulation of immune effector process

Table 7. TABF-FAGO terms upregulated in PD for antimetabolites (LUAD)

TABF Terms	FAGO terms Upregulated in PD for Antimetabolite
RNA Processing	RNA processing, RNA 3'-end processing
ncRNA processing	ncRNA metabolic process

Table 8. TABF-FAGO terms upregulated in PD for antimetabolite (COAD)

TABF Terms	FAGO terms Upregulated in PD for Antimetabolite
JAKSTAT cascade	positive regulation of JAK-STAT cascade, tyrosine phosphorylation of STAT protein, regulation of tyrosine phosphorylation of STAT protein
G1/s , G2/M phase	cell cycle G1/S phase transition, cell cycle phase transition, mitotic cell cycle phase transition
Interleukin-1 production	interleukin-1 beta production, regulation of interleukin-1 secretion, positive regulation of interleukin-1 production
Inflammtion	inflammatory response, acute inflammatory response, production of molecular mediator involved in inflammatory response, regulation of inflammatory response
ERK cascade	regulation of ERK1 and ERK2 cascade
IFN production	response to interferon-gamma, interferon-gamma production
cytokine production	regulation of cytokine production, response to cytokine, cytokine secretion, cellular response to cytokine stimulus, cytokine-mediated signaling pathway
cellular response to tumor necrosis factor	Tumor necrosis factor superfamily cytokine production

Table 9. TABF-FAGO terms upregulated in CR for mitotic inhibitor (LUAD)

TABF Terms	FAGO terms Upregulated in CR for Mitotic Inhibitor
KEGG signaling pathway	signaling receptor activity, MAPK signaling pathway, Calcium signaling pathway, NF-kappa B signaling pathway, Phosphatidylinositol signaling system, Jak-STAT signaling pathway, cell surface receptor signaling pathway, G-protein coupled receptor signaling pathway
B cell activation	B cell proliferation, B cell activation,B cell differentiation
Innate immune response	positive regulation of immune response, activation of immune response, regulation of innate immune response
regulation of T cell and lymphocyte activation	T cell activation,T cell differentiation, alpha-beta T cell activation, alpha-beta T cell differentiation,CD4-positive, alpha-beta T cell differentiation,T cell selection, activated T cell proliferation,T cell differentiation in thymus, regulation of T cell activation, positive regulation of T cell activation, regulation of T cell differentiation, positive regulation of T cell differentiation, regulation of alpha-beta T cell activation, lymphocyte proliferation, lymphocyte activation involved in immune response, regulation of alpha-beta T cell differentiation, positive regulation of alpha-beta T cell activation, response to interferon-gamma, positive regulation of immune effector process, T cell proliferation, regulation of T cell proliferation, positive regulation of T cell proliferation, lymphocyte costimulation, T cell activation involved in immune response
Regulation of cAMP	cyclic nucleotide metabolic process, cyclic purine nucleotide metabolic process, cAMP metabolic process, cAMP biosynthetic process, regulation of cAMP biosynthetic process, positive regulation of cyclic nucleotide metabolic process, positive regulation of cAMP metabolic process
Inflammatory response	Inflammatory response
Cytokine Production	response to cytokine, Cytokine-cytokine receptor interaction, regulation of cytokine production, macrophage cytokine production, positive regulation of macrophage cytokine production, macrophage cytokine production, positive regulation of macrophage cytokine production

Table 10. TABF-FAGO terms upregulated in PD for mitotic inhibitors (LUAD)

TABF Terms	FAGO terms Upregulated in PD for Mitotic Inhibitor
Negative regulation of cell death	negative regulation of cell death, negative regulation of apoptotic process, negative regulation of programmed cell death
Mitotic Checkpoint and DNA damage response	DNA damage checkpoint, mitotic DNA damage checkpoint, negative regulation of cell cycle G1/S phase transition, mitotic G1 DNA damage checkpoint

Table 11. TABF-FAGO terms upregulated in CR for Anti-angiogenesis (LUAD)

TABF Terms	FAGO terms Upregulated in CR for Anti-angiogenesis
Negative regulation of inflammatory response	regulation of inflammatory response, negative regulation of inflammatory response
macrophage activation	macrophage differentiation
Negative regulation of angiogenesis	regulation of vasculature development, regulation of angiogenesis, negative regulation of angiogenesis
myeloid cell activation	myeloid cell differentiation , myeloid leukocyte differentiation, regulation of myeloid cell differentiation
regulation of T cell and lymphocyte activation	T cell activation, regulation of T cell activation, T cell differentiation, lymphocyte differentiation, regulation of lymphocyte activation, regulation of lymphocyte activation, lymphocyte proliferation
Apoptosis	cell death, apoptotic process
Cell signaling	cell surface receptor signaling pathway, cell-cell signaling, cell-substrate adhesion, cell-cell adhesion
Organ morphogenesis	skeletal system development, cardiovascular system development, heart development, embryo development, organ morphogenesis
B cell proliferation and activation	B cell activation, B cell differentiation
Regulation of interleukin 6	interleukin-6 production, tumor necrosis factor production, regulation of interleukin-6 production
Regulation of cytokines	cytokine activity, cytokine receptor binding, cytokine metabolic process, Cytokine-cytokine receptor interaction
Wounding	Response to wounding
Hemopoiesis	hemopoiesis, regulation of hemopoiesis, negative regulation of hypoxia
Regulation of leukocyte	myeloid leukocyte differentiation, regulation of myeloid cell differentiation, regulation of myeloid leukocyte differentiation, negative regulation of leukocyte differentiation, regulation of leukocyte chemotaxis

Table 12. TABF-FAGO terms upregulated in PD for anti-angiogenesis (LUAD)

TABF Terms	FAGO terms Upregulated in PD for Anti-angiogenesis
Angiogenesis	Angiogenesis, anatomical structure morphogenesis
Growth Signaling Pathway	epidermal growth factor receptor signaling pathway, vascular endothelial growth factor receptor signaling pathway
Negative regulation of apoptosis	negative regulation of cell death, negative regulation of programmed cell death, cell division, cellular macromolecular complex assembly, chromosome organization, cellular macromolecule catabolic process, cell cycle, cell cycle process, chromosome segregation, cellular developmental process, cell migration, regulation of cellular process
Antigen Processing & Presentation	antigen processing and presentation of endogenous peptide antigen, antigen processing and presentation of exogenous peptide antigen via MHC class I, TAP-independent, antigen processing and presentation of exogenous peptide antigen, antigen processing and presentation of exogenous peptide antigen via MHC class II, antigen processing and presentation of peptide antigen via MHC class I, antigen processing and presentation of exogenous peptide antigen via MHC class I, antigen processing and presentation of endogenous peptide antigen
Regulation of innate immune response	activation of immune response, immune response-regulating signaling pathway, regulation of innate immune response, positive regulation of innate immune response, activation of innate immune response, organ or tissue specific immune response
DNA repair	DNA ligation, DNA metabolic process, DNA repair, translation
Defense Response	response to radiation, cellular response to stress, defense response to Gram-positive bacterium, response to other organism
KEGG pathways in cancer	Pathways in cancer

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 788-795 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Interdisciplinarity to Integrate Knowledge in Engineering

Stella Abreu[1,3], Amélia Caldeira*[1,3,4], Alexandra R. Costa[2], Tiago Gomes[2], Luís A. C. Roque[1,5]

[1]*School of Engineering Polytechnic of Porto (ISEP-IPP), Mathematics Department,* 4249 - 015*, Portugal*

[2]*School of Engineering Polytechnic of Porto (ISEP-IPP), Planning and Management Department,* 4249 - 015*, Portugal*

[3]*LEMA – Engineering Mathematical Laboratory– ISEP,* 4249 - 015*, Portugal*

[4]*SYSTEC- ISR – Research Center for Systems and Technologies University of Porto, 4200-464, Portugal*

[5]*LIADD-INESC-TEC Campus da FEUP, 4200-465, Portugal*

ARTICLE INFO

Article history:
Received: 05 April, 2017
Accepted: 22 May, 2017
Online: 16 June, 2017

Keywords:
Interdisciplinarity
soft skills
team work
mathematics teaching
working methods

ABSTRACT

This paper is an extension of work originally presented at the 2nd International Conference of the Portuguese Society for Engineering Education and aims to describe an interdisciplinarity teaching experiment involving three subjects of the scientific area of Mathematics and a fourth one in the area of Management. Using only one project, the students developed skills, in an integrated way, in the fields of the subjects involved. The structure of the project is described in detail. It is shown how the knowledge obtained in the different subjects is needed and how it connects together to answer the proposed challenges. We report the progress of the students' work, the main difficulties and the skills developed during this process. We conclude with a reflection on the main problems and gains that may arise in similar projects.

1. Introduction

This paper is an extension of work [1], originally presented at the 2nd International Conference of the Portuguese Society for Engineering Education.

The Bologna reform aims to make the European Higher Education more comparable and compatible, unifying three different levels of study, for Bachelors, Master and Doctoral awards, across all member countries [2, 3].

This reform seeks to promote the student-centred approaches and the effective implementation of more active learning practices [4].

According to these authors, the Bologna process emphasizes that efforts should be made to include learning activities with "meaning" for students, providing them additional motivation. These authors state that one of the methodologies to achieve these objectives is based on interdisciplinary project approaches [4].

This paper aims to describe an interdisciplinarity teaching experiment involving three subjects of the scientific area of Mathematics and a fourth one in the area of Management, in a Portuguese school of engineering.

Societal, environmental, economic, and philosophical problems are often so complex that it is impossible to fully understand them from a single perspective. Multiple viewpoints can help to solve these complex problems and challenges by converging synergistic team efforts [5].

In line with this idea, many employers recognize the need to hire graduates that are able to work as members of a team as well as being able to understand certain core disciplinary competencies and to adapt to different contexts. Because of that, in recent times the theme of interdisciplinarity has gained popularity in different circles. Despite the skepticism in some education environments, the supporters of its application to school contexts have been increasing, especially with its introduction in university curricula and research agenda [5, 6].

In this field, Klein, quoted by Jacobs [5], makes an important clarification of concepts, presenting a distinction between multi- inter- and trans-disciplinarity that we highlight below: The multidisciplinarity is made by a juxtaposition of knowledge,

*Amélia Caldeira, Rua Dr. Bernardino de Almeida, 431, 4249-015 – Porto (Portugal), acd@isep.ipp.pt

methodologies or information from different disciplines even though the knowledge structure and the identity of the original disciplines remains unquestionable; on the other hand, interdisciplinary approaches emphasize integration and interaction, promoting the disciplinary transformation at methodological and theoretical levels; finally, in transdisciplinary approaches, research questions and practices are framed by problems arising from the life-world and studied in different points of view [5, 7].

"*Interdisciplinarity is a means to integrate knowledge and methods in the interest of problem solving*" [6]. Crossing disciplinary boundaries is particularly important for a future engineer because she/he will be called to solve complex problems. This requires not only crossing boundaries horizontally (across subjects) but also vertically (across experts, policymakers, practitioners, and the public) [8, 9, 10].

The advantages of interdisciplinarity studies are widely accepted by teachers and researchers. Some authors refer that interdisciplinary works use the knowledge that comes from different curriculum areas that offer different perspectives on a particular problem, making the curriculum more compact and more consistent. Furthermore, interdisciplinary works provide the students with relevant, challenging and enjoyable learning experiences [5, 6].

Chettiparamb [6] refers that "*Interdisciplinary pedagogy fosters in students a sense of self-authorship and a situated, partial and perspectival notion of knowledge that they can use to respond to complex questions, issues or problems. While it necessarily entails the cultivation of the many cognitive skills such as differentiating, reconciling, and synthesizing […] it also involves much more, including the promotion of student's interpersonal and intrapersonal learning.*"

Despite this, some authors believe that collaboration is often more truly interdisciplinary in social science and humanities disciplines than in technological ones because these disciplines require that researchers work more closely to agree on methods and interpretations [11]. Notwithstanding this, interdisciplinary approaches are critical to solving the most pressing technological challenges. An evidence of this is the fact that interdisciplinary is progressively more valued by the courses' accrediting agencies [7].

In alignment with these ideas, some universities have developed some interdisciplinary design courses with the goal of improving students' abilities to operate across disciplines and improve their preparation for work [12].

This paper aims to present the results of an interdisciplinary teaching experiment performed in the Electrical Engineering – Power Systems degree of the School of Engineering – Polytechnic of Porto (ISEP). The goal of this experiment was to carry out an interdisciplinary project involving four subjects of the first year of the degree, three of which from the scientific area of Mathematics - Mathematics I (MATE1), Mathematics Laboratory I (LMAT1), Linear Algebra and Analytic Geometry (ALGAN) - and one subject in the area of Management: Working Methods in Engineering (MTENG).

The subjects concerning the first semester of the first year program as well as the respective European Credit Transfer and Accumulation System (**ECTS**) [13,14] are listed in Table I.

TABLE I. COURSE REQUIREMENTS (1ST YEAR/1ST SEMESTER)

SUBJECTS	ECTS	Contact hours
Experimental Physics	5	4
Introduction to Electrotechnics	6	4
Linear Algebra and Analytic Geometry	5	4
Mathematics I	5	4
Mathematics Laboratory I	3	2
Working Methods in Engineering	6	4

Using only one project work, we intended to create the opportunity for students to develop skills, in an integrated way, in the fields of the involved subjects. Students were challenged to perform a set of tasks by applying concepts and skills developed in the four aforementioned subjects. Specific objectives were proposed for this work in each subject.

The experiment took place during twelve weeks in the first semester of the first year of the degree.

Students were assessed by three outcomes: one report, one oral presentation with discussion and a MatLab© script. All teachers were involved in the assessment of the report and the oral presentation. The MatLab© script was mainly assessed by the Mathematics Laboratory I (LMAT1) teacher.

Students faced initial difficulties in dealing with the interdisciplinary challenge, and with developing the soft skills required in the project. The main challenges were:

- to gain experience in team work as means to solve the interdisciplinary problems;
- to acquire insight into their own behaviour and understand how it influences collaboration in the team;
- to acquire and understand their competence and realize how it may be used for the benefit of the team.

This led to considerable gains in learning, so we will make a reflection on these gains and the benefits of carrying out a single project involving different subjects. We explore the main objectives of this interdisciplinary work, as well as the methodology and its main results.

We conclude this text with a reflection on the main lessons learned as well as with some suggestions for future interdisciplinary work involving these or other scientific areas.

2. The project

Interdisciplinary projects are encouraged and very welcome in the degrees of ISEP. Using some past experiences, a group of four teachers – responsible for three subjects of the scientific area of Mathematics and a fourth one in the area of Management – proposed to the students of the Electrical Engineering – Power Systems degree an interdisciplinary work.

2.1. Organizing the project

The purpose of this project was to give an integrated view of the several subjects of Mathematics, as well as to develop in students a number of soft skills that are pointed out as fundamental by the employers of graduated students.

In interdisciplinary research, success depends largely on the personality and attitudes of researchers. Useful characteristics are:

- curiosity about, and willingness to learn from other subjects;
- flexibility, adaptability, creativity;
- an open mind to ideas coming from other subjects and experiences;
- good communication and listening skills;
- an ability to bridge the gap between theory and practice;
- a good team worker.

The following additional abilities are very important:

- understanding (not necessarily in depth) across the project's main subject domains;
- understanding application areas, namely, Electrical Engineering, for project outcomes;
- respect for other subjects and some understanding of their general principles;
- focus on practical results, to overcome differences among subjects.

These abilities are highly relevant to achieve a good project management.

In addition to this, specific objectives were proposed in each subject.

In the subject of Working Methods in Engineering, the project had as objective to develop skills in the areas of:

- Project management competences, including planning and project control;

- Write technical reports using an appropriate structure and technical language;

- Communicate in public effectively, using technical data to support the message;

- Communication. Namely developing the capacity to reach consensus, make decisions and manage conflicts.

Mathematics Laboratory I is inserted in the first semester of the first year of the Degree in Electrical Engineering - Electric Energy Systems' syllabus. It is a scientific subject of the Mathematics area that intends to provide the skills for solving typical problems in engineering algebraically and numerically, using numerical methods (see, for example, [15] and [16]) and computer algebra systems.

As main goals of this subject, it is intended that students obtain reasoning and abstraction skills in order to:

- implement methods of algebraic and numeric resolution of problems;

- model engineering problems related to the degree mathematically;

- acquire critical sense, deductive reasoning, and graphic interpretation;

- be critical in the use of computer algebra software and numeric tools;

- do the integration between the nuclear subjects of Mathematics and Electrical Engineering.

As specific goals it is intended that a student should be able to:

- implement methodologies for solving problems using the algebraic and numeric tools (Matlab, Python and/or Matlab clone);

- understand the benefits and limitations of the algebraic and numeric tools (Matlab software, Python and/or Matlab clone);

- find and implement alternative methodologies to solve the problems;

- understand the several ways in which error can be introduced in the solution of the problem;

- solve equations of one variable numerically;

- use iterative methods in finding the solution sets of linear systems;

- evaluate derivatives and integrals numerically.

The subject contents were: the software introduction and error analysis; numerical methods for solving equations; methods for solving systems of equations and numerical differentiation and integration.

Mathematics I is a traditional one variable analysis subject (see, for example, [17] and [18]) with the following objectives:

- to characterize and manipulate real-valued functions in one variable;

- to use the techniques of differential calculus;

- to understand the notion of primitive of a function and to compute the primitive of functions using several methods;

- to understand the concept of definite integral and to use it to compute areas;

- to work with series of real numbers and with Taylor series, realizing its importance and its application to Engineering problems.

Linear Algebra and Analytic Geometry aims to complement and consolidate the mathematics training of students gained during their academic life and develop the abstraction and reasoning ability of students. The subjects of ALGAN also aims the development of a mathematical language and the acquisition of critical spirit. It intends to give basic training in linear algebra and analytic geometry in order to enable the creation and solving mathematical models applied to Engineering (see [19]).

After attending the subject of ALGAN, students should be able to:

- perform basic operations on matrices, compute determinants and solve matrix equations;

- use matrices and determinants in the resolution and discussion of systems of linear equations;

- identify and generate vector fields and check linear dependence on the vectors; identify linear transformations and compute the associated matrices, eigenvalues, and eigenvectors;

- give analytical expressions for lines and planes; study intersections of lines and planes and identify geometrical spaces that satisfy some pre-determined conditions.

Teams of four students were formed in the beginning of the scholar year, in the classes of Mathematics Laboratory I (LMAT1). Students were invited to form the teams according to their preferences. There was a consensus among the students in the formation of the groups and there was no need for direct intervention of teachers in this process. Students without LMAT1, but attending one of the other three subjects, would not have an interest in participating in this project, as it needed a lot of support given in the classes of LMAT1. For these students, an alternative assessment was defined.

A total of eighty one students participated in this project. Each team received a project worksheet. The worksheets were not equal for all teams, but they proposed similar tasks.

2.2. Planning for implementation of the tasks(weekly)

An overview of the project milestones and tasks is shown in Table II. Task 1 included the study of a real-valued function of one real variable and solving some equations numerically. This was performed during the four weeks of October, the first month of the project.

A preliminary report was delivered at the end of the second week of November. This was followed by the task of solving systems of linear equations using algebraic and numerical methods. The preliminary report of this task was delivered at the end of the first week of December.

Finally, the problem involving sequences, Taylor series and symbolic and numerical integration was addressed in the second and third weeks of December. The preliminary report about this task was delivered at the end of the third week of December. The final report was delivered in the last week of December.

TABLE II. TASKS SCHEDULING (WEEKLY)

Tasks	October	November				December				January (2016)	
	During 4 weeks	1st	2nd	3rd	4th	1st	2nd	3rd	4th	1st	2nd
MATE 1 Task1: Real-valued function of a real variable											
LMAT 1 Task1: Solve equations numerically											
ALGAN Task2:Solve systems of linerar equations											
LMAT 1 Task2:Solve systems of linerar equations numerically											
MATE 1 Task3: Sequences/ Series/Integration.											
LMAT1 Task 3: Numerical integration											
MTENG Preparation of the Report											
Results/ display resolution: task 1											
Results/display resolution: task 2											
Results / display resolution: task3											
Report delivery											
Presentation/Discussion											

2.3. Main tasks developed in the project

The tasks proposed to the students include three groups of questions. The first task to be developed in this project involved real-valued functions of one real variable. The voltages across the capacitor component were studied. Thus, the charge/discharge of the capacitor, $q(t)$, in the RLC circuit (Fig. 1) was given. The analytic expression involved various exponential functions, for example,

$$q(t)=\frac{57024\left(e^{\frac{\sqrt{43}\,t}{6}}+e^{-\frac{\sqrt{43}\,t}{6}}\right)+\frac{374112}{\sqrt{43}}\left(e^{\frac{\sqrt{43}\,t}{6}}-e^{-\frac{\sqrt{43}\,t}{6}}\right)}{6}e^{-\frac{7t}{6}}$$

$$-(192\,t+19008)\,e^{-\frac{t}{12}}.$$

$$v_C(t)=\frac{q(t)}{C} \qquad v_L(t)=L\frac{di}{dt}.$$

$$i=\frac{dq}{dt}.$$

The values for resistor, inductor, and capacitor are given, respectively, by

$$R = 7\Omega \;\; L = 3H \qquad C = 2F.$$

Some questions concerning real-valued functions were addressed at the beginning:

$$v_{LC}(t) = v_L(t) + v_C(t)$$

60%

0.05

0.0005

In the second task, an electrical circuit was given (see Figure 2). The circuit has different meshes. Using the Kirchhoff's laws for electrical circuits the students formulated the linear equations system to determine the currents in the circuit meshes.

A linear equations system was obtained. Then, the students were challenged to solve the system using algebraic methods such as Gaussian elimination and the Cramer's rule. After the use of these two algebraic methods, the students solved numerically the system of linear equations, using the iterative Jacobi and Gauss-Sidel methods. Here, it was provided the opportunity of comparison between the results obtained from algebraic and numerical methods. Furthermore, this approach gave to the students the opportunity to compare the performance of the two numerical methods.

The last task concerned an integral calculus problem where symbolic integration was not possible. Therefore, the problem should be solved using Taylor series/polynomials, in order to obtain an approximation of the integral. In addition, the Simpson and trapezoidal rules were proposed to obtain the integral, using numerical quadrature. We focus that these tasks led to the simultaneous use of algebraic and numerical methods. The function was $l(x) = 3\, e^{-\frac{x^2}{4}}$ and the problem was to calculate the definite integral

$$I = \int_{-2}^{2} l(x)dx.$$

At first, it was computed the Taylor series of l and, next, the numerical series that allows computing the value of I. The convergence of the numerical series was also studied.

After that, other two approximations of I were calculated:

- on the one hand, approximating $l(x)$ by its Taylor polynomial of degree n (n specified) and the value of the corresponding integral;

- on the other hand, using the trapezoidal and Simpson's rules.

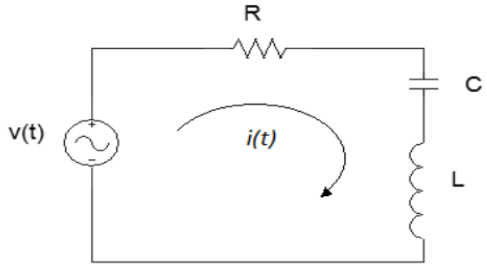


Figure 1: RLC circuit analyzed in Task 1.

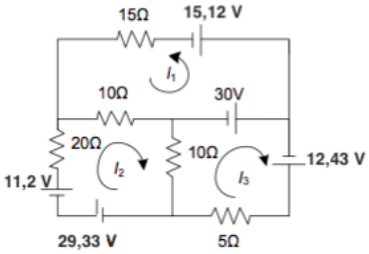


Figure 2: Electrical circuit solved in Task 2.

2.4. Interdisciplinarity

The interdisciplinarity present in the project proposed to students is shown in the flowchart in Fig.3. The linear equation system obtained from Kirchhoff's voltage law application is commonly addressed on LMAT1 and ALGAN. Algebraic methods, such as Gaussian elimination and Cramer's rule, had been studied in ALGAN. Numerical methods, such as Jacobi and Gauss-Sidel iterative methods, were implemented in LMAT1. The issue of real-valued functions of one real variable, including the voltages across the capacitor component or the charge/discharge of the capacitor in the RLC circuit, was addressed in MATE1 and in LMAT1.

The study of real-valued functions was done analytically in the MATE1 subject. Some equations were solved numerically in LMAT1. The numerical methods used were the bisection and Newton–Raphson methods. Another topic involves the integration of functions without any symbolic primitive. This problem was addressed in MATE1 and LMAT1. Taylor series/polynomial were used to find an approximation of the integrand function and so to obtain an approximation of the integral. In LMAT1, the Simpson and trapezoidal rules were used for numerical integration. The management and guidance for performing the reports and presentations were carried out in the MTENG subject. This general issue had also the contribution of all participant subjects.

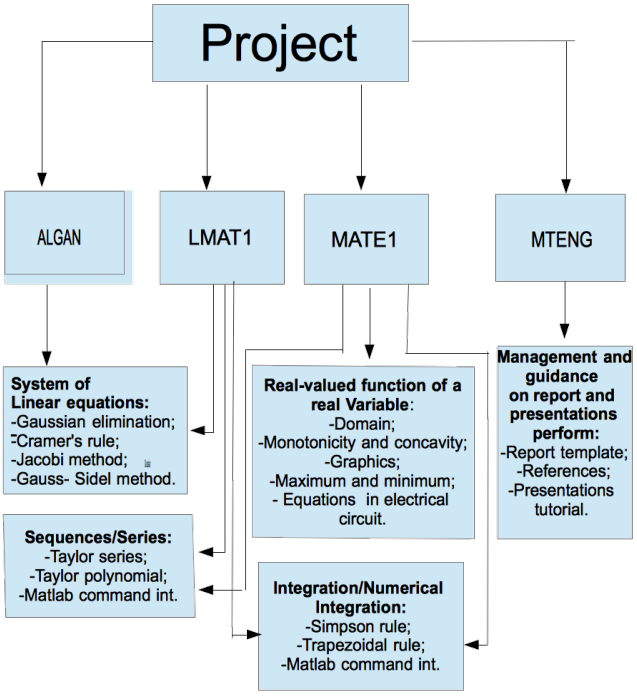


Figure 3: Topics and problems split into four different subjects

Prerequisite skills needed for the execution of the project are rather minimal. Students are required to have Mathematics education of secondary school. They should have some knowledge of basic electrical circuits. Some programming experience in Matlab or other higher-level language is helpful, but not required.

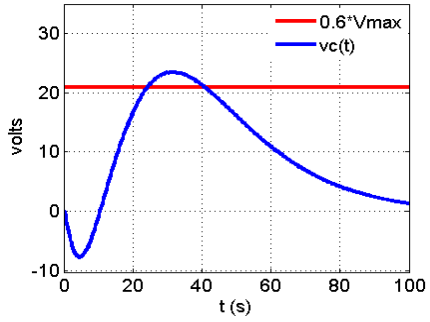


Figure 4: Graphic interpretation used in Task 1.

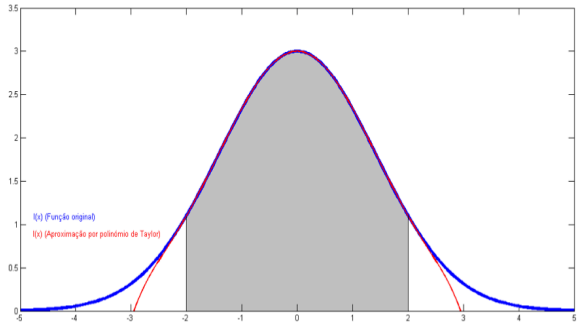


Figure 5: Numerical Integration.

We also highlight the use of the software MatLab© and its symbolic and numeric toolboxes to support the study. Furthermore, the graphical interpretation done by the students was improved by exploiting the graphical capabilities of the software. For instance, the graph presented in Fig. 4 (obtained with MatLab©) provides a better insight of the equation to be solved numerically while the graphic presented in Fig. 5 provides an insight concerning the integral to be computed.

The students were faced with questions where it is necessary to choose between numeric and algebraic approaches. They should analyze the advantages and disadvantages of each other.

2.5. Using the Moodle platform

The use of a communication platform between teachers and students is essential for work coordination. There are several platforms available, but at ISEP it is implemented the Moodle platform, so we used it.

The *Moodle* platform was used for communication with students, providing supporting material and for the delivery of the final work. A proper space, accessible to all persons involved, was created in this platform to help the management of the process of delivery of the final work and its logistics.

2.6. Transferability to other interdisciplinary projects

This work emerged from a group of teachers' desire and the conviction that this kind of project brings significant gains for students. These are challenged to integrate the knowledge that is often presented to them in a disintegrated way giving the idea of knowledge fragmentation. But this type of work presents challenges, not only for students, teachers are also challenged to think globally, to overcome the limits of their subject, to integrate its specific knowledge with other areas and to coordinate their work with other teachers.

Everyone involved in this experience - teachers and students – shares the opinion that it should be repeated and, furthermore, there should be conditions to transfer it to other contexts and to other areas. In order to facilitate this transferability, we leave some practices that we consider essential to the success of future experiments in this context:

1. Project Planning - Phase essential to the success of the project. It should involve communication between teachers from different subjects in order to i) describe goals for each subject, ii) define tasks to be developed by students, iii) define moments and assessment tools and respective criteria, iv) prepare the project worksheet and v) set a plan of regular control meetings.

2. Project presentation to students - The project must be presented to students in detail, in the context of class, through the delivery of a project worksheet which should include i) project goals, ii) detailed tasks and deadlines iii) materials and tools necessary for the tasks implementation and iv) assessment moments, tools and respective criteria.

3. Teams definition - teams should be defined at the beginning of the work by the students.

4. Teachers monitoring - Teachers should accompany the group throughout the implementation of the project, acquainting themselves from problems that arise and supporting students to solve it.

5. Coordination of the teaching staff - It is essential that teachers talk to students well-coordinated. For this purpose, project control meetings are very important.

6. Communication Platform for teachers / students - The use of a communication platform (in this case, the Moodle) seems to us essential to the success of the project.

7. Final Assessment - It is important that, in the end of the project, teachers and students have time to evaluate the process in which they were involved and reflect on the key lessons learned and define adjustments for the project replication.

8. The presence of all teachers in the oral presentation of projects – If we want to convey the message of interdisciplinarity, it is important that teachers are an example of commitment and networking.

3. Results and Conclusions

The Bologna reform brought new challenges for teachers and students of higher education. More student-centered teaching and crossing the frontiers of knowledge through interdisciplinary projects are examples of these new challenges [2, 3, 4]. In this paper, we intend to present the results of an interdisciplinary project carried out in an engineering school in the north of Portugal.

The results obtained by the participating students were generally positive. Students' involvement in the project was notorious. This involvement may be justified by the fact that, as Jacobs [5] and Chettiparamb [6] refer, interdisciplinary works provide the students with relevant, challenging and enjoyable learning experiences which may influence their degree of motivation and commitment in performing the tasks. In fact, the percentage of relative involvement of each student in the project was evaluated by his partners in the group. This information was asked to students in the moment of presentation and discussion of each team work.

As mentioned by Chettiparamb Interdisciplinary projects force students to integrate subjects that were apparently independent. The interdisciplinarity has implied a change of students' attitudes in the search for the development of integrated knowledge [6].

With this kind of challenge, students had the opportunity to develop a reflexive thought to overcome difficulties and, in this way, to develop knowledge and skills [5, 6].

In fact, many different skills have been developed by students throughout the execution of this project. As an example, we can point out the ability to work with others, to solve problems and conflicts, to integrate knowledge from different fields, to write a report using the citation rules and also the ability to make public presentations. As some authors have argued this kind of skills are very important for a future engineer that will be called to solve complex problems that require crossing boundaries [7, 11].

This type of work also puts great demands on the teachers involved in the project, particularly in terms of flexibility to integrate knowledge. Nevertheless, we think that the results in terms of benefits to the students justify all efforts that can be done to develop interdisciplinary work.

The results of the assessment of this project were used to evaluate students in the different subjects. The weight given to this project by each subject was different from subject to subjects. As each subject has established different goals to achieve with this interdisciplinary project, it has also defined its own assessment criteria as well as the weight of the project in the overall evaluation of the subject. This weight was 10% of the overall evaluation of the subject for ALGAN, 20% for MATE1 and LMAT1, and 35% for MTENG.

At the beginning of the oral presentation and discussion of the teamwork, each element of the team was asked to answer an auto assessment question. This question had to do with the perception that each element of the team had about the percentage of contribution of each element to all the work done.

In the report and in the oral presentation, students answered, in generally, correctly to what was asked. We can mention that the function studied on the first task was not an ``academic" one (easy to analyse). They had to do research work and discuss it with the MATE1 teacher.

The exception was the last task (subsection 2.3) where students were asked to approach a definite integral by two different processes (numerical integration and using Taylor polynomials). The students' weakness was that they did not have the critical spirit to comment on the results obtained. For example, in one team of students, the results obtained by the two processes were very different. They did not realize that they should have compared the results and make conclusions.

By the end of the project, students' opinions regarding the challenge they had faced were assessed using a questionnaire. The students were asked to give their opinion about the nature of the project and its objectives (question 1); about their satisfaction concerning their personal involvement in the project (question 2); about the impact of the project in motivating them to learn mathematics related subjects (question 3); about the opinion on the overall assessment of the project (question 4), and about the project contribution to develop personal and team work skills (question 5). Students were also asked to provide general feedback on the project as well as to identify strengths and weaknesses. Except for this last open question, all the other questions were assessed using a Likert scale, being 1 "somewhat satisfied", and 5 "very satisfied". Thirty-six students answered the questionnaire. The average score was over 3.9 considering all the questions, reaching 4.1 in those questions where students assessed the objectives of the project, and 4.2 in the ones concerning students' opinion about the project contribution to the development of personal and team work skills (Fig. 6). As for strengths, students highlighted the relevance of the project for developing those skills and also for learning and consolidating mathematics-related subjects.

Although many of the respondent students did not mention any weakness, some pointed out the workload, and the short time period to develop the work as aspects to be improved in future editions. With that information in mind, we may conclude that students recognized the importance of this project not only to develop personal and teamwork skills but also to consolidate knowledge of the several subjects involved in the project.

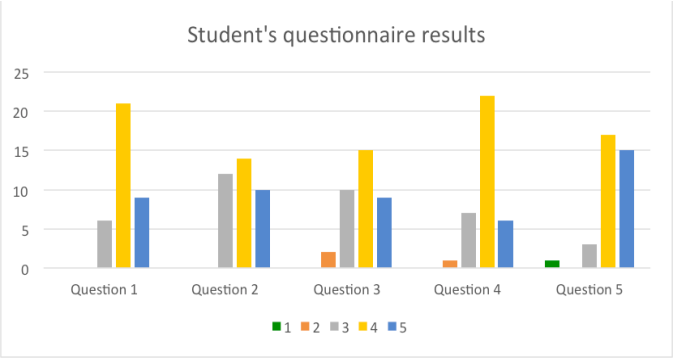


Figure 6: Students' questionnaire results

Despite the major benefits that this project has brought to the students, we are aware of the need to improve some aspects regarding planning and execution. At this stage, the coordination between the different involved teachers, the support provided to students during the project execution, and the management of the time students have dedicated to the project are the most challenging aspects and those with major impact on the success of the final project.

Taking into account the characteristics of this project and the context of the degree in which it is integrated (Electrical Engineer) it is important, in future experiences, to involve the subjects of Experimental Physics and Introduction to Electrotechnics. Their absence was one of the weaknesses of this experiment.

Other aspects to improve in new editions of the project are:

- to implement changes in the Moodle platform configuration to ensure a greater dynamic on the interaction between the members of each group, a greater dynamic on the interaction between students/teachers during the period of development of the work, and also to facilitate team work in online learning;
- to provide better conditions in online learning experiences, including team work will thus prepare students as effective members of teams in a virtual workplace of the future;
- to include a final report from every member identifying their contribution to the project;
- to provide some more class time for meetings because at the nowadays it is very hard for students to manage their schedules;
- to make the workload reasonable and the goals clear.

To conclude, it should be referred the intention of the authors to implement this project in other ISEP first degree courses (such as Electrical and Computer Engineering or Mechanical Engineering).

Conflict of Interest

The authors declare no conflict of interest.

References

[1] S. Abreu, A., Caldeira, A. Costa, T. Gomes, L. Roque, *Interdisciplinary team work: applying working methods to a math project*. In: Proceedings of the 2nd International Conference of the Portuguese Society for Engineering Education (CISPEE) (paper n.26). IEEE *Xplore* Digital Library. ISBN 978-1-5090-3912-8 IEEE catalog number CFP16CIK-ART, 2016.

[2] M. Borrego, J. Bernhard, The Emergence of Engineering Education Research as an Internationally Connected Field of Inquiry. *Journal of Engineering Education*, *100*(1), 14–47. https://doi.org/10.1002/j.2168-9830.2011.tb00003, 2011.

[3] R. Deem, K. H. Mok, K. H., L. Lucas, Transforming Higher Education in Whose Image? Exploring the Concept of the "World-Class" University in Europe and Asia. Higher Education Policy, 21(1), 83-97, 2008

[4] R. M. Lima, D. Carvalho, R. M. Sousa, A. C. Alves, Management of Interdisciplinary Project Approaches in Engineering Education: a Case Study. in First Ibero-American Symposium on Project Approaches in Engineering Education – PAEE2009 (D. Carvalho, N. v. Hattum-Janssen & R. M. Lima), Guimarães - Portugal, 149-156, 2009

[5] W. J. Jacobs, Interdisciplinary trends in higher education. *Palgrave Communications* 1:15001 doi: 10.1057/palcomms, 2015.

[6] A. Chettiparamb, Inter-disciplinarity in teaching: probing urban studies. *Journal for Education in the Built Environment,* 6:1**,** 68-90, 2011.

[7] M. Borrego, S. Cutler, Constructive alignment of interdisciplinary graduate curriculum in engineering and science: an analysis of successful IGERT proposals. Journal of Engineering Education, 99(4), 355–369, 2010. https://doi.org/10.1002/j.2168-9830.2010.tb01068.

[8] S. G. Clark and R. L.Wallace, Integration and interdisciplinarity: concepts, frameworks, and education. *Policy Sciences,* 48**,** 233-255, 2015.

[9] S. Lélé and R. B. Norgaard, Practicing interdisciplinarity. *BioScience,* 55**,** 967, 2005.

[10] I.-B. Păvăloiu, I. Petrescu and C. Dragomirescu, Interdisciplinary project-based laboratory works. *Procedia - Social and Behavioral Sciences,* 180**,** 1145-1151, 2015.

[11] M. Borrego, L. K. Newswander, Characteristics of successful cross-disciplinary engineering education collaborations. Journal of Engineering Education, 97(2), 123–134, 2008. https://doi.org/10.1002/j.2168-9830.2008.tb00962

[12] G. P. Harrison, D. Ewen Macpherson, D. A. Williams, Promoting interdisciplinarity in engineering teaching. European Journal of Engineering Education, 32(3), 285–293, 2007. https://doi.org/10.1080/03043790701276775.

[13] European Commission/EACEA/Eurydice. The European Higher Education Area in 2015: Bologna Process Implementation Report. Luxembourg: Publications Office of the European Union, 2015.

[14] A. Veiga, A. Amaral, Survey on the implementation of the Bologna process in Portugal. *Higher Education*, *57*(1), 57-69, 2009.

[15] L. Richard, J. Burden, Douglas Faires. Numerical Analysis, 9th edition, Cengage Learning, 2011.

[16] J. Stoer, R. Bulirsch, Introduction to Numerical Analysis. 3ª.ed. Springer, 2002.

[17] Larson, Hostetler and Edwards, Calculus, 8th edition, vol. 1, McGrawHill, 2011.

[18] J. Stewart, Calculus, 7th edition, vol. I, Cengage Learning, 2013.

[19] G. Strang, Introduction to Linear Algebra", 4th edition, Wellesley-Cambridge Press, 2010.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 796-804 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Treadmill and Vision System for Human Gait Acquisition and Analysis

Paulo A. Ferreira[1], João P. Ferreira[1,2], Manuel Crisóstomo[*,1], A. Paulo Coimbra[1]

[1]*Institute of Systems and Robotics, Dept. of Electrical and Comp. Engineering, Univ. of Coimbra, 3030-290 Coimbra, Portugal*

[2]*Dept. of Electrical Engineering, Superior Institute of Engineering of Coimbra, 3030-199 Coimbra, Portugal*

ARTICLE INFO

Article history:
Received: 05 April, 2017
Accepted: 18 May, 2017
Online: 16 June, 2017

Keywords:
Human gait
Image processing
Human joint angles
Gait characterization

ABSTRACT

This paper presents a developed low cost system for human gait analysis. Two web cameras placed in opposite sides of a treadmill are used to acquire images of a person walking at different speeds on a treadmill, carrying a set of passive marks located at strategic places of its body. The treadmill also has passive marks with the color chosen to contrast with the ambient dominant color. The body joint angle trajectories and 3D crossed angles are obtained by image processing of the two opposite side videos. The maximum absolute error for the different joint angles acquired by the system was found to be between 0.4 to 3.5 degrees. With this low cost measurement system the analysis and reconstruction of the human gait can be done with relatively good accuracy, becoming a good alternative to more expensive systems to be used in human gait characterization.

1. Introduction

This paper is an extension of the work originally presented in 2016 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM2016) [1]. The study of the human gait has been done in medical science [2-5], psychology [6, 7], and biomechanics [8-15] for more than five decades. Recently it has generated much interest in fields like robotics [16], biometrics [17] and computer animation [18]. In computer vision, recognizing humans by theirs gaits has recently been investigated [19]. The human gait is a pattern of human locomotion and can be described by kinetic or kinematics characteristics [20]. Gait signatures are the most effective and well defined representation methods for kinematic gait analysis. Gait signatures can be extracted from motion information of human gaits and have been used in computer graphics, clinical applications, and human identification [21-23].

Furthermore, gait data can be used to assess pathologies in a variety of ways. For example, stride parameters such as walking speed, step length and cadence provide an overall picture of gait quality.

Given the growing need for such analysis equipment, this paper presents a low cost system developed to characterize human gaits as an alternative to much more expensive solutions [24, 25], taking the advantage of requiring much less space for its installation (about 5 times less).

The organization of this paper is as follows: Section 2 describes the developed acquisition system and its setup. The static and dynamic performance tests are described in Section 3. The results of repeatability tests are presented in Section 4 and the conclusions in Section 5.

2. Acquisition System

The developed acquisition system is based on computer vision and tracks passive marks (circles with 4 cm of diameter) of a color chosen to contrast with the surround colors.

The system is mainly composed by one treadmill and two web cameras, one located on the right, and another on the left side of the treadmill, see Figure 1. Each one covers one side of the walking person.

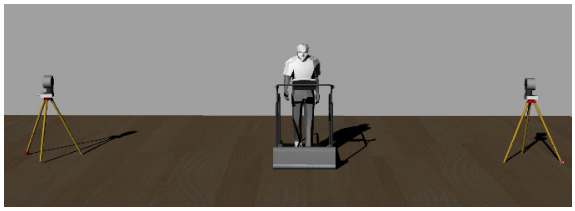


Figure 1. Layout of the developed system.

[*]M. Crisóstomo, Inst. of Syst. and Rob, Dep. of Elect. and Comp. Eng, Univ. of Coimbra, 3030-290 Coimbra, Portugal, Tel. +351 9130520256, mcris@isr.uc.pt

The web cameras used have the following characteristics: CMOS 640×480 (VGA) sensor, maximum of 30 frames per second, USB 2.0 interface.

The cameras are placed at a height of 1.15 m and at a distance of 2.35 m from the treadmill. Each camera is aligned and centred to the treadmill and to the opposite camera.

After laying out the system components, the following procedures need to be followed to extract human joint angles trajectories:

A – Calibrate and align both cameras;

B – Place detection marks on the person;

C – Calibrate both sides of the pelvis;

D – Synchronize both side videos;

E – Process video frames, correct the depth of the marks, and extract joint angles.

A. Cameras Calibration and Alignment

Next it is described the cameras calibration and alignment procedures.

A.1. Cameras Calibration

When using low cost video cameras, image distortions must be considered and it is mandatory to calibrate the cameras using standard calibration procedures.

Image distortions are constant and with a calibration and remapping they can be corrected [26]. Next, it is explained the basic topics of camera calibration: radial and tangential distortion, and the perspective transformation model.

A.1.1. Radial Distortion

The radial distortion is corrected using:

$$x_{corrected} = x(1 + k_1 r^2 + k_2 r^4 + k_3 r^6) \quad (1)$$

$$y_{corrected} = y(1 + k_1 r^2 + k_2 r^4 + k_3 r^6) \quad (2)$$

$$r^2 = x^2 + y^2 \quad (3)$$

where k_1, k_2, and k_3 are the radial distortion coefficients. Therefore, for an old pixel point at (x, y) coordinates in the input image, its position on the corrected output image will be $(x_{corrected}, y_{corrected})$. The presence of the radial distortion manifests in the form of the "barrel" or "fish-eye" effect.

A.1.2. Tangential Distortion

Tangential distortion occurs because the camera lenses are not perfectly parallel to the imaging plane. It is corrected by:

$$x_{corrected} = x + \left[2 p_1 xy + p_2 (r^2 + 2x^2)\right] \quad (4)$$

$$y_{corrected} = y + \left[p_1 (r^2 + 2y^2) + 2 p_2 xy\right] \quad (5)$$

where p_1 and p_2 are the tangential distortion coefficients. All the five distortion parameters were calculated using the Matlab camera calibrator app.

A.1.3. Perspective Transformation Model

For the units conversion the following formula (in homogeneous coordinate system) is used:

$$\begin{bmatrix} x \\ y \\ w \end{bmatrix} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \quad (6)$$

where the intrinsic camera parameters f_x and f_y are the camera focal lengths and (c_x, c_y) is the optical center expressed in pixels coordinates.

The process of determining all these nine parameters is the calibration. Calculation of these parameters is done through basic geometrical equations, and it depends on the chosen calibrating object, which was a classical black-white chessboard.

A.1.4. Camera Calibration Parameters Calculation

To do the calibration was used the Matlab camera calibrator app (Figure 2). The first step is to input images of a checkerboard calibration pattern. For accurate results it is recommended to use between 10 and 20 calibrating images, from each camera, and, in this case, 11 images were used (Figure 3). The checkerboard was used as calibrating object because its regular pattern make it easier to be detected automatically.

After inputting the checkerboard square size, the app will detect the checkerboard in all calibrating images. Then it is possible to inspect the results. This is helpful to find incorrect detections, and remove bad images for calibration.

After this, the calibration parameters can be calculated and if necessary images with biggest errors can be rejected and substituted by another ones to try to minimize the overall mean error.

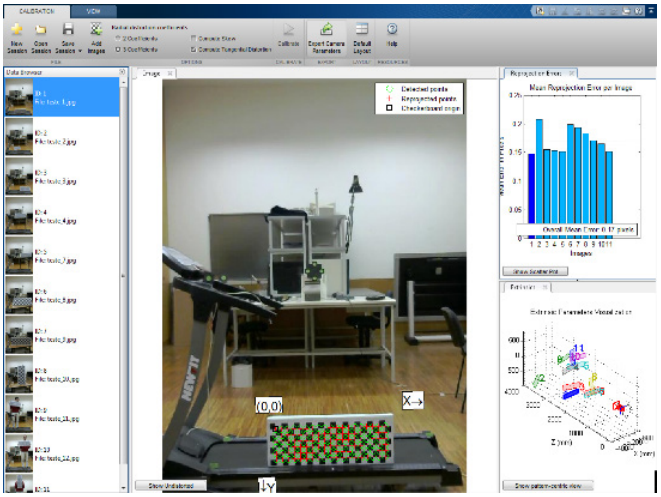


Figure 2. Matlab camera calibrator app.

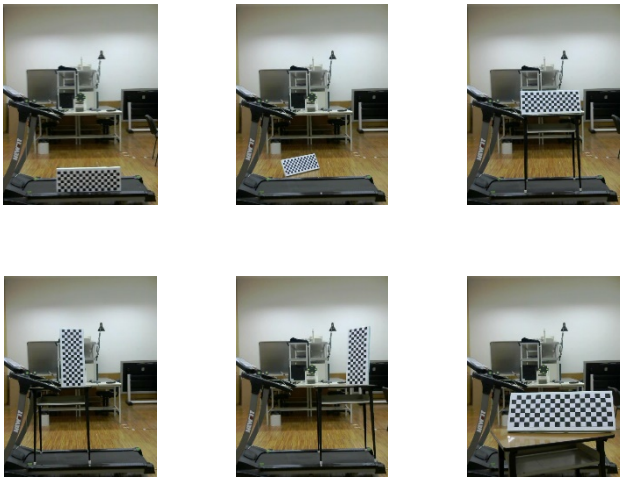


Figure 3. Some of the images used for camera calibration.

The obtained distortion coefficients for left camera are

$$Radial_{coef.} = [-0.0171 \quad 0.4642 \quad -1.4012]$$

$$Tangential_{coef.} = [-0.0048 \quad -0.0010]$$

and the intrinsic parameters (in pixels) are

$$focal_{length} = [f_x \quad f_y] = [809.5672 \quad 811.7467]$$

$$optical_{center} = [c_x \quad c_y] = [247.9749 \quad 296.3430]$$

For the right camera, the obtained distortion coefficients are

$$Radial_{coef.} = [0.1228 \quad -1.3171 \quad 6.2424]$$

$$Tangential_{coef.} = [-0.0054 \quad -0.0068]$$

and the intrinsic parameters (in pixels) are

$$focal_{length} = [f_x \quad f_y] = [819.1826 \quad 818.3655]$$

$$optical_{center} = [c_x \quad c_y] = [232.5183 \quad 332.3357]$$

A.2. Alignment of Cameras

Since there are two cameras and a treadmill between them, it is required to align each system component with each other to guarantee a good performance [27]. For that, it was developed a software module that measures the distance, in pixels, between six reference marks on the treadmill, and those marks viewed by the camera. Since the focal distance of the cameras and the distances between the treadmill and each camera are known, the location where the six marks should appear in the image can be calculated.

Now, it is necessary to do a manual fine alignment of the cameras in order to overlap each calculated mark position with its respective detected position in the image. Then, it is calculated the larger distance between these marks, and the goal is to ensure that the maximum distance is less than four pixels.

B. Placing Detection Marks on the Person

After the alignment of both cameras, it is necessary to attach 10 marks on each side of the walking person. Marks 1, 2, 3, A, 5, 6 and 8 correspond to joints in the human body: shoulder, elbow, wrist, pelvis, knee, ankle and finally finger toes, see **Error! Reference source not found.**.

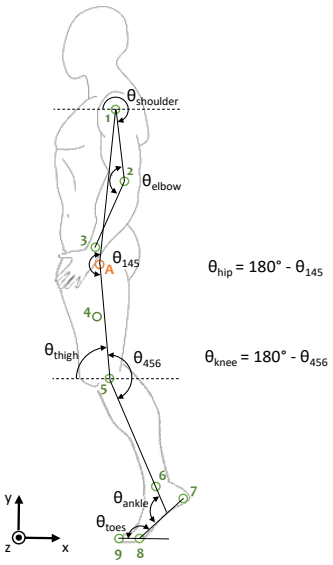


Figure 4. Joint angles definition and marks placement on the left side of a walking person.

C. Pelvis Calibration

Initially, the walking person must carry twenty marks (ten on each side) placed on him, Figure 5, but after the *"pelvis calibration"* procedure, the pelvis marks (A in Figure 4) are removed, and the pelvis joint marks will be inferred by the extra marks placed on the upper legs (mark 4 in Figure 4).

Considering the marks do not move during data acquisition, marks 4 can be removed and their positions inferred from the positions of the other two marks. This procedure is necessary because the arm and hand occlude the pelvis marks (marks A) during gait data acquisition. Figure 5 shows the software module window for the pelvis calibration.

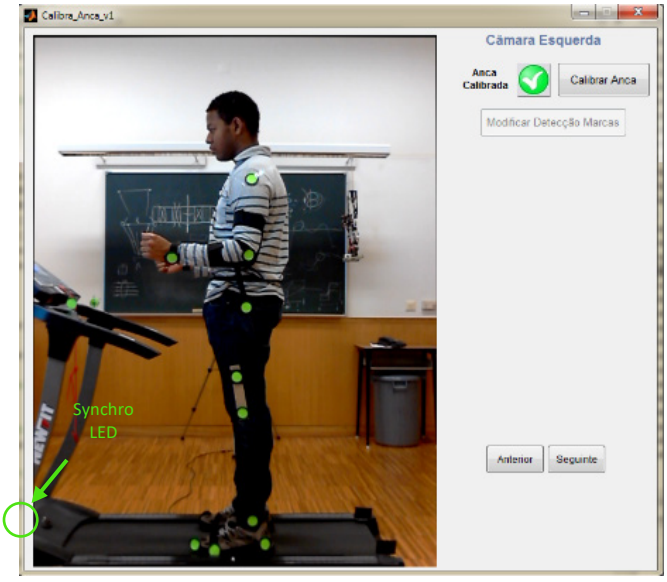


Figure 5. Left side pelvis calibration software module and synchronization LED position.

D. 3D Construction and Synchronization

To obtain 3D human gait data, it is required to cross synchronized both side's images [28]. To synchronize cameras, there is a white synchronization LED that blinks and that is visible by both cameras (Figure 5). The video acquisition starts first, and then the synchronization LED starts blinking. The synchronization procedure is automatic and is done when processing the acquired videos, and is necessary because cameras do not start exactly at the same time and the image acquisition rate is not exactly the same for both cameras and has fluctuations.

So, it is necessary to detect on both side videos when the synchronization LED turns ON and OFF, checking if the number of frames between LED ON and OFF cycles in both videos is the same. If they are not, the marks trajectories of the side with less frames are interpolated so that both side trajectories have the same number of data. When the number of frames in deficit is equal to or greater than three, the stride where this deficit occurs is considered not okay, and is discarded. This way, both side's videos can be crossed to extract 3D gait data.

To reference both camera images in space, there are two marks in the treadmill that are visible for both cameras (Figure 7). Both side images have their reference relative to these marks. This way it is ascertained not only the sagittal angles but also transversal and frontal rotation angles of the trunk, shoulders and pelvis.

It is necessary to add depth to the person's right and left planes, so, since the distances between all marks are measured (Figure 6), it is possible to convert those metric distances to pixels. To do this the required measures are: person height; shoulders width; arm and forearm lengths; hip width; thigh and leg lengths; distance between both ankles, and between both knees when in normal walking; great toe length; foot hinge finger to heel distance; and ankle to foot base distance. Other data are collected too, like age, gender and weight. Therefore the calculation of Body Mass Index (BMI) can be done.

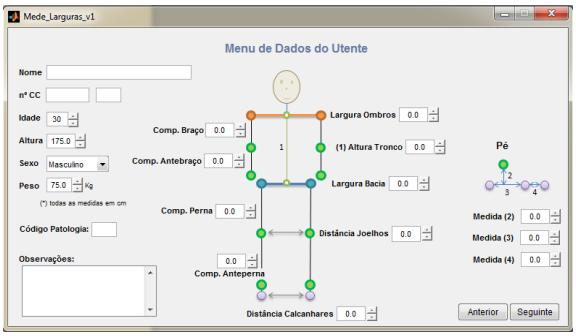


Figure 6. Menu to input the walking person data.

After these measurements and pelvis calibration the person walks about 6 minutes on the treadmill to be accustomed to it [29], and only after the person feels at ease with the treadmill use, the gait data acquisition can start (Figure 7).

Figure 7. Right side view of a person during gait data acquisition showing the 9 green marks on the person and the 2 on the treadmill.

E. Processing of Video Frames, Correcting the Depth of the Marks, and Extracting Joint Angles

After gait data acquisition, all saved data is loaded, and computed to extract the desired angle trajectories.

In this phase all frames are loaded to pass through a color segmentation process where all marks are extracted from the original videos, and the respective image coordinates are determined, [30, 31]. This segmentation is adjustable for every mark color.

The trajectories of all marks have been smoothed by using a zero lag Butterworth filter with a cut-off frequency of 6 Hz [9, 32]. Since for speeds below 5.5 Km/h, 99.7 % of the signal power is contained in the lower seven harmonics (below 6 Hz), the use of a 6 Hz cut-off frequency is considered to be adequate for the tests.

After being determined the coordinates of each mark for all video frames and after the noise reduction is done, each joint angle trajectory can be obtained.

Afterwards it is possible to visualize, for the entire duration of the test, the person 3D animated skeleton with all joints and links.

Stride detection makes possible stride grouping. Each stride is defined when the right foot starts going from up to down. After, the same foot stops moving forward and begins to move backward. I.e., forward to backward movement is first flagged, and after that, when upward to downward movement is detected, this instant corresponds to a new stride detection. Each stride detection ends the previous stride and starts a new one.

Figure 8 shows a window where the saved data can be viewed and analyzed.

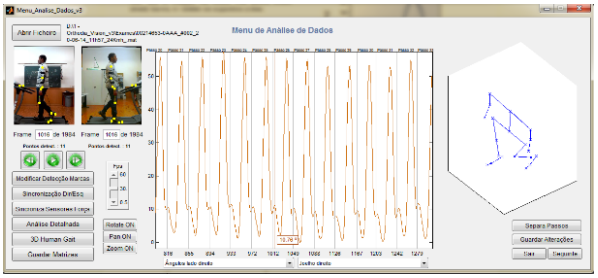


Figure 8. Menu to display collected data and respective joints angles.

In this window it can be seen: both sides videos, the location of each detected mark (yellow circles in video frames), the angle values for all recorded person strides, a 3D animated skeleton of the person for each frame, and other options, like zoom, pan, rotate, RGB levels for mark detection, intensity threshold for synchronization LED detection, export data, etc.

At the end, all collected data can be exported and opened in another window (Figure 9), where all strides are grouped and the 17 mean joint angles profile, and their respective standard deviation (in green) are seen.

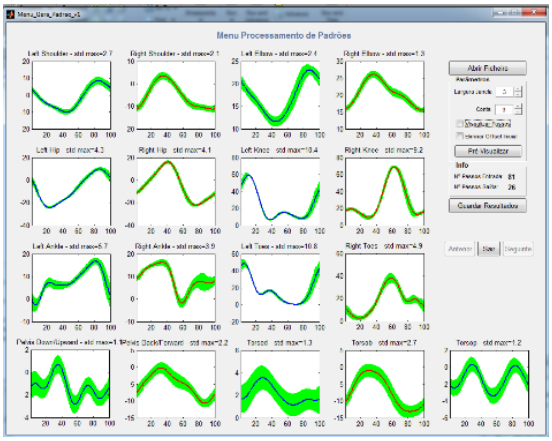
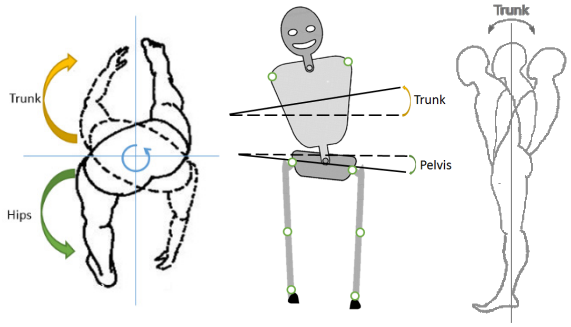


Figure 9. Window showing the mean, the standard deviation and other indicators from recorded data.

Beyond the angles referred to in **Error! Reference source not found.** it is also possible to extract other angles: backward/forward pelvis and trunk, downward/upward pelvis and trunk, and posterior/anterior trunk, as can be seen in Figure 10.



a) b) c)

Figure 10. a) - Backward/Forward Trunk/Pelvis angle.
b) - Downward/Upward Trunk/Pelvis angle.
c) - Posterior/Anterior Trunk angle.

The one side marks placed on the arm and leg are in different planes (see Figure 11). Instead of considering every mark in the same plane, it was made a depth correction [33]. I.e., using the physical characteristics of the person in test (Figure 6), the difference distances between the planes were calculated. Combining the values calculated before with the physical system setup distances, the points outside of the reference plane β (like knee, ankle and every foot points) were projected to it, permitting a more accurate calculation of the angles.

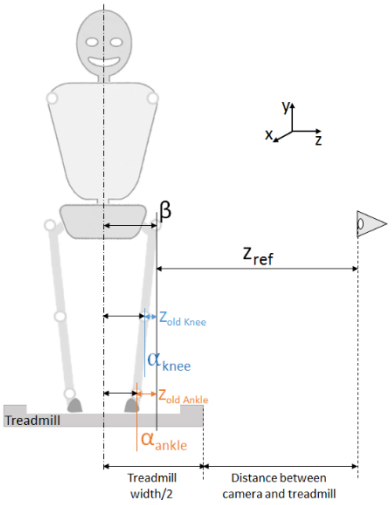


Figure 11. Position and metric relationship between reference plane and marks planes.

Involved variables can be viewed in Figure 12.

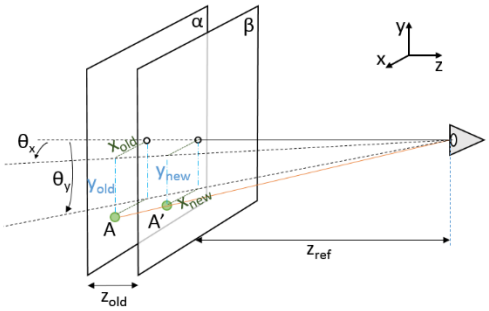


Figure 12. Variables used to correct the depth of certain points.

The reference plane β contains the pelvis mark point (see Figure 12).

Consider a point A with coordinates (x_{old},y_{old}) on the plane α parallel to the reference plane. The new coordinates of the point A (x_{new},y_{new}), on the reference plane β, A', are obtained by

$$\begin{bmatrix} x_{new} \\ y_{new} \end{bmatrix} = \frac{z_{ref}}{z_{ref} + z_{old}} \cdot \begin{bmatrix} x_{old} \\ y_{old} \end{bmatrix} \quad (7)$$

It must be mentioned that the image referential has to be at the centre of the image, and not at a corner of the image, like usually is. I.e., all of the points have to be referenced to the centre of the image.

The above procedure is executed for all points, except for pelvis point (since it is the reference point), and finally has to be done their re-reference for previous origin (original referential).

With these transformations, more accurate results are achieved.

3. Performance Tests

Static tests were performed using a static planar dummy (with known joint angles) located in three different positions on the treadmill (see Figure 13). Those angles were measured using the

software and compared with the real ones to determine system static errors. Also, a dynamic test was done, where a person walked on the treadmill at different speeds. The trajectories of the joint angles were then analysed and their standard deviations were calculated to check if they were as expected, having literature trajectories as reference. A comparison between these angle profiles and the literature normal profiles was also done.

3.1 Static Tests

To determine the maximum static system error, some tests were made using a planar dummy placed at three different positions upon the treadmill: at the front, at the center, and at the rear (see Figure 13).

Figure 13. Static dummy placement, for static error determination (right view).

The planar dummy consists of two styrofoam plates with marks located as the joint positions of a human. The plates were carefully positioned in order to achieve a vertical inclination identical to that of a human (see Figure 14b). The traced and measured angles are shown in Table 1 and Table 2. These angles are defined in Figure 14.

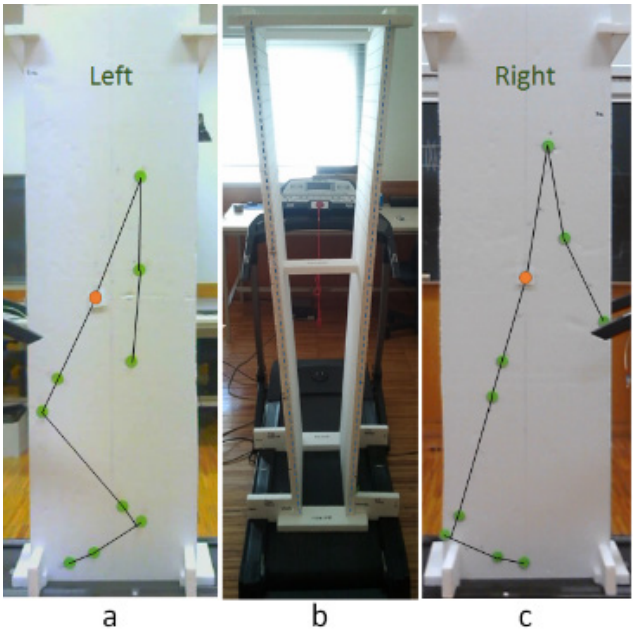


Figure 14. Planar dummy with known joint angles: a) left view; b) front view; c) right view.

The tests were made at the three referred positions upon the treadmill to be clear that there is neither a significant error at each location, nor a relation between the location and the error.

Each test lasted 3 minutes at 30 fps, which resulted in 5400 frames. In the total of the 3 tests this corresponds to 16200 analysed frames, which are more than enough to represent accurately the precision of the system in steady state.

After analysing the results it is concluded that there was no relation between the position of the dummy on the treadmill and the error of the measurements. Table 1 and Table 2 present the maximum errors obtained in all tests.

Table 1. Measured left and right sagittal angles, and respective errors, in degrees.

Angle (degrees)	Left				Right			
	Real value	Mean meas. value	Mean absolute error	Max. absolute error	Real value	Mean meas. value	Mean absolute error	Max. absolute error
Shoulder	20	20.4	0.4	0.7	-20	-19.6	0.4	0.6
Elbow	-5	-4.3	0.7	1.0	-15	-15.5	0.5	0.7
Trunk	-20	-21.0	1.0	1.2	10	9.7	0.3	0.5
Hip	-5	-3.7	1.3	1.8	5	6.1	1.1	1.5
Knee	65	64.3	0.7	1.0	0	-0.3	0.3	0.6
Ankle	-95	-97.2	2.2	2.3	-85	-83.0	2.0	2.3
Toes	-10	-11.2	1.2	1.7	-10	-9.4	0.6	0.9

Table 2. Measured crossed angles and respective errors, in degrees.

Angle (degrees)	Real value	Mean meas. value	Mean absolute error	Max. absolute error
Downward/ Upward Trunk	11	8.8	2.2	2.2
Backward / Forward Trunk	23	24.1	1.1	1.2
Downward / Upward Pelvis	9	7.3	1.6	1.8
Backward / Forward Pelvis	-9	-5.6	3.4	3.5
Posterior / Anterior Trunk	-5	-5.3	0.3	0.4

These tests show a relatively low static absolute error, with a maximum of 3.5 degrees obtained on crossed angles and 2.3 degrees on sagittal plane angles.

3.1 Dynamic Tests

A healthy person did a gait data acquisition, walking at different speeds. This person is a 30 year old male, 1.85 m tall, and weighs 87 kg.

Table 3 shows the maximum standard deviation of all angles for the different speeds of the treadmill.

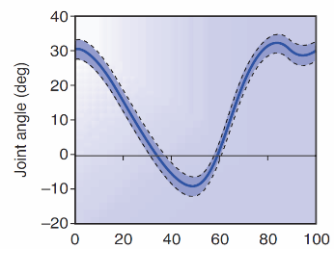
Table 3. Maximum standard deviation of all angles for all dynamic tests, in degrees.

Speed [Km/h]	1.1	2.7	4.1	5.4
R. Shoulder	2.0	1.8	2.2	2.9
L. Shoulder	3.4	1.8	2.8	2.0
R. Hip	2.3	1.9	2.5	3.6
L. Hip	1.8	1.3	2.2	2.3
R. Knee	3.0	2.3	2.7	3.2
L. Knee	3.4	4.3	**6.1**	5.2
R. Ankle	2.0	1.3	1.6	2.4
L. Ankle	2.0	1.7	3.0	3.0
Pelvis bf [(1)]	1.8	1.4	1.1	1.7
Pelvis du [(2)]	0.8	0.7	0.5	0.8
Trunk bf [(3)]	2.6	1.9	1.6	2.1
Trunk du [(4)]	1.4	1.0	0.7	1.1
Trunk pa [(5)]	0.8	0.7	0.8	0.8

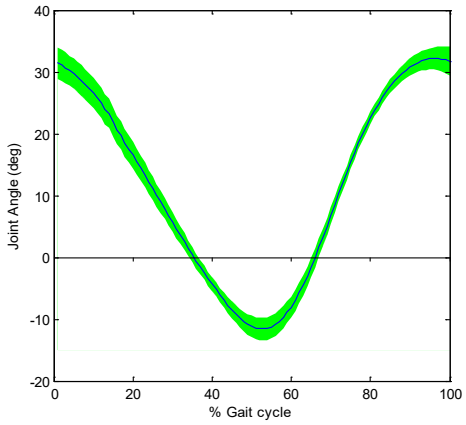
[(1)] Pelvis backward/forward, [(2)] Pelvis downward/upward, [(3)] Trunk backward/forward, [(4)] Trunk downward/upward, [(5)] Trunk posterior/anterior.

The maximum standard deviation of these tests is 6.1 degrees, which is similar to the value registered in [34].

Figure 15a shows the patterns of hip angles for the right side of a healthy person [35]. Figure 15b shows the measured values of the hip angle for the right side of a healthy person using the developed system.



a)



b)

Figure 15. a) Normal pattern of the right hip angle, and its standard deviation [35]; b) Measured right hip angle, and its standard deviation (in green) for a 4.1 km/h gait of a healthy person.

From Figure 15 it is seen that the measured values for the right hip have similar appearance to the pattern of this angle seen in the literature [35], except at the final phase of the gait cycle, where the pattern has a small nuance and our measurements do not. This nuance may be justified by the walking speed to obtain those pattern graphs, which is not mentioned in [35].

4. Repeatability Tests

Two more tests were done, where a person walked at the treadmill at different speeds and in two distinct sites. The whole system was completely installed and uninstalled in two different sites, and the same person did the gait acquisition twice, at the same speeds. The data were compared with each other, to check if both results are similar.

Figure 16 shows the mean measured values for both tests (solid lines) and their standard deviations (dashed lines). The curves obtained in both tests are similar, and their offsets are very low. The existing difference may be derived from the manual placement of each mark, which can vary, depending, for instance, on the dressing of the person.

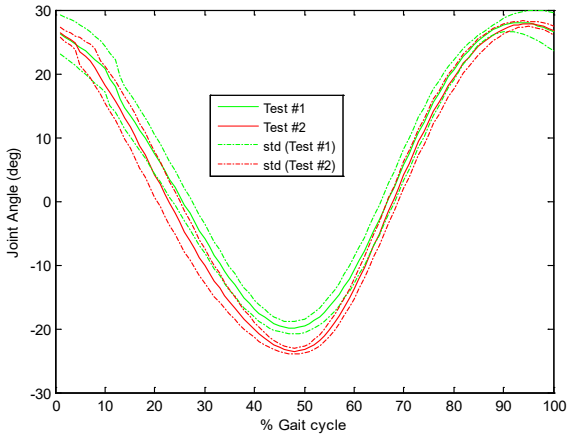


Figure 16. Right hip mean angle trajectories (solid lines) for two independent tests at 5.4 km/h gait speed, and their respective standard deviations (dashed lines).

To quantify the similarity of the results of these two tests, the following differences D_i were defined as

$$D_i = \frac{\min(\alpha_{2i} - (\alpha_{1i} \pm \sigma_{1i})) + \min(\alpha_{1i} - (\alpha_{2i} \pm \sigma_{2i}))}{2} \quad (8)$$

where

– i varies from 1 to 100, and represents the current percentage of the stride phase;
– α_{1i} is the angle i value of test #1;
– α_{2i} is the angle i value of test #2;
– σ_{1i} is the standard deviation of test #1 with index i;
– σ_{1i} is the standard deviation of test #2 with index i.

The Root Mean Square of D is then

$$RMS_D = \frac{1}{n}\sqrt{\sum_{i=1}^{n=100}(D_i)^2} \quad (9)$$

Table 4 presents the RMS_D values calculated with (9).

Table 4. Root mean square of the difference, in degrees, between two similar test sets, with the same walking person, at different sites.

Speed [km/h]	Shoulder		Hip		Knee		Ankle	
	Right	Left	Right	Left	Right	Left	Right	Left
1.1	0.6	0.7	0.5	0.9	0.3	1.6	0.2	0.3
2.7	0.0	0.4	1.8	1.9	1.0	1.3	0.4	0.0
4.1	0.9	0.0	0.6	**2.1**	1.1	0.1	0.2	0.1
5.4	0.2	0.9	1.3	1.2	1.5	1.1	1.1	0.1

Speed [km/h]	Pelvis bf [1]	Pelvis du [2]	Trunk bf [3]	Trunk du [4]	Trunk pa [5]
1.1	2.3	0.9	0.4	0.4	0.2
2.7	**5.0**	1.9	0.3	0.2	1.2
4.1	3.0	1.9	0.1	0.2	0.0
5.4	4.1	1.4	0.5	0.5	0.9

[1] Pelvis backward/forward, [2] Pelvis downward/upward,
[3] Trunk backward/forward, [4] Trunk downward/upward,
[5] Trunk posterior/anterior.

It is seen that the maximum RMS difference on sagittal plane angles is 2.1 degrees for the left hip at 4.1 km/h, which is a reasonable repeatability error that is compatible with the maximum absolute error measured on static tests. For crossed angles the maximum error is 5.0 degrees at backward/forward pelvis angle at 2.7 km/h, which is a maximum error of about twice the maximum static absolute error. This is acceptable since these crossed angles are determined by crossing right side points with left side points.

With these repeatability tests it is confirmed that precision was not affected by the portability of the system.

5. Conclusions

A system for acquisition and analysis of the human gait was developed and can be used for various applications such as physiotherapy, pathology identification and rehabilitation. It presents a mean error of about two degrees, what is perfectly acceptable for human gait characterization, once the human gait presents deviations above six degrees relatively to mean values [34]. The limitation of this system is the precision on crossed angles, which, because of their side cross nature, have an error of about five degrees. However, this is not relevant for the proposed purpose of human gait characterization.

A concurrent system has a maximum error of about 0.3º, seven times lower than the presented system. However, the presented system costs about 2,500 €, is significantly cheaper than concurrent systems (20 to 40 times), presenting a high quality/price ratio.

The presented system is a simple effective solution for human gait data acquisition and characterization, with a good precision, costing much less than current systems, and having the advantage of requiring much less space for installation setup.

Acknowledgment

The authors acknowledge Fundação para a Ciência e a Tecnologia (FCT) and COMPETE 2020 program for the financial support to the project "Automatic Adaptation of a Humanoid Robot Gait to Different Floor-Robot Friction Coefficients" (PTDC/EEI-AUT/5141/2014).

References

[1] Paulo A. Ferreira, João P. Ferreira, Manuel Crisóstomo, A. Paulo Coimbra, "Low Cost Vision System for Human Gait Acquisition and Characterization", 2016 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM2016), pp. 291-295, Bali, Indonésia, December 4-7, 2016.
[2] Michael W. Whittle, "Clinical gait analysis: A review", USA Human Movement Science 15, 1996 Elsevier Science, 369-387.
[3] M. P. Murray, "Gait as a Total Pattern of Movement", American Journal of Physical Medicine, 46(1), pp.290- 333, 1967.
[4] J. Perry, "Gait Analysis: Normal and Pathological Function", Slack, 1992.
[5] Yuki Ishikawa, Qi An, Junki Nakagawa, Hiroyuki Oka, Tetsuro Yasui, Michio Tojima, Haruhi Inokuchi, Nobuhiko Haga, Hiroshi Yamakawa, Yusuke Tamura, Atsushi Yamashita & Hajime Asama "Gait analysis of patients with knee osteoarthritis by using elevation angle: confirmation of the planar law and analysis of angular difference in the approximate plane", Advanced Robotics, 31:1-2, 68-79, 2017.
[6] G. Johansson, "Visual Perception of Biological Motion and a Model for Its Analysis", Perception and Psychophysics, 14(2), pp.201-211, 1973.
[7] S. V. Stevenage, M. S. Nixon, and K. Vince, "Visual Analysis of Gait as a Cue to Identity", Applied Cognitive Psychology, 13, pp.513-526, 1999.
[8] V. T. Inman, H. J. Ralston, and F. Todd, "Human Walking", Williams & Wilkins, 1981.
[9] D. A. Winter, "The Biomechanics and Motor Control of Human Movement", 4nd Eds., John Wiley & Sons, 2009.
[10] T. P. Andriacchi and D. E. Hurwitz, "Gait Biomechanics and the Evolution of Total Joint Replacement", Gait and Posture, 5(3), pp.256-264, 1997.
[11] Ferreira, J. P., Crisóstomo, M., Coimbra, A. P., "Human Gait Acquisition and Characterization", IEEE Transactions on Instrumentation and Measurement, Vol. 58, nº9, September 2009.
[12] A. Sona, R. Ricci and G. Giorgi, "A measurement approach based on micro-Doppler maps for human motion analysis and detection", Proc. IEEE International Instrumentation and Measurement Technology Conference, pp. 354–359, May 2012.

[13] G. Panahandeh, N. Mohammadiha, A. Leijon, and P. Handel, "Continuous hidden Markov model for pedestrian activity classification and gait analysis", IEEE Trans. Instrum. Meas., vol. 62, no. 5, pp. 1073–1083, May 2013.
[14] A. Sona, R. Ricci, "Experimental validation of an ultrasound-based measurement system for human motion detection and analysis", in Proc. IEEE International Conference on Instrumentation and Measurement Technology (I2MTC), May 2013
[15] A. Balbinot, Schuck, A.,Bruxel, Y., "Measurement of transmissibility on individuals", in Proc. IEEE International Conference on Instrumentation and Measurement Technology (I2MTC), May 2013.
[16] Long Y., Du Z., Dong W., Wang W. "Human Gait Trajectory Learning Using Online Gaussian Process for Assistive Lower Limb Exoskeleton". In: Yang C., Virk G., Yang H. (eds) Wearable Sensors and Robots. Lecture Notes in Electrical Engineering, vol 399. Springer, Singapore, 2017.
[17] Kateřina Sulovská, Eva Fišerová, Martina Chvosteková, Milan Adámek, "Appropriateness of gait analysis for biometrics: Initial study using FDA method", Measurement (Elsevier), Volume 105, pp 1-10, July 2017.
[18] Mengyuan Liu, Hong Liu, Chen Chen, "Enhanced skeleton visualization for view invariant human action recognition", Pattern Recognition (Elsevier), Volume 68, pp 346-362, August 2017.
[19] M. S. Nixon, et al., "Automatic Gait Recognition", in BIOMETRICS - Personal Identification in Networked Society, A. K. Jain, et al. Eds., pp.231-249, Kluwer, 1999.
[20] J. K. Aggarwal, Q. Cai, W. Liao, and B. Sabata, "Nonrigid Motion Analysis: Articulated and Elastic Motion", Computer Vision and Image Understanding, 70(2), pp.142-156, May 1998.
[21] Ferreira, J. P., Crisóstomo, M., Coimbra, A. P., Carnide, D., and Marto, A., "A Human Gait Analyzer", 2007 IEEE International Symposium on Intelligent Signal Processing-WISP'2007, Madrid, Espanha, 3-5 Outubro, 2007.
[22] Bux A., Angelov P., Habib Z., "Vision Based Human Activity Recognition: A Review". In: Angelov P., Gegov A., Jayne C., Shen Q. (eds) Advances in Computational Intelligence Systems. Advances in Intelligent Systems and Computing, vol 513, Springer, 2017.
[23] S. K. Natarajan, X. Wang, M. Spranger and A. Gräser, "Reha@Home - a vision based markerless gait analysis system for rehabilitation at home", 2017 13th IASTED International Conference on Biomedical Engineering (BioMed), pp. 32-41, Innsbruck, Austria, 2017.
[24] Vicon: Bonita. Internet, May 8, 2017, available at: http://www.vicon.com/System/Bonita.
[25] NDI Measurement Sciences: Optotrak-Certus. Internet, May 8, 2017, available at: http://www.ndigital.com/msci/products/optotrak-certus.
[26] OpenCV: Camera calibration with OpenCV. Internet, May 8, 2017, available at:http://docs.opencv.org/doc/tutorials/calib3d/camera_calibration/camera_calibration.html.
[27] Miles H., Seungkyu L., Ouk C., Radu H., "Time of Flight Cameras: Principles, Methods and Applications", Springer, 2013.
[28] Andrea F., Juergen G., Helmut G. Xiaofeng R., Kurt K., "Consumer Depth Cameras for Computer Vision: Research Topics and Applications ", Springer, 2013.
[29] Angelo Matsas, Nicholas Taylor, Helen McBurney, "Knee joint kinematics from familiarized treadmill walking can be generalized to over ground walking in young unimpaired subjects", Elsevier, Gait and Posture 11 p46–53, 2000.
[30] J.J.H.López, A.L.Q.Olvera, J.L.L.Ramirez, F.J.R.Butanda, M.A.I. Manzano, D.Luz A.Ojeda, "Detecting objects using color and depth segmentation with Kinect sensor", The 2012 Iberoamerican Conference on Electronics Engineering and Computer Science, 2012.
[31] Michael Van den Bergh, Luc Van Gool, "Combining RGB and ToF Cameras for Real-time 3D Hand Gesture Interaction", IEEE Workshop on Applications of Computer Vision, 2011.
[32] Holden J.P, Chou G, Stanhope S.J, "Changes in Knee Joint Function Over a Wide Range of Walking Speeds", Clinical Biomechanics Vol. 12(6): 375-382, 1997.
[33] P.Howard, Brian J. Rogers, "Perceiving in Depth: Stereoscopic Vision", Vol. 2. Oxford, 2012.
[34] A. A. Biewener, C. T. Farley, T. J. Roberts and Marco Temaner, "Muscle mechanical advantage of human walking and running: implications for energy cost", J Appl Physiol 97:2266-2274, 2004.
[35] Margareta Nordin, Victor H. Frankel, "Basic Biomechanics of the Musculoskeletal System", 4th Edition, Wolters Kluwer, 2012.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 805-811 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Assessment of the heat sinking effect of a human hand that holds a flexible phototherapy device for use in Kangaroo Mother Care

Luis Jimenez[*1], Luis Vilcahuaman[2], Jorge Galdos[3]

[1]*Department of Electronics Engineering, Universidad Nacional de San Antonio Abad del Cusco, Cusco, Perú*

[2]*Engineering Section, Pontificia Universidad Católica del Perú, Lima, Perú*

[3]*Pediatrics Department, Hospital Regional Cusco, Cusco, Perú*

ARTICLE INFO

Article history:
Received: 05 April, 2017
Accepted: 03 June, 2017
Online: 19 June, 2017

Keywords:
Human hand heat transfer
Thermal models
Kangaroo mother care
Heat flow sensors
Phototherapy devices

ABSTRACT

The heat transfer process from a 6.4 Watt blue light flexible phototherapy mattress to a human hand has been studied. The intended use of the mattress is the provision of neonatal jaundice phototherapy during Kangaroo Mother Care (KMC) or skin-to-skin care. The heat transfer process has been studied with temperature and heat flow sensors inside an expanded- polystyrene spheroid 25mm of thickness where the phototherapy mattress's non emitting surface is in contact and folded around the hand while the light radiates outwards, to the polystyrene spheroid. The effective thermal capacitance of the phototherapy mattress was calculated as 3.4 J/ºK, the mattress-to-hand thermal conductance was found to be 0.25 W/ºC, the maximum temperature gradient between the mattress and the palm of the hand was 10ºC and the maximum power absorbed by the hand, 2.5W. The palm of the hand skin-to-core temperature gradient was 0.5ºC. It is expected that when used in KMC only one half of the hand skin surface would be in contact with the phototherapy mattress so that the effective thermal conductance to the hand in this conditions would be 0.125 W/ºC. Further study is suggested to include sweating effect of the hand.

1. Introduction

This paper is an extension of the work originally published in 2016 Healthcare Innovation Point of Care Technologies Conference [1], that paper was about the development and thermal assessment of a 1 Watt flexible blue light emitting diode (LED) array designed to apply neonatal jaundice phototherapy simultaneously with Kangaroo Mother Care (KMC) or skin-to-skin care, to benefit from both, well stablished, medical interventions in neonatal healthcare, KMC improves temperature stability of the neonate [2] while blue light irradiance is effective for treating neonatal jaundice [3], some studies like Samra's [4] showed the effectiveness of the simultaneous application. Before starting to use the device on neonates, the authors assessed the thermal performance of this form of phototherapy using two silicone polymer, thermally controlled mannequins, one of a neonate and the other of a mother. Results showed that the power transferred from the phototherapy mattress to the neonatal mannequin was 0.395 Watt, and the temperature increase in the contact surface (back of the neonatal mannequin) was 2.75ºC above its axillari temperature. The average blue light phototherapeutic irradiance obtained over the neonatal mannequin's back was 350 uW/cm^2, even if this irradiance is not considered intensive phototherapy; it approximately doubles that of white light fluorescent phototherapy devices.

Figure 1 shows the experimental setup with mannequins in [1], it is important to remark that the phototherapy mattress was held over the neonate's back by means of a long bandage acting as a belt around the mother and the baby, with no intervention of the mother's hand. In fact, a real mother would hold the phototherapy device to the baby's back with one hand to make sure the baby don't fall back.

Figure 1: The intended use of the flexible phototherapy mattress in [1]

Luis Jimenez, Av. Huayna Capac 184, Cusco – Peru. Tel +51 984988933
Email: wl_jimenez@yahoo.com

The present study's aim is to understand how the mother's hand would exchange heat with the phototherapy device; at first one might expect that the hand would act as an extra heat source contributing to the temperature increase of the phototherapy mattress and also that of the neonate. A closer look would make us realize that the mother temperature is physiologically controlled to 37ºC and that it would have a heat sinking effect when the phototherapy mattress temperature is above the mother temperature.

It is expected that with the hand's sinking effect, the power of the phototherapy mattress could be incremented safely to obtain a higher phototherapeutic irradiance over the neonate's skin. In this work we built a 6.4 Watt phototherapy mattress that has the same dimensions and flexibility as the one described in [1]. Experiments with this mattress around a volunteer's hand and the whole arrangement inside an expanded polystyrene spheroid are performed. Also, Simulink models are developed to obtain the effective thermal conductance from the phototherapy mattress to the human hand.

The thermal conductance we will obtain will allow us to have a design parameter to boost phototherapy irradiance keeping skin temperatures of the mother and the baby in a safe range.

2. Review of related work

An example of a medical device emitting light to a tissue, used in contact with the skin is the pulse oximeter; it is for measuring oxygen saturation non invasively, the difference with the phototherapy device of this work is that it only has two diodes and the power dissipated by the device is low compared to the phototherapy mattress of [1]. Medical device regulation agencies test the thermal safety of pulse oximeters over a person's finger that is inside an incubator at 37ºC [5]. Finally the temperature reached by the pulse oximetry sensor is measured, it has to be at most 4 degrees above body core temperature to get the authorization. The same temperature safety criteria has been applied in [1].

Researchers like Sienkiewicz [7] have quantified the amount of heat that is transferred from a hand submerged in water that is maintained between 37 to 40 degrees. In our case, the phototherapy mattress will not be used underwater, so the conditions for our study are different. Ducharme [6] studies the effect of blood circulation over heat transfer from body limbs to the environment. A review of Taylor [8] provides various thermal heat transfer parameters for hands and feet. An application of hands heat transfer is developed by Heller and Grahn [9], they develop a device that allows to reduce body core temperature rapidly cooling the hands by means of a glove that facilitates perfusion and allows a person to recover faster after exercise and to be able to exercise again.

Heat transfer between a hand and objects has been studied for human-computer interfacing in [10] as short contact time transients. The present work has to consider much longer times because Kangaroo Mother Care sessions last at least two hours as recommended by WHO in [11].

Heat flow from fingers to objects and Peltier pumps have been measured with thin film sensors as in the work of Guiatni and Kheddar [12], we use a similar sensor type as will be described below.

Also, thin film heat flow sensors have been used to measure heat transfer in processes that reached steady state, requiring more than 30 minutes to settle as in [13] where heat flow from human skin or from mannequins to water has been measured, they recommended to apply a thermal compound between the sensor and the material.

In the present work we don't require steady state to occur because modeling will consider heat capacity of the materials. The human hand will be treated as a constant temperature object with a thermal conductance to be determined experimentally.

3. Materials and Methods

3.1. Materials

The phototherapy mattress of this work is an arrayo of blue light emmitting diodes (LEDs) over an aluminium, duct reparing adhesive tape. Five flexible LED strips are placed in parallel separated 1 cm from each other over the adhesive tape. The phototherapy mattress dimensions are 10 x 25 cm and it weights 50 g. A 2110 Keithley desk multimeter has been used to measure DC current and voltage resulting that the mattress requires 6.4 Watt from a 12 Volt power supply. To assess the heat sinking effect of the hand that holds the phototherapy mattress, a semi-adiabatic experimental setup has been designed, it uses a hollow expanded polystyrene spheroid formed by two symmetric halves of 25mm of thickness hold together by adhesive tape for the experiments as shown un figure 2.

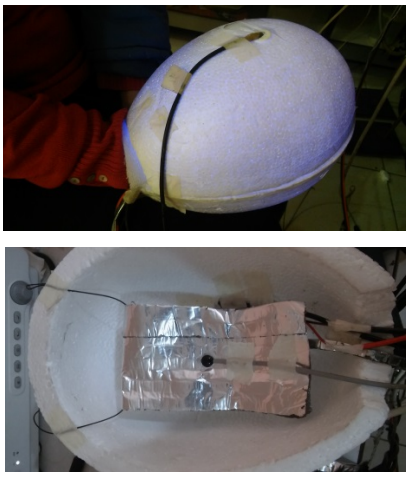


Figure 2: The prolate spheroid for the experiments with phototherapy matress inside

The polystyrene spheroid is prolate , c measures 14 cm, the perpendicular axis a measures 10.25 cm; for the internal surface of the spheroid c is 11.5 cm and a is 7.75 cm due to the thickness of the spheroid (2.5 cm). For the experiments, the phototherapy matress is folded in two at the middle of its longest size and sensor are placed inside the spheroid. The weight of the spheroid is 50 g.

The temperature of the phototherapy mattress is measured with a physiological monitor (Star 8000 –Comen Medical), its range is 0 to 50ºC with 0.1 ºC resolution; the metallic flat surface of the temperature sensor has been fixed to the aluminium tape back surface of the phototherapy matress with a drop of cyanoacrylate adhesive.

A temperature logger based on 5 LM35 sensors records the polystyrene spheroid inner wall temperature in two opposite points, another sensor is inserted in the outer wall of the spheroid as shown in the upper photograph of figure 2. Annother sensor is at the room air temperature. The last temperature sensor is for the experiment with the hand, it is fixed to the skin of the palm of the hand with a drop of cyanoacrylate adhesive. The resolution of this logger is the same as the physiological monitor and its temperature values have been calibrated with the physiological monitor. To measure the heat flow (in Watts/m^2) we use a thin film heat flow sensor, model HF4 from Omega Engineering Inc.; it has a sensitivity of 2 uV/(Watt/m2). The microvolt output is measured with a 2110 Keithley multimeter connected to a computer as a voltage logger. In the experiment without the hand, the heat flow sensor has been installed over the inner surface of the spheroid with adhesive tape at the borders, caring not to cover the thermocouple array with the tape as shown in figure 3. For the experiment with the hand, the heat flow sensor has been fixed over the palm of the hand with a drop of cyanoacrylate adhesive in each corner of the flow sensor. In both cases, a thin layer of thermally conductive cream has been applied between the heat flow sensor and the surface of interest.

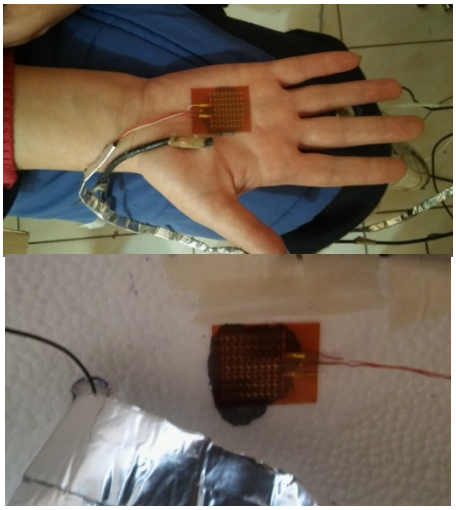


Figure 3: The heat flow sensor HF4 and temperature sensor LM35 fixed to the palm of a hand and to the polystyrene spheroid internal surface.

3.2. Methods

We use a thermal modeling method common in power electronics design, specifically the continued fraction or Cauer model [14].

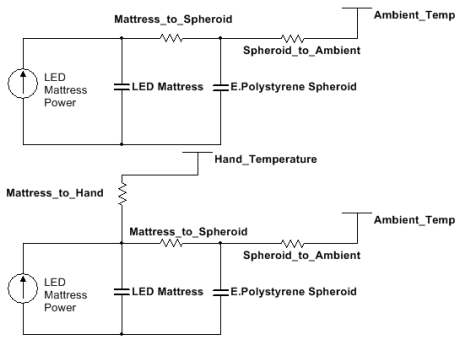


Figure 4: Simplified thermal models for the experiments.

In figure 4 the upper circuit represents the simplified thermal models for the first experiment (without the hand) and the lower circuit represents the experiment with the hand ; the main objective is to determine the thermal conductance between the LED mattress and the human hand.

The first is experiment is to study the thermal behavior of the phototherapy mattress inside the polystyrene spheroid without the hand inside, this is to find the effective thermal specific heat of the mattress, the effective thermal conductance (convective plus radiative) between the phototherapy mattress and the polystyrene spheroid and the effective thermal conductance between the polystyrene spheroid and the surrounding ambient air. The mattress has been kept hanged up inside the polystyerene spheroid and folded in two as shown in figure 5 with the two folded light emmiting surfaces parallel to the hemispheres that form the spheroid that is also kept vertically hanged up by a cotton thread as shown.

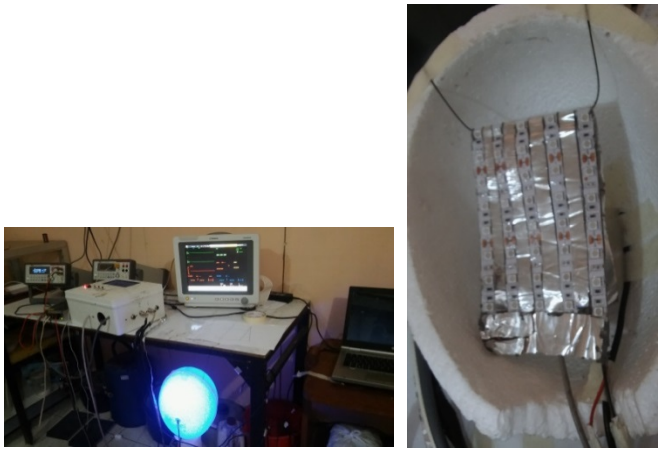


Figure 5: Phototherapy mattress hanged up and folded in two for the experiment without the hand.

The second experiment is the experiment with the hand of a healthy volunteer 48 years old and 70 Kg weight, the total surface area of his hand was measured as 0.024 m^2; the two halves of the polystyrene spheroid have been cut in the southern pole to allow the forearm to pass remain outside while the hand is inside the spheroid as shown in figure 6.

Figure 6: Experimental setup for the experiment with the hand.

3.3. Conductive, convective and radiating heat transfer:

- The effective thermal capacity of the phototherapy mattress will be determined in the experiment without the hand; as soon as the mattress is turned on, electrical power minus the radiative power emitted as light by the mattress begins to heat it, increasing the mattress temperature with approximate constant slope the first two minutes.

- The thermal specific heat of expanded polystyerene spheroid is assumed as 1500 J/(Kg ^{o}K). With this value and the known weight (50g), the spheroid's thermal capacity is 75J/^{o}K so the temperature slope related to power is 0.0133 ^{o}K/(Watt second) that will be used in the to model the experiments.
- Thermal conductance of the spheroid is calculated knowing the thermal conductivity of expanded polysterene assumed 0.035W/m^{o}K . The approximate surface of the spheroid is calculated from its dimensions: averaged internal and external dimensions: c_av=12.75cm, a_av=9cm. As the thickness of the spheroid is 25mm, the thermal conductance Gspher= 0.354 W/^{o}C; the heat loss of the spheroid to the surrounding air is assumed purely convective assuming low radiative losses, it will be proportional to the temperature difference between the core of the spheroid and the surrounding air temperature. Experimentally, the core of the spheroid temperature is calculated as the average of the inner and outer wall temperatures of the polystyrene spheroid given by the temperature sensors.
- The convective heat loss from the phototherapy mattress to the surrounding polystyrene spheroid is assumed proportional to the temperature difference between the mattress and the spheroid core temperature. The unknown convective thermal conductance will be adjusted by trial and error to make the Simulink model match the results of the experiment without the hand.
- The human hand and the phototherapy mattress are in contact in the second experiment, so, power transfer from the mattress to the hand can be modeled by a constant heat conductance so Pcond= Gcond*(Tmattress-Thand) where power will be in Watts, temperatures in degrees celsius and thermal conductance Gcond in Watt/oC. The effective value for Gcond will be determined so the model matches the results of the experiment with the hand.

3.4. Simulink model of the experiment without the hand

Two Simulink models are proposed to compare and adjust some unknown model parameters to match modeling results to the experimental results. The two models use an integrator to model temperature of the phototherapy mattress (output of the integrator); the input to the integrator is the net power received followed by a gain representing the inverse of the heat capacity of the mattress (initially unknown but this parameter will be set according to the results of the experiment without the hand as described in the previous section). Another integrator is used to model the core temperature of the polystyrene spheroid. The Simulink model for the experiment without the hand is shown in figure 7.

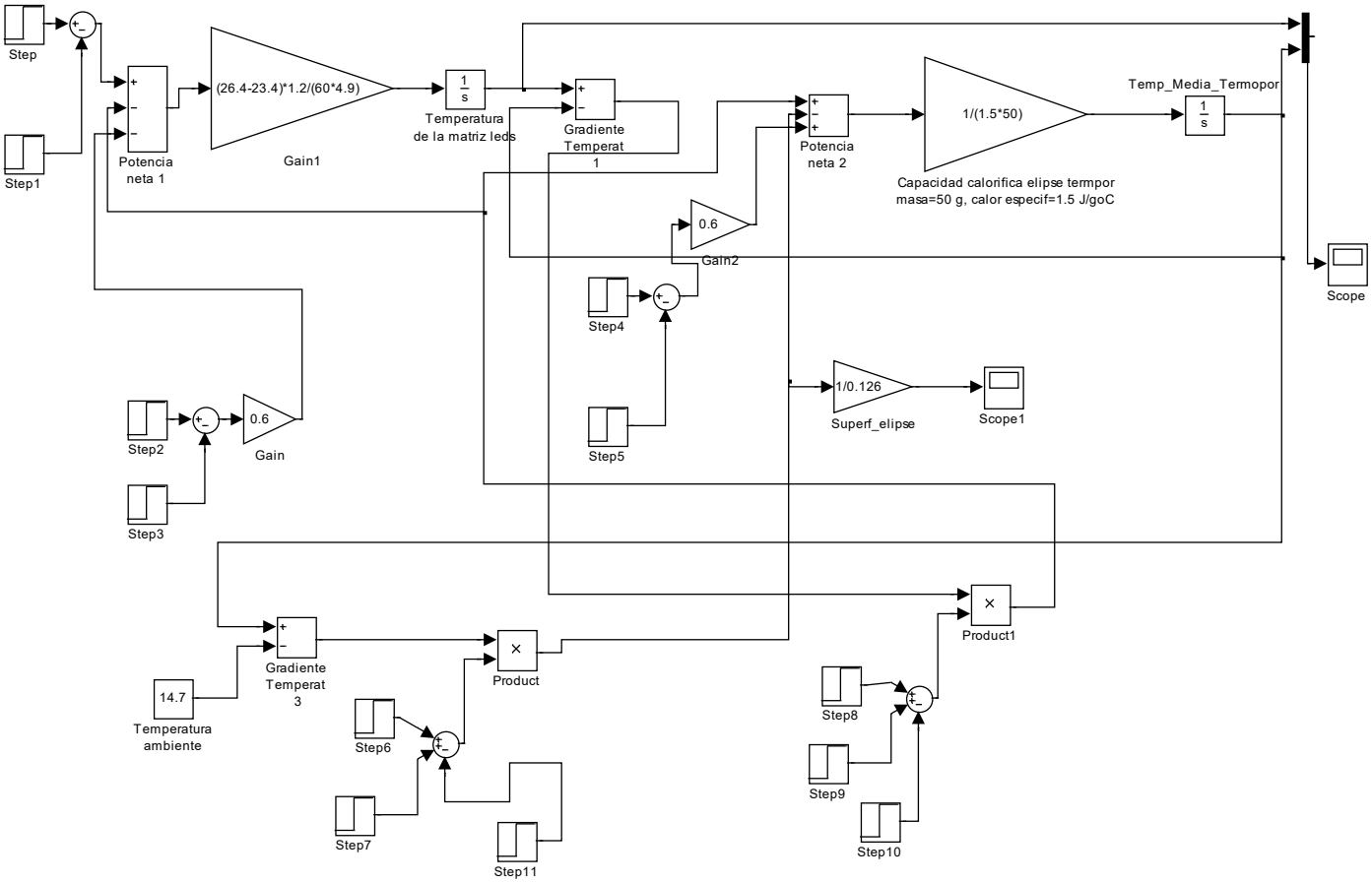


Figure 7. Simulink model for the experiment without the hand.

Step and Step1 sources model the electric power applied to the mattress (6.4 Watt) that starts in zero seconds and finishes in 960 seconds. Part of this power converts to blue light irradiance that goes directly from the LEDs to the polystyrene spheroid. The power corresponding to this irradiance is modeled by Step2 and Step3 sources and has to me substracted from the electrical power because it does not contribute to heating the phototherapy mattress; inversely the same irradiance power modeled by Step4 and Step5 sum directly as power gain to the spheroid. Gain 1 models the inverse of the thermal capacity of the phototherapy mattress, this parameter is adjusted later to match experimental results. The blue light irradiance and the corresponding power are calculated with experimental data from the heat flow sensor described above and the known interior surface area of the spheroid. The upper left integrador is for the mattress temperature and the upper right integrator is for the polystyrene spheroid whose gain at the left represents the inverse of its heat capacitance.

The power loss from the spheroid to the surrounding air is modeled by the temperature gradient with respect to the constant air temperature of 14.7 ºC (in the lower left of figure 7) and the Simulink block product, that models the thermal convective conductance together with Step6, Step7 and Step11 sources (a time variant approach is used to increase the convective conductance that is 0.354 W/ºC from 0 to 480seconds to 0.454 W/ºC from 480 to 1440 seconds to match experimental data, after this time, conductance returns to the original value of 0.354 W/ºC until the end of the simulated experiment time. Two values of conductance have been necessary because with only one conductance the difference between simulation and experiment mattress temperatures was too high.

The same approach of time variant convective conductance from the phototherapy mattress to the spheroid is implemented by the block Product1 and the Step8, Step 9 and Step10 sources. The power loss from the phototherapy mattress to the spheroid is proportional to the temperature gradient between the mattress and the spheroid (the substraction of the outputs of both integrators) and the conductance that is adjusted to match the experimental data. It is necessary to state that the phototherapy mattress has been turned off when its temperature approached 50oC because of the temperature maximum range of the physiological monitor used to record the mattress temperature. The Simulink model has been adjusted to reflect what happened in the experiment at this respect.

3.5. Simulink model of the experiment with the hand

The Simulink model for the experiment with the hand is shown in figure 8. The range of temperatures observed in the experiment with the hand allows the model to be accurate enough with a constant conductance (0.15 W/ºC) for the mattress to spheroid and a constant conductance (0.454 W/ºC) for the spheroid to ambient; these values are the same that were obtained in the model without the hand. Also, the same thermal capacity of the mattress is used (Gain block entering the integrator). The main difference here is the "hand effect" that is represented by a heat transfer proportional to the temperature gradient between the mattress and the hand, the hand is being represented by constant temperature of 36.5 ºC according to the experiment data. The gain block following the gradient represents the thermal conductance between the phototherapy mattress to the palm of the hand.

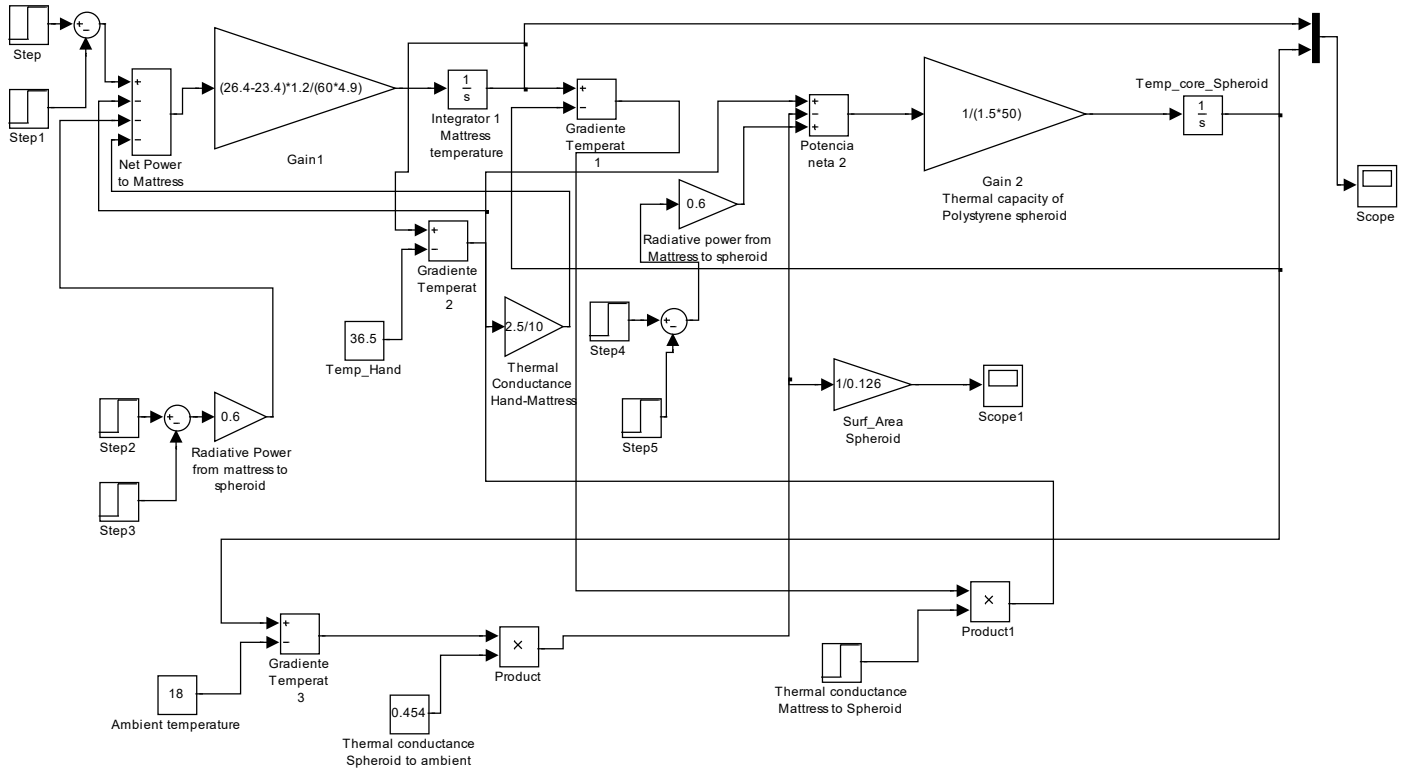


Figure 8. Simulink model for the experiment with the hand.

4. Results

4.1. Results of the experiment without the hand.

The initial temperature of the mattress and polystyrene spheroid were 18.2 ºC and 14.8ºC respectively, the electrical power provided to the phototherapy mattress was 6.4W during 16 minutes until its temperature achieved 49.5 ºC and power was turned off manually to zero. The exponential curve of the mattress temperature is of the charging capacitor type, the same occurs to the polystyrene spheroid core calculated as described in section 3.3.

Figure 9 shows the experimental data obtained with the first experiment (wihout the hand): Tmattress, Tcore_spheroid, Heat flow in W/m^2 , measured in the inner wall of the polystyrene spheroid, and also, the simulated mattress temperature with the Simulink model with the parameter values adjusted by trial and error to obtain a match between the experiment and simulation.

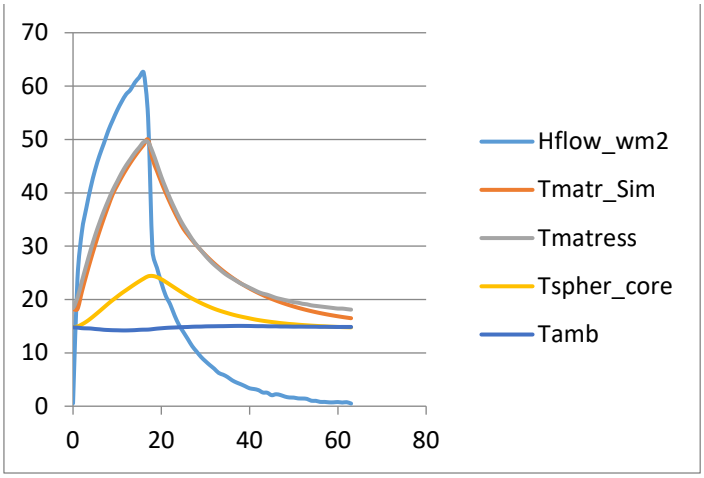


Figure 9. Simulink model for the experiment without the hand.

The first interesting result is the heat flow waveform at the spheroid inner wall showing an immediate transfer of radiating power from the phototherapy mattress because of the step increase from 0.5 to 23 W/m^2 , this justifies the radiating power sources in the Simulink model. There is the same type of variation when the phototherapy mattress is turned off.

Next, the phototherapy mattress temperature is showed in grey color overlaps closely the orange color simulated temperature of the mattress, this has been obtained adjusting by trial and error the model parameters to the next values: The thermal conductance of the mattress to the spheroid is not constant trough the simulation time with a value of 0.12 W/ºC the first 8 minutes, then 0.15 W/ºC until the 24th minute and returning to 0.12 until the end of the experiment. An average thermal conductance value of 0.135W/ºC from the mattress to the spheroid could be used for design purposes in future studies.

Also in the simulation, the thermal conductance from the polystyrene spheroid to the ambient is not constant neither, with values: 0.354 W/ºC and 0.454 W/ºC the same intervals of time indicated above. The thermal capacity for the phototherapy mattress in the simulation is obtained from the first minute of the mattress temperature ramp of the experiment giving a value of 3.4 J/ºK and knowing the 50g mass of mattress, the effective specific heat of the mattress results 68 J/KgºK. For reference, the calculated thermal capacity of the polystyrene spheroid is 70 J/ºK.

The temperature of the mattress during the experiment is limited only because the power has been turned off before it reached 50ºC. It is expected that the hand in the next experiment would be a limiting factor for the phototherapy mattress temperature.

4.2. Results of the experiment with the hand.

Figure 10 shows the results of the experiment with the hand with time in minutes. The first minutes are with the mattress turned off, so The heat flow sensor has been fixed with adhesive tape in the borders to the palm of the hand. The negative values mean heat is flowing from the hand to the outside and positive values that heat flows into the hand. The light yellow curve is the heat flow in W/m2 and show that while the matress is off, an average heat flow of 3.5W/m2 goes from the hand to the mattress that is folded around the hand, during this time the hand is hotter than the mattress by about 1.5ºC (34.5ºC-33ºC).

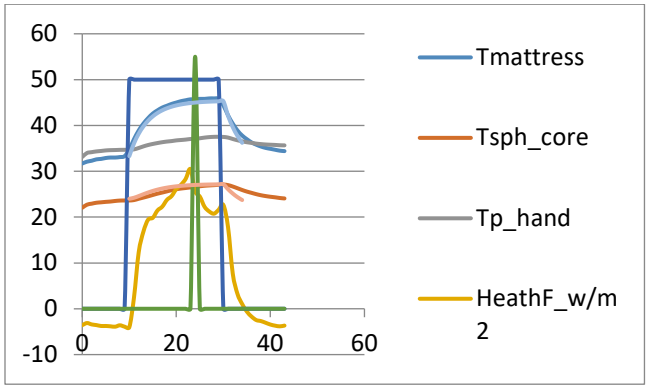


Figure 10. Results for the experiment with the hand.

When the mattress is turned on, the heat flow rapidly changes sign while the mattress's temperature rises but, this time, limited by the temperature of the hand that remains below 37.5ºC so, the phototherapy mattress reaches 45.8 ºC asymptotically. In the later part of the temperature rising, an occlusion of the brachial artery is produced with the help of the physiological monitor by means of the non invasive blood pressure measurement (NIBP) procedure, this causes a temporal descent of the heat flow to the hand that recovers just before the phototherapy mattress is turned off for the latest part of the experiment. The waveform of the heat flow to the hand during the heating phase is irregular, peaking before the occlusion. The Simulink model for this experiment takes most of the parameters from the previous model without the hand. To match the results of the simulated mattress temperature to the experiment with the hand, the thermal conductance between the mattress and the hand has been found by trial and error to be 0.25 W/ºC, there is a close match between simulation and experimental results (curves Tmattress and Tmat_simul). The volunteer reported he felt his hand started to sweat when the mattress temperature rose above 39ºC and felt thermal discomfort after the NIBP occlusion.

5. Discussion

Regarding neonate-mother mannequin pair in kangaroo mother care, in [1] the contact point between the 1Watt phototherapy mattress and the back of a thermally regulated neonatal mannequin was 2.75 ºC above the mannequin

temperature, but now for a 6.4W device, the temperature of the palm of a real hand incremented only 0.5ºC above core temperature (37ºC), so, the real hand has a much higher ability to sink power from the phototherapy device due to blood perfusion in the hand compared to a silicone rubber mannequin. It can be expected that the back of a real neonate as a blood perfused tissue that can sweat would be more able to absorb heat than the mannequin in [1].

Comparing the temperature of the phototherapy mattress in both experiments (Figures 8 and 9) , without the hand it rises up to 50ºC with no limit, contrary to this, in the experiment with the hand, it acts as a heat sink that limits the mattress temperature below 45ºC. The effective thermal conductance of 0.25 W/ºC and the approximate 10 ºC temperature gradient between mattress and hand gives about 2.5W absorbed by the hand. But the measured heat flow to the palm of the hand with the heat flow sensor was 25 W/m^2; when considering a hand with 0.024 m^2 surface, a total power of 0.6Watt would be absorbed by the hand, the difference (2.5W- 0.6W) might be explained by the sweating of the hand.

Our results are in agreement to Taylor [8] that indicates a maximum theoretical transfer of 12W to both hands with 1ºC skin to core temperature gradient (we obtained 2.5W to one hand with 0.5ºC skin to core temperature). We suggest to extend this study to understand the sweating effect, measuring relative humidity inside the expanded polystyrene spheroid. Finally with this results, in Kangaroo Mother Care (KMC) position, the hand of a mother holding the phototherapy mattress would have a contact surface with the mattress of approximately 50% of the surface in this experiment (only the palm, not the hand's dorsum), so the expected conductance would be 0.125 W/ºC in KMC.

6. Conclusions

The effective thermal capacitance of a 6.4 W phototherapy mattress and the thermal conductance from it to a human hand have been calculated from confined experiments inside a polystyerene spheroid, this thermal model parameters will be used in a future clinical study of heat transfer processes involving a similar phototherapy mattress and real mothers providing Kangaroo Mother Care to neonates requiring jaundice phototherapy.

The ability of a thin film heat flow sensor based on thermocouple arrays to measure radiating heat transfer has been confirmed, it opens the way to use it to study heat transfer processes where clinical radiant warmers are used.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgments

The authors wish to thank Universidad Nacional de San Antonio Abad del Cusco and Hospital Regional Cusco for supporting this study through the project "Effectiveness of blue light irradiance adapted to Kangaroo Mother Care in neonatal jaundice treatment"

References

[1] Jimenez, L. Vilcahuaman, J. Galdos "Development and thermal assessment of a blue light emitting diode phototherapy device for neonatal jaundice treatment in Kangaroo mother care" in Healthcare Innovation Point of Care Technologies (HI-POCT), 2016 IEEE, Cancún Mexico, 2016.

[2] Ibe O.E., Austin T., Sullivan K., Fabanwo O., Disu E.,Costello A.M. "A comparison of kangaroo mother care and conventional incubator care for thermal regulation of infants <2000 g in Nigeria using continuous ambulatory temperature monitoring" Annals of Tropical Paediatrics (2004) 24, 245-251.

[3] Maisels MJ, Kring EA, DeRidder J. "Randomized controlled trial of light-emmiting diode phototherapy" Journal of Perinatology 2007, 27; 565-567

[4] Samra NM1, El Taweel A, Cadwell K. "The effect of kangaroo mother care on the duration of phototherapy of infants re-admitted for neonatal jaundice". J. Matern Fetal Neonatal Med. 2012. Aug; 25(8):1354-7.

[5] JG Webster, Editor. Design of Pulse Oximeters, IOP Publishing Ltd 1997.

[6] Ducharme MB, Tikuisis P. "Role of blood as heat source or sink in human limbs during local cooling and heating" Journal of Applied Physiology 76(5):2084-94, 1994.

[7] Sienkiewicz ZJ1, O'Hagan JB, Muirhead CR, Pearson AJ "Relationship between local temperature and heat transfer through the hand and wrist" Bioelectromagnetics. 1989;10(1):77-84

[8] Taylor NA, Machado-Moreira CA, van den Heuvel AM, Caldwell "Hands and feet: physiological insulators, radiators and evaporators" JN Eur J Appl Physiol. 2014 Oct;114(10):2037-60

[9] Heller H. Craig, Grahn Dennis A. "Enhancing Thermal Exchange in Humans and Practical Applications " Disruptive Science and Technology. Volume 1, Number 1, 2012. Mary Ann Liebert, Inc.

[10] Boron K, Kos A. "Thermal model of selected parts of human hand and thermal touch screen for the blind" Metrol. Meas. Syst., Vol. XIX (2012), No 3, pp. 593-602.

[11] World Health Organization "Kangaroo Mother Care, A practical guide" Department of Reproductive Health and Research, 2003

[12] Guiatni M, Kheddar A. "Theoretical and Experimental Study of a Heat Transfer Model for Thermal Feedback in Virtual Environments" 2008 IEEE/RSJ International Conference on Intelligent Robots and Systems.

[13] Mak, L.; Farnworth, B.; Ducharme, M.; Kuczora, A.; Sweeney, D.; Uglene,W.; Hackett, P.; Potter, P. "Thermal protection measurement of immersion suit comparison of two manikins with humans pilot study report" Technical Report 2010-04 . National Research Council Canada.

[14] Kunzi R. "Thermal design of power electronic circuits" Proceedings of the CAS-CERN Accelerator School. CERN, Geneva 2015

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 812-818 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Network Intrusion Detection System using Apache Storm

Muhammad Asif Manzoor*, Yasser Morgan

Faculty of Engineering and Applied Sciences, University of Regina, SK, Canada

A R T I C L E I N F O

Article history:
Received: 05 April, 2017
Accepted: 27 May, 2017
Online: 19 June, 2017

Keywords:
Network Intrusion Detection
Support Vector Machine
Apache Storm
LibSVM
KDD 99

A B S T R A C T

Network security implements various strategies for the identification and prevention of security breaches. Network intrusion detection is a critical component of network management for security, quality of service and other purposes. These systems allow early detection of network intrusion and malicious activities; so that the Network Security infrastructure can react to mitigate these threats. Various systems are proposed to enhance the network security. We are proposing to use anomaly based network intrusion detection system in this work. Anomaly based intrusion detection system can identify the new network threats. We also propose to use Real-time Big Data Stream Processing Framework, Apache Storm, for the implementation of network intrusion detection system. Apache Storm can help to manage the network traffic which is generated at enormous speed and size and the network traffic speed and size is constantly increasing. We have used Support Vector Machine in this work. We use Knowledge Discovery and Data Mining 1999 (KDD'99) dataset to test and evaluate our proposed solution.

1. Introduction

Network security is of increasing importance than ever with increased usage of network-based computing resources. Several network security mechanisms like encryption, firewalls, cryptography, access control, authentication, and intrusion detection are used to provide network security. These techniques are used to detect and prevent the malicious activities and network attacks. Network intrusion detection is an essential task for any network based computing environment. Network Intrusion Detection system continuously monitors the network traffic and perform in-depth packet analysis for malicious activities and attacks in real time. Once the intrusions are detected, appropriate actions can be taken to mitigate the threat against the computing and network resources. Currently, computer networks are generating a huge amount of data traffic at enormous speed and this data generation speed is continuously increasing. Network traffic also contains different types (variety) of data. Network traffic satisfies all the three V's of big data; volume, velocity, and variety. The focus of this paper is to propose a network intrusion detection mechanism that can deal with all the 3 V's of big data.

This paper is an extension of work originally presented in 2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON) [1]. In original work, we have used anomaly-based approach for network intrusion detection. Anomaly based approaches use machine learning algorithms to identify the abnormal behavior from the normal data traffic. We have used statistical approaches for feature reduction and support vector machine with the linear kernel for the classification process. We also have proposed an Apache Storm topology for the real-time big data streaming application. The current work uses Radial Basis Function (RBF) kernel for the support vector machine. This paper discusses the class imbalance problem and its possible solutions. We also discuss multiple techniques to reduce the false positive and false negative rates in the context of network intrusion detection system.

Mostly, network intrusion detection system follows one of the two major detection mechanism; Anomaly-based network intrusion detection and Signature based network intrusion detection. Some researchers also have proposed hybrid approaches. Each detection approach has its own strengths and weaknesses.

1.1. Signature Based Network Intrusion Detection

Signature based or misuse intrusion detection systems have better detection rate as compared to anomaly based systems in case of known network attacks. This type of intrusion detection systems scans predetermined signatures in network traffic packets and recognizes predefined patterns of network intrusions and attacks.

*Muhammad Asif Manzoor, Faculty of Engineering and Applied Sciences, University of Regina, SK, Canada, AsifManzoor@uregina.ca

The signature based system cannot detect new potential threats until their signature is added to the database. During the time, when a new threat is detected and its signature is applied, these systems cannot detect the new threats. Malicious strings in TCP/IP packets, abnormal flag combination, request for non-existing services, and intrusion attack signature are few of the approaches [2], [3] used to design signature based network intrusion detection systems.

1.2. Anomaly Based Network Intrusion Detection

Anomaly based system try to capture the behavior of incoming network traffic. These systems establish a baseline model for normal user/network behavior against the abnormal behavior. This model is created by analyzing the network traffic over a period of time and train the system to classify the normal traffic and malicious traffic. The baseline model reflects the valid activities of applications, hosts, users, and network. All the incoming network traffic is scanned using the defined model to classify it into normal and intrusion network packet. Alarms are raised for all the packets which are classified as intrusions. Anomaly based detection systems are good to detect new network threat at the cost of increased false alarm rate (both false positive and false negative detection). Despite this problem, anomaly based detection is an indispensable method due to increase in new type intrusion attacks.

The rest of paper is organized as follow, anomaly based intrusion detection systems are review in section II. Section III provides the detailed description of KDD 99 data set and also discusses the pre-processing techniques used in this work. Implementation details and methodology is discussed in Section IV. Experimental results are discussed in section V. Class imbalance problem and false positive/negative problem for network intrusion detection are discussed in Section VI and the paper is concluded in Section VII.

2. Literature Review

Machine learning algorithms are used to design anomaly based intrusion detection system. Both unsupervised and supervised learning methods are used in this domain. Although many intrusion detection methods are proposed using various different machine algorithms we will review Artificial Neural Network (supervised learning) and K-Mean Nearest neighbor (unsupervised learning) based approaches.

Artificial Neural Networks are used by many researchers as supervised learning algorithm to train the intrusion detection technique. Rapake et al. [5] proposed to use numbers of system calls that are executed on the host machine. These numbers are utilized for neural network training. Han and Cho [6] proposed to use evolutionary neural networks to detect anomalies based on learning the behavior of the program. Authors have used 1999 DARPA IDEVAL data set to evaluate their proposed method. Liang et al. [4] proposed to apply Fisher feature selection algorithm on KDD 99 dataset for performance enhancement. Authors used artificial neural networks to develop their intrusion detection system.

K-Mean Nearest Neighbor (KNN) is a clustering algorithm based on unsupervised learning method. Wang et al. [7] used KDD 99 dataset and defined new features based on density, cluster centers, and nearest neighbors. KNN is used to classify the normal traffic and network intrusion. Li et al. in [8] combined KNN algorithm with Particle Swarm Optimization (PSO) to design hybrid system. The K-mean algorithm suffers from premature convergence. PSO algorithm helps to avoid premature convergence in the proposed method. Xian et al. [9] proposed to unite fuzzy KNN algorithm with clonal selection algorithm to design network intrusion detection. Jiang et al. [10] used incremental KNN algorithm to detect intrusions. Authors used outlier factor to calculate the deviation degree of the cluster in this work.

Multiple network intrusion detection systems are proposed using data mining algorithms. Han et al. [11] used data mining approach to detect intrusions. They have analyzed attributes of network traffic protocol to identify misuse signature. They also have analyzed network packet contents for intrusion signatures. Qin and Hwang [12] proposed internet trace technique which discards few non-functionary frequent episodes rules dynamically. These episode rules are utilized to recognize abnormal sequences in network traffic connections. Otey et al. [13] develop a general purpose tunable algorithm to detect outlier/anomalies. The proposed algorithm can work with mixed attribute dataset as well dynamic and streaming data sets. KDD 99 intrusion dataset is one of the dataset used to test and evaluate the proposed tunable algorithm.

Gondal et al. [14][14] determined center of clusters on basis of diversity. These Diversity-based centroids are used to develop network intrusion detection system. Xiao et al. in [15] used Bayesian Network Model Averaging in their proposed work and compared it with Naïve Bayes classifier and Bayesian Network classifier. Authors have evaluated their work over NSL-KDD dataset.

Jeong et al. [16] proposed in their work to used Discrete Wavelet Transform to extract the features from network traffic. Principal Component Analysis (PCA) is applied to the extracted features to identify principal components and a visual analytics tool is proposed for intrusion detection. Le et al. [17] used deep learning approach in their network intrusion detection system. Authors explored various optimizers with Long Short-Term Memory Recurrent Neural Network.

Table I: Details of Dataset

	Training	**Testing**
Normal	97280	60623
Probe	4107	4166
DoS	391458	229853
U2R	52	13939
R2L	1124	2478
Total	494021	321026

Table II: Network Attacks and Categories

Category	Network Attacks
Denial of Service (DoS)	back, land, neptune, pod, smurf, teardrop, **apache2, mailbomb, processtable**
Probe	ipsweep, nmap, portsweep, satan, **mscan, saint**
User to Root (U2R)	ftpwrite, guespasswd, imap, multihop, phf, spy, warezclient, Warezmaster, **httptunnel, named, sendmail, snmpgetattack, xlock, xsnoop**
Remote to Local (R2L)	bufferiverflow, loadmodule, perl, rootkit, **ps, snmpguess, sqlattack, worm, xterm**

3. Knowledge and Discovery 1999 Dataset (KDD 99)

Knowledge and Discovery 1999 Network Intrusion Detection dataset (KDD 99) [18] is used in this work to test and evaluate the performance of the proposed method. KDD 99 is publicly available dataset and widely used by various researchers. The complete KDD 99 dataset contains around five million records for training and 3 million records for testing tasks. Due to the large size of the original dataset, smaller version of KDD 99 dataset is frequently used for evaluating intrusion detection systems. This smaller version contains 10% of testing and training records. We also have tested our intrusion detection system using smaller dataset. KDD 99 dataset contains normal records as well as malicious records. Twenty two types of attacks are included in training dataset along with normal network packets. Seventeen additional attacks are added to the testing dataset and can be used to determine whether the proposed algorithm can detect new attacks or not. Network attacks are divided into four categories; Denial of Service (DoS), Probe, User-to-Root (U2R), and Remote-to-Local (R2L). Table I provides the distribution of training and testing dataset into normal and intrusion packets. KDD 99 dataset consists of 494021 training records and 311029 testing records. The division of network attacks into categories is given in Table II. Network attacks in bold are new attacks added to testing dataset only.

Each record has 41 attributes; description of these attributes is given below:

- 9 basic and single connection derived (SCD) header features,
- 9 time based multiple connection derived (MCD) header features,
- 10 host based multiple connection derived (MCD) header features, and
- 13 content based features collected from traffic payloads.

3.1. Pre-Processing

KDD 99 data set is pre-processed in order to make it appropriate for the machine learning algorithm. The second reason for the pre-processing is to enhance the intrusion detection rate. Pre-processing is performed in three steps.

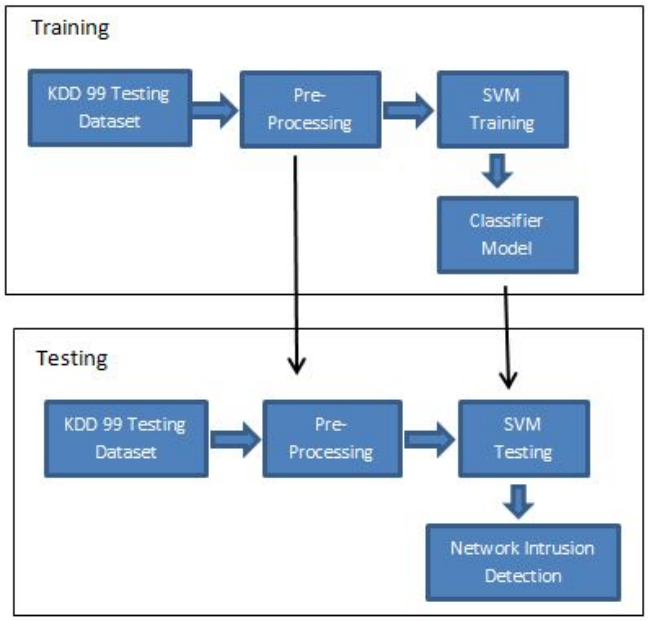


Figure 1: Network Intrusion Detection System Architecture

1. Each record in the dataset consists of numeric as well as categorical attributes. Textual data is used for categorical attributes. Support vector machine algorithm requires numeric data (either discrete or continuous). The first step in pre-processing is to convert this categorical attributes to numeric attributes. The dataset contains three categorical attributes while rest of the thirty eight attributes are numeric. Every category of an attribute is assigned a specific number.

2. Next, we have normalized all the numeric attributes in the training dataset. Testing dataset attributes are normalized using ratios determined for training dataset, so both (training and testing) subsets are normalized with same ratios. Normalization process converts all attributes to the same range; hence each attribute has a similar effect during the training process. The normalization process also significantly reduces the training time for support vector machine algorithm.

3. We applied statistical operations on the training dataset to reduce the features. Same features are dropped from testing dataset. These dropped attributes had a negative impact on the performance of network intrusion detection system. Reduce the number of attributes have a positive effect on the computation time.

4. We have used LibSVM library for SVM implementation; it requires training and testing data in a specific format. Lastly, we have converted the dataset to LibSVM compatible format.

4. Methodology

The architecture of Support Vector based Network Intrusion Detection system is given in Figure 1. KDD 99 training data set is pre-processed and Support Vector Machine (SVM) algorithm is applied to build a model. Pre-processing information determined for the training dataset is used to pre-process the training dataset. SVM model built during the training process is used to predict the category during the testing process.

4.1. Support Vector Machine

Support Vector Machine algorithm [19] is developed to solve binary classification problems. SVM is supervised learning algorithm which can be used for classification and regression problems and it is based on statistical learning approach. The fundamental idea of SVM is to identify hyperplanes (decision boundaries) between two categories based on the training data and one providing the maximum separation between two categories is selected as shown in Figure 2. This hyperplane provides the optimal separation between two categories. The identification of optimal hyperplane is considered as an optimization problem. Testing examples are projected and the category is predicted based on their position with respect to the optimal hyperplane. Cover's theorem [20] states; if linearly non-separable dataset is transformed using non-linear kernel function then it is probable that the transformation process will generate linearly separable dataset in higher dimensional space. Support vector machine also uses different types of non-linear kernel function to create linearly separable dataset and increase classification performance.

Let we have a training dataset S= $\{(x_1, y_1), (x_2, y_2),\ldots, (x_n, y_n)\}$, where $x_i \in R^n$ and it represents input feature vector and $y_i \in \{-1, 1\}$ is the category. Let weight and bias of hyperplane is given by w and b. The non-linear Kernel function is applied on training dataset to transform it into linearly separable categories and $\varphi(x)$ is used to represent it. The hyperplane between the two categories can be defined as:

$$w.\varphi(x) + b = 0 \qquad (1)$$

The optimization problem for calculation of w and b is:

Minimize

$$\varphi(w) = \tfrac{1}{2} \|w\|^2 \qquad (2)$$

Subject to

$$y_i(w.\varphi(x) + b) \geq 1$$

New variables ζ_i (slack variable) and C (regularization constant) are included in above optimization problem:

Minimize

$$\varphi(w,\zeta) = \tfrac{1}{2} \|w\|^2 + C\sum_{i=1}^{N} \zeta_i \qquad (3)$$

Subject to

$$y_i(w.\varphi(x) + b) \geq 1 - \zeta_i, \zeta_i \geq 0$$

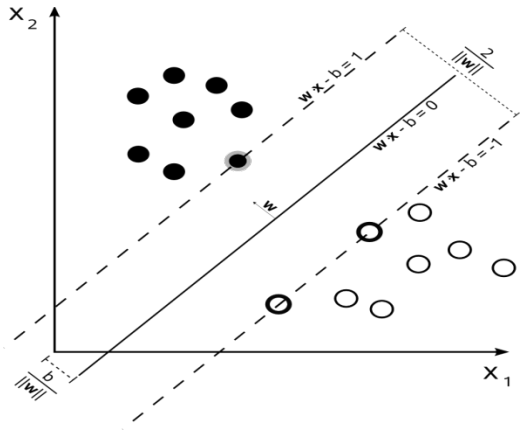


Figure 2: Decision Boundary with maximum margin

Where ζ is used to relax the hard margin constraint and C is used to manage the trade-off between classification error and maximal margin of separation.

Multiclass SVM is often constructed using basic binary SVMs to deal with real world problems like pedestrian detection [21], medical diagnosis [22], sentiment analysis [23] and many others. For multiclass SVM, 1-vs-1 [24] and 1-vs-all [25] are two approaches used to build basic binary SVMs.

We have used LibSVM [26] Java API in this work for SVM implementation. LibSVM provides both regression and classification implementations. We have used C-SVM approach to classifying the network traffic data into normal and intrusions. Radial Basis Function (non-linear kernel function) is used in this work for transformation into higher dimensional space. The value of gamma used here is 1 and the value of C is 6. As kernel implementation is part of LibSVM library; hence we are considering it a part of SVM training process and not a part of pre-processing. Finally, we have used the 1-vs-1 multiclass approach in this work.

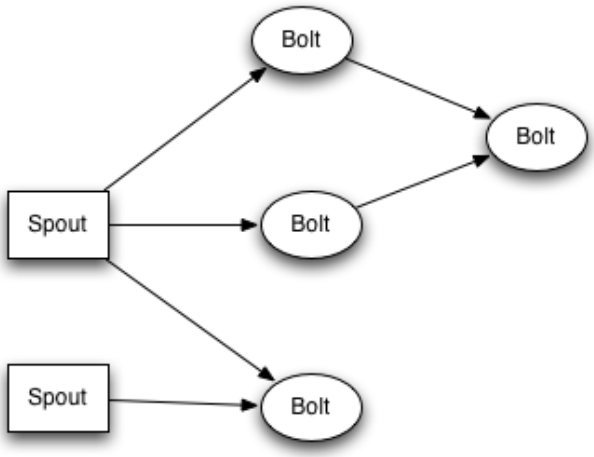


Figure 3: Apache Storm Topology

4.2. Apache Storm

Apache Storm is a scalable distributed framework that can be easily configured on cloud services or clusters. It is developed to process real-time big streaming data. Spouts and bolts are the basic building blocks for Apache Storm applications. Spouts are used as data sources; data is forwarded in form of tuples to the bolts. Bolts are the computation unit of the Apache Storm. Bolts process and evaluate the incoming tuples and can forward the results in form of tuple to other bolts for further processing. Spouts and bolts are combined to form a topology. Topologies are directed acyclic graphs and used to develop Apache Storm applications. An Apache Storm topology is shown in Figure 3. The edges of the topology show the flow of data. Bolts and spouts are vertices in the topology. We can create many instances of every bolt and spout to speed up the processing.

Apache Storm framework offers distributed and fault tolerant computing. An Apache Storm cluster is shown in Figure 4. The cluster consists of one or multiple worker nodes and only one master node. Worker nodes are named as supervisors while the master node is named as nimbus. Apache Storm also requires ZooKeeper cluster to provide coordination between the master node and worker node(s).

The proposed Apache Storm topology is given in Figure 5. This topology has one spout and three bolts. The detail of the bolts and spout is given below:

1. **Input Reader (Spout)** is the only data source in this topology. KDD 99 dataset is stored in a text file. This spout reads the KDD 99 dataset and each record is forwarded as a tuple to the next bolt.
2. **Data Pre-Processer (Bolt)** receives the tuples from Input Reader and performs pre-processing as explained in Section 3. It converts the categorical data to numeric data, perform normalization and feature reduction. Finally, it converts the data into LibSVM compatible format.

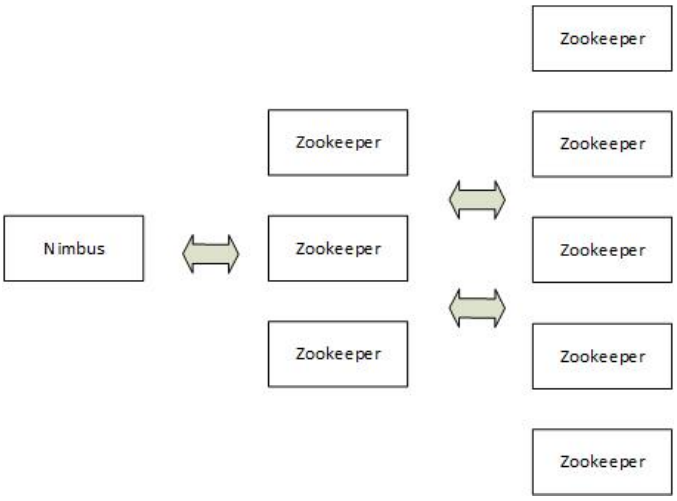


Figure 4: Apache Storm Cluster

Figure 5: Apache Storm Topology

3. **Support Vector Machine (Bolt)** performs the classification process for the network intrusion detection system.
4. **Result Aggregator (Bolt)** is the last processing unit in this topology. It aggregates the classification results. It stores these results on a text file for further analysis.

5. Experimental Results and Discussion

5.1. Experimental Setup

Apache Storm 0.9.6 is configured on Ubuntu 12.04. All the experiments are performed on 3.4 GHz Intel Core i7 processor with 16.00 GB of memory.

5.2. Evaluation and Analysis

We have used KDD 99 intrusion detection dataset to test our SVM based method. The testing dataset contains 18729 records belonging to new network intrusion attacks (17 new network attacks included) which are not included in the training dataset. Whereas 292300 records belong to network intrusion attacks which are also part of the training dataset. The accuracy for intrusion detection of the proposed method is 98.03% when only known attacks are considered (Network attacks used during training process). This intrusion detection rate drops to 92. 60% when all the network attacks are considered during the testing process. All the discussion, following this point, will be based on the complete testing dataset. The classification results for complete dataset (new network attacks included) of the proposed intrusion detection method are given in Table III. A simplified confusion matrix is given in Table IV; which only divides the classification results into normal and malicious network traffic.

Table III: Confusion Matrix for Network Intrusion Detection System

	Normal	Dos	Probe	R2L	U2R
Normal	60294	66	225	7	1
DoS	6571	223266	16	0	0
Probe	760	441	2965	0	0
R2L	12849	91	72	927	0
U2R	2462	3	0	7	6

Table IV: Simplified Confusion Matrix

	Normal	Attack
Normal	60294	299
Attack	22642	227794

The performance of the can also be determined using False Positive Rate (FPR), False Negative Rate (FNR), Precision and Recall. These four performance evaluators are calculated below using simplified confusion matrix (Table IV).

FPR = FP / (FP + TN) = 22642 / (22642 + 227794) = 0.09

FNR = FN / (FN + TP) = 299 / (299 + 60294) = 0.005

Precision = TP / (TP+FP) = 60294 / (60294+22642) = 0.73

Recall = TP / (TP+FN) = 60294 / (60294 + 299) = 0.995

For a good classification rate, FPR and FNR must be close to zero while Precision and Recall must be close to one. Other than Precision, all other performance parameters indicate good classification performance.

Normal traffic and DoS attacks are recognized correctly mostly. Whereas Remote to Local (R2L) and User to Root (U2R) are the two categories which are recognized incorrectly most of the time. Class imbalance is the main reason behind this shortcoming. R2L and U2R have the least number of training samples. Similarly, probe attack category also has relatively less training examples. While remaining two categories have the majority of the training examples; around 98.9% of the dataset. This class imbalance creates biased results during the training process. We will discuss few techniques to solve class imbalance problem in the next section.

The proposed network intrusion detection system is supposed to handle real-time streaming data traffic. The training process is done offline and it is not a real-time task. This system can process roughly 13,600 packets in a second for testing/prediction purposes. This processing speed is achieved with a single general purpose computer. The processing speed can be further increased by using a cluster of dedicated servers. This processing speed is a good indicator for real-time big streaming data scenario.

Support Vector Machines are machine learning algorithm used for supervised classification tasks whereas Apache Storm is development platform used to develop real-time big streaming data processing applications. Apache Storm supports multiple languages for coding. We have used Java in our work. The proposed system achieves good classification results along with a good processing speed.

6. Performance Enhancement for Network Intrusion Detection Systems

Class imbalance refers to the uneven representation of classes in training dataset. The skewed distribution causes performance degradation for many machine learning algorithms [27]. The methods to solve class imbalance problem are generally divided into two categories; data level and algorithm level solutions.

6.1. Data Level Approach

Data level approaches apply different techniques to the dataset before training process. It can be considered as pre-processing step to accommodate the class distribution. Sampling, under-sampling, and over-sampling, is a data-level approach to managing class distribution. Over-sampling duplicates the records from the minority category whereas under-sampling removes records from the majority category to minimize the imbalance ratio between different categories in training data.

Yen et al. [28] proposed to use cluster based approach to solve the class imbalance problem by selecting representative data. Clusters are used to identify the representative data to achieve better prediction rate for minority categories. Yu et al. [29] used vector quantization to under-sample the majority class and constructed representative local models. Synthetic Minority Over-Sampling Technique (SMOTE) [30] is an adaptive data-level approach for over-sampling. It computes the probability distribution to model minor categories and use this probability distribution to add new examples to the smaller classes. Guo and Viktor [31] proposed to combine boosting and data generation to over-sample the minority classes. This approach is named as DataBoost-IM algorithm.

6.2. Algorithm Level Approach

In algorithm level approach, specialized machine learning algorithms are used to tackle class imbalance problem.

One of the techniques used is to develop modified version of the basic learning algorithm which can manage imbalanced classes. z-SVM [32] and GSVM-RU [33] are modified version of SVM to manage imbalanced class distribution. z- SVM tries to maximize g-mean value by moving hyperplane using a parameter z. GSVM-RU uses granular computing to enhance the classification accuracy.

Another algorithm level approach is to use cost sensitive learning; whenever a misclassification occurs, an expensive cost is imposed on the classifier. Cao et al. [34] proposed Particle Swarm Optimization (PSO) based cost sensitive artificial neural network for imbalanced data classification.

7. Conclusion

Network intrusion detection is an important component of network management and it is a defense mechanism for network security. Anomaly based intrusion detection detects anomalous behavior or new network attacks which signature-based intrusion detection cannot. Higher processing speed is required to handle incoming real-time data traffic. The proposed Network Intrusion detection system can process 13,600 packets in a second on a single general purpose computer. Apache Storm based network intrusion detection system can achieve higher processing by deploying a cluster of computers. Data pre-processing techniques are implemented to reduce the number of feature; which reduces the training and testing time and increased the classification accuracy.

In this paper, we outline the machine learning based method for anomaly based intrusion detection. We propose to utilize support vector machine as supervised learning algorithm to classify the incoming network traffic into normal and network intrusion attacks (DoS, Probe, R2L, U2R). KDD 99 intrusion detection dataset is used in this work. The data set is pre-processed in order to make incompatible with support vector machine

algorithm as well as with LibSVM library. The experimental results show the feasibility of the proposed intrusion detection method. The performance can be further enhanced by solving class imbalance problem as discussed in section 6. This paper suggests using Support Vector Machine for classification process and Apache Storm as development platform to develop network intrusion detection system to handle big data characteristics of the network data traffic.

References

[1] Manzoor, Muhammad Asif, and Yasser Morgan. "Real-time Support Vector Machine based Network Intrusion Detection system using Apache Storm." Information Technology, Electronics and Mobile Communication Conference (IEMCON), 2016 IEEE 7th Annual. IEEE, 2016.

[2] Bro, Bro Intrusion Detection System 2002.

[3] SNORT, SNORT: The open source network intrusion detection system 2002.

[4] Hu, Liang, et al. "An improved intrusion detection framework based on Artificial Neural Networks." Natural Computation (ICNC), 2015 11th International Conference on. IEEE, 2015.

[5] A. Rapaka, A. Novokhodko, and D. Wunsch, "Intrusion detection using radial basis function network on sequences of system calls," in Proc. Int. Joint Conf. Neural Netw., 2003, vol. 3, pp. 1820–1825.

[6] S. J. Han and S. B. Cho, "Evolutionary neural networks for anomaly detection based on the behavior of a program," IEEE Trans. Syst., Man, Cybern. B, Cybern., vol. 36, no. 3, pp. 559–570, Jun. 2006.

[7] Xiujuan Wang, Chenxi Zhang and Kangfeng Zheng, "Intrusion detection algorithm based on density, cluster centers, and nearest neighbors," in China Communications, vol. 13, no. 7, pp. 24-31, July 2016.

[8] Z. Li, Y. Li, and L. Xu, "Anomaly intrusion detection method based on k-means clustering algorithm with particle swarm optimization," in Proceedings of the 2011 International Conference of Information Technology, Computer Engineering and Management Sciences. Washington, DC, USA: IEEE Computer Society, 2011, pp. 157–161.

[9] J. Xian, F. Lang, and X. Tang, "A novel intrusion detection method based on clonal selection clustering algorithm," in Proc. Int. Conf. Mach. Learn. Cybern., 2005, vol. 6, pp. 3905–3910.

[10] S. Jiang, X. Song, H.Wang, J. Han, and Q. Li, "A clustering-based method for unsupervised intrusion detections," Pattern Recognit. Lett., vol. 27, no. 7, pp. 802–810, May 2006.

[11] H. Han, X. L. Lu, and L. Y. Ren, "Using data mining to discover signatures in network-based intrusion detection," in Proc. Int. Conf. Mach. Learn. Cybern., 2002, vol. 1, pp. 13–17.

[12] M. Qin and K. Hwang, "Frequent episode rules for Internet anomaly detection," in Proc. IEEE Int. Symp. Netw. Comput. Appl., 2004, pp. 161–168.

[13] M. E. Otey, A. Ghoting, and S. Parthasarathy, "Fast distributed outlier detection in mixed-attribute data sets," Data Min. Knowl. Discov., vol. 12, no. 2/3, pp. 203–228, May 2006.

[14] Gondal, M.S.; Malik, A.J.; Khan, F.A., "Network Intrusion Detection Using Diversity-Based Centroid Mechanism," in Information Technology - New Generations (ITNG), 2015 12th International Conference on , vol., no., pp.224-228, 13-15 April 2015

[15] Liyuan Xiao; Yetian Chen; Chang, C.K., "Bayesian Model Averaging of Bayesian Network Classifiers for Intrusion Detection," in Computer Software and Applications Conference Workshops (COMPSACW), 2014 IEEE 38th International , vol., no., pp.128-133, 21-25 July 2014

[16] Jeong, Dong Hyun, Bong-Keun Jeong, and Soo-Yeon Ji. "Designing a hybrid approach with computational analysis and visual analytics to detect network intrusions." Computing and Communication Workshop and Conference (CCWC), 2017 IEEE 7th Annual. IEEE, 2017.

[17] T. T. H. Le, J. Kim and H. Kim, "An Effective Intrusion Detection Classifier Using Long Short-Term Memory with Gradient Descent Optimization," 2017 International Conference on Platform Technology and Service (PlatCon), Busan, South Korea, 2017, pp. 1-6.

[18] The UCI KDD Archive University of California, Irvine, 1999: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

[19] Cortes, C., Vapnik, V.: Support vector networks. Mach. Learn. 20(3), 273–297 (1995)

[20] Devroye, Luc, László Györfi, and Gábor Lugosi. A probabilistic theory of pattern recognition. Vol. 31. Springer Science & Business Media, 2013.

[21] J. Baek, J. Kim and E. Kim, "Fast and Efficient Pedestrian Detection via the Cascade Implementation of an Additive Kernel Support Vector Machine," in *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 4, pp. 902-916, April 2017.

[22] H. Ma, T. Tan, H. Zhou and T. Gao, "Support Vector Machine-recursive feature elimination for the diagnosis of Parkinson disease based on speech analysis," *2016 Seventh International Conference on Intelligent Control and Information Processing (ICICIP)*, Siem Reap, Cambodia, 2016, pp. 34-40.

[23] Ye Fei, "Simultaneous Support Vector selection and parameter optimization using Support Vector Machines for sentiment classification," *2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, Beijing, China, 2016, pp. 59-62.

[24] Kreßel, U.: Pairwise classification and support vector machines. In: Schölkopf, B., Burges, C., Smola, A. (eds.) Advances in Kernel Methods: Support Vector Learning, pp. 255–268. MIT Press, Cambridge (1999)

[25] Vapnik, V.: Statistical Learning Theory. Wiley, New York (1998)

[26] Chang, Chih-Chung, and Chih-Jen Lin. "LIBSVM: a library for support vector machines." ACM Transactions on Intelligent Systems and Technology (TIST) 2.3 (2011): 27.

[27] Wang, Shuo, and Xin Yao. "Multiclass imbalance problems: Analysis and potential solutions." *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 42.4 (2012): 1119-1130.

[28] Yen, Show-Jane, and Yue-Shi Lee. "Cluster-based under-sampling approaches for imbalanced data distributions." *Expert Systems with Applications* 36.3 (2009): 5718-5727.

[29] Yu, Ting, et al. "A hierarchical VQSVM for imbalanced data sets." *Neural Networks, 2007. IJCNN 2007. International Joint Conference on*. IEEE, 2007.

[30] Chawla, Nitesh V., et al. "SMOTE: synthetic minority over-sampling technique." *Journal of artificial intelligence research* 16 (2002): 321-357.

[31] Guo, Hongyu, and Herna L. Viktor. "Learning from imbalanced data sets with boosting and data generation: the databoost-im approach." *ACM Sigkdd Explorations Newsletter* 6.1 (2004): 30-39.

[32] Imam, Tasadduq, Kai Ting, and Joarder Kamruzzaman. "z-SVM: an SVM for improved classification of imbalanced data." *AI 2006: Advances in Artificial Intelligence* (2006): 264-273.

[33] Tang, Yuchun, et al. "SVMs modeling for highly imbalanced classification." *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 39.1 (2009): 281-288.

[34] Cao, Peng, Dazhe Zhao, and Osmar R. Zaïane. "A PSO-based cost-sensitive neural network for imbalanced data classification." *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer Berlin Heidelberg, 2013.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 819-838 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Towards Optimality in Online Learning – The OLeCenT Approach

Carl Beckford*, Ezra K. Mugisa

Department of Computing, University of the West Indies, Jamaica

ARTICLE INFO

Article history:
Received: 06 April, 2017
Accepted: 19 May, 2017
Online: 19 June, 2017

Keywords:
Learning Management Systems
Course Delivery
Learner-centric
Online Learning

ABSTRACT

Higher Education Institutions (HEIs) employ Learning Management Systems (LMSs) primarily for greater efficiency, profitability, technological advancement or survival. The predominantly used LMSs, Moodle and Blackboard account for in excess of 60% usage by the top HEIs. However, the individual international regions do not necessarily bear the percentages of the overall total. Gaps are identified in optimality in course delivery within online learning when one studies LMSs and their functionalities. Advanced Distributed Learning (ADL) Initiative which was established to standardize and modernize training and education management and delivery, developed and recommended usage of Sharable Content Object Reference Model (SCORM) 2004 and later versions. SCORM 2004 which provides for flexibility in sequencing and navigation for learner-centric course delivery is not supported in any version of the more prevalently used LMSs. It is believed that most people have a preferred way in processing information. We propose codifying one or more Learning Style Instruments (LSIs), diagnosing the preferred teaching approach(es) and dominant/existing learning styles within a batch of learners, then providing course delivery as a best-fit per learner. As a proof of concept, OLeCenT allows the input of one or more course learning paths with real-time learning and automatic reconfiguration of the course path where a new trend or pattern is identified. OLeCenT identified disparity in teaching-learning and provided a mechanism towards improving online learner-centric course delivery. OLeCenT also identified comparative levels of similarities among learners and instructors even where they are deemed to be of different teaching-learning styles/mechanisms.

1. Introduction

For the formative years, the primary mode of knowledge transfer was through the face-to-face mode of delivery. Eventually knowledge was captured in writing which provided an alternative to the usual face-to-face delivery. As technology developed, new ways were found to transfer knowledge. This paper is an extension of work originally presented in Future Technologies Conference 2016 which describes a tool to enhance learning [1]. The development of technology also made available the combination of different forms of learning [2-3]. Amidst these, the need for the initial and most established mode of delivery still has its significance [4-6].

Whether through face-to-face or alternative means of course delivery, differences between the learners' approach to learning or styles of learning and the teacher's teaching mechanism or approaches within a batch of learners and teachers may lead to various rates of learning and a non-optimal level of learning in the overall batch of learners [7-9]. Identifying where there is a teaching-learning style mismatch and solving same can help achieve a resultant increase in the level of learning, and should free a teacher to more ably deal with special cases such as learners with special needs or develop varied course material and course delivery mechanisms to assist additional learners.

We examine teaching-learning in Higher Education Institutions, discuss theories of teaching and learning, and analyze the usage of learning mechanisms or styles in Learning Management Systems. These are used to propose OLECENT, an approach in view of learner-centric course delivery. We describe

*Carl Beckford, Department of Computing, University of the West Indies, Mona, Kingston 7, Jamaica, 876-702-4455, carl.beckford@uwimona.edu.jm

the features, results and benefits of the derived artifact, OLeCenT, an e-Learning Tool towards optimising the teacher-learner interaction [10-11].

2. Teaching-Learning in Higher Education Institutions

2.1. Teaching-Learning using Learning Management Systems

Learning Management Systems (LMSs) are used worldwide by Higher Education Institutions to assist with efficiency in course delivery or management [12-13]. There are different types of LMSs, which all work slightly differently but ultimately perform the same or similar functions and can deliver the same learning materials with different levels of efficiency [14]. Higher Education institutions tend to have one of the following three categories of LMSs in operation: off-the-shelf or proprietary, open source, or bespoke [15].

In an analysis of top universities in each international region, home-grown self-used systems accounted for about 17%, Sakai accounted for about 10% and Canvas accounted for only 5%. Usage of 40% and 21% by Moodle and Blackboard respectively confirm the dominant players in the LMS market. However, the individual regions do not necessarily bear the percentages of the overall total; for example, top universities of Latin American reflect a 100% usage of Moodle [16].

A detailed review of the four named systems was done based on features related to direct or indirect relevance to the level of learning, the finite number of human teachers, and resources or system design which may affect positively or negatively the teaching-learning process. In view of [17-30], Table 1 reflects the comparison of LMS features that are important in mitigating the problem of course delivery being static or erroneously less learner-centric.

Noteworthy are the features of Learning Management Systems that are not fully offered by any of the predominantly used learning environments. These are:

1. Provide for assessment of learning and teaching styles
2. Offer course delivery path/workflow to each learner to match learner's dominant learning style
3. Provide options for learner to receive course delivery in his non-dominant or other learning styles
4. Allow input and usage of text unique to courses
5. Allow for integration of activity simulation where practical knowledge is to be transferred
6. Allow instructor to modify course path according to students' expectations and/or indications from course beginning to end, in terms of transfer of knowledge and necessities for emphasis for repeats
7. Provide for assessment which equates to or exceeds the traditional mode but matches the style of learning
8. Provide for learner-centric scheduling of synchronous activities where such include online activities

Other LMSs or tools had some features but were significantly lacking in others. E.g. Intelligent Tutoring Systems (ITS) represents some of the knowledge and reasoning of good one-to-one human tutors, as it attempts to simulate a human tutor's behaviour and guidance [31].

Table 1: Comparison of LMS Features that are important in mitigating the Problem of Course Delivery being Static or becoming less Learner-Centric

Feature	Learning Management Systems			
	Black board	***Canvas***	***Moodle***	***Sakai***
Improvement for lack in the traditional with use of tools in the online environment				
Provide diagnostic assessment of learning/teaching styles				
Restrict and/or provide sections of course path based on pre-entered conditions or complex access criteria	√		√	
Modify course delivery path/workflow to match dominant learning style in batch of learners	√	√	√	√
Offer course delivery path/workflow to each learner to match learner's dominant learning mechanism / style				
Provide options for learner to receive course delivery in his non-dominant or other learning styles				
Automatically determine loss of learning of learners prior to assessment(s)	√	√		
Alert instructor and/or learner of loss of learning	√	√	√	√
Allow instructor to instantly assess student comprehension	√	√	√	√
Provide instructor with a view of the student's perspective	√		√	
Allow students to understand their own progress towards mastering key outcomes		√		
Equitable benefits to all types of courses and maintain/improve on tools in the traditional				
Allow input of all types of course material	√	√	√	√
Allow input and usage of text unique to courses	●		●	
Provide for use of external tool seamlessly within the learning environment	√	√	●	●
Allow for integration of activity simulation where practical knowledge is to be transferred	●	●	●	●
Align content items to objectives and goals	√			√
Allow students to add personal course artefacts that improved their learning	√	√		
Provide for Collaboration	√	√	√	√
Allow for student feedback outside of those pre-determined by instructor	●	√	√	●
Allow instructor to modify course path according to students' expectations and/or indications	●			
Provide for assessment which equates to or exceeds the traditional mode but matches the style of learning				
Integration of administrative information or functionalities that may aid in learning				
Allow input of teaching-learning schedule of course	√	√	√	√
Use teaching-learning schedule for notifications	√	√		
Provide for learner-centric scheduling of synchronous activities where such include online activities				●

Feature	Learning Management Systems			
	Black board	***Canvas***	***Moodle***	***Sakai***
Mitigate falsehood especially in assessment	√			
Allow HEI to offer course to learners worldwide through MOOC		√		
Permit customization so that HEIs can add according to holistic or long-term learner profile			√	√

Key: √ - feature is fully offered
● - feature is partially offered
blank - feature is not offered

2.2. Limitations of Learning Management Systems

There are documented common complaints about Learning Management Systems. Whereas all are genuine complaints, the only complaint that is identified to be of significant interest for this research is that LMSs do not fit in the existing administration workflows [32]. A key reason to implement a LMS is to support efficient training administration. So, if this isn't working for you, then it is a real problem [32]. We concur with [32] that when selecting a LMS, the HEI should be sure to map the existing administration and end user workflows, and use these maps to communicate your LMS implementation requirements. However, we purport that the LMS implementation requirements should be guided by best practices as well as best-case possibilities available in view of existing skills, resources, and achievements.

Blackboard's market strategy over the years (acquisition), Moodle's significant growth, and the arrival of Canvas (the first significant new product in this space in at least 8 years) tell interesting tales [33]. The various Learning Management Systems have delved into course management and delivery, and some have aimed for improved learning. Not many have designed or implemented a system to manage learning in view of the learner's behavior or learning style. The emergence of MOOC platforms, which for the high-profile MOOC initiatives such as Coursera and edX are homegrown products may be the ideal platform to have self-paced managed learning where learning is tracked and attempts made during the learning process to enhance the level of learning.

It is possible that Learning Management Systems (LMS) are not equivalently suited for all types of courses [34]. Where a course's primary aim is to transfer knowledge with respect to a practical achievement of a task, as opposed to a description or narrative view of an issue. This view could be used to support the view that LMSs may be more advantageous to certain types of courses. Among respondents who were engaged in a programming course delivered with use of a LMS, the communication and discussion aspects of the course were rated among the highest and guidance to solving problems among the lowest [34]. Although LMSs are deemed beneficial, they fall short in some aspects specific to computer science education, in particular programming. A Learning Management System lends itself to web-based learning environments by enabling the management of courses, delivery of content and lectures, assessment and tracking of student engagement with the course. The implementation of a worldwide networking course by Cisco Networking Academy which seeks to transfer the knowledge about a practical task should provide some encouragement for what is possible for other course content such as computer programming.

We are in the midst of learner-centric shifts in education, and the drivers for these changes are social and economic, not just technological [35]. A distinction is made between competency based learning (CBL) which focuses on learning outcomes rather than competency based education (CBE) which is based on time. The emphasis on the knowledge transfer to the learner being the guide for success is an essential aspect of old, existing or new Learning Management Systems.

A Delta Initiative analysis points out that most of the core Learning Management Systems were developed between 1997 and 2004, and there really has not been many new pedagogical capabilities added since that time. Instead, innovation has been driven by a horizontal expansion of business areas, e.g. interoperability with peripheral systems for portals, portfolios, ERP integration, and the ERP systems themselves [33]. As interoperability is achieved, improvements in pedagogical capabilities are achieved. The advancement of technology and the use of computing resources could assist in this way.

In referring to examinations or other forms of assessments, students have a strong tendency not to learn for life, but "to the test" [36]. A mismatch or misalignment [37] in form or content between examination tasks and learning objectives will lead to students engaging in the wrong learning activities and acquiring the wrong competences [36]. Multiple-choice tests are perhaps the most infamous example of such mismatch and the associated detrimental effects, decried for fostering surface learning and student disengagement [38]. Such a mismatch is likely to alienate students and erode their learning motivation [36]. The possibility of mismatch or misalignment suggests the need for LMSs that ensure that learning motivation is not eroded, course delivery is competence-oriented [39], as well as a LMS that mitigates the mismatch between examination tasks or other course delivery activities and learning objectives.

There are three functional areas the LMS is designed to support: Course Administration, Communication and Collaboration, and Custom Educational Content. One of the most important things an academic technologist can deliver to faculty is to show them how the LMS can be used for more than just a platform for course administration but to teach in new ways [33]. This implies that (1) there may be ways provided by the LMS to teach or improve teaching exists but are not being employed, (2) there may be new ways of teaching that may be inspired and developed with use of a LMS that have not been explored, (3) there may be new ways of teaching that are inspired by external factors to the LMS, maybe through the face-to-face environment, that may be developed as new features to the LMS.

Amidst the many Learning Management Systems and even the smaller number of those that are used predominantly by HEIs, Columbia University has used aspects of at least six of these, namely Sakai, Moodle, Angel, Canvas, eCollege and Coursera [33]. The change from one LMS to another suggests a feature, functionality, improvement or enhancement in at least one aspect of course administration, communication and collaboration, and providing educational content, which was in one system above another. Therefore as technology develops and the area of e-learning or online learning expands, newer functionalities and enhancements may become possible.

3. Consideration of Learning Styles in Learning Management Systems

Education has always had new ideas about learning and teaching. Educators are asked to use new curricula, new teaching strategies, and new assessments. They are directed to prepare students for the new state standardized test or to document and assess students' work through portfolios and performance assessments [40]. While classroom teaching and management strategies are well documented, the online learning environment presents different challenges and benefits [41]. Teaching in an online environment requires a special set of teaching skills since many of the strategies and tactics associated with best teaching practices are somewhat constrained by the primarily text-based environment [41]. Implicitly, teaching could be taking place in the online environment yet learning is not taking place or certainly not at the level at the perception of the teacher or instructor.

3.1. Theories of Teaching and Learning

Many theories of teaching and learning have been purported and/or established. Some of these are Cognitive Load Theory of Multimedia Learning, Maslow's Hierarchy of Needs, Cognitivism, Experiential Learning (Kolb), Constructivism, ADDIE Model, ARCS Model of Motivational Design, and Multiple Intelligences Theory [42].

Cognitive Load Theory of Multimedia Learning discusses that there are three types of cognitive load: extraneous cognitive load, intrinsic cognitive load, and germane cognitive load. The intrinsic cognitive load embodies the idea that all instruction has an inherent difficulty associated with it. This inherent difficulty may not be altered by an instructor. However many schemas may be broken into individual "subschemas" and taught in isolation, to be later brought back together and described as a combined whole [43, 44]. It is important that systems of learning provide for segmentation or combining of learning units to facilitate the intrinsic cognitive load. Extraneous cognitive load, is under the control of instructional designers. Where possible, the ideal LMS provide for the setup and automation of learner paths. Extraneous cognitive load and intrinsic cognitive load are not ideal; they result from inappropriate instructional designs and complexity of information. Germane cognitive load is coined as "effective' cognitive load, caused by successful schema construction. Each of the cognitive loads are additive, and instructional design's goal should be to reduce extraneous cognitive load to free up working memory [43]. The ideal LMS should model aspects of the human cognitive architecture, and provide for ease of use by course designers in their endeavor to apply sound instructional design principles.

Abraham Maslow's theory, Hierarchy of Needs is a motivational theory in psychology that argues that while people aim to meet basic needs, they seek to meet successively higher needs in the form of a hierarchy. Maslow presented the idea that human actions are directed toward goal attainment. Any given behavior could satisfy several functions at the same time; for instance, going to a pub could satisfy one's needs for self-esteem and for social interaction [45]. Maslow's Hierarchy of Needs has often been represented in a hierarchical pyramid with five levels. The four levels (lower-order needs) are considered physiological needs, while the top level is considered growth needs. The lower level needs need to be satisfied before higher-order needs can influence behavior [45]. Development and implementation of a learning system or tool must consider that the body of learners as well as instructors and teachers would tend to follow Maslow's Hierarchy of Needs.

The theory of Cognitivism says that the learner is viewed as an information processor (like a computer). Cognitivism focuses on the inner mental activities – opening the "black box" of the human mind is valuable and necessary for understanding how people learn [46]. The ideal learning system should provide for some capture of each learner's learning mechanism/style/intelligence. In view of the theory, ensuring that stimuli exist within a Learning Management System will contribute to the system being more learner-centric.

The theory of Constructivism says that learning is an active, constructive process. The learner is an information constructor. People actively construct or create their own subjective representations of objective reality. New information is linked to prior knowledge, thus mental representations are subjective. Knowledge is constructed based on personal experiences and hypotheses of the environment [46]. This theory is used to support the proposal that the system should provide for basic information access needs such as system user-friendliness and ease of use, to allow for ease of transition in a learner using the system to increase the level of learning. The theory says that learners continuously test the hypotheses of the environment through social negotiation. Each person has a different interpretation and construction of knowledge process. The learner is not a blank slate but brings past experiences and cultural factors to a situation [46]. Past experiences and cultural factors is used to support for a diagnostic mechanism that ascertains learner preferences or other factors that may influence learning. Constructivism assumes that all knowledge is constructed from the learner's previous knowledge, regardless of how one is taught [46]. Thus, even listening to a lecture involves active attempts to construct new knowledge.

Constructionism builds up the theory of Constructivism. Constructionism is 'learning by making'. Constructionism shares constructivism's connotation of the learning. It then adds the idea that this happens especially felicitously in a context where the learner is consciously engaged in constructing a public entity [47]. The theory of constructionism proposes that learning may be enhanced through kinesthesia. A Learning Management System should facilitate kinesthetic learning objects.

The theory of Multiple Intelligences identifies seven distinct intelligences. The theory says that we are all able to know the world through language, logical-mathematical analysis, spatial representation, musical thinking, the use of the body to solve problems or to make things, an understanding of other individuals, and an understanding of ourselves. Our educational system is heavily biased toward linguistic modes of instruction and assessment and, to a somewhat lesser degree, toward logical-quantitative modes as well [42]. It is argued that the broad spectrum of students would be better served if disciplines could be presented in a numbers of ways and learning could be assessed through a variety of means. It may seem impossible to teach to all learning styles [42]. However, as we move into using a mix of media or multimedia, it becomes easier. As we understand learning styles, it becomes apparent why multimedia appeals to learners and why a mix of media is more effective [48, 42]. The theory of multiple intelligences satisfies the many types of learning preferences that one person may embody or that a class embodies.

3.2. Teaching and Learning Styles

Learning styles are ways of learning presumed to allow individuals to learn best [49,50]. It is believed that most people have a preferred way in processing information [51]. There are a number of instruments developed for determining learning styles. Some of these are Witkin's Field-Dependence/Field-Independence, Kolb's Experiential Learning Model (ELM) and Learning Style Inventory (LSI), Hunt et al.'s Conceptual Level Model, Dunn et al.'s Learning Styles Inventory (LSI), Fleming et al.'s VARK Model, C.I.T.E Learning Styles Instrument, and Gregorc Learning Style Model [52]. Learning styles has a bad press. It seems that they are lauded and then attacked on an almost cynical basis [53]. This is probably because it is very difficult to measure learning (in part because it is difficult to define learning in useful ways), especially if one wants to know when learning happens or to what it can be ascribed. The critics of learning styles say things like 'Knowing ones learning style does not improve learning.' However knowing one's learning style can be beneficial if learners take the next step, and consider how and when they learn as part of a reflective meta cognitive process, with action to follow [53].

Predominantly, it was stated that there were three learning styles, namely auditory, tactile/kinesthetic and visual [54,55]. Currently it is more accepted that there are more than three learning styles [48,31]. There are discussions whether intelligences are styles or whether they should be considered as learning styles [56]. Gardner says that intelligences are not the same as learning styles. Intelligences are mental computers of varying strength, while styles are the ways in which individuals putatively approach diverse tasks [42]. The prevalently used instruments were chosen for the research, with consideration given to the level of their accessibility. Consequently, the learning styles models discussion includes Fleming et al.'s VARK Model, Kolb's Experiential Learning Model (ELM) and Learning Style Inventory (LSI). C.I.T.E Learning Styles Instrument was incorporated primarily due to its accessibility.

3.2.1 VARK Model of Learning Style

The VARK Model evolved from the VAK Model. The Visual, Auditory and Kinesthetic (VAK) learning style model is a common and widely-used model of learning style. According to this model, most people possess a dominant or preferred learning style; however some people have a mixed and evenly balanced blend of the three styles of visual learners, auditory learners and kinesthetic learners. In summary visual learners tend to learn through seeing and think in pictures, auditory learners tend to learn through listening and have highly developed auditory skills, and kinesthetic learners tend to learn through moving, doing and touching and expressing themselves through movement [38]. With use of an appropriate instrument, the ideal Learning Management System should provide for ascertaining the difference in styles of the learners; this should aid in the system's provision of a learner-centric course delivery.

VARK is an acronym for Visual, Aural, Read/Write and Kinesthetic. The VAK inventories had been around for years until expanded by the inventor of the model. The inclusion was a second 'visual' modality for Read/Write learners. From observation some students have a distinct preference for the written word whilst others preferred symbolic information as in maps, diagrams, and charts [53]. These two preferences were not always found in the same person. He also said that there is more acceptance of this distinction today than in the 1980s [53]. Fleming's VARK Model is suggesting that the visual and auditory/aural aspects of the previous VAK model did not capture ideally the read/write practices or preferences that were exemplified by some learners. The Learning Management System that is best suited for the learner should be able to distinguish a learner who is visual, aural or other or combinations of these styles.

3.2.2 Kolb's Learning Style Model

The Kolb's Experiential Learning Model and Learning Style Inventory (LSI) has been called the Kolb Learning Style Model, a 9-fold classification of learning styles with combinations of feeling, watching, thinking and doing/acting scored as a 12-item self-report questionnaire [51]. Kolb's learning theory sets out four distinct learning styles (or preferences), which are based on a four-stage learning cycle. The model offers both a way to understand individual people's different learning styles, and an explanation of a cycle of experiential learning that applies to us all. Kolb says that ideally (and by inference not always) this process represents a learning cycle or spiral where the learner 'touches all the bases', i.e., a cycle of experiencing, reflecting, thinking, and acting. Immediate or concrete experiences lead to observations and reflections. These reflections are then assimilated (absorbed and translated) into abstract concepts with implications for action, which the person can actively test and experiment with, which in turn enable the creation of new experiences [51]. The current Kolb Learning Style Model reflects an enhancement to a former version of the model [57]. As knowledge concerning learning styles increases and learning style theory, mechanism and instruments improve, there should be flexibility in a Learning Management System to facilitate ease of adaptability to the changes.

Kolb explains that different people naturally prefer a certain single different learning style. Various factors influence a person's preferred style, notably in his experiential learning theory model (ELT). The discussion on experiential learning theory is not included as it should not affect the approach for Learning Management Systems and the use of this approach [51]. The determination of a learning style is done by the following of a protocol or rules that govern the assessment process. The ideal learning environment that provides mechanism for assessing learners should provide for the seamless creation and modification of rules per learning style and per learning style instrument that may vary with an institution's selection of learning style approach. Knowing a person's learning style enables learning to be orientated according to the preferred method. Everyone responds to and needs the stimulus of all types of learning styles to one extent or another - it's a matter of using emphasis that fits best with the given situation and a person's learning style preferences [51, 58].

Whereas there has been support for the Kolb Model, there has been some criticism. The concept of learning styles has been criticized by many and experts suggest that there is little evidence to support the existence of learning styles at all. One large scale study looked at more than 70 different learning style theories and

concluded that each lacked enough valid research to support its claims [59]. In a 2008 article, educator Mark K. Smith argued that Kolb's model is supported only by weak empirical evidence and that the learning process is actually far more complex than the theory suggests. He also noted that the theory fails to fully acknowledge how different experiences and cultures may impact the learning process [59]. Additionally, [60] says that despite the various refinements of Kolb's inventory, the instrument still appears to possess several weaknesses which limit its use, including low reliability, questionable validity, and low predictive powers. The instrument presupposes that individuals can only possess one learning style. Their study says that although the four modes suggested by Kolb may be valid, little evidence is provided to confirm Kolb's two dimensions. They suggest three unidimensional factors, where the resulting learning style, will be based on one's relative position on each of the three dimensions [60].

Amidst the criticisms, [60] states that in 2005 amidst a list of 71 learning style models, Kayes chose Kolb's model as the most influential learning style model. Kolb Model has still been in usage across the world [61]. Kolb's Experiential Learning Model (ELM) and Learning Style Inventory (LSI) is the preferred model for inclusion in a Learning Management System. Notwithstanding, the ideal learner-centric Learning Management System should provide for setup of multiple learning style instruments and capability to select or activate the one(s) to be used per institution, per course or per period.

3.2.3 Teaching Style

By experimenting with innovative strategies intended to help students learn who formerly had not been successful, [62] found out that most methods and/or resources appeared to work with some students but not others. It was said further that in many cases, those who were successful with a particular method or set of materials had essentially similar characteristics. Repeatedly, data revealed that when taught through methods that complemented their learning characteristics, students at all levels became increasingly motivated and achieved better academically. If it is true that 90 percent of all instruction is conducted through either lecture or lecture discussion, it is no wonder that so few students achieve as well as we believe they should [62]. The research by Dunn & Dunn suggests that similar to how learners may favour a particular learning style or mechanism, a teacher or instructor may favour a particular style, mechanism or approach in providing knowledge or information to the learner.

Dunn & Dunn says that for decades, supervisors have been evaluating faculty in an effort to isolate those characteristics that produce effective instruction. Their efforts have been misdirected by weaknesses both in their assumption and their basic designs [62]. The attitude teachers hold towards various instructional programs, methods, and resources as well as kinds of youngsters they prefer working with constitute part of their "teaching style". Teachers can assess themselves with an instrument that simultaneously identifies their teaching style and reveals the areas that need to be expanded to respond to additional characteristics [62].

In summarizing the paper, Learning Styles/Teaching Styles: Should They… Can They Be Matched? [62] answers the question "Why not just match teachers with students?". The paper purports that the fallacy is this (a) learning style and teaching style characteristics do not always cluster into such neat packages; (b) students are not consistently one way or the other – nor are teachers; (c) neither traditional nor informal teachers are necessarily excellent, and it is possible to match a student's learning style and a teacher's teaching style and still not provide that youngster with an effective teacher; and (d) given the practical, "how-to" skills for teaching students through their individual learning styles, most teachers can become effective with most students and, simultaneously, provide a humanistic care, nurturing atmosphere [62].

We have concurrence with the views of [62]. However, within a Higher Education Institution, where the human teacher may be the primary source of course delivery, he/she is not necessarily provided with the resources of time to ensure that all avid learners being taught at a particular time are at their highest level of learning.

The analysis was done on a BTEC Applied Science course to determine whether student-centred learning activities improve learning outcomes [63]. The research states that deeper learning was significantly improved after instructional teaching as opposed to student-centred teaching. In the research, preferences for teaching style were also obtained from the students [63]. This research indicates that teaching style is a contributing factor to the level of learning. Where a Learning Management System is able to provide a teaching style – learning style match, the learning process may be aided. However, within a Higher Education Institution, where the human teacher may be the primary source of course delivery, he/she is not necessarily provided with the resources of compensation or training to ensure that all avid learners being taught at a particular time are at their highest level of learning.

In the discussion of teaching style, one may imagine the teacher or instructor being a computer software program or general form of Computer-Assisted Instruction (CAI) [64]. Of note to this research is the Intelligent Tutoring System (ITS) in Learning. We make note of [65] in their work for argument-making in higher education. The best result of 81.74% in classification correct rate was obtained when all grade classes were used [65]. It was said that while CAI may be somewhat effective in helping learners, they do not provide the same kind of personalised attention that a student would receive from a human tutor. One-to-one tutoring allows learning to be highly individualized and consistently yields better outcomes than other methods of teaching. Unlike previous CAI systems, ITS represents some of the knowledge and reasoning of good one-to-one human tutors, and consequently can coach in a much more detailed way than CAI systems. ITS does not just provide feedback such as correct or wrong answers for the student's mathematics solution but it also provides details of the misconception that the student face. The Student Model assesses student's knowledge and makes a hypothesis on the conceptions and reasoning strategies employed by the student to achieve his/her knowledge state [66]. Findings suggest participants' learning capacity of the introductory statistics could be improved successfully when CAI used as a supplement to regular lecture in teaching introductory statistics course [64].

The mechanism without the rules of Kolb's Model was attainable without a cost. For the C.I.T.E. Learning Styles

Instrument, both the mechanism and detailed rules for implementation were available.

3.3. Current Status of LMSs incorporating Learning Styles and Sequencing

The Advanced Distributed Learning (ADL) Initiative was established in 1997 to standardize and modernize training and education management and delivery. To meet its mission, the ADL Initiative developed the Sharable Content Object Reference Model (SCORM) and the ADL Registry. The release in 2009 is SCORM 2004 4th Edition. ADL supplies resources for SCORM 1.2, SCORM 2004 3rd Edition, and SCORM 2004 4th Edition. ADL recommends use of SCORM 2004 4th Edition [67]. Some consider the Experience API (Tin Can API) released in 2013 with currently in excess of 170 adopters, to be the successor to SCORM [68].

Further, the IMS Global Learning Consortium (IMS GLC) is a global, nonprofit, member organization that strives to enable the growth and impact of learning technology in the education and corporate learning sectors worldwide. Consequently, IMS develops Learning Tools Interoperability (LTI) to allow remote tools and content to be integrated into a Learning Management System (LMS). In addition to LTI, IMS GLC developed another standard, Common Cartridge. IMS Common Cartridge (IMS CC) provides a standard way to represent digital course materials for use in online learning systems so that such content can be developed in one format and used across a wide variety of learning systems, as well as enable new publishing models for online course materials and digital books that are modular, web-distributed, interactive, and customizable.

The three standards described, namely Sharable Content Object Reference Model (SCORM), Learning Tools Interoperability (LTI) and Common Cartridge have been employed in learning management systems in varied levels. The main systems in use which provide such functionality are Blackboard which is proprietary, and Moodle, Canvas, Sakai which are open source.

SCORM 2004 is not supported in any version of Moodle. Parts of the API have been implemented, but sections such as Navigation and Sequencing, which will assist in learner progress and assessment detailed tracking, have not been implemented. Moodle has announced that development on native SCORM 2004 support in the system has stopped. They further suggest that where a fully certified SCORM 2004 Player in Moodle is desired, Rustici Software have a Moodle plugin which connects to their commercial SCORM Cloud service turning Moodle into a fully compliant SCORM 2004 LMS [42].

Where there is a human teacher as the primary source of course delivery, he/she is not necessarily provided with the requisite resources to ensure that all avid learners being taught at a particular time are at their optimal level of learning. From the concept of levels of learning by [10] and the discussion of optimal learning by [11], the optimal level of learning is the highest level of learning achievable in a given time and nature of the uptake function; we consider the nature of the uptake function to include the learner, learning environment and learning style [69]. Learners are expected to comprehend the material within the mode made available by the teacher. All learners in a particular batch tend to be governed by the path of learning carved out by the teacher in his/her expertise and experience. Whereas the teacher has the knowledge to be transferred and is better able to determine the path to be taken based on prior batches of students, the flexibility does not exist at the point of learning for a change of path or delivery mode. There is also often disparity between the mode employed for course delivery and assessment, such as oral versus written or electronic versus hardcopy. Hence, as the number of learning styles may increase as the population within a batch of learners increases, there is an increasing tendency away from the optimal level of learning, amidst the finite or a limited number of course delivery paths and human teachers, and the current design of learning management systems.

4. OleCenT – The Online Learner-Centric Tool

4.1. The OLeCenT Approach to Course Delivery

In the interest of a positive effect on the level of learning, we propose OLeCenT, a tool for learner-centric course delivery in the online environment. OLeCenT may be integrated with a Learning Management System for enhanced course administration. We embrace the integration of learning styles to achieve a maximal matching [70] with the teaching styles. Teaching-learning in higher education institutions is examined with an analysis being done on course delivery in view of learning styles. We suggest how instructional design may be applied amidst a standard and specification for web-based e-learning with emphasis on how learning takes place.

4.1.1 The Underlying Theories of the OLeCenT Approach

We are influenced by [71] that computing impacts the three necessities of teaching-learning, namely basic skills instruction, advanced skills instruction and assessment of student progress, as well as the types of learning. We discuss the learner-centric tool by examining teaching and learning styles in a graph theory context.

4.1.1.1 OLeCenT and Graph Theory

By examining teaching and learning styles in a graph theory context, we propose an approach for course delivery, and describing elements and standards of the learner-centric model. The learner-centric tool may be designed for any of the sets of learning styles. We propose [51] Learning Styles Inventory 9-fold classification of concrete experience (feeling), reflective observation (watching), abstract conceptualization (thinking), active experimentation (doing), diverging (feel and watch), assimilating (think and watch), converging (think and do), accommodating (feel and do) and perception continuum (feel, watch, think, do). This model provides flexibility in terms of basic learning styles and diverse combinations of these styles. It is believed that this 9-fold classification engulfs others such as the seven stated in the classification by [48].

We propose that in the context of all teaching and learning styles, the relation for teaching-learning styles is a bijective function [70]. A teaching-learning style is a teaching style that provides greatest learning for (matches) a particular learning style, and vice versa. Considering a teaching-learning unit such as a course or a session of learning where knowledge transfer is the aim, there is a set of teachers and a set of learners. Correspondingly, there are sets of teaching styles, learning styles, and teaching-learning styles (the matching between a teaching style and a learning style).

Let G be a directed, bipartite graph with disjoint vertices T (the set of teaching styles, t_1, t_2, ...) and L (the set of learning styles, l_1, l_2, ...) in which the edges are directed from T to L (Figure 1). We propose that in any teaching- learning unit, optimal learning occurs where there is a maximal matching [70] for G with all elements of L having a pre-mapping (maximal matching with surjectivity) (Figure 1a, Figure 1b.) even where a complete matching does not exist (Figure 1b). Optimal learning may exist even where there is no surjective (onto) mapping as the same may exist where a teaching style exists in a teaching learning unit which has no mapping to a learning style in that unit (Figure 1b). Where there is a complete matching, that is with all teaching styles having a matching (Figure 1a, Figure 1c), non-optimal learning may still occur (Figure 1c) as all learning styles may not be catered to. For large learning groups with the number of learning styles tending towards maximum, invariably there are learning styles not mapped and far from optimal learning (Figure 1d). For learner-centric course delivery in e-learning, we propose that maximal matching with surjectivity for the teaching-learning styles is attainable with use of the learner-centric tool and how the LMS is implemented.

4.1.1.2 OLeCenT and the Push-Based Model

The push-based (build-to-stock) model of supply chains is built on the basis that the supplier or provider of goods or service knows best as to what goods or service are needed [72]. This build-to-stock model schedules production based on best guesses of demand and the availability of raw material in inventory. Considering the teachers' "raw material" or availability of teaching styles, some HEIs operate in a push-based model where the course content is provided in a mode as a best guess of demand. We propose that the pull-based approach (demand-driven) be used in course delivery where the customer orders trigger events in the supply chain [72]. The desire for learning should not be what triggers what is taught (course content) but rather the learning style (for the course content) should trigger events for the material to be provided in a particular style.

The pull-based model is similar to the self-examination mentioned by [71] in the PRECALC project or may be likened to diagnostic and formative assessments with benefits for the current year or summative assessments with benefits for the ensuing year. Research provided for the three initially proposed as the existing learning styles, namely visual, tactile, and kinesthetic [71]. The learner-centric tool also provides for learning styles which are combination of these as well as other forms of learning styles [51]. Further, formative and summative assessments primarily tend to be used to assess for knowledge gaps or knowledge acquisition respectively, rather than for an assessment of teaching style(s) to be employed. Even where these are used for this purpose, it must be noted that benefits can only accrue subsequent to the time of assessment.

4.1.1.3 OLeCenT and Supply-Demand Theory

The learner-centric tool should cause a lessening of the gap between teaching and learning styles within a batch of learners. Some learners should be more comfortable with the course material as it is received in a manner closer to the type of learning to which they are adept. The avid learner is not first required to "translate" course material from another teaching-learning style to ensure his/her optimal learning. Learners within the same batch benefit during the teaching-learning process as opposed to the trend where changes are put in place for subsequent batches

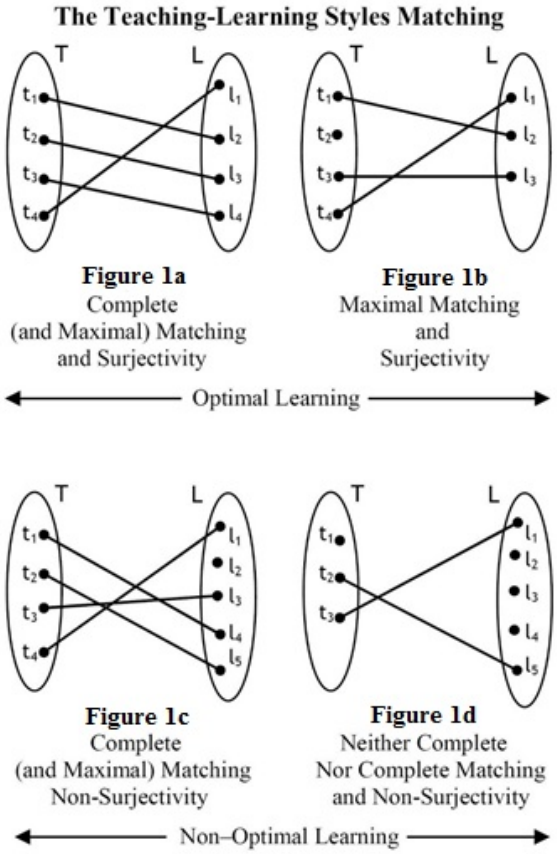


Figure 1: The Teaching Styles (T) and Learning Styles (L) Matching reflecting when optimal learning (7a and 7b) or non-optimal learning (7c and 7d) is achieved. The figure shows various combinations of complete matching, maximal matching and surjectivity

of learners. This is because the tool implements discovered trends or learning patterns of the current learners. The provision of learning objects specific to an individual's learning style should minimize the need to complement or supplement their learning with non-LMS sources. There is therefore a seamless integration required for fluidity of learning and a mitigating of the teachers' concerns for the use of material without approval.

In examining market mechanism, [73] states that the supply curve tells us how much producers are willing to sell at various prices, and the demand curve tells us how much consumers are willing to buy at these prices. The two curves intersect at the equilibrium point [73] where quantity demanded is equal to quantity supplied. In our study the aim is for optimal learning collectively and individually. Applying the supply and demand curve principle within a course and considering how much (teaching styles) teachers are willing, able or allowed to provide, and how much (learning styles) learners are willing to adopt, we propose that there exists a teaching-learning equilibrium (A) where teaching-learning tends to operate. With the learner-centric tool the level and speed of learning should increase and there should be an upward movement of the teaching-learning equilibrium point (A → B) (Figure 2). Figure 2 is constructed based on the following assumptions: (1) as the number of teaching styles increases, ceteris paribus, the level of learning increases because there is a higher likelihood that a learner will find a matching teaching style (T_1), and (2) similarly as the number of learning styles increases,

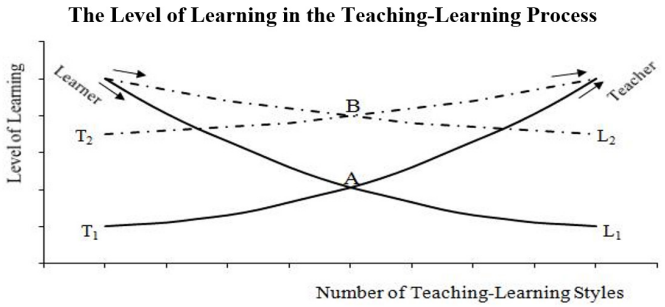


Figure 2: Relation between the Level of Learning and the Number of Teaching-Learning Styles for the human Teacher and the Learner in the Teaching-Learning process

ceteris paribus, the level of learning decreases (L_1). The tool provides a shift in both teaching (T_2) and learning (L_2) curves.

The graph shows expectations before (solid) and after (dashed) the use of a learner-centric course delivery tool. The teaching-learning styles axis represents the teaching styles for the teacher graphs (T_1 and T_2), and the learning styles for the learner graphs (L_1 and L_2).

4.1.2 Standards of the Learner-Centric Tool

One of the essential elements of the learner-centric tool is the emphasis placed on the learners and the determined learning style(s). We propose that in instructional design, learning style should drive course delivery and correspondingly course assessment (Learning Style → Course Delivery → Assessment). We apply Bloom's taxonomy and suggest examples of measurable learner outcomes to assess the learner-centric tool. The Instructional System Development model is adopted with specifics stated for the tool. We then state not only that the learner-centric tool is SCORM compliant but also how the SCORM categories should be used with the learner-centric tool.

4.1.2.1 Educational Objectives

Bloom's Taxonomy of Educational Objectives is widely used for organizing levels of expertise. Reference [10] states the three overlapping domains of educational objectives as cognitive, psychomotor and affective and based on the expected knowledge transfer within a course, the specific domain is identified. We propose the application of Bloom's Taxonomy to determine the learner's expertise in using the learner-centric tool. The learners' values, attitudes or interests are affected by the course delivery technique embracing teaching-learning styles, therefore the domain of affective goals is the most relevant to assess the learners' expertise and the effectiveness of the learner-centric tool. We determine each learner's level of expertise by assessing which measurable learner outcomes have been attained. The levels of expertise are listed in the order of complexity (Table 2). The collective summary of the learners' expertise suggests the effectiveness of the tool.

4.1.2.2 Instructional Design

Instructional Design models are used to define instructional systems. There are more than one hundred instructional design models, but almost all including the Instructional System Development (ISD) Model are based on the generic ADDIE model [74]. The ISD model is the most well-developed [71, 75] and contains similar phases as the ADDIE model, namely Analysis, Design, Development, Implementation and Evaluation. The phases are sequential sets of activities where the output of each serves as input to the subsequent phase. The learner-centric tool is defined below using the Instructional System Development model. The activities in the various phases for the learner-centric tool do not replace the actual activities for the instructional system but rather complement same. Consequently, good instructional design is expected of the system and these are only superseded where there is a conflict with the activities of the learner-centric tool.

Table 2: Measurable Learner Outcomes for Assessing Learners' Levels of Expertise in Using the Learner-Centric Tool, based on Bloom's Taxonomy for Affective Goals [10]

Level of Expertise	Measurable Learner Outcome
Receiving	Given an option, the learner chooses to use the learner-centric tool.
Responding	The learner uses optional features of the learner-centric tool such as some of the formative assessments.
Valuing	The learner seeks out in other learning environments, similar techniques employed by the learner-centric tool.
Organization	The learner finds differences with the method of learning (such as with the learning outcome) within the learner-centric tool, and the normal face-to-face mode of learning or the LMS without the learner-centric tool. The learner attempts to resolve the mental conflict(s).
Characterization by a value or value complex	The learner decides to encourage other learners to make use of the learner-centric tool.

4.1.2.2.1 The Analysis Phase

For the analysis phase the aim is to identify existing deficiencies before the teaching-learning exercise between what is desired and what is existing. The analysis is needed not only for the first delivery of the course content but for each delivery of the same. In addition to the analysis concerning desired outcome and existing knowledge and skills, the tool aids in the discovery of gaps between teaching skills and learning skills. Both teacher and learner are tested to determine teaching-learning skills matching. The human teachers for the course are tested and the teaching styles noted for each teacher as being preferred and/or practiced. Learners are tested with the aim of determining the prevalent learning style and the list of all learning styles in the set of learners. These are grouped into categories of dominant and existing learning styles. The teaching and learning styles resulting from the tests are put in a matrix and used to extract the levels of matching. An analysis document and file are produced with a listing of levels of each teaching-learning style match, levels of each unmatched learning style or combination of learning styles, and unmatched teaching style(s). The results from the analysis may be used by the HEI without the tool but these are generated specifically for and form the basis of the next phase for the design of the tool.

4.1.2.2.2 The Design Phase

In the design phase, the design document is produced to describe the learning objects to be created based on the results from the analysis phase. For each teaching-learning unit (such as material relating to a particular topic of a course), there are specific learning objectives including assessment(s) and corresponding learning objects. For each learning objective, it is determined which teaching-learning style learning objects are to be created to compensate for each teaching-learning style match and unmatched learning style or combination of learning styles. Some learning styles may require more learning objects than others for optimal learning. It is noted which learning objects per learning objective and learning style, already exist in the repository. Priorities are stated for all non-existent learning objects with greater values being given to those for the unmatched learning styles combinations. In LMSs, course designers indicate the workflow of teaching-learning units and the learning objects which relate to the teaching-learning units as well as the direction and timing for display. This exercise continues for the learner-centric tool. Additionally, the designer indicates these workflows per learning style; the actual order of teaching-learning units may vary based on the learning style. The course designer also documents rules as to which learning objects should be omitted based on other options chosen. The rules also indicate the conditions on which to allow the learner-centric tool, based on its learning, to assist in the workflow of the teaching-learning process.

4.1.2.2.3 The Development Phase

The creation of the learning objects is done in the development phase. Within this phase, based on the priorities indicated in the design document, learning objects are created. HEIs may choose not to create learning objects for teaching-learning style matches based on the lack of time before delivery and/or the lack of compensatory resources. However, this decreases the effectiveness of the learner-centric tool. For unmatched learning styles, where the levels indicate that the majority of the learners fall in this grouping, the learning objects are loaded into memory cache. The memory cache is a specialized memory storage used to optimize data transfers resulting in faster and more efficient data retrieval. Based on the workflow, direction and timing of the teaching-learning units and the learner's current place in the garnering of the course material, the loading of objects into cache is done at run-time as the learner accesses the course environment. The loading of objects into cache is important primarily where learning objects are large or may significantly increase retrieval time. The setting up of the rules given in the design document in the learner-centric tool is done at this juncture. The developer also ensures that the tool has the facility for the learner to set the timing of their learning sessions with the possibility for teacher override. The learning objects created are ensured to be SCORM 2004 Compliant.

4.1.2.2.4 The Implementation and Evaluation Phase

As it relates to instructional design, the implementation phase speaks to the delivery of the course material to the learner. All developments are put into operation and are expected to follow whatever rules were setup. Within the evaluation phase, one ensures greater efficiency and effectiveness based on the results garnered from implementation. The usage of the learning objects per learning objective per learning style is compared to the expected levels indicated in the analysis document. Where the learner-centric tool was allowed to assist in the workflow based on its learning, a determination is made whether the tool appears consistent with the learners' desires. This information influences decisions in subsequent analyses. Some of the other control mechanisms are the usage pattern of the objects loaded into memory cache, and the level of reuse of the objects.

4.1.2.3 Standards for Web-based e-Learning

The learner-centric tool is SCORM 2004 4th Ed. Compliant but may be developed with less effectiveness to be compliant with other editions of SCORM 2004. The Sharable Courseware Object Reference Model (SCORM) is a collection of standards and specifications for web-based e-learning [76]. The parts of SCORM 2004 are the Run-Time Environment, the Content Aggregation Model (CAM) and the Sequencing and Navigation (SN). The CAM has three parts: the Content Model, the Metadata and Content Packaging. The Content Model describes the relationships between course modules and provides for Sharable Content Objects (SCOs) and assets; the Metadata is used to describe the content into some common vocabulary. The SN governs how navigation between SCOs is handled by the LMS through the definition of rules known as the Sequencing Definition Model [67].

We state only the aspects for which there is uniqueness for the learner-centric tool. We recommend that the learner-centric course delivery tool uses SCORM 2004 with the following:

1. The teaching-learning style learning objects (the learning objects per learning objective and learning style) are represented as the (SCORM) assets. [Content Aggregation Model – Content Model]
   The SCO is the combination of these assets (learning objects) for a learning objective. A recursive definition would be more appropriate for accomplishing the needs of the learner-centric tool but to achieve reuse, SCORM defines that a SCO is not allowed to reference other SCOs.
2. In a grouping of learning objects per learning objective, one or more of these learning objects may or may not be used in an instance of accomplishing the learning objective, based on the learning style(s) of the learners. [Content Aggregation Model – Content Model]
3. The Meta-metadata category stores the teaching-learning styles relevant to the object. Its usage is therefore mandatory. [Content Aggregation Model – Metadata]
4. The usage of the General, Lifecycle, Relation, and Classification categories are optional but recommended. [Content Aggregation Model – Metadata]

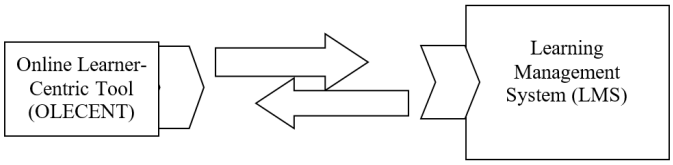


Figure 3: The Online Learner-Centric Tool (OLeCenT) - Learning Management System (LMS) Interface

5. The Sequencing Rules category outlines the workflow between the teaching-learning style learning objects. [Sequencing and Navigation Specification – Sequencing Definition Model]
6. The following categories are also mandatory: Rollup Rules, Rollup Controls, Objectives, Delivery Controls and Navigational Controls. Completion Threshold Controls is mandatory for SCORM 2004 4th Edition only. [Sequencing and Navigation Specification – Sequencing Definition Model]
   The categories of Sequencing Control Modes, Constraint Choice Controls and Limit Conditions are not mandatory but their usage are advantageous based on the implementation and application of the learner-centric tool.

All other aspects of or necessities for SCORM which tend to be common to all SCORM compliant tools are applied: the LMS launches all content in a web browser and is required to implement an API, all of the relevant course data is stored in a XML file, and all content is to be transferred in a folder or a ZIP file called a package interchange file (PIF) [76, 67].

The Online Learner-Centric Tool (OLeCenT) is designed to interface with a Learning Management System (LMS) (Figure 3), thereby course delivery and administration is still managed by the LMS, but enhanced by the learner-centric delivery provided through OLeCenT.

4.2. Components of OLeCenT

The tool has four components, namely Diagnostic Analysis, Repository and Workflow Setup, Learning Administration, and Learner-Centric Assessment and Evaluation (Figure 4). Diagnostic Analysis allows for the setup of one or more Learning Style Inventories as well as the diagnostic assessments of the teacher(s) and learner(s). Repository and Workflow Setup receives and updates learning objects of different types for a single unit of learning as well as setup a designated workflow of how the learning objects are ordered for delivery of the course content per learning style. Learning Administration generates learner course paths and Assessment and Evaluation provides for formative and summative assessments in consideration of each student's teaching-learning style.

4.2.1 Diagnostic Analysis

The Diagnostic Analysis component incorporates two (2) major processes, namely Setup Learning Style Inventory (LSI)

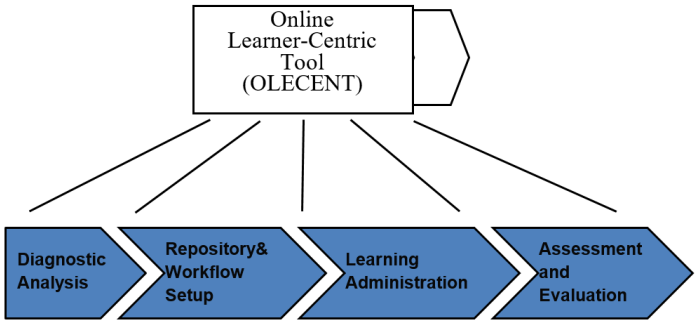


Figure 4: Components of OLECENT

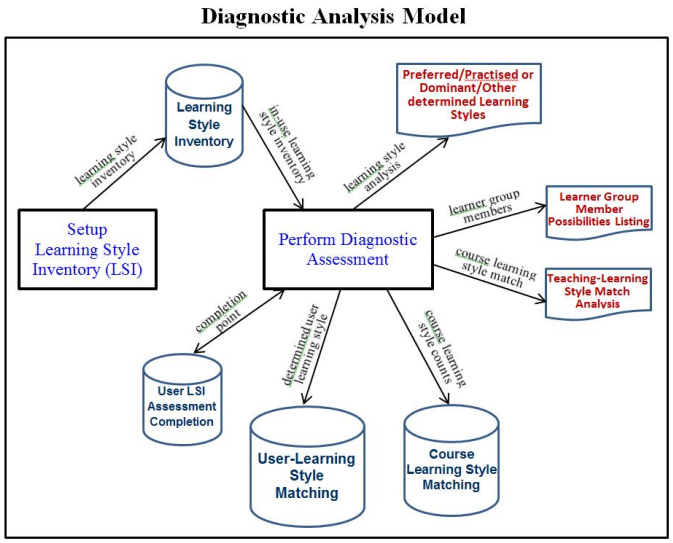


Figure 5: Diagnostic Analysis Model

and Perform Diagnostic Assessment (Figure 5). The tool is designed to allow for any learning style mechanism that has measurable notations. OLeCenT uses learning style indexes throughout its implementation and therefore translates the terms, codes and notations of specific learning style instruments to the OLeCenT learning style indexes.

The tool provides each teacher and learner with a diagnostic assessment to ascertain the teaching-learning style(s) and preferences; it uses this information to inform the human teacher (where existent) of the learning style make-up of the class and teaching-learning style match or gaps as displayed in Table 3.

The learner-centric tool suggests a learning style or combination of learning styles for each student. The learner is allowed to choose another teaching-learning style in which to provide the course content but the system maintains a record of the determined learning style. In Higher Education Institutions (HEIs) there are group assessments, and learning by collaboration is encouraged. By analyzing the collection of student learning styles within the batch of learners, the tool determines for each student or teacher, suggested lists of group members based on the computation of a compatibility factor with a maximum of 100. The suggested lists are based on similar or dissimilar learning style and preferred group make-up. Figure 6 details the 4-part algorithm for determining teacher or learner compatibility factor. The four (4) parts are: Determine Main User Compatibility Factor, Initialize Selection Arrays per Similarity, Determine Other User Compatibility Factors, and Generate List as per Compatibility Ratios.

Table 3: OLeCenT Diagnostic Teaching-Learning Style Analysis

No.	Learning Style	Dominant		Existing		Teaching Match	
Course: COMP1005							
1	Visual Numerical	14	37%	17	22%	Preferred	1
2	Social Individual	8	21%	13	17%	Practiced	1
3	Visual Language	7	18%	4	5%	Preferred	1
4	Auditory Numerical	3	8%	13	17%	Practiced	1
5	Kinesthetic – Tactile	3	8%	6	8%		
6	Oral Expressiveness	2	5%	5	6%	Practiced	2
7	Written Expressiveness	0	0%	6	8%		
8	Social Group	0	0%	6	8%		
9	Auditory Language	0	0%	4	5%		

4.2.1.1 Determine Main User Compatibility Factor

The main user compatibility factor is first determined as a benchmark for assessing the other users. Dominant and existing learning styles, and learning assessment scores for each style is obtained ensuring that the dominant learning style is procured with highest priority (1). Three (MAX_LS_PRI) learning style records are procured. The calculated factor is a summation of a maximum score per learning style with a higher weighting (MAX_LS_PRI/i) being allocated for the dominant learning style. The factor of 3 is used as the other users will be compared based on the 3VAK elements.

4.2.1.2 Initialize Selection Arrays per Similarity

A compatibility factor holder will maintain the compatibility factors of all other users. Based on whether the grouping was desired by Similarities or Dissimilarities, the calculated value would be tending away from zero or not respectively. The initialization values ensure that the calculated user values have higher precedence.

4.2.1.3 Determine Other User Compatibility Factors

The dominant and existing learning styles, and learning assessment scores for each style is obtained for each user, ensuring that the dominant learning style is procured with highest priority (1). Three (MAX_LS_PRI) learning style records are procured. The calculated factor is a summation of a maximum score per learning style with a higher weighting (MAX_LS_PRI/i) being allocated for the dominant learning style. A factor of 1 to 3 is used to multiply the learning style score based on the level of similarities of the 3VAK elements of the main user's learning style and that of the user being determined. As record 3 (MAX_LS_PRI) of each user is ascertained, an insertion sort is applied only maintaining the NOOFDISPLAYNEEDED.

4.2.1.4 Generate List as per Compatibility Ratios

A compatibility factor ratio is determined by considering the variance of the user factor to the main factor, as a ratio of the main factor. The NOOFDISPLAYNEEDED users are shown with compatibility factors in comparison to 100.

Both teacher and learner are tested to determine teaching-learning skills matching. The human teachers for the course are tested and the teaching styles noted for each teacher as being preferred and/or practiced. Learners are tested with the aim of determining the prevalent learning style and the list of all learning styles in the set of learners. These are grouped into categories of dominant and existing learning styles. The teaching and learning styles resulting from the tests are put in a matrix and used to extract the levels of matching.

The learner-centric tool may be designed for any of the sets of learning styles with any Learning Style instrument (LSI). We propose Kolb & Kolb's Learning Styles Inventory 9-fold classification [51] based on its widespread usage, flexibility and a classification that seems to engulf other LSI. The learning style units are classified based on the Visual- Auditory-Kinesthetic (VAK) content thereby providing ease of relation to learning objects. OLeCenT allows rules for learning style determination to be codifed. Learning Style rules are entered as text; the text is parsed on input to determine a formula for the specific learning style calculation. The parsed calculation is stored. An example of the text inputs is (1):

```
[DETERMINE MAIN USER COMPATIBILITY FACTOR]
MainUserArray <- GET Styles and Scores for Main User [M]
MAX_LS_PRI = 3 (or number of learning styles per user stored)
main_compatibility_factor = 0
for i = 1 TO MAX_LS_PRI {
    Get Next Record from MainUserArray;
    Mls_calc_max = Mls_calc_major[i] + (Mls_calc_major[i] –
                    Mls_calc_minor[i])
    main_compatibility_factor += (MAX_LS_PRI/i) Mls_calc_max
                    + 3  Mtlstylescore[i]
}
[INITIALIZE SELECTION ARRAYS PER SIMILARITY]
for i = 1 TO NOOFDISPLAYNEEDED {
    if SIMILAR USERS NEEDED
        Compatibility_Factors[i] = 0;
    elseif DISSIMILAR USERS NEEDED
        Compatibility_Factors[i] = 1000;
}
[DETERMINE OTHER USER COMPATIBILITY FACTORS]
CompUserArray <- GET Styles etc for Others To Be Compared [C]
For ALL in CompUserArray {
    Get Next Record from CompUserArray;
    compatibility_factor = 0;
    Cpri = priority
    Cls_calc_max = Cls_calc_major + (Cls_calc_major –
                    Cls_calc_minor)
    comp_factor[Cpri] = (MAX_LS_PRI / Cpri)
        (teachlearnstyle[Cpri] == Cteachlearnstyle)  Cls_calc_max
    + ((Mls_vak_elem_V[Cpri] == Cls_vak_elem_V) +
            (Mls_vak_elem_A[Cpri] == Cls_vak_elem_A) +
            (Mls_vak_elem_K[Cpri] == Cls_vak_elem_K))  Ctlstylescore
    if (Cpri == MAX_LS_PRI)
    {
        for i = 1 TO MAX_LS_PRI
            compatibility_factor += comp_factor[i]
        for i = 1 TO NOOFDISPLAYNEEDED
        {
            if (((compatibility_factor > Compatibility_Factors[i])
                    and SIMILAR USERS NEEDED) or
                ((compatibility_factor < Compatibility_Factors[i])
                    and DISSIMILAR USERS NEEDED) or
                ((compatibility_factor == Compatibility_Factors[i])
                  and (MainUserTeachStyle[1] == teachlearnstyle[1])
                    and SIMILAR USERS NEEDED) or
                ((compatibility_factor == Compatibility_Factors[i])
                 and (MainUserTeachStyle != teachlearnstyle[1])
                    and DISSIMILAR USERS NEEDED))
            {
                for j = NOOFDISPLAYNEEDED-1 TO i
                {
                    Compatibility_Factors[j+1] =
                                        Compatibility_Factors[j]
                }
                Compatibility_Factors[$i] <- Store
                    compatibility_factor, User ID and other data
            }
        }
    }
}
[GENERATE LIST AS PER COMPATIBILITY RATIOS]
for indx = 1 TO NOOFDISPLAYNEEDED
    if SIMILAR USERS NEEDED
        compatibility_factor = (main_compatibility_factor –
            abs(main_compatibility_factor –
                    Compatibility_Factors[indx])) /
                            main_compatibility_factor  100
    elseif DISSIMILAR USERS NEEDED
                compatibility_factor = (abs(main_compatibility_factor
                        - Compatibility_Factors[$indx])) /
                                main_compatibility_factor  100
```

Figure 6: Algorithm for Determining Compatibility

$$(Q5+Q13+Q21+Q29+Q37)2 \quad (1)$$

4.2.2 Repository and Workflow Setup

The tool is able to (1) receive and update learning objects of different types for a single unit of learning as well as (2) setup a designated workflow of how the learning objects are ordered for delivery of the course content. Instructors are allowed to setup a single course learning path per learning style or for all learning styles. Figure 7 reflects these two processes as Maintain Learning Object Learning Style Index(es) and Setup Learning Style Course Path(s) respectively. Within the latter process, the tool also provides an option for setting learning sessions based on input of date, time or date-time ranges.

In view of Learning Object Repositories which exist [77], the Maintain Learning Object Learning Style Index(es) process is not designed as a priority to create new learning objects. In the final implementation, the Repository and Workflow Setup component may work with any SCORM 2004 4th Edition Learning Object Repository. The component ensures that pre-existing learning objects may be viewed, the learning style indexes may be updated, and the user may be informed of the state of each learning object. The state of the learning object includes its SCORM compliance, its OLeCenT readiness in view of learning styles, and other general learning object parameters such as id, title, and objectives. Learning objects (LOs) are tagged for general VAK features and therefore provide the flexibility of being used with multiple LSI without having these tagged to various instruments. Consequently, an institution is provided with the ease of switching between LSIs without having to again setup learning paths or retagging objects.

A design document is produced from the Maintain Learning Object Learning Style Index(es) process to describe the learning objects to be created based on the results from the analysis phase. For each teaching-learning unit (such as material relating to a particular topic of a course), there are specific learning objectives including assessment(s) and corresponding learning objects. For each learning objective, it is determined which teaching-learning style learning objects are to be created to compensate for each teaching-learning style match and unmatched learning style or combination of learning styles. Some learning styles may require more learning objects than others for optimal learning. It is noted which learning objects per learning objective and learning style, already exist in the repository. Priorities are stated for all non-existent learning objects with greater values being given to those for the unmatched learning styles combinations.

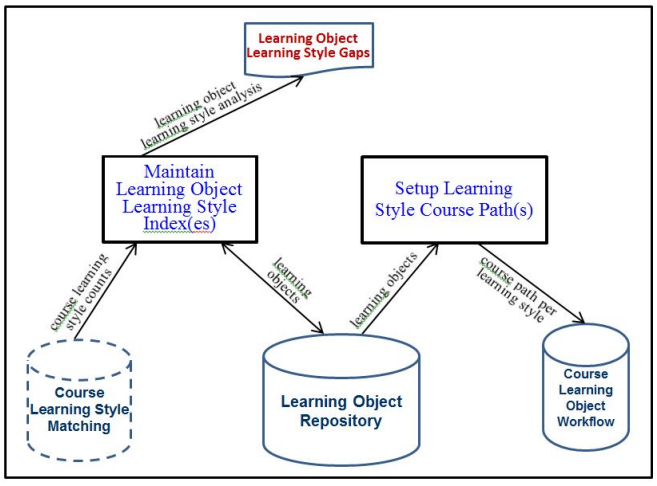


Figure 7: Repository and Workflow Setup Model

Each type of learning object is specific to a basic teaching-learning style. LMSs have the mechanism for teachers to provide learning content in reusable learning objects. The course designer is able to state the specific unit of course content and the related objects specific to the teaching-learning styles. There is a one-to-many relationship between each unit of course content and the teaching-learning style learning object. The teaching-learning style learning object forms part of the learning process and is specifically designed to support one or more learning styles. The objects are placed in a repository for retrieval at course compilation subject primarily to the learning style course path that has been setup.

The Course Learning Object Workflow is initially setup by the course designer or teacher /instructor but may be updated as the system learns the norms of various users per learning style. Consequently, in the Setup Learning Style Course Path(s) process, the user is provided with an option to enter/modify the Course Path Update Factor and Course Path Automatic Update. The Course Path Update Factor is the numeric ratio where during Learning Administration (OLeCenT Component 3) if a different learner workflow for a particular learning style is followed by the learners that many times when compared to the teacher-entered course learning object workflow, the teacher should be notified/alerted concerning updating the Course Learning Object Workflow for the course for the particular learning style. When alerted, the teacher may manually update the workflow. The Course Path Automatic Update may be set during Setup Learning Style Course Path. If the preferred option is to have automatic update, this may be done and an alert/notification sent to the instructor; the instructor is not provided with the option not to receive notifications in the automatic configuration instance.

4.2.3 Learning Administration

For each learner the tool generates a learner course path (the set of teaching-learning style learning objects designated by the workflow for delivery of the course content), learns which other teaching-learning styles are favoured by specific types of learners and uses this acquired knowledge to enhance the formulated learner course paths (Figure 8).

The learner course path is determined from the learning style for the learner determined from Diagnostic Analysis, as well as the course learning object workflow for each learning style setup by the course designer or teacher during Repository and Workflow Setup. Learning administration is barred from access to the student user without a determined learning style. Having determined the learner course path, the tool provides the teaching-learning style learning objects based on the teacher-designed/approved and/or system-determined workflow and where the learner is at in the scope of learning the course content.

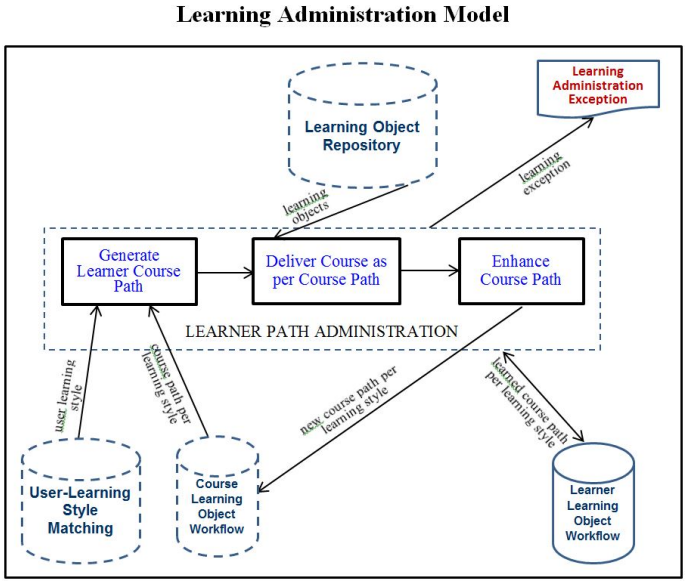


Figure 8: Learning Administration Model

OLeCent attempts to provide a best-fit learning object based on course learning path setup by the instructor and learner's chosen learning style or determined dominant leaning style. The learner may however choose to view or undertake another path or other relevant but not best fit objects on the assigned course learner path. For fluidity of learning, the tool provides the objects in hypermedia with more than one learning object related to the determined or chosen learning style. It is where the learner chooses to undertake other path(s) or other learning objects on the same path, that the tool learns which other learning styles or path are favoured, preferred or taken by specific types of learners in the delivery of certain sections of the course.

For unmatched learning styles, where the levels indicate that the majority of the learners fall in this grouping, the learning objects are loaded into memory cache. The memory cache is used to optimize data transfers resulting in faster and more efficient data retrieval. Based on the workflow, direction and timing of the teaching-learning units and the learner's current place in the garnering of the course material, the loading of objects into cache is done at run-time as the learner accesses the course environment. The loading of objects into cache is important primarily where learning objects are large or may significantly increase retrieval time.

The Learning Administration component maintains the path of learning objects, duration and completion percentage for each learner. This is stored within the Learner Learning Object Workflow and is used as the source of data to determine whether an alert should be generated for the course designer or teacher. The Course Learning Object Workflow setup by the course designer or teacher may be updated as the system learns the norms of various users per learning style. Where the Course Path Update Factor and/or Course Path Automatic Update was entered/modified during Repository and Workflow Setup, and where a new learner path was undertaken by the learner, the Update Factor is used to determine whether the new path is used that many times by learners with similar learning style when compared to the teacher-entered course learning object workflow. If these conditions are met, the teacher is notified/alerted concerning updating the Course Learning Object Workflow for the course for the particular learning style.

```
AlertRatio =
        NoOfConsistentLSVariance / NoOfDiagnosedLSUsers
ActiveLSRatio =
        NoOfConsistentLSVariance / NoOfActiveLSUsers
ChangeEnrolmentRatio =
        NoOfConsistentLSVariance / NoOfEnrolledUsers
if (AlertRatio  100) > CourseUpdateFactor
        If AlertRatio >
                        Average (ActiveLSRatio, ChangeEnrolmentRatio)
                If AutomaticUpdateIsOn
                        AutoConfigureLearningPath
        AlertInstructor
```

Figure 9: Alert Notification and Automatic Configuration Formulae

When alerted, the teacher may manually update the workflow. Where the Course Path Automatic Update is set to the affirmative, the system performs the update automatically. For the update, the teacher-entered course learning object workflow is put in a dormant state, and the new course learning object workflow for the particular learning style is added to Course Learning Object Workflow and made active.

The alerts to enhance the Course Learning Object Workflow, shown in Figure 9 as Learning Administration Exception may be in report form, web page, email or another form of notification. Alert Notifications and Automatic Configuration is determined based on calculated alert, active learning style and change enrolment ratios. The formulae in Figure 9 reflect that the instructor is alerted once the alert percentage exceeds the course update factor. However, when automatic update is on, only where the alert ratio exceeds the average of the active learning style ratio and the change enrolment ratio, is the automatic configuration done. This mitigates the system flopping between decisions where the magnitude of the variance is not significant.

4.2.4 Assessment and Evaluation

The tool will provide for formative and summative assessments in consideration of each student's teaching-learning style. LMSs have the mechanism for ongoing assessments and some allow the teacher to indicate the level of difficulty for each assessment unit. This feature represents only a modification based on the teaching-learning style; the learner undertakes an assessment based on the determined teaching-learning style.

The Assessment and Evaluation Model (Figure 10) diagrammatically shows that similar to Learning Administration, the learner assessment path is determined from the learning style for the learner determined from Diagnostic Analysis, as well as the course learning object workflow for each learning style setup by the course designer or teacher during Repository and Workflow Setup.

Having determined the learner assessment path, the tool provides the teaching-learning style learning objects based on the workflow and where the learner is at in the scope of learning the course content. The path of learning (assessment) objects, duration and completion percentage for each learner is stored. Similar to Learning Administration, this is stored within the Learner Learning Object Workflow and is used as the source of data to determine whether an Assessment Exception (alert) should be generated for the course designer or teacher. The Assessment and Evaluation component uses the Course Path Update Factor to determine whether the assessment grades is that many times by any learning style when compared to the other learning styles.

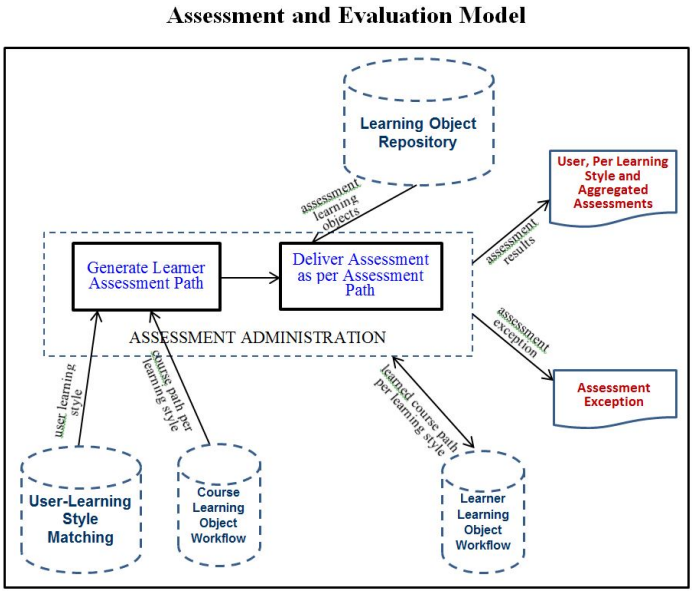


Figure 10: Assessment and Evaluation Model

5. Implementation and Results

5.1. Development of Tool

OLeCenT is developed using HTML 5, XML, CSS and Javascript for the graphical user interface (client) and PHP for all backend (server) transactions. The database employed is MySQL. All components of OLeCenT use fields of SCORM 2004 4th Edition. Additionally, the Diagnostic Analysis component was developed and tested to be fully SCORM 2004 4th edition compliant.

5.2. Implementation Instance

OLeCenT provides for the setup of a Learning Style Inventory (LSI) or Model by learning style indexes. OLeCenT captures information about each learner and teacher and suggests the dominant and other learning/teaching styles. There is a matching which assigns the learners' learning styles and the teachers' teaching styles to a number of the learning styles indexes which matches to the learning styles of the LSI in use. Learning objects are stored in the repository and also given a learning style index or a number of indexes as determined by the course administrator / deliverer / teacher / lecturer. The learners are provided with learning objects in the form of course paths which vary based on the course administrator's projected path, the learners' learning styles, and other assigned or chosen course paths of similar learners.

The current instance of OLeCenT allows a user to be defined as one of five possible user types. The user types are Student, Instructor/Lecturer, Course Designer, Administrator and System Administrator. Each user is assigned optionally to a specific department and/or institution. Where a user of type Student or Instructor is not provided with a department or institution, that user is available to be assigned to any course as in the use of OLeCenT for Massive Open Online Courses (MOOCs).

The first three components of OLeCenT were tested with an instantiation including The C.I.T.E. Learning Styles Inventory (http://www.d11.org/Transition/Assessments/Documents/C.I.T.E.%20Learning%20Styles%20Instrument.pdf) as the Learning Style Instrument (LSI). The C.I.T.E. instrument diagnosed one or more of nine (9) learning styles by use of forty-five questions. For testing purposes only, access to OLeCenT was successfully provided via the Moodle environment, but implementation was not done for full integration which allows for the ease of demographic and course data between the two systems.

Learning Objects were setup as types V - Visual, A - Auditory, K - Kinesthetic, R – Reading/None Specified, and O - Other. Video Files (WMV, MP4, AVI) and Powerpoint Presentation (PPT, PPTX) and Show (PPSX, PPS) without Audio were setup as Visual. Audio Files (MP3,WAV,M4A) and Powerpoint Presentation (PPT, PPTX) and Show (PPSX, PPS) with Audio and Timing were setup as Visual. Articulate Storyline v. 2 Web-enabled Files (HTML) and Powerpoint Presentation (PPT, PPTX) and Show (PPSX, PPS) with Audio and Transition but without Timing were setup as Kinesthetic. PDF, TXT,RTF files were set as Reading. Where learning objects of an auditory type were not readily available, the Windows version of Natural Reader v. 14 was employed for creation of such objects. Natural Reader is a free text to speech software with naturally sounding male and female voices. For the Computing or Math-based courses, Natural Reader was deemed useful primarily for course outlines, rather than actual course content or assessments. Subsequent to the Learning Objects Setup, the Learning path (per learning style or for all learning styles) is also setup by the course designer/instructor.

5.3. Implementation Results

OLeCenT was used in four (4) instances including a focus group of ten (10) students of a private institution administered a MOOC of secondary-level Information Technology, and three computing qualitative and quantitative courses at a tertiary-level institution administered at first year, second year and third year. The first year course COMP1005 Software Engineering Essentials was qualitative and contained over thirty-five (35+) students; the second year course COMP2005 Discrete Mathematics was quantitative and included five (5) students; the third year course INFO3005 Information Systems was qualitative and was administered to 3 students.

5.3.1 Diagnostic – Teaching-Learning Style Match Analysis

The teaching style (TS) - learning style (LS) match may be determined based on the learners' dominant or practiced styles and the instructors' teaching approaches. The teaching-learning style matches determined by OleCenT were:

For the first year qualitative course COMP1005

- 38% match - preferred TS and dominant LS with main instructor (Figure 11)
- 57% match - preferred TS and dominant LS with all instructors (Figure 11)
- 90% match - total of preferred and other TS and group's dominant LS

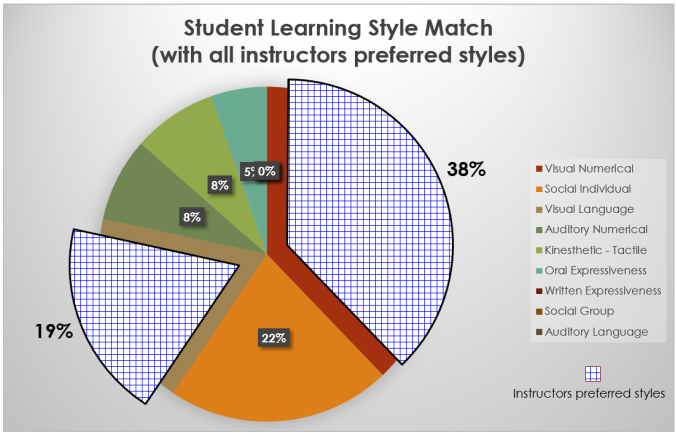


Figure 11: Learning Style Match with all instructors preferred styles

For the small qualitative course INFO3005

- 33% match - preferred teaching style (TS) and group's dominant learning style (LS)
- 66% match - total of preferred and other TS and group's dominant LS

The results also showed in the instance of the quantitative course COMP2005 that the preferred teaching style was matched by only 30% of the group's dominant learning style. In view of the total of preferred teaching style and other teaching styles, there was a 49% match with the group's dominant learning style.

The data for all courses proved useful both to the face-to-face and online instructors who were favoured with the information as to how to possibly better reach the batch of learners in teaching.

5.3.2 *Diagnostic – Teacher and Learner Group Member Possibilities*

Similar and dissimilar learning methods of students may be determined based on a student's dominant or existing styles, and a comparative review with other dominant and existing styles and scores from the learning style instrument. In view of teaching approaches, similar and dissimilar approaches may also be determined; these are based on a teacher's preferred or practiced approaches and other similar approaches and comparison of scores from the learning style instrument.

Figure 12 displays a student with id number 620099999 with his/her top 5 students with similar styles. OLeCent may also display a teacher with his/her top teacher with dissimilar styles. Top teacher dissimilarity possibilities may be used to ensure that there is greater coverage of the learning styles that are reached in a batch of learners. The similarity check is not restricted to learners/teachers of a similar learning-teaching style. Figure 12 highlights that student 620099999 of dominant learning style Social Individual and other styles of Visual Numerical and Auditory Numerical was deemed to have the fourth highest compatibility factor 80 (of 100) with student 415001546 of dominant style Oral Expressiveness and other styles of Visual Numerical and Auditory Numerical.

The similarity and dissimilarity comparisons are based on the relative difference between the main student/teacher being

STUDENT GROUP MEMBERS LISTING, SEMESTER 2, 2016-2017
Members of **SIMILAR** Type

Profile for User ID: 620099999 Student_FN Student_LN
Dominant Learning Style: Social Individual
Other Existing Learning Styles: **Visual Numerical and Auditory Numerical**

User ID: **620099256 (Compatibility Factor: 88 of 100)**
Dominant Learning Style: **Social Individual**
Other Existing Learning Styles: Visual Numerical and Auditory Language

User ID: 413002757 (Compatibility Factor: 84 of 100)
Dominant Learning Style: Social Individual
Other Existing Learning Styles: Visual Numerical and Kinesthetic - Tactile

.
.

User ID: **415001546 (Compatibility Factor: 80 of 100)**
Dominant Learning Style: **Oral Expressiveness**
Other Existing Learning Styles: Visual Numerical and Auditory Numerical

User ID: 416000954 (Compatibility Factor: 76 of 100)
Dominant Learning Style: Social Individual
Other Existing Learning Styles: Written Expressiveness and Auditory Numerical

Figure 12: Listing of Top Students with Similarities as Student 620099999

evaluated and each other student/teacher in the batch of learners/teachers. A compatibility factor for each learner is computed from the relative differences noted among the dominant learning styles, existing styles, and learning styles instrument scores 1, 2 and 3 (Figure 13). For the teacher, compatibility factors are computed from the relative differences noted among the dominant preferred styles/approaches, practiced styles or approaches, and learning styles instrument scores 1, 2 and 3.

The teacher-learner group member possibilities determined by OLeCenT were:

For a student of a particular learning style

- Similar and dissimilar styles
- 70% or greater similarity was found even where the individuals were of different dominant learning styles

For a teacher of a particular teaching approach

- A 39% dissimilarity was identified with other teachers/instructors.

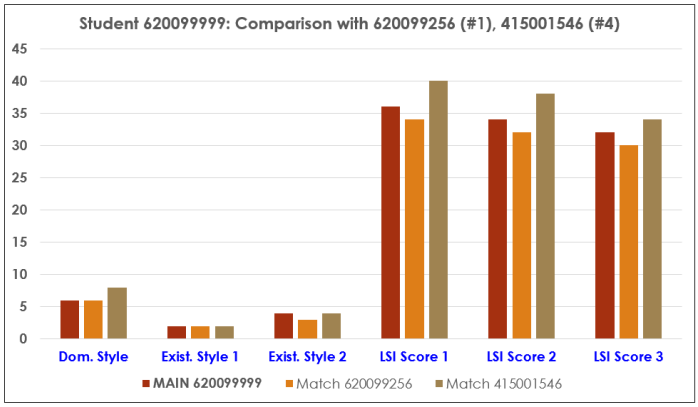


Figure 13: Relative Comparison for Compatibility Factor Determination

5.3.3 Repository – Learning Object Learning Style Gaps

Learning Object Learning Style gaps may be identified per course and per VARK containment. VARK containment describes whether a learning object is deemed to be appealing to the Visual, Auditory, Reading or Kinesthetic learner [53]. The instructor and/or course designer may be immediately notified where learners of a particular style may be required to access the course material in the mode not deemed for his/her highest level of learning.

5.3.4 Learning Path Consistency and Disparity

Chosen learning paths per student per learning style may be identified. The learning path setup by the teacher / course designer may be setup uniquely per leaning style. Subject to the flexibility setup with use of pre-requisites, OLeCenT determines the level of compatibility or disparity between the teacher designed path and the student chosen path. Additionally, the favoured learning VARK objects per learning style are ascertained. For each unit within the learning path, a number of learning objects per visual, auditory, reading and kinesthetic containment may be available. OLeCenT determines the most suitable VARK learning object per unit of the learning path to be displayed dominantly with all other available learning objects for the same unit (of the learning path) subtly displayed.

OLeCenT may generate a learning path consistency and disparity analysis for each course in view of one or more teaching-learning styles, activities (including A-Assessment, B-Lab, F-Field work, G-General, L-Lecture, O-Other, S-Seminar, T-Tutorial), disparity allowance time (length of time the learning object was accessed) for 0 or more minutes, the diagnosed or chosen learning style, and consistency check criterion of learning path or learning object. Figure 14 is the learning path consistency and disparity analysis for COMP1005 for all teaching-learning styles, all activities, where learning objects were accessed for 0 or more minutes, for the diagnosed learning style, and using consistency check criterion being the learning object.

Figure 15 is the learning path consistency and disparity analysis for COMP1005 for all teaching-learning styles, activity type lecture, where learning objects were accessed for 0 or more minutes, for the diagnosed learning style, and using consistency check criterion being the learning object. This diagram may be used to easily identify the level of consistency and disparity for the various learning styles.

LEARNING PATH CONSISTENCY AND DISPARITY, SEMESTER 2, 2016-2017

Teaching/Learning Style(s): ALL
Learning Object Activity/ies: 'A','B','F','G','L','O','S','T'
Disparity Allowance Time (mins.): 0
Assigned/Chosen Learning Style: Diagnosed
Disparity/Consistency Check Criterion: Learning Object

No.	Learning Style	Consistency		Disparity		Total	
		Course: COMP1005					
1	Visual Language	1	14%	6	85%	7	100%
2	Visual Numerical	5	33%	10	66%	15	100%
4	Auditory Numerical	0	0%	3	100%	3	100%
5	Kinesthetic - Tactile	2	66%	1	33%	3	100%
6	Social Individual	0	0%	8	100%	8	100%
8	Oral Expressiveness	0	0%	2	100%	2	100%
Course Totals:		8	21%	30	78%	38	100%

Figure 14: Learning Path Consistency and Disparity Check for All Activities

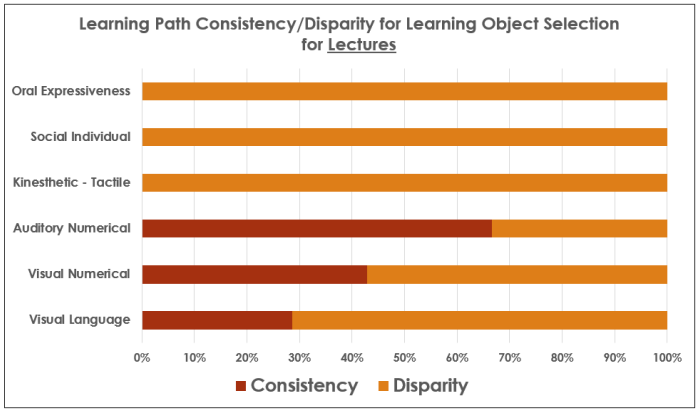


Figure 15: Learning Path Consistency/Disparity for Learning Object Selection for Lectures

New student learning paths may be auto-configured subject to a minimum threshold and the instructor's approval. Where there is auto configuration and the minimum threshold met for a particular style, subsequent students of the learning style(s) reconfigured are administered the course content according to the newly learnt path. The minimum threshold is met where the threshold percentage/factor of the students of a particular style chooses a consistent path (contrary to the teacher designed path). For the analysis including all activities, chosen learning path variances met the minimum threshold criterion of 75% (of students within similar learning style) for updates in learner course path; this was evident for Visual Language, Auditory Numerical, Social Individual, Oral Expressiveness (Figure 14). Where the minimum threshold is not met but there is significant disparity for a particular learning style, the notifications to the teacher(s)/instructor(s) were ideal as they could more easily manage the process in view of what learning objects were favoured. There were instances where the assigned teacher paths were followed, especially for students that were of a Kinesthetic-Tactile learning style.

Patterns of path selection per learning style were identified for the larger courses (COMP1005, COMP2005). In the instance for reviewing OLeCenT in which the users were selected similar to that of a focus group, the data showed more inconsistency with unidentified patterns. The patterns were more identifiable where OLeCent was used in the actual class environments.

5.3.5 Usage of OLeCent

Comparing COMP1005 which provides course delivery to 39 students through OLeCenT, and a Moodle instance of INFO3005 which provides similar delivery to 44 students, each environment was setup by the same teacher and course designer with a similar number of learning objects subject to the limitation of the Moodle environment in terms of managing VARK learning objects. OLeCenT reflected 1934 views in the course or 50 views per student, whilst Moodle reflected 3492 views in its course or 70 views per student.

Across two (2) campuses, in view of all students and excluding the maximum and minimum length of time spent accessing OLeCenT, the usage over eleven (11) weeks reflected an average of between twenty-two (22) and fifty-seven (57) hours. The findings are significant as students from the same campus as the

main course teacher did not reflect lower usage figures where the main teacher was more easily physically accessible. Considering all students in the course, the average usage for 11 weeks per student at the campus without lecturer was 3,449 minutes or 57 hours. This was comparatively more than the average for students on the same campus as the main lecturer of 2,255 minutes or 38 hours. Where students with the maximum and minimum averages were excluded, the pattern remained the same with average usage of 2,923 minutes or 49 hours versus 1,313 minutes or 22 hours respectively.

A survey of students of the tertiary computing first year course COMP1005 was administered to a class of 39 students. Of the 31 responses (or 79% of all responses) received, 41% of students have used 1 other learning management tool, e.g. Moodle or Blackboard and 55% have used 2 or more other LM tools.

The summary of the survey reflecting OLeCenT usage are included as strengths, weaknesses and other points of concern:

Strengths of OLeCenT

- 93% used OLeCent at least once per week. Internationally: 56% (of students in a course) use LMS daily
- 93% recognized the learning object of different formats and used opportunity to view other formats
- 87% felt that OLeCent was effective for course delivery
- Only 9% felt that OLeCent did not improve learning performance
- Only 3% felt that OLeCent was inefficient
- 58% chose the VARK Learning Object assigned by system

Weaknesses of OLeCenT

- 36% undecided or agreed that OLeCenT was frustrating to use
- 29% said that OLeCenT was not easy to use

Other Points of Concern with respect to OLeCent usage

- 52% undecided whether a system should modify path based on learning (45% agreed to system recalculation of learners' paths)

29% of respondents were undecided whether OLeCenT could be considered higher in course delivery methodology (48% agreed that OLeCenT was above other learning management tools).

5.4. Benefits and Drawbacks

OLeCenT determines learning course paths based on the determined learning styles. The learning styles are determined through the diagnostic assessment in its first module, Diagnostic Analysis. OLeCenT uses the input of any accepted learning style assessment or inventory model. Whereas this may be a benefit to the flexibility of OLeCenT, the drawback is that two of the other three modules, namely Learning Administration, and Learner-Centric Assessment and Evaluation are dependent on the correct determination of the learners' learning styles. Course developers may develop only some of the teaching-learning style learning objects recommended by the learner-centric tool and still increase the level of learning. For this to happen, developing learning objects for the teaching-learning style matches is omitted but learners make full use of the teaching provided through the human teacher(s).

Other benefits include:

- Allow codifying of multiple Learning Style instruments
- Allow for learning objects to be tagged for single or multiple learning types and can be used across multiple Learning Style instruments
- Identify disparity in teaching and learning and aids face-to-face teacher in knowing percentage of students with greatest benefit in course delivery
- Allow learner flexibility of receiving course content deemed to be best fit as first priority
- Reshape course path per learning style as it learns of learning preference
- Should cause a skew primarily of the teaching curve and a shifting of the equilibrium point [70] of the teaching-learning process if the knowledge gained is used appropriately.

Amidst these benefits, the learner-centric course delivery tool has a number of drawbacks. These include the extensive possibly additional work for the course developer in providing learning objects for each learning style for each course objective to meet the Visual-Auditory-Kinesthetic (VAK) standard. Providing an electronic solution sometimes requires more resources in terms of time, energy, or skill. There is also always the concern of determining the level of granularity for the teaching-learning style learning objects to ensure maximum usability [77]. Another drawback is that eventually all components of OLeCenT will be SCORM 2004 compliant. Many LMSs and tools especially in some international regions are compliant with SCORM 1.2 [16] and there would be decreased efficiency of the tool if designed to be compliant with SCORM 1.2.

Other drawbacks include:

- Learning and auto-configuration takes place during course delivery and therefore provides greater benefit to latter users
- Where a user's learning style does not change, as a precautionary measure, that user has to periodically (once per term, per year or greater as deemed fit by the administrator) re-diagnosed for learning style.

6. Conclusion

Higher Education Institutions use Learning Management Systems (LMSs) worldwide by to assist with improvements in course delivery or management. Moodle and Blackboard are predominantly used throughout Higher Education Institutions. The Learning Management System of greater usage has limitations and do not foresee immediate usage of features including sequencing and navigation. Learning styles have been purported by some to have relevance and usage in assisting learning. Researches [10-11] suggest that there may be an optimal level of learning in a given time and nature of the uptake function; we consider the nature of the uptake function to include the learner, learning environment and learning style.

The OLECENT approach considers drawbacks of some of the Learning Management Systems (LMSs). OLeCenT, a proof of

concept of the OLECENT approach, is developed to be integrated within these LMSs, thereby enhancing functionalities that were not obtained. OLeCent incorporates teaching approaches / styles and learning mechanism / styles with an aim of ensuring that learner-centric course delivery may be administered thereby increasing the level of learning. Instantiations of OLeCenT have shown benefits to teachers and learners including identifying disparity in teaching-learning and providing a mechanism towards improving online learner-centric course delivery, and allowing the input of one or more course learning paths with real-time learning and automatic reconfiguration of the course path where a new trend or pattern is identified. The OLECENT approach may be adopted within Learning Management Systems or the tool incorporated to enhance learning where there may be physical disabilities.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

Thanks to the University of the West Indies which made the resources possible for this research.

References

[1] C. Beckford, E. Mugisa, "OLeCenT, ATool towards Improving Online Learner-Centric Course Delivery" Proceedings of Future Technologies Conference 2016. IEEE (FTC 2016), San Francisco, USA, 2016.

[2] D. Laurillard, "Rethinking university teaching: A conversational framework for the effective use of learning technologies" Routledge, 2013.

[3] I. Forsyth, "Teaching and learning materials and the Internet" Routledge, 2014.

[4] B. Asheim, L. Coenen, J. Vang, "Face-to-face, buzz, and knowledge bases: sociospatial implications for learning, innovation, and innovation policy" Environment and Planning C, 25(5), 655, 2007.

[5] S. R. Aragon, S. D. Johnson, N. Shaik, "The influence of learning style preferences on student success in online versus face-to-face environments" The American Journal of Distance Education, 16(4), 227-243, 2002.

[6] I. E. Allen, L. Seaman, "Changing Course: Ten Years of Tracking Online Education in the United States" Sloan Consortium. PO Box 1238, Newburyport, MA 01950, 2013.

[7] M. Bar-Yam, K. Rhoades, L. Booth Sweeney, J. Kaput, Y. Bar-Yam, "Changes in the Teaching and Learning Process in a Complex Education System" Complex Systems Perspectives on Education and the Education System, New England Complex Systems Institute, 2002.

[8] S. Kara, "Learning Styles And Teaching Styles: A Case Study In Foreign Language Classroom" Conference of the International Journal of Arts and Sciences 1(20): 77 – 82, 2009.

[9] G. Sabeh, R. Bahou, N. N. Bacha, M. Nabhani, "A Match or a Mismatch between Student and Teacher Learning Style Preferences" International Journal of English Linguistics, Vol. 1,No., 2011.

[10] D. R. Krathwohl, B. S. Bloom, B. B. Masia, "Taxonomy of Educational Objectives, the Classification of Educational Goals. Handbook II: Affective Domain" New York: David McKay Co., Inc, 1973.

[11] L. K. Son, R. Sethi, "Metacognitive control and optimal learning" *Cognitive Science, 30*(4), 759-774, 2006.

[12] H. Coates, R. James, G. Baldwin, "A critical examination of the effects of Learning Management Systems on University Teaching and Learning" Tertiary Education and Management 11: 19-36,2005, Springer, 2005.

[13] A. S. Sife, E. T. Lwoga, C. Sanga, "New technologies for teaching and learning: Challenges for higher learning institutions indeveloping countries" International Journal of Education and Development using ICT > Vol.3, No. 2, 2007.

[14] D. A. Falvo, B. F. Johnson, "The Use of Learning Management Systems in the United States" TechTrends – March/April 2007, Volume 51, Number 2. 2007.

[15] P. M. Drinkwater, C. M., Adeline, S. French, K. N. Papamichail, T. Rickards, "Adopting a web-based collaborative tool to support the Manchester method approach to learning" *Electronic Journal on e-Learning*,*2*(1), 61-68, 2004.

[16] C. Beckford, E. Mugisa, "Learning Management Systems – The Current Picture" Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education, AACE (E-Learn 2014), New Orleans, USA, 2014.

[17] Hamline University, "Blackboard at Hamline" www.hamline.edu/blackboardhelp/, 2004. Retrieved 2016-01-10.

[18] Princeton University, "Blackboard Overview" The Trustees of Princeton University. www.princeton.edu/bb/, 2016. Retrieved 2016-06-10.

[19] S. Overly, "The Download: Blackboard makes first acquisition under CEO Jay Bhatt" The Washington Post. www.washingtonpost.com/business/capitalbusiness/the-download-blackboard-makes-first-acquisition-under-ceo-jay-bhatt/2014/01/16/35e7dc1e-7d5d-11e3-9556-4a4bf7bcbd84_story.html, 2014. Retrieved 2014-09-09.

[20] Saudi Gazette "Mideast adapts to e-learning technologies" Saudi Gazette. www.saudigazette.com.sa/index.cfm?method=home.regcon&contentid=20130429163468, 2013. Retrieved 2014-09-09.

[21] Blackboard Inc., "Blackboard Learn Platform – There's More to Learn" www.blackboard.com/about-us/index.aspx, 2016. Retrieved 2016-06-14.

[22] R. Riddell, "12 Leaning Management System providers and What They Bring to Classrooms" EducationDIVE. www.educationdive.com/news/12-learning-management-system-providers-and-what-they-bring-to-classrooms/97613/, 2013. Retrieved 2014-09-07.

[23] Moodle, "Moodle Features" docs.moodle.org, 2016. Retrieved 2016-06-14.

[24] D. Nagel, "Instructure Creates Free 'Canvas Network' for Online Courses" campustechnology.com/articles/2012/11/01/instructure-creates-free-canvas-network-for-online-courses.aspx, 2012. Retrieved 2014-09-10.

[25] Harvard University, "New School Year, New Course Platform. The President and Fellows of Harvard College" tlt.harvard.edu/news/new-school-year, 2016. Retrieved 2016-06-12.

[26] Instructure, "Canvas Features" www.instructure.com, 2016. Retrieved 2016-06-01.

[27] IBL Studios Education "Is Canvas The Next Dominant LMS?" IBL Studios Education, http://iblstudios.com/is-canvas-the-next-dominant-lms/, 2014. Retrieved 2016-06-15.

[28] K. Winkler, "Instructure Canvas – The Next Dominant Education Platform?" Edukwest, 2014.

[29] Columbia University, "The Sakai and Canvas Comparison" https://canvas.gsb.columbia.edu/courses/361092/pages/lms-selection-process, 2012. Retrieved 2014-09-09.

[30] Sakai "SAKAI: Collaboration and Learning Environment For Education" sakaiproject.org, 2016. Retrieved 2016-03-07.

[31] E. Cheng, R. Kleinberg, S. Kruk, W. Lindsay, D. Steffy, "A Strictly Combinatorial Approach To A University Exam Scheduling Problem" Congress, Numerantium, Vol 167. Winnipeg; Utilitas Mathematica, 2004.

[32] A. Bickford, "12 Common Complaints about Learning Management Systems" Connect Thinking Pty Ltd., 2013.

[33] P. F. Bergen, "The State of the Learning Management at The University of Chicago and Beyond" The University of Chicago, https://itservices.uchicago.edu/connect/article/state-learning-management-system-university-chicago-and-beyond, 2013. Retrieved 2014-09-11.

[34] C. Govender, D. W. Govender, "An Exploratory Study: The effectiveness of a Learning Management System (LMS) in the delivery of a face-to-face programming course" Academia.edu 2014.

[35] D. Everhart, "Competency Based Learning and Learner-Centric Shifts in Education" Blackboard Inc., 2014.

[36] T. Halbherr, K. Reuter, D. Schneider, C. Schlienger, T. Piendl, "*Making Examinations More Valid, Meaningful And Motivating: The Online Exams Service At Eth Zurich"* European University Information Systems (EUNIS), 2014.

[37] J. Biggs, C. Tang, "*Teaching for quality learning at university*" McGraw-Hill International, 2011.

[38] K. Struyven, F. Dochy, S. Janssens, "Students' perceptions about evaluation and assessment in higher education: A review 1. Assessment & Evaluation in Higher Education" 30(4), 325-341, 2005.

[39] J. T. Gulikers, T. J. Bastiaens, P. A. Kirschner, "A five-dimensional framework for authentic assessment" *Educational Technology Research and Development*, 52(3), 67-86, 2004.

[40] S. M. Wilson, P. L. Peterson, "*Theories of learning and teaching: what do they mean for educators?*" Washington, DC: National Education Association, 2006.

[41] J. Savery, "BE VOCAL: Characteristics of Successful Online Instructors" Journal of Interactive Online Learning, Volume 4, Number 2, 2005.

[42] H. Gardner, "The Unschooled Mind: How Children Think and How Schools Should Teach" Basic Books, 2013.

[43] J. Sweller, J. Van Merriënboer, F. Paas, "Cognitive load during problem solving: Effects on learning" *Cognitive Science (*2): 257–285, 1998.

[44] S. D. Sorden, "The cognitive theory of multimedia learning" Handbook of educational theories, Charlotte, NC: Information Age Publishing, 2012.

[45] A. H. Maslow, "*Motivation and Personality*" 2nd. Ed., New York, Harper & Row, 1970.

[46] D. Hung, "Theories of learning and computer-mediated instructional technologies" *Educational Media International*, *38*(4), 281-287, 2001.

[47] S. Papert, I. Harel, "Constructionism" Ablex Publishing Corporation, 1991.

[48] H. Gardner, "Frames of Mind" NY: Basic Books, 1983.

[49] S. Kara, "Learning Styles And Teaching Styles: A Case Study In Foreign Language Classroom" Conference of the International Journal of Arts and Sciences 1(20): 77 – 82, 2009.

[50] A. P. Gilakjani, "A Match or Mismatch Between Learning Styles of the Learners and Teaching Styles of the Teachers" I. J. Modern Education and Computer Science, 11, 51-60, 2012.

[51] A. Kolb, D. Kolb, "Learning styles and learning spaces: Enhancing experiential learning in higher education" Academy of Management Learning and Education, 4(2):193–212, 2005.

[52] S. Cassidy, "Learning Styles: An overview of theories, models, and measures" Educational Psychology Vol. 24, No. 4, Carfax Publishing, 2004.

[53] N. Fleming, D. Baume, "Learning Styles Again: VARKing up the right tree!", Educational Developments, SEDA Ltd, Issue 7.4, 4-7, 2006.

[54] A. Gilakjani, A. S. Masoumeh, "The Effect of Visual, Auditory, and Kinaesthetic Learning Styles on Language Teaching" International Conference on Social Science and Humanity IPEDR vol.5, IACSIT Press, Singapore, 2011.

[55] N. Chick, "Learning Styles" Vanderbilt University, Centre for Teaching, cft.vanderbilt.edu/guides-sub-pages/learning-styles-preferences/, 2011. Retrieved 2016-06-04.

[56] H. F. Silver, R. W. Strong, M. J. Perini, "*So each may learn: Integrating learning styles and multiple intelligences*" Association for Supervision and Curriculum Development, 1703 North Beauregard Street, Alexandria, VA 22311-1714, 2000.

[57] D. A. Kolb, R. E. Boyatzis, C. Mainemelis, "Experiential learning theory: Previous research and new directions" *Perspectives on thinking, learning, and cognitive styles*, *1*, 227-247, 2001.

[58] J. J. Koob, J. Funk, "Kolb's Learning Style Inventory: Issues of Reliability and Validity", Sage Publications, 2011.

[59] K. Cherry, "Kolb's Learning Styles" About Education, About, 2014.

[60] C. Manolis, D. Burns, J. David J, R. Assudani, R. Chinta, "Assessing experiential learning styles: A methodological reconstruction and validation of the Kolb Learning Style Inventory" Learning and Individual Differences, Volume 23. Elsevier, 2013.

[61] E. das Chagas Lemos, L. A. M. do Amaral, L. R. M. Oliveira, "Using Virtual Environments Adapted to Individual Capacities of Build Knowledge: an Approach to Distance Learning Courses" In *Conference Proceedings. The Future of Education* (p. 97), libreriauniversitaria.it Edizioni, 2016.

[62] R. Dunn, K. J. Dunn, "Learning Styles/Teaching Styles: Should They… Can They Be Matched?" Association for Supervision and Curriculum Development Journal, 1979.

[63] D. V. Dear, "Do student-centred learning activities improve learning outcomes on a BTEC Applied Science course in FE?" *Journal of Further and Higher Education*, 1-10, 2016.

[64] R. Basturk, "The Effectiveness of Computer-Assisted Instruction in Teaching Introductory Statistics" Educational Technology & Society, 8 (2), 170-178, 2005.

[65] C. H. Chuan, D. Dinsmore, J. Schmuller, T. Morris, " An Intelligent Tutoring System for Argument-Making in Higher Education: A Pilot Study" 13th International Conference on Machine Learning and Applications (ICMLA), 2014.

[66] T. C. Chen, A. S. MdYunus, W. W. Ali, A. Bakar, "Utilization of Intelligent Tutoring Systems (ITS) in mathematics learning" *International Journal of Education and Development using ICT, Vol. 4, No. 4*, 2008.

[67] Advanced Distributed Learning (ADL), "SCORM Users Guide for Instructional Designers. Sharable Content Object Reference Model (SCORM®)" 2004 4th Ed. Version 8, Advanced Distributed Learning (ADL), 2011.

[68] Rustici Software, "Tin Can API: Current Adopters" tincanapi.com/adopters/, 2016. Retrieved 2016-06-15.

[69] C. Beckford, E. Mugisa, "A Tool for Learner-Centric Course Delivery in e-Learning" In J. Sanchez & K. Zhang (Eds.), Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education (pp. 2475-2484), Chesapeake, VA: AACE (E-Learn 2010), Orlando, USA, 2010.

[70] R. Johnsonbaugh, "*Discrete Mathematics*" Upper Saddle River, NJ: Pearson Prentice Hall, 2009.

[71] J. C. de Lemos, A. Pelin, J. Quiroz, "The Relationship between Learning Styles and the Computer-based Instruction System PRECALC" Florida International University, 1995.

[72] K. Laudon, J. Laudon, "*Management Information Systems, Managing The Digital Firm*" Upper Saddle River, NJ: Pearson Prentice Hall, 2014.

[73] R. S. Pindyck, D. L. Rubinfeld, "*Microeconomics*" Needham Heights, MA: Simon & Schuster Custom Publishing, 2009.

[74] Herridge Group, "Use of traditional Instructional Systems Design Models for eLearning" Herridge Group Inc., 2004.

[75] R. M. Gagne, L. J. Briggs, "*Principles of instructional design*" NY: Holt, Rinchart and Winston, Inc., 1974.

[76] S. Xu, S. Li, "The SCORM Standard and its Application in Web-based Education Resources Building" *IEEE International Symposium on IT in Medicine and Education*, 2008.

[77] C. Innis-Allen, E. Mugisa, "A flexible taxonomy of learning objects based on content and media centric approaches to granularity" *Seventh IASTED International Conference on Web-based Education,* Austria. 275, 2008.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 839-844 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

An efficient model to improve the performance of platelet inventory of the blood banks

Annista Wijayanayake*, Manisha Dandunna

Department of Industrial Management, University of Kelaniya, 094, Sri Lanka

ARTICLE INFO

Article history:
Received: 05 April, 2017
Accepted: 16 May, 2017
Online: 19 June, 2017

Keywords:
Platelets
Shortages
Wastage
Inventory management
Demand forecasting

ABSTRACT

Platelet transfusions are vital for the prevention of fatal hemorrhage. Therefore, a stable inventory of platelets is required for an efficient and effective delivery of services in all the hospitals and medical centers. However, over the past decades, the requirement for platelets seems to be continuously increasing, while the number of potential donors is decreasing. Moreover, due to its very short life span of just five days, a large volume of platelets expires while they are on the shelves, resulting unnecessary shortages of platelets. Furthermore, it is very costly and difficult to get platelets from another blood bank in a short notice. Hence, these unexpected shortages put the life of patients at risk.

This study is focused on addressing the issues discussed, by developing an efficient blood inventory management model to reduce the platelet shortages, and wastages, while reducing the related inventory costs. Currently, the blood banks are managing platelet inventory according to their own instincts, which result to shortages and wastages. As a solution, we propose a model to manage the daily supply of platelets by forecasting the daily demand. Three different algorithms were developed using lower bound, average and upper bound values and tested to find the optimal solution that best fits to manage platelet inventory. These models were tested using data for 60 days obtained from two different levels of blood banks in Sri Lanka, namely a General Hospital blood bank and a Base Hospital blood bank. In General hospitals, the demand for blood components including platelets is very high when compared to the Base hospitals.

The study was able to come up with two different inventory management models for the two different types of blood banks. The model that best fits the General Hospital blood bank where the demand is high and was able to reduce the shortages by 46.74%, wastage by 89.82% and total inventory level by 39.10% and, the model that best fits the Base Hospital blood bank was able to reduce the shortages by 100%, wastage by 81.48% and total inventory level by 37.95%. It can be concluded that the same models can be adopted by other blood banks with similar characteristics to manage their blood inventory efficiently.

1. Introduction

Blood transfusions are generally required in surgeries, childbirths, organ transplants and for patients who are receiving treatments for diseases such as cancers and anemia. In Sri Lanka, total annual blood requirement exceeds 350,000 units and over the past decade the demand seemed to be increasing continuously. Since there are many voluntary donors as a result of the national culture and beliefs of Sri Lankans being a South Asian country, the demand was balanced throughout the history.

However, the blood units are outdated and blood is wasted in some areas while there are shortages in other areas. Blood shortages can put a human life at risk as the operations may have to be postponed, and moreover blood is an invaluable product and its wastage is highly undesirable.

Furthermore, with the increase of diseases and the increase in the number of patients who required surgery and transplants, over the past two decades, the demand for blood is expected to show an exponential growth in the future. On the other hand, with the increase in diseases transmitted through blood such as HIV and Hepatitis B, and with the increase in the aging population, the

* Annista Wijayanayake, Department of Industrial Management, University of Kelaniya, 094, Sri Lanka | Email: anni@kln.ac.lk

number of potential donors will be reduced. Hence there would be a challenge to fulfill the increasing demand with the reducing supply in the future.

1.1 The National Blood Transfusion Service

National Blood Transfusion Service (NBTS) is engaged in the collection of blood, screening the collected blood and making different blood components, and distribution. Also, NTBS deals with any issues of whole blood and blood component etc. It is a nationally coordinated system with the National Blood Center (NBC). This has many branches/clusters in every province with several main cluster centers in each province. The National Blood Centre, which is the operational headquarters for the entire Blood Transfusion Service, has 96 bloods Banks Island wide. These 96 blood banks are incorporated into 19 cluster centers according to their geographical locations. Apart from the NBC cluster, in all other clusters, the main cluster center is a General Hospital blood bank and the blood banks under them are Base Hospital blood banks.

The inefficient inventory management methods in each cluster and lack of coordination among different clusters lead to wastages as well as shortfalls of blood. Furthermore, despite the fact that blood is a costly product, the national blood transfusion service in Sri Lanka is not cost effective and the expenses are increasing day by day.

1.2 Platelets

The whole blood is processed into several blood components or products including plasma, platelets, red blood cells and, another hundred and fifteen other components. Blood plasma is a straw colored liquid component of blood that normally holds the blood cells in whole blood in suspension. Red blood cell**,** also called erythrocyte, cellular component of blood millions of which in the circulation of vertebrates give the blood its characteristic color and carry oxygen from the lungs to the tissues. Platelets are the blood cells that help the blood clot, stop bleeding when blood vessels are injured. Furthermore, when patients are receiving chemotherapy or suffering bleeding complications require platelet transfusions to prevent a fatal hemorrhage. Normally platelets are produced from whole blood in the initial separation process; on the other hand, platelets can also be collected through apheresis. However, in Sri Lanka only the national blood center has an apheresis machine.

Among all the blood components platelets have the shortest life span of just five days making it difficult to handle the inventory without experiencing both wastage and shortages. Hence this study is focused merely on platelets and all the solutions provided are customized to the special characteristics of platelets. Moreover, component processing and testing for transmissible diseases take an average of one day to complete and platelets are generally available for transfusion only for four days. However, platelets need not to be cross-matched as red blood cells.

2. Literature Review

Only very few research was done on the topic on ‘management of blood bank supply chain. Research on the management aspects of the blood supply chain were started in the 1960s, peaked in the late 1970s and early 1980s and then dropped off continuously to the present time. In 70s and 80s, many researches were done regarding blood bank management policies [1], and regarding theories of perishable inventories [2]. Thereafter, it shows a decline and experts argue that this was caused by the collapse of funding for studies in the area and in the increasing difficulty of the remaining problems in this area [3]. Since the mid-1980s, research on this topic has mostly been in the development of information systems to support donor screening, inventory management, blood ordering, blood usage review and compatibility testing [4]. Furthermore, few other studies were done during the past two decades as well.

From 1960s to 2000s William P. Pierskalla has done several researches in this area addressing different issues. In 1970s a transportation and location allocation model was developed for regional blood banking in the hope of decreasing overall costs. The objective was to decide how many central blood banks to be set up, where to locate them, how to allocate hospitals to the banks and how to route the periodic supply operations [5].

In 2001 the Nova Scotia government funded a project to investigate mechanisms to coordinate the platelet supply chain between CBS (Canadian Blood Services) and local hospitals. This study was done in three phases and in the first step a platelet repository was developed to store information necessary for an integrated planning model. In the second step a dynamic programming model was developed to identify optimal ordering policies for both the CBS and hospitals, and finally a simulation model was developed to evaluate integrated order policies. To simplify the problem by reducing the problem search space, orders and demands were aggregated into larger units and that approach was not used in the previous studies [6].In 2004 a detailed analysis was published, addressing the issues; regional structures and economies of scale, optimal target inventory levels, cross-match release policies, allocation, transshipment and issuing policies, vehicle routes and demand forecasting [3]. Cross-match releasing policies and other issues such as dealing with mismatching and keeping rare blood products in the inventory were addressed later by Katsaliaki and Brailsford. [7].

Apart from the statistical approaches, few studies were conducted on applying lean techniques to improve the performance of the blood supply chain. A recent study was done in Serbia to identify the most appropriate lean six sigma improvement methods and hence develop a framework to improve main blood inventory management processes that will minimize the wastage of blood [8].

Though different models were developed to address different issues, only few models have integrated more than one blood product or varieties of products, due to the complexity of modeling the different shelf lives and of computing the remaining shelf life. Furthermore, in all the above cases, collection of blood is a centralized activity, whereas in Sri Lanka, blood is collected by each and every blood bank to satisfy their own demand. On the other hand, the national blood center is distributing blood to all blood banks in the country if a request was made. Apart from that hospital blood banks can share their excess volumes with other blood banks where there are shortfalls. Therefore, policies regarding inventory management as well as distribution are completely different from the above studies and hence there was

a need to conduct a study specific to the Sri Lankan blood transfusion service.

3. Methodology

In the current method, blood inventory management was handled manually. However, there is no proper mechanism used to handle them efficiently, but based on the previous demands and experience of the staff and their own instincts on the inventory of the blood components are handled and managed. Therefore, the current practices cause many issues and lead toward shortages and wastages at many occasions. In this research, a customized spreadsheet based on a statistical analysis was developed to manage the platelet inventory level in hospital blood banks. To develop the spreadsheet, several mathematical models were tested and using the results from each model, the best model was selected using a multiple criteria decision analysis method (MCDA).

The models were developed in three phases; in the first phase different methods for forecasting the daily demand for platelets were tested and the best models for each technique was selected, in the second phase using the selected forecasting models different lower bound, mean and upper bound values were defined and in the third phase three different algorithms were developed to determine the optimal target inventory levels for blood banks that will reduce both shortages and wastages, hence the cost of operations.

In 2007 Katsaliaki and Brailsford [7] found that in the UK the demand for platelets follow a Log Normal distribution and all previous researchers have used the same approach in their research. Hence it can be assumed that the demand is normally distributed and moreover as a pilot test, some demand values that match the Sri Lankan conditions were randomly generated and tested for normality. From the pilot test it was found that the platelet demand is normally distributed.

3.1 Phase I: Forecasting the daily demand for platelets

Since the daily platelet/blood supply needs to be balanced, it is needed to forecast the daily demand for platelets/blood. Four different time series techniques were used to forecast the demand. Since the daily demand is stochastic, the Moving Average (MA), Weighted Moving Average (WMA) and Exponential Smoothing (ES) methods were used to forecast the demand. In addition, a Trend Analysis was conducted to check whether the daily demand shows any linear trend. According to the experience and knowledge of the staff of the blood banks there is no upward or downward trend in the demand of platelets. However, in the months of December and April, due to the festival season, which experiences the high rate of accidents, an increase in the demand for blood components can be observed and experienced. Nonetheless, according to the experts there is no possibility of observing a trend in daily demand and yet the trend analysis was conducted to verify that fact.

Sampling frame for this study includes all the blood banks of the NBTS except the National Blood Center cluster. When selecting the samples, the level of the blood bank was taken into consideration and one general hospital blood bank (a main cluster center) and one base hospital blood bank was selected for data collection. In order to make the sample more representative, from two blood banks that were selected, one was with a comparatively high level of demand and the other was with a low level of demand and yet the level of supply is equivalent in both hospitals. Two months (60 days) data taken from the blood banks were used for forecasting and analyzing demand, under the all four techniques.

3.2 Phase II: Defining mean, lower bounds and upper bounds for daily demand

Under the normality assumption, using the daily demand values, the selected forecasting methods MA, WMA and ES, computed the, mean, lower bound (LB) and upper bound (UB) values for the demand for each model. These mean, UB, LB values were considered as the minimum inventory level that should be maintained in order to avoid the shortages. The confidence level of 95% was used to calculate the different upper and lower bounds.

3.3 Phase III: Managing the daily supply and total platelet inventory level

Several mathematical models were developed in this final phase. The objectives were to find the best model that minimizes wastages, experience no shortages, keep the inventory level as low as possible by managing the supply and ultimately reduce the cost.

Considering the three different levels of inventory; lower bound, mean and the upper bound, three algorithms were developed and those algorithms were used to develop the different models. In all these algorithms, it was assumed that the entire intake of blood is usable despite the fact that a certain number of blood units will be rejected and disposed if they were infected with diseases transmitted through blood. However, in Sri Lanka still it is a nominal negligible amount, yet it seems to be continuously increasing. Since the figures of rejected blood are difficult to obtain due to confidentiality issues and as they were small in values according to the experts, it was not considered in these algorithms.

Once the blood is donated, it will be separated into different components and then it is tested for infectious diseases. It almost takes one day and therefore platelets only have four more days left on which they can be used. However, blood banks only records the date a blood unit is donated, but not the time it was donated. Therefore, the remaining shelf life is calculated on a daily basis but not on the time scale.

Three algorithms were developed around the same basic theory as described below. Currently, blood banks receive all the blood units donated by donors and yet the main objective of algorithms, is to manage the supply in an efficient manner. All the algorithms are updating remaining shelf life on a daily basis, however data can be input and outputs can be viewed at any time. In these algorithms the platelets are issued to the wards based on "First In First Out" basis. However, platelets with a remaining shelf life of one day are not given away or taken from another blood bank as those units have a high possibility of expiring the next day, before they are used. Also, cost reduction is one indirect objective of the algorithms, dispatching or ordering such platelet units is highly undesirable.

In all algorithms, before taking blood from donors, the inventory levels are checked to find if the desired reorder levels are met. If the current inventory level is above the reorder level, then the donors were asked to donate it some other day or direct them to the nearest cluster location where the blood is needed. This would minimize the medical expenses used for testing the blood components unnecessarily, which may avoid wastages of blood components. If the blood inventory goes below the reorder level, orders are placed to get platelets from other blood banks, to fill the gap. Here it is assumed that the blood is available at all times and the lead time to receive ordered blood as zero days. Therefore, the reorder level and the safety stock levels will be the same.

It was assumed that the blood banks will perfectly manage the model and however, extremely rare or unexpected situations arises where demand could go very high or supply could go very low continuously for a few days. These situations were excluded as in such situations the NBTS intermediate to control such situations. Therefore, such situations were not taken into consideration when developing the model.

3.4 Mathematical model to manage the Blood inventory system

Variables in the inventory management algorithms

LB – Lower bound for the forecasted daily demand for algorithm 1

M – Mean for the forecasted daily demand for algorithm 2

UB – Lower bound for the forecasted daily demand for algorithm 3

S – Actual platelet supply during the day

N – Days of stocks to be kept (Each algorithm was tested for both N=1, 2, 3 and 4)

Xi – Amount of platelet units in the inventory with a remaining shelf life of i days

Yi – Amount of platelet units with a remaining shelf life of i days, which need to be dispatched from the inventory

Zi – Amount of platelet units with a remaining shelf life of i days, which need to be ordered

D – Actual demand for platelets during the day

TI – Total inventory level in a day
TI = X1 + X2 + X3 + X4 + X5

WS – Wastage in a day

SH – Shortages in a day

From a pilot test done using computer generated data it was found that the best values for the days of stocks to be kept N are 2 and 3, among the four values tested 1, 2, 3 and 4, in order to minimize the wastages and shortages.

In all three algorithms, it was assumed that the orders will be delivered within the same day as the order is placed. Furthermore, while checking for the desired inventory levels, X1 value was not taken into consideration. The reason that the algorithms were developed in that way is to see if any, remaining units (after satisfying today's demand) with a shelf life of only one day that is in today, will be expired tomorrow, and platelets taken today become usable only from the next day onwards. Therefore, ignoring X1 value reduces the risk associated with fulfilling the actual demand next day. However, algorithms were designed to ultimately minimize the total inventory level which is equal to X1 + X2 + X3 + X4 + X5.

The algorithm 1 presented below, which maintains the value X2 + X3 + X4 + X5 in the lower bound value is shown below. However this can be easily replaced by algorithm 2 and algorithm 3 by replacing the LB values by the mean and UB values

Algorithm 1

Objectives: - Min SH
Min WS
Min TI

Step 1: Checking the inventory
If X2 + X3 + X4 < LB*N Then
X5 = L*N - (X2 + X3 + X4)
Else
X5 = 0

Step 2: Amount to be dispatch or receive
If S >= X5 Then
Y5 = S - X5
Z5 = 0
Else
Y5 = 0
Z5 = (LB*N - (X2 + X3 + X4) - S

Step 3: Shortages
If D > X1 + X2 + X3 + X4 Then
SH = D - (X1 + X2 + X3 + X4)
Else
SH = 0

Step 4: Wastage
If X1> D Then
WS = X1 - D
Else
WS = 0

Total inventory level
TI = X1 + X2 + X3 + X4 + X5

Step1: The amount of platelet required on that day will be calculated
Step2: If the supply of platelet is greater than the required amount, then compute the dispatch amount Yi. If it is less than the required amount then compute the amount needed to be ordered Zi
Step3: Compute the amount of shortages
Step4: Compute the amount of wastages

3.5 Selecting the model that best fits the objectives

With the objective of minimizing the wastage, shortages and the total inventory level, all three algorithms explained above were tested using the different LB, Mean and UB values and using MA, WMA, and ES techniques under different confidence levels. There were altogether eighteen different models were developed and given different codes for each model which is shown in the below table (Table 1). For each model, the total wastage, total shortages and the total inventory level were calculated to compare the results and choose the best inventory management model, associated with the best forecasting technique using 95% confidence interval.

Forecasting methods	LB Algorithm 1		Mean Algorithm 2		UB Algorithm 3	
	** N=2	N=3	N=2	N=3	N= 2	N=3
MA	MA 12	MA 13	MA 22	MA 23	MA 32	MA 33
WMA	WMA 12	WMA 13	WMA2 2	WMA 23	WMA 32	WMA 33
ES	ES 12	ES 13	ES 22	ES 23	ES 32	ES 33

Table 1:- Alternative Models for Platelet Inventory Management

**N=2, 3, 4, 5 days were considered but when N=2 and 3 produced better results.

In each testing model, the first two to three characters represent the forecasting technique used, the first digit represents the algorithm model number and the last digit represents the number of days that the stock needed to be considered. Since all the objectives are measured in the same unit, Weighted Sum Model (WSM) was used as the MCDA method to determine the best model among the eighteen models (given in Table 1). Delphi method was used to rank the three objectives and determine a weight for each objective. This was done merely consulting the experts and the staff of blood banks. The weights that were given can be changed according the experience and knowledge of importance. In our model the importance or weights for shortages given as 0.4, wastages the weight given was 0.375 and the total inventory management level the weights given were 0.275 while the sum of weights become 1. The model that yields the minimum total performance value was selected as the best model for managing inventory level.

4. Results

Data on demand from both blood banks was tested for normality to find whether they follow a normal distribution. Shapiro - Wilk test was used as the test statistic and it was found that demand for both hospital blood banks follows a normal distribution.

4.1 Outcomes of the different forecasting techniques

Using different number of days where n=2 to 12 the demand was forecasted in the Moving Average method and for each n value, Mean Absolute Deviation (MAD) was calculated to find the best value which minimize the forecasted Vs the actual error values. For General hospital blood bank, the least MAD value can be observed when n = 3, whereas for Base hospital blood bank, n = 9 gives the least MAD value. Therefore, n = 3 and n = 9 is selected as the best values for forecasting the platelet demand using the moving average method, for General hospital blood bank and Base hospital blood bank respectively.

Using weighted moving average method, for n = 2 to 15, the daily demand for platelets was forecasted and rank order centroid method was used to calculate the weighting factor. The least MAD value of General hospital blood bank can be observed when n = 5 and for Base hospital blood bank it is n = 12 hence for General hospital blood bank, n = 5 was selected and for Base hospital blood bank, n = 12 was selected to forecast the daily platelet demand using the weighted moving average method.

Using different values for the smoothing factor α, the daily demand for the platelets were forecasted and, MAD values for each α were calculated to find the best α for each blood bank. For General hospital blood bank, the least MAD value can be observed when α = 0.2 and for Base hospital blood bank, the least MAD value can be observed when α = 0.3. Therefore, α = 0.2 and α = 0.3 were selected for General hospital blood bank and Base hospital blood bank respectively to calculate the forecasted daily demand.

It was found that the demand shows no linear trend, exponential trend or polynomial trend in the daily platelet demand at both General hospital blood bank and Base hospital blood bank, throughout the two months considered. Hence it can be concluded that the daily demand for platelets shows no underlying trends and the demand is random for both General hospital blood bank and Base hospital blood bank.

Using the all three forecasting methods selected, from samples of seven days, the mean daily demand was calculated for each day and, using 95% confidence levels, the respective lower bound and upper bound values were calculated for each day considered.

When the forecasted demand and the actual demands were analyzed the difference was a negligible amount. The actual and forecasted demands were shown below for ES33 model in Fig 1

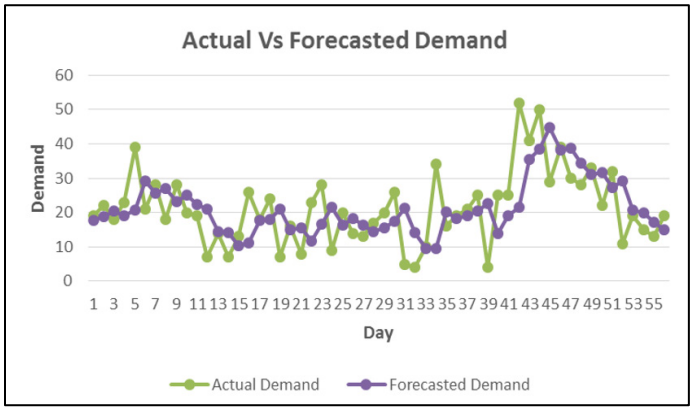


Figure 1:- Actual vs Forecasted Demand for model ES33

4.2 Analysis of the platelet inventory management models

For each blood bank, all the eighteen models shown in Table1 were applied and the percentage of reduction in shortages, the percentage of reduction in wastages and percentage reduction of the total inventory level was calculated for each of those models.

Thereafter, using the weighted sum model as explained in section 3, the total performance for each model was calculated.

The models were tested separately for 30 days and for 60 days using the new forecasted demand values. Thereafter the average performance for each model was calculated using the actual values. The difference between the actual values and the forecasted values were considered to evaluate the best model.

The table below shows the total performance of best three models for General hospital blood bank.

Model	Reduction of Shortages %	Reduction of Wastage %	Reduction of TI %	Total Performance %
WMA32	46.74	89.82	39.10	61.18
MA32	46.47	89.76	37.32	60.64
ES33	86.61	55.24	9.82	57.57

Table 2:- Best three Models for General hospital Blood Bank

As it shows the model WMA32 which use the weighted moving average method for forecasting the demand and use 95% confidence interval upper bound values and where the days of stock to be kept is two, gives the highest total performance for the General hospital blood bank. In average model WMA32 reduces shortages by 46.74%, wastage by 89.82% and total inventory level by 39.10%, resulting in a total performance of 61.18% achievement of objectives.

The below table shows the total performance of best three models for the Base hospital blood bank.

Model	Reduction of Shortages %	Reduction of Wastage %	Reduction of TI %	Total Performance %
ES33	100.00	81.48	37.95	79.09
WMA33	100.00	75.68	33.51	75.92
MA33	100.00	72.90	31.99	74.53

Table 3:- Best three Models for Base hospital blood bank

As the above table shows, the model ES33 which uses exponential smoothing as the forecasting method and use 95% CI upper bound values for the forecasted demand and where the days of stock to be kept is three, gives the best total performance for Base hospital blood bank. In average model ES33 reduces shortages by 100%, wastage by 81.48% and total inventory level by 37.95%, resulting in a total performance of 79.09% achievement of objectives.

When we analyze the efficiency of all 18 models that we proposed, model ES33 produce better results for both General hospitals and base hospitals to manage blood. Here the reduction of shortages is 86% and 100% respectively.

It can be concluded that the above model can be adopted by the other blood banks to improve their inventory performance while minimizing the shortages. The same algorithm can be applied to other products with similar characteristics, such as food industry, pharmaceutical items and other perishable goods.

As discussed and proven above the proposed model could reduce the wastages and manage the shortages, while maintain a minimum amount in the inventory of blood banks. This would save most of the medical expenses unnecessary spend on testing and managing the inventory of blood components and could be used to uplift the medical services in any hospitals.

5. Further research

The area is fruitful for future research as there are so many areas that need to be improved, in the blood supply chain. This study has focused only on platelets, however the scope can be increased by adding other blood components such as red blood cells and plasma. Another area that needs to be addressed in a further research is clustering the blood banks island-wide, in a more effective way based on the transportation cost and time that takes to deliver a blood product/s to another blood bank. It should be noted that the blood banks are clustered currently based on the district they belong to and no other valid logic was used to define those clusters.

Further research can be focused to develop algorithms for scheduling donors. A database should be maintained for existing and potential donors and it is needed to develop algorithms for scheduling those donors based on the requirements of each day. Therefore many of the practical issues can be overcome by developing a web based management information system incorporating our proposed algorithm. This would enable the staff of the blood bank to find out when there is a scarcity or excess of blood. According to the need the registered donors can be informed in advance and manage the blood supply chain smoothly and effectively. Therefore the wastages and shortages of the blood component can be reduced further.

References

[1] Prastacos, G., Blood Inventory Management: An overview of theory and practice, Management Science, vol. 30, pp. 777-800, 1984.

[2] Nahmias, S., Perishable inventory theory: An overview, Operations Research, vol. 30, pp. 680-708, 1982.

[3] Brandeau, M.L., Sainfort, F. and Pierskalla W.P., Supply chain management of blood banks in Operations research and health care, Boston: Kluwer Academic Publishers, pp. 103 – 145, 2004.

[4] Kern, D.A. and Bennett, S.T., Informatics applications in blood banking, Clinics in Laboratory Medicine, vol. 16, pp. 947-960, 1996.

[5] Or, I. and Pierskalla W.P., A transportation location – allocation model for regional blood banking, AIIE Transactions, vol. 11, no. 2, pp. 86 – 94, 1979.

[6] Blake, J.T., Thompson, S., Smith, S., Anderson, D., Arellano, R. and Bernard, D., Optimizing the platelet supply chain in Nova Scotia, in Proceedings of 29th meeting of the EURO Working Group on Operational Research Applied to Health Services, Prague, Czech Republic, pp. 47 – 65, July 27 – August 1, 2003.

[7] Katsaliaki K. and Brailsford S.C., Using simulation to improve the blood supply chain, Journal of the Operational Research Society, vol. 58, pp. 219-227, 2007.

[8] Jovanovic, R., Delic, M., Kamberovic, B., Vulanovic, S. and Radlovacki, V., Planning the use of lean six sigma as a framework for blood bank management improvements, International Journal of Industrial Engineering and Management, vol. 4, no. 4, pp. 237-244, 2013.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 845-854 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Performance Evaluation of Associative Classifiers in Perspective of Discretization Methods

Zulfiqar Ali[1,2*], Waseem Shahzad[1]

[1]Department of Computer Science, National University of Computer & Emerging Science, Islamabad, ZIP Code #44000, Pakistan.

[2]Department of Computer Science and IT, The University of Lahore, Lahore, ZIP Code#54000, Pakistan.

ARTICLE INFO

Article history:
Received: 22 April, 2017
Accepted: 16 May, 2017
Online: 19 June, 2017

Keywords:
Associative Classification
Discretization Methods
KEEL
Data Mining

ABSTRACT

Discretization is the process of converting numerical values into categorical values. Contemporary literature study reveals that there are many techniques available for numerical data discretization. The performance of classification method is dependent on the exploitation of the data discretizing method. In this article, we investigate the effect of discretization methods on the performance of associative classifiers. Most of the classification approaches work on the discretized databases. There are various approaches exploited for the discretization of the database to compare the performance of the classifiers. The selection of the discretization method greatly influences the classification performance of the classification method. We compare the performance of associative classifiers namely CBA and CBA2 on the selective discretizing methods i.e. 1R Discretizer (1R-D), Ameva Discretizer (Ameva-D), Bayesian Discretizer (Bayesian-D), Discretization algorithm based on Class-Attribute Contingency Coefficient (CACC-D), Class-Attribute Dependent Discretizer (CADD-D), Distribution-Index-Based Discretizer (DIBD-D), Cluster Analysis (ClusterAnalysis-D), Chi-Merge Discretizer (ChiMerge-D) and Chi2 Discretizer (Chi2-D) in terms of accuracy. The main object of this study is to investigate the impact of discretizing method on the performance of the Associative Classifier by keeping constant other experimental parameters. Our experimental results show that the performance of the Associative Classifier significantly varies with the change of data discretization method. So the accuracy rate of the classifier is highly dependent on the selection of the discretization method. For this comparative performance study, we use the implementation of these methods in KEEL data mining tool on public datasets.

1. Introduction

Discretization methods have played a great role in data mining and knowledge discovery. The discretization process makes learning more accurate and faster. There are various emerging classification problems in various domains of knowledge like image processing, medical science, business analytics and data mining etc. data, images, audio, video and textual data. The rapid growth in the data reservoirs in the fields of business, basket analysis, Engineering sciences, social networks, stock exchange and geological data is very high due to the cheaper storage resources. The high growth rate and huge data volume create a challenging problem i.e. knowledge discovery from the huge databases in the field of Data Mining. For the appropriate, effective and comprehensive knowledge discovery for the managers and decision makers; researchers are proposing continuously more efficient knowledge mining approaches.

The field of artificial neural networks, expert systems, medical science, bioinformatics, machine learning is example areas where extensively classification approaches have been studied. There are various approaches exploited for the building of associative classifiers. This comparative study provides the extension of the work presented in [1]. We provide the performance analysis of Associative Classifiers (CBA and CBA2) with the variation of the data discretizing method.

Thabtahand Fadi Abdeljaber provided the review of associative classification in [2]. Ranjana Vyas et al. describe the application of Associative Classifiers for Predictive analytics in [3]. The exploitation of associative classifiers for predictive analysis in the field of health care is surveyed in [4] by Sunita Soni and O.P.Vyas. Huan Liu et al. provide the extensive survey of the

*Zulfiqar Ali, National University of Computer & Emerging Science, H-11/4, Islamabad, Pakistan, +923335262172 & zulfiqar.ali@nu.edu.pk

discretization techniques in[5]. The focus of this article was an exploration of discretization methods with prospective of their historic development, the trade-off between speed and accuracy. Authors provided the hierarchical framework to categorize the exiting discretizing methods. Sotiris Kotsiantis and Dimitris Kanellopoulos surveyed the discretization techniques applied for the data discretization in [6].The theoretical and empirical perspective of the discretizing methods is very in [7] by the Salvador Garcıa. In [8] Pancho et al. provided the analysis of fuzzy association rules with Fingrams in KEEL.

In this article, we investigate the effect of discretizing methods on the performance of Associative Classifiers CBA and CBA2. We use the selective data discretizing methods i.e. 1R Discretizer (1R-D) [9], Ameva Discretizer (Ameva-D) [10], Bayesian Discretize r(Bayesian-D) [11], Discretization algorithm based on Class-Attribute Contingency Coefficient (CACC-D) [12], Class-Attribute Dependent Discretizer (CADD-D) [13], Distribution-Index-Based Discretizer (DIBD-D) [14], Cluster Analysis (ClusterAnalysis-D) [15], Chi-Merge Discretizer (ChiMerge-D) [16] and Chi2 Discretizer (Chi2-D) [17] for continuous data discretization purpose by exploiting the implementation of these methods in KEEL [18], a data mining tool by using the public datasets.Our experimental results reveal that the performance of the Associative Classifier significantly varies with the change of data discretization method in terms of accuracy. So the accuracy rate of the classifier is highly dependent on the selection of the data discretizing method. Our comparative study reveals that the performance of CBA (Associative Classifier) is better on the Ameva Discretizer than the other discretizing methods in terms of accuracy. The main object of this study is to investigate the impact of discretizing method on the performance of the Associative Classifier at the same other experimental parameters.

The Section 2 of the paper discusses the associative classification and describes the selective associative classification methods that are the focus of our study for the comparative analysis. Section 3 describes the data discretization process and selective methods for the data discretization i.e. 1R-D, Ameva-D, Bayesian-D, CACC-D, CADD-D, DIBD-D, ClusterAnalysis-D, ChiMerge-D, and Chi2-D. Section 4 explains the experimental Set-up exploited for this study, data sets and KEEL tool used for the experimentation. Section 5 describes the comparative performance results achieved by various discretizing methods used for datasets discretization by using Associative Classifiers CBA and CBA2. In section 6 more results discussion is provided and finally, the last Section concludes the study.

2. Associative Classification

The Associative Classification (AC) is a classification approach which integrates the classification rules mining and association rules mining that are two important data mining tasks. The Association Rule Mining (ARM) is unsupervised learning method in which no class attribute involved during the discovery of rules. The aim of the association rule mining is to discover associations between items in a transaction database. The attributes in the consequent of a rule could be more than one in association rule mining. The associative classification is a supervised leaning where a class must be given for the discovery of classification rules. For the construction of a classifier that can forecast the classes of test data objects is the main objective of associative classification. The consequent of a rule is an only class attribute. The over fitting is a considerable issue in the associative classification rule discovery. The over view of the selective Associative Classification approach CBA which is exploited to investigate the impact of discretizing method on the performance of the classification approach is given in the following sections.

2.1. CBA

Bing Liu, Wynne Hsu and Yiming Ma proposed a new hybrid classification approach by integrating the concept of association rule mining and classification rule mining in [19] that is named Classification Based on Associations (CBA). In this associative classification approach, the integration is done by focusing on the discovery of a special subset of association rules that are known as class association rules (CARs).

All class association rules are discovered those satisfy the minimum support and minimum confidence by using an existing association rule mining algorithms[20].The CBA associative classifier consists of two parts 1) a rule generator (CBA-RG) and 2) a classifier builder (CBA-CB).This approach possesses various advantages like the discretization of continuous attributes based on the classification pre-determined class target. The Data Mining task in CBA consists of the three steps;1) discretization of continuous attributes if any;2) class association rules;3) classifier building based on the generated class association rules.

2.2. CBA2

Bing Liu, Yiming Ma and Ching-Kian Wong proposed the enhancement and improvements in an associative classifier CBA. The new improved associative classification approach is named CBA2 developed in [21]. In this paper, theauthor tried to coup up with weaknesses of an exhaustive search based classification system CBA. The authors proposed two new techniques to deal with the observed weaknesses of the classification approaches. The first weakness observed is that as the traditional association rule mining exploits only a single minsup in rule generation which results inadequate for unbalanced class distribution. Secondly, classification data often contains a huge number of rules, which may cause a combinatorial explosion. For various databases, the rule generator is unable to generate rules with many conditions while such rules may be important for accurate classification. The first problem with this approach is tackled by using multiple class minsups in rule generation instead of single minsup as in CBA. The second problem which is caused by the exponential growth of the number of rules is dealt indirectly. The decision tree method [22] is exploited. The main working concept of the CBA2 is to use the rules of CBA2 to segment the training data and then select the classifier. These improvements in CBA improved the accuracy and lower error rate of the classification.

3. Data Discretization

Discretization is a data preprocessing technique used in many knowledge discoveries, machine learning and data mining tasks. Discretization process converts the continuous data into discrete form as most of the knowledge discovery and data mining algorithms work on discrete data. The discretization technique transforms a set of continuous attributes into discrete ones. By associating categorical values to intervals discretizing approach transforms quantitative data into qualitative data. The data discretiztion techniques are exploited to enhance the performance of the many knowledge discoveries and data mining approaches. We have used selective discretization method for our study to investigate the performance of Associative Classifier CBA by

using public data sets. The discretization methods used in this study are described in the following.

3.1. Discretization Methods

Gonzalez-Abrilet al. proposed a new discretization method named Ameva Discretizer (Ameva-D) in [10] that is proposed for supervised learning algorithms. The Ameva discretization approach maximizes a contingency coefficient based on Chi-square statistics. It helps in generating a potentially minimal number of discrete intervals. The most distinguishing feature of Ameva with respect to other discretizing approaches is that it does not need the user to indicate the number of intervals.

Xidong Wu proposed a new discretizing algorithm namely Bayesian Discretizer (Bayesian-D) in [11]. The Bayesian-D discretization approach exploits the Bayes formula. Cheng-Jung Tsai et al. propose a discretization algorithm based on Class-Attribute Contingency Coefficient named (CACC-D) in [12]. The CACC-D discretizing approach is motivated by the contingency coefficient. The CACC algorithm is a static, global and incremental discretizing approach. The CACC-D is supervised and top-down discretization algorithm which is based on Class-Attribute Contingency Coefficient. J.Y. Ching et.al proposed anew method for continuous data discretization named Class-Attribute Dependent Discretizer (CADD-D) in [13]. The class-dependant discretizing is performed in this approach for inductive learning from continuous and mix-mode data. The CADD-D is a discretizing method optimized for supervised learning based on information theoretic discretization method. The interdependence redundancy between the discrete intervals and the class labels is measured in the CADD-D method tom maximize the mutual dependence.

L.A. Kurgan et.al proposed a discretization method namely known as Class-Attribute Interdependence Maximization (CAIM-D) in [23]. The CAIM-D method is proposed for the supervised data classification. The main objective of this proposed approach is to increase the class-attribute interdependence to maximum level and to produce a minimal number of discrete intervals. The classification results in terms of accuracy are more promising with CAIM discretization method with respect to other discretization approaches. Huan Liu and Rudy Setiono proposed a method for converting numeric data discrete named Chi2 Discretizer (Chi2-D) in [17]. This approach (Chi2-D) takes data sets with numeric attributes as an input. This approach can intelligently and automatically discretize the numeric attributes as well as remove irrelevant ones as output. The Chi2 algorithm applies the X 2 statistic which conducts a significance test on the relationship between the values of an attribute and the categories.

In [16] Randy Kerber proposed a method for the discretization known as Chi-Merge Discretizer (ChiMerge-D) which is a general, robust algorithm that uses the X 2 statistic to discretize numeric attributes. The ChiMerge approach provides a useful and reliable summarization of numeric attributes. The number of intervals needed is determined according to the characteristics of the data. Michal R. Chmielewski and Jerzy W. Grzymala-Busse proposed discretizing method based on Cluster Analysis named ClusterAnalysis-D in [15]. The hierarchical cluster analysis is used for discretizing attributes in ClusterAnalyusis-D.

The ClusterAnaysis-D can be classified as either locally discretizing method or globally discretizing method. The methods that are characterized by operating only one attribute are called local method while the methods considering all attributes are called global methods. QingXiang Wu et.al proposed a new discretizing approach named Distribution-Index-Based Discretizer(DIBD-D) in [14]. The DIBD-D approach is based on the definition of dichotomic entropy and a compound distribution index. This criterion is applied to discretize continuous attributes adaptively. The DIBD-D can discretize any continuous attribute adaptively according to the simple adaptive rules. The adaptive rule is based on maximal compound decrement and minimal dichotomic entropy.

4. Experimental Set-Up

In this section, we conduct experiments to evaluate the performances of the associative classification systems. For the comparative performance analysis of the selective associative classifiers, we exploited the implementations of these algorithms included in Knowledge Extraction based on Evolutionary Learning (KEEL) [18]. The overview of the Data Mining and machine learning tool KEEL is given in the following section. In this section, we describe the datasets used for the comparative analysis of the associative classifier (CBA) and CBA2 in terms of accuracy and error rate by using the different discretizing methods. The parameters set for the experiments and the experiment graph designed for these experiments in the KEEL tool are described in this section.

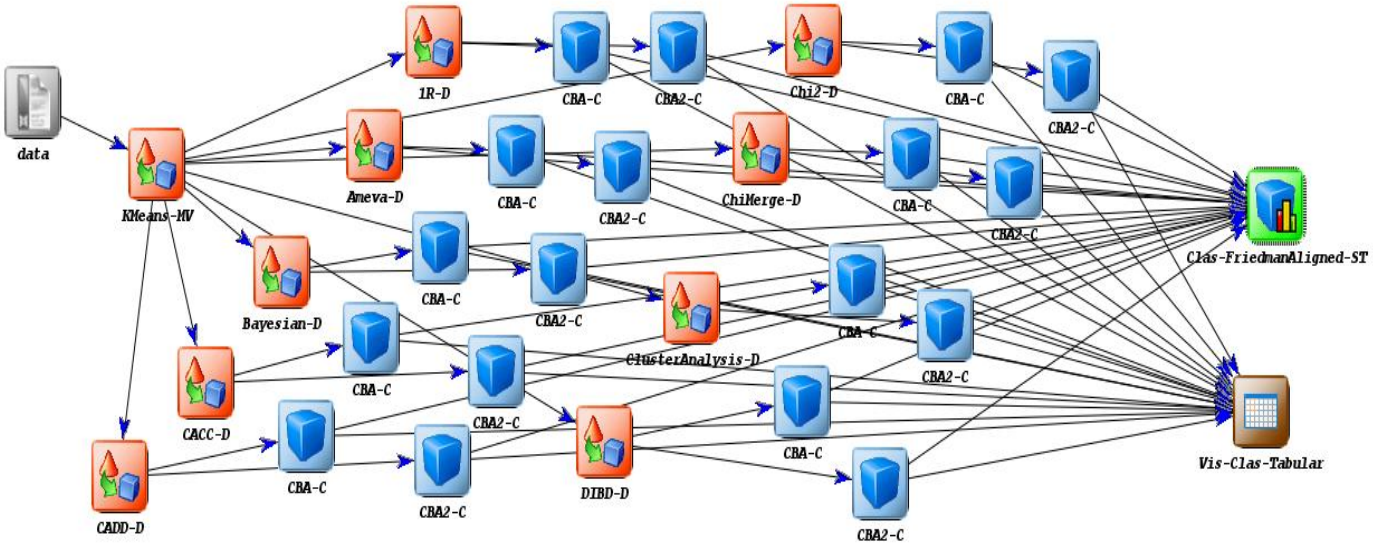
4.1. Data Sets

The description of datasets used for the comparative performance analysis of the selective Associative Classifiers under this study is given in Table2. The number of attributes (#Attributes), number of instances in the database (#Examples) and thenumber of classes (#Classes) are shown in the table. The missing values (Missing_V) in the dataset are representing by “Yes” (missing values present)/ “No” (missing values not present). The missing values of the datasets are imputed with the KMean-MV module implemented in KEEL. The datasets are discretized with the Ameva-D module included in KEEL as the associative classifiers accept the discretized form of datasets. We use the 10-fold cross-validation model for the datasets provided in KEEL. Table I summarizes the main characteristics of the 12 datasets which are given at Knowledge Extraction based on Evolutionary Learning (KEEL)-dataset repository[18].

Table 1 Data Sets Considered For The Experimental Stu

Dataset Name	#Attributes	#Examples	#Classes	Missing_V
Bupa	6	345	2	No
Cleveland	13	297	5	yes
Ecoli	7	336	8	No
Glass	9	214	7	No
Harberman	3	306	2	No
Iris	4	150	3	No
Monks	6	432	2	No
Newthyroid	5	215	3	No
Pima	8	768	2	No
Vehicle	18	846	4	No
Wine	13	178	3	No
Wisconsin	9	683	2	Yes

Figure 1 Experiment Graph Generated in KEEL



4.2. Experiment Graph

The experiment graph shows the components of the experiment and describes the relationships between them. The experimental graph of the comparative study is given in Figure1. The first component of the experimental graph is data which enables to select the datasets given in the KEEL Tool as well as to load user datasets. In our study, we selected standard KEEL datasets. The second component of the graph is KMeans-MV which is a module to impute the missing values in the database. The third component of the experiment graph is amodule for data discretization. In our case, we use the selective nine modules i.e.1R-D, Ameva-D, Bayesian-D, CACC-D, CADD-D, DIBD-D, ClusterAnalysis-D, ChiMerge-D and Chi2-D for the discretization of continuous data values. The fourth stage of the experiment graph is Associative Classification methods (CBA and CBA2) which are the focus of our study. The last stage of the experiment graph is the modules for the representation of the results of theclassifier and astatistical module for the analysis of the results produced by the algorithms used in the experiment. The module Vis-Class-Tabular provides the facility of representation of results of multiple classification methods in the form tabular representation.

4.3. Parameters of the Methods

The parameters of the associative classifiers (CBA and CBA2) under the focus of this comparative study are shown in the Tabel.2. The parameters of the method are selected according to the recommendation of the corresponding authors which are the default parameters settings included in the KEEL software tool [18]. In the Table 2,*Minsup* stands for minimum support, *Minconf* for minimum confidence, and *RuleLimit* for maximum candidate rules limit in the corresponding method.

Table 2 Parameters of the Methods for Experiment

Methods	Parameters
CBA-C	*Minsup = 0.01, Minconf = 0.5, Pruned = yes, RuleLimit = 80,000*
CBA2-C	*Minsup = 0.01, Minconf = 0.5, Pruned = yes, RuleLimit = 80,000*

5. Experimental Results

Table3 shows the comparative performance of the selected Associative Classifiers. We use the implementation of the corresponding algorithms in KEEL 3.0 for our comparative performance analysis of various discretizing methods by using CBA and CBA2 Associative Classifiers on the public datasets. We used the selective 9 discretization methods i.e 1R-D, Ameva-D, Bayesian-D, CACC-D, CADD-D, DIBD-D, ClusterAnalysis-D, ChiMerge-D and Chi2-D for this comparative study. The performance of both Associative Classifiers CBA and CBA2 is investigated on 12 public datasets for 9 discretizing methods. In table 3, the fold face red values show the best performance of CBA classifier for a specific dataset at a particular discretizing method with respect to other discretizing methods under the focus of this study. The fold face blue values in table 3 represents the best performance of the CBA2 classifier on a specific dataset with the exploitation of a specific discretizing method with respect to other discretizing methods. The last row of table 3 shows the average performance of the corresponding classifier for all datasets for a specific discretizing method. Overall the best average performance of classifier for a specific increasing method is presented with bold face red and blue for CBA and CBA2 respectively in table 3.The performance of the CBA classifier is better on Ameva-D discretizing method as compare to other discretizing methods but the average performance is better on the ChiMerge-D. The performance of CBA2 is promising with ChiMerge-D discretizing method. The performance of each discretizing method varies with respect to change in the database. The values bold face shows the wining of the corresponding discretizing methods for the CBA and CBA2 on the corresponding datasets.

Table 3 The Comparative Performance Results Of Associative Classification Methods

	1R-D		Ameva-D		Bayesian-D		CACC-D		CADD-D		DIBD-D		ClusterAnalysis-D		ChiMerge-D		Chi2-D	
Datasets	**CBA**	**CBA2**	**CBA**	**CBA2**	**CBA**	**CBA2**	**CBA**	**CBA2**	**CBA**	**CBA2**	**CBA**	**CBA2**	**CBA**	**CBA2**	**CBA**	**CBA2**	**CBA**	**CBA2**
Bupa	59.3746	59.2388	**66.9423**	**67.3425**	64.2971	60.2366	66.2854	64.4565	57.8924	57.8242	65.6888	65.6911	61.1710	62.6642	62.2117	62.2369	63.2472	65.0482
Cleveland	51.4624	45.3568	54.3978	53.7048	**58.0968**	47.1554	54.1613	54.0469	52.8172	48.6315	54.4409	53.7341	54.8065	50.7429	52.0968	**54.3304**	54.7957	53.1574
Ecoli	72.6114	67.8334	**78.0214**	72.1925	72.6292	66.9989	77.3886	**76.7704**	49.1176	48.3795	76.4973	76.4868	71.4082	66.5046	75.9002	72.1682	77.1658	74.6070
Glass	59.0015	57.9856	50.3569	49.7675	54.2565	54.8018	35.5661	35.4949	52.7213	53.1945	65.4916	60.8186	54.3821	52.6976	67.4854	**66.3038**	**70.3353**	65.6928
Haberman	74.1720	73.8807	**74.7742**	**74.1349**	71.5376	71.4663	74.1505	72.9814	73.5269	73.2942	73.4839	73.8416	73.1935	72.1017	73.4731	73.5386	72.2043	68.2111
Iris	91.3333	90.9091	92.6667	92.7273	93.3333	93.9394	92.6667	92.7273	74.0000	76.9697	91.3333	89.6970	88.0000	86.6667	92.0000	91.5152	**95.3333**	**95.7576**
Monks	52.6608	52.6255	**97.2674**	**97.5159**	51.2972	51.3858	**97.2674**	**97.5159**	52.6608	52.6255	58.5273	59.4050	80.6451	80.3385	**97.2674**	**97.5159**	**97.2674**	**97.5159**
New-Thyroid	**94.0043**	93.2900	93.9610	94.0968	92.5758	92.4242	93.0303	93.2507	87.3377	86.7965	92.1212	91.5978	93.9827	**94.5494**	93.0303	93.6836	92.1212	92.8571
Pima	73.1810	70.0638	71.2276	72.1981	**73.7160**	**73.3807**	72.3948	70.6555	65.1081	65.0924	72.4051	72.1998	69.6657	71.8346	72.7980	72.7916	72.9212	72.4235
Vehicle	67.3768	63.0685	**70.7955**	70.1350	68.4398	66.2605	68.5672	69.7059	56.9636	60.1401	64.7745	64.8752	62.5364	61.7571	67.0168	68.7395	68.4412	**70.8913**
Wine	91.4706	91.2359	94.9020	**94.8604**	93.8235	93.3749	93.7255	93.7908	80.3268	77.0648	92.0915	91.2953	**95.4575**	94.3553	93.8562	92.8699	82.5163	80.0654
Wisconsin	81.6936	82.0403	96.1366	95.9627	95.1304	95.3096	96.1366	95.9627	65.5238	65.6277	96.1387	95.7030	94.5652	94.6640	**96.4244**	**96.4860**	**96.4224**	96.3523
Average	**72.3619**	70.6274	**78.4541**	77.8865	**74.0944**	72.2278	**76.7784**	76.4466	**63.9997**	63.8034	**75.2495**	74.6121	**74.9845**	74.0730	**78.6300**	**78.5150**	**78.5643**	77.7150

Table 4 Win/Draw/Lose Record of Discretization Methods for CBA and CBA2 Classifiers

Discretizing Methods	**1R-D**		**Ameva-D**		**Bayesian-D**		**CACC-D**		**CADD-D**		**DIBD-D**		**ClusterAnalysis-D**		**ChiMerge-D**		**Chi2-D**	
Classifiers	**CBA**	**CBA2**	**CBA**	**CBA2**	**CBA**	**CBA2**	**CBA**	**CBA2**	**CBA**	**CBA2**	**CBA**	**CBA2**	**CBA**	**CBA2**	**CBA**	**CBA2**	**CBA**	**CBA2**
Win	1	0	**4**	**3**	2	1	0	1	0	0	0	0	1	1	0	**3**	2	2
Draw	0	0	**1**	**1**	0	0	1	1	0	0	0	0	0	0	2	**1**	2	1
Loss	11	12	**7**	**8**	10	11	11	10	12	12	12	12	11	11	10	**8**	8	9

Table 4 shows the comparative performance record of CBA and CBA2 in terms of Win/Draw/Lose of the Discretizing Methods. The 1R-D, Ameva-D, Bayesian-D, CACC-D, CADD-D, DIBD-D, ClusterAnalysis-D, ChiMerge-D, and Chi2-D. The performance of CBA classifier is better on Ameva-D and Chi2-D discretized. The performance of CBA2 remains same for Ameva-D and ChiMerge-D discretizing methods as shown in table 4.

6. Results Discussion

6.1. Comparative Performance Analysis in terms of Accuracy

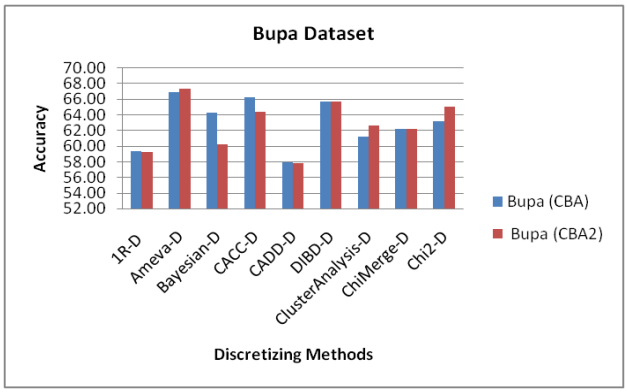
Figure2 shows the comparative performance of the CBA and CBA2 on different discretizing methods for the selective datasets. With the critical observation of the graphs in figure 2, the performance behavior of CBA and CBA2 is symmetric in terms of accuracy for various discretizing methods. The performance of Associative Classifiers (CBA and CBA2) significantly changes with the change of discretizing method as shown in figure 2. The performance variation of CBA and CBA2 at datasets Glass, Monks, Ecoli and Wisconsin at the usage of different discrete methods is given in figure2. At the dataset Glass in figure 2, there is significant variation in performance of Associative Classifiers (CBA and CBA2). The performances of CBA and CBA2 are highest for Chi-D and lowest for CACC-D discretizing methods. The experiments for the dataset Ecoli shows that the performances of CBA and CBA2 lower significantly at the usage of CADD-D discretized while remaining almost same for the other discretizing methods. For dataset Monks in figure 2, the performances of Classifiers are drastically decreasing for 1R-D, Bayesian-D and CADD-D discretizing methods. The performance of CBA2 significantly decreased on dataset Haberman for Chi2-D discretizing method. Finally, the experimental results at Wine and Wisconsin datasets, reveals that the performances of CBA and CBA2 significantly down for the CADD-D discretized in terms of accuracy.

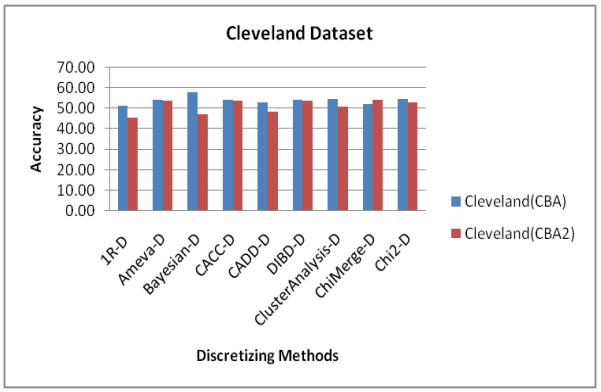
6.2. Comparative Performance Analysis in terms of Variance

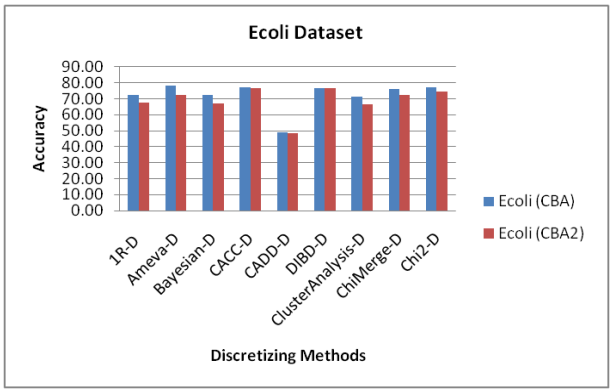
This subsection describes the impact of discretizing methods on the performance of Associative Classifiers i.e. CBA and CBA2 in terms of variance on public datasets. The variance depicts the consistency of the classification approach. The lower value of the variance indicates that the classifier is more consistent for the corresponding dataset. The discretizing methods producing a smaller value of variance for a specific dataset for the specific classification approach are more promising and provide more robustness for the classifier. The variation of performance of

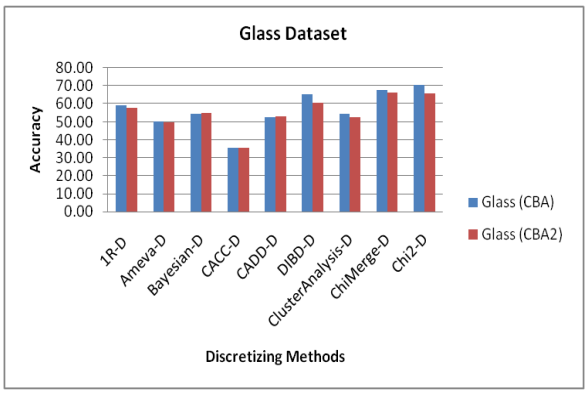
Associative Classifiers in terms of variance is elaborated in Figure 3. The performance of both classifiers is almost same for the datasets Bupa and Cleveland for on all the discretizing methods.Variance is significantly high for Chi2-D and Ameva-D discretizing methods for CBA2 and CBA classifiers on datasets Bupa and Cleveland respectively. The performance associative classifiers (CBA and CBA2) for Iris, New-thyroid, Wine and Wisconsin datasets is very promising and consisting in terms of variance for all the discretizing methods except CADD-D, Chi2 and 1R-D. The variance is very high of CBA and CBA2 on Iris and New-thyroid for CADD-D discretizing method.

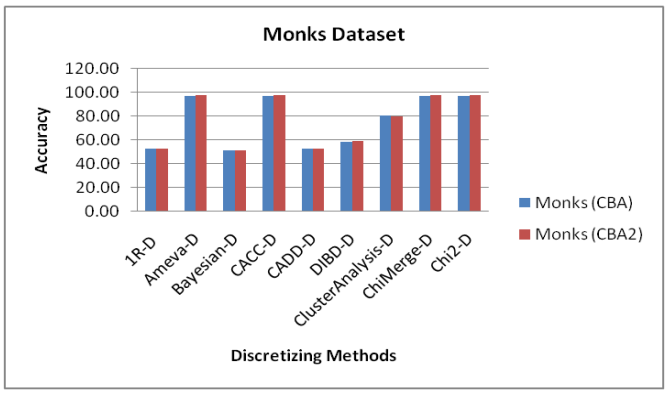
With the critical observation of Figure 3, it concluded there is a significant difference in the performance of associative classifiers in terms of variance as well as for the various discretizing methods on the pubic datasets under the focus of this study. Mostly of the discretizing methods produce promising results for some datasets while on the other datasets their performance is lower. The performance of classification approach significantly is dependent on the discetzing method used for the discretizing the continuous datasets in discrete form.

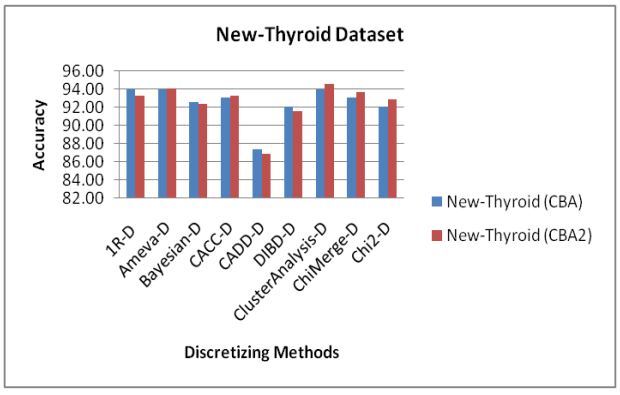












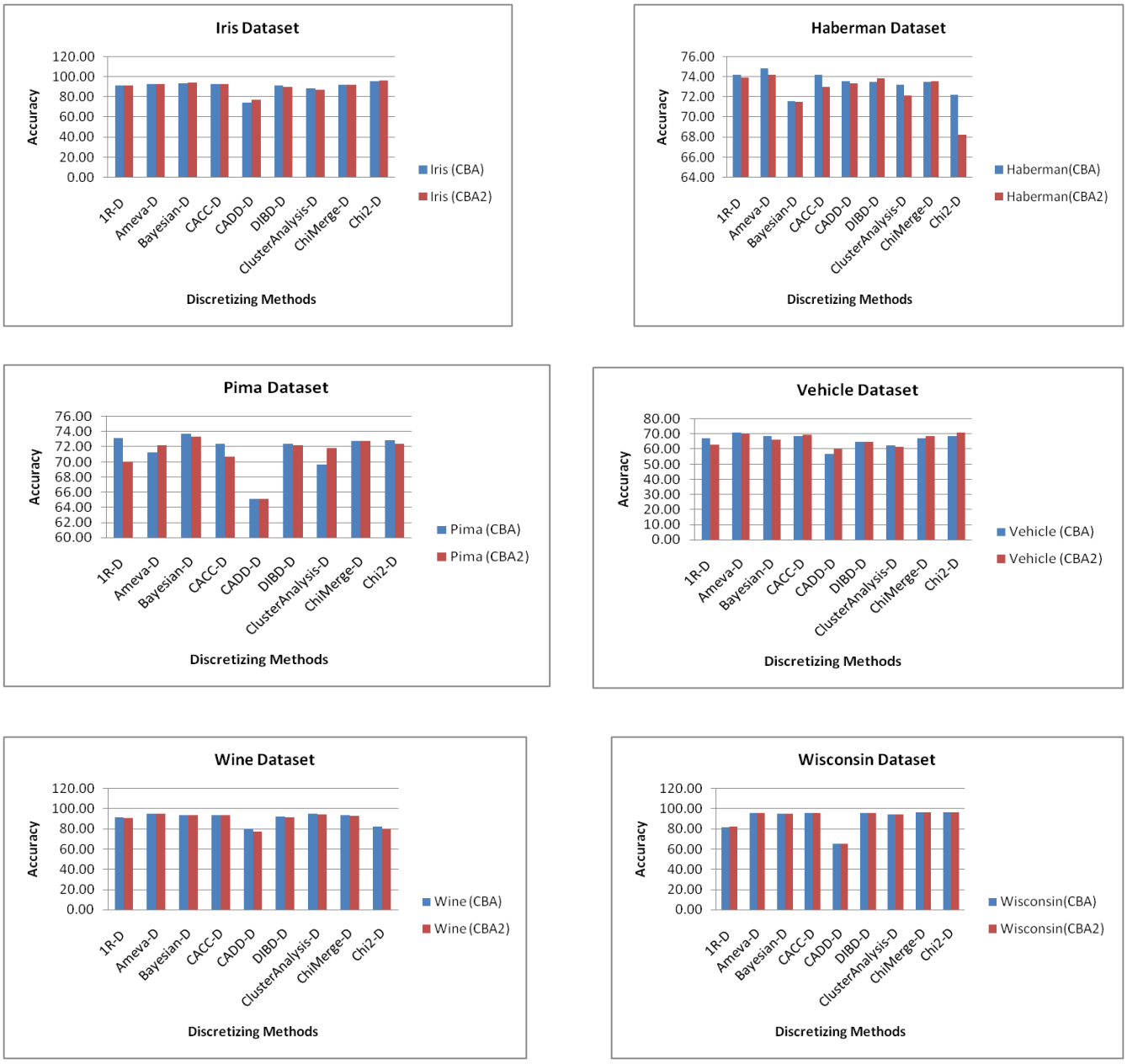
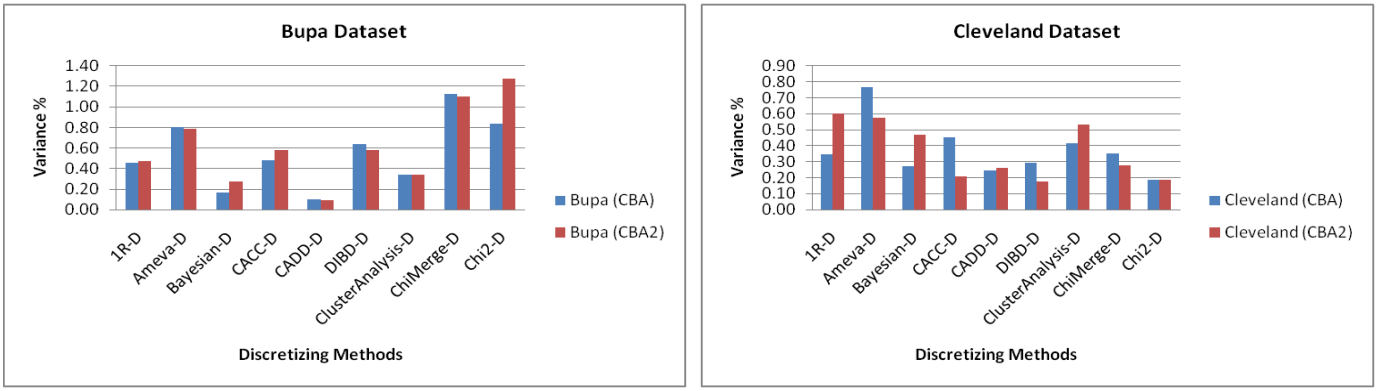
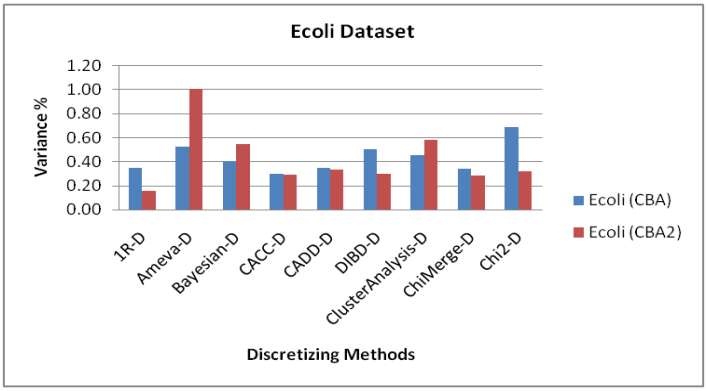
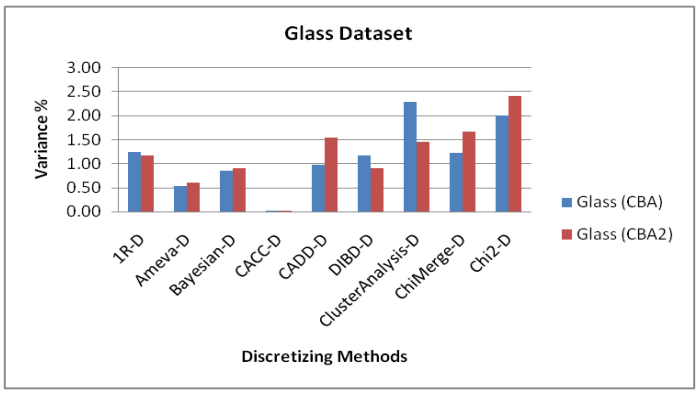
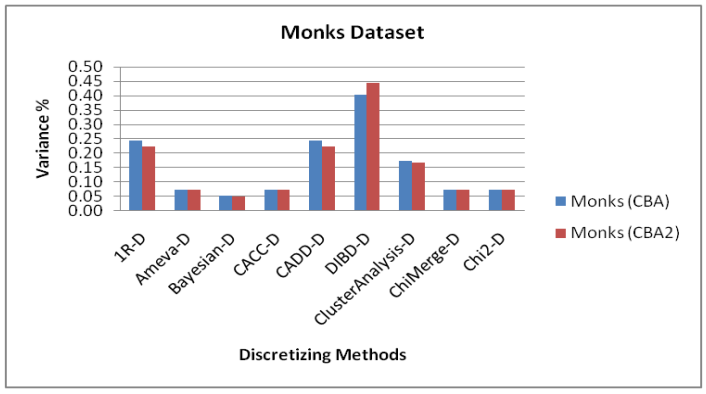


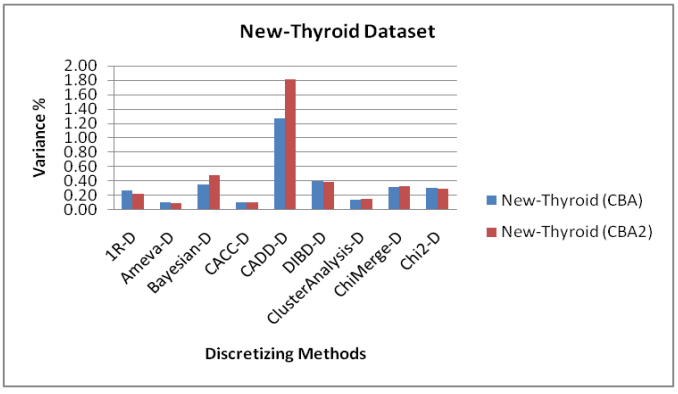
Figure 2 Impact of Discretization Methods on the Performance Associative Classifiers (CBA and CBA2) on Various Datasets

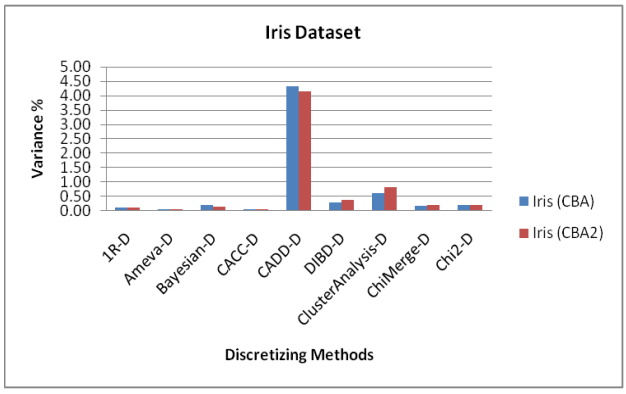


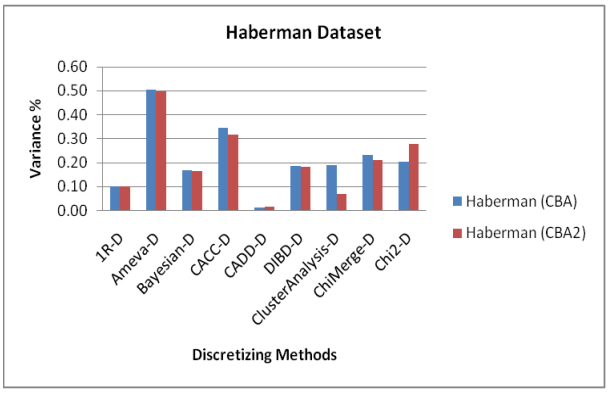

Ecoli Dataset
Variance %
1.20
1.00
0.80
0.60
0.40
0.20
0.00
1R-D
Ameva-D
Bayesian-D
CACC-D
CADD-D
DIBD-D
ClusterAnalysis-D
ChiMerge-D
Chi2-D
Ecoli (CBA)
Ecoli (CBA2)
Discretizing Methods

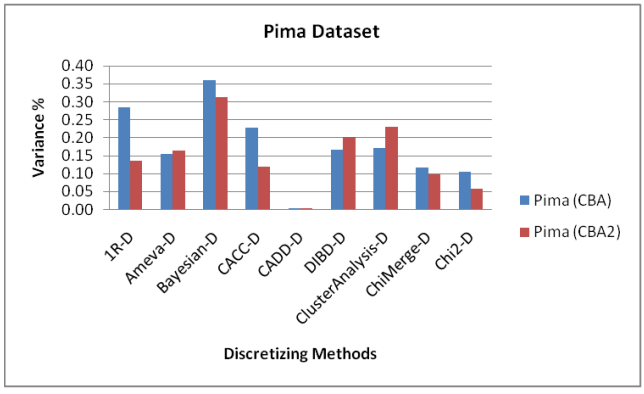

Glass Dataset
Variance %
3.00
2.50
2.00
1.50
1.00
0.50
0.00
1R-D
Ameva-D
Bayesian-D
CACC-D
CADD-D
DIBD-D
ClusterAnalysis-D
ChiMerge-D
Chi2-D
Glass (CBA)
Glass (CBA2)
Discretizing Methods

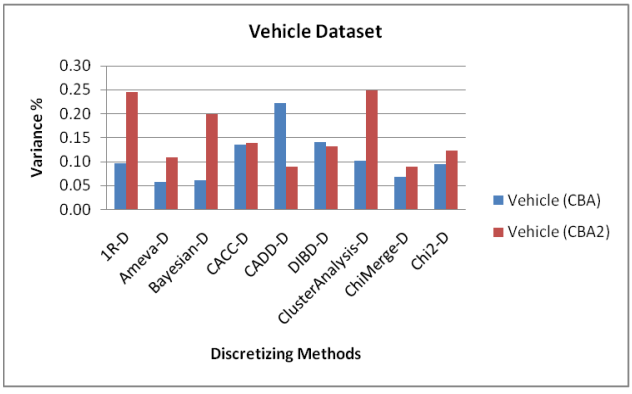

Monks Dataset
Variance %
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
1R-D
Ameva-D
Bayesian-D
CACC-D
CADD-D
DIBD-D
ClusterAnalysis-D
ChiMerge-D
Chi2-D
Monks (CBA)
Monks (CBA2)
Discretizing Methods


New-Thyroid Dataset
Variance %
2.00
1.80
1.60
1.40
1.20
1.00
0.80
0.60
0.40
0.20
0.00
1R-D
Ameva-D
Bayesian-D
CACC-D
CADD-D
DIBD-D
ClusterAnalysis-D
ChiMerge-D
Chi2-D
New-Thyroid (CBA)
New-Thyroid (CBA2)
Discretizing Methods


Iris Dataset
Variance %
5.00
4.50
4.00
3.50
3.00
2.50
2.00
1.50
1.00
0.50
0.00
1R-D
Ameva-D
Bayesian-D
CACC-D
CADD-D
DIBD-D
ClusterAnalysis-D
ChiMerge-D
Chi2-D
Iris (CBA)
Iris (CBA2)
Discretizing Methods


Haberman Dataset
Variance %
0.60
0.50
0.40
0.30
0.20
0.10
0.00
1R-D
Ameva-D
Bayesian-D
CACC-D
CADD-D
DIBD-D
ClusterAnalysis-D
ChiMerge-D
Chi2-D
Haberman (CBA)
Haberman (CBA2)
Discretizing Methods


Pima Dataset
Variance %
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
1R-D
Ameva-D
Bayesian-D
CACC-D
CADD-D
DIBD-D
ClusterAnalysis-D
ChiMerge-D
Chi2-D
Pima (CBA)
Pima (CBA2)
Discretizing Methods


Vehicle Dataset
Variance %
0.30
0.25
0.20
0.15
0.10
0.05
0.00
1R-D
Ameva-D
Bayesian-D
CACC-D
CADD-D
DIBD-D
ClusterAnalysis-D
ChiMerge-D
Chi2-D
Vehicle (CBA)
Vehicle (CBA2)
Discretizing Methods

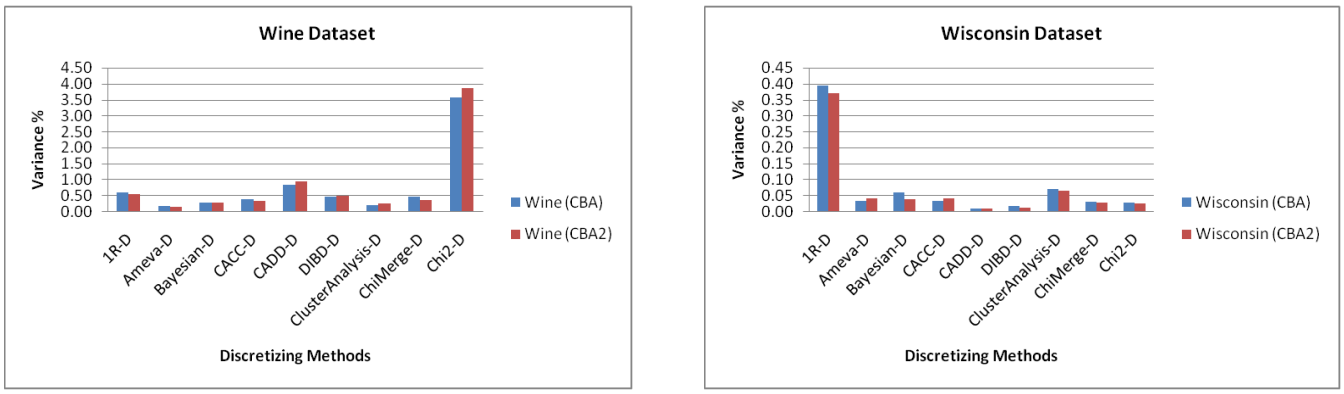


Figure 3 Comparative Impact of Discretizaing Methods on the Performance of Associative Classifiers (CBA and CBA2) in terms of Variance

7. Conclusion

Discretization algorithms have played a great role in the performance of classification techniques. We investigate the effect of discretization Methods on the Performance of Associative Classifiers. Most of the classification approaches work on the discretized databases. There are various approaches exploited for the discretization of the database to compare the performance of the classifiers. The selection of the discretization method influences the classification performance of the classification method. We compare the performance of associative classifiers namely CBA and CBA2 on the selective discretizing methods i.e. 1R-D, Ameva-D, Bayesian-D, CACC-D, CADD-D, DIBD-D, ClusterAnalysis-D, ChiMerge-D and Chi2-D in terms of accuracy and error rate. Our experimental results show that the performance of the Associative Classifiers significantly varies for different discretization methods for the same classifier. So the accuracy rate and variance in results of the classifier is highly dependent on the choice of the discretization method. For this comparative performance study, we use the implementation of these methods in KEEL data mining tool on public datasets.

In future, we will analyse the impact of discretizing methods for other classifiers by considering other parameters and to derive the significance of results by using statistical methods.

Conflict of Interest

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

The authors wish to thanks Higher Education Commission of Pakistan. This work is supported in part by a grant from Higher Education Commission of Pakistan.

References

[1] Ali, Z. and W. Shahzad. *Comparative Study of Discretization Methods on the Performance of Associative Classifiers*. in *Frontiers of Information Technology (FIT), 2016 International Conference on*. 2016. IEEE.

[2] Thabtah, F., *A review of associative classification mining.* The Knowledge Engineering Review, 2007. **22**(01): p. 37-65.

[3] Vyas, R., et al. *Associative classifiers for predictive analytics: Comparative performance study*. in *Computer Modeling and Simulation, 2008. EMS'08. Second UKSIM European Symposium on*. 2008. IEEE.

[4] Soni, S. and O. Vyas, *Using associative classifiers for predictive analysis in health care data mining.* International Journal of Computer Applications, 2010. **4**(5): p. 33-37.

[5] Liu, H., et al., *Discretization: An enabling technique.* Data mining and knowledge discovery, 2002. **6**(4): p. 393-423.

[6] Kotsiantis, S. and D. Kanellopoulos, *Discretization techniques: A recent survey.* GESTS International Transactions on Computer Science and Engineering, 2006. **32**(1): p. 47-58.

[7] Garcia, S., et al., *A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning.* IEEE Transactions on Knowledge and Data Engineering, 2013. **25**(4): p. 734-750.

[8] Pancho, D.P., et al. *Analyzing fuzzy association rules with Fingrams in KEEL*. in *2014 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. 2014. IEEE.

[9] Holte, R.C., *Very simple classification rules perform well on most commonly used datasets.* Machine learning, 1993. **11**(1): p. 63-90.

[10] Gonzalez-Abril, L., et al., *Ameva: An autonomous discretization algorithm.* Expert Systems with Applications, 2009. **36**(3): p. 5327-5332.

[11] Wu, X., *A Bayesian discretizer for real-valued attributes.* The Computer Journal, 1996. **39**(8): p. 688-691.

[12] Tsai, C.-J., C.-I. Lee, and W.-P. Yang, *A discretization algorithm based on class-attribute contingency coefficient.* Information Sciences, 2008. **178**(3): p. 714-731.

[13] Ching, J.Y., A.K.C. Wong, and K.C.C. Chan, *Class-dependent discretization for inductive learning from continuous and mixed-mode data.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 1995. **17**(7): p. 641-651.

[14] Wu, Q., et al., *A distribution-index-based discretizer for decision-making with symbolic ai approaches.* IEEE transactions on knowledge and data engineering, 2007. **19**(1): p. 17-28.

[15] Chmielewski, M.R. and J.W. Grzymala-Busse, *Global discretization of continuous attributes as preprocessing for machine learning.* International journal of approximate reasoning, 1996. **15**(4): p. 319-331.

[16] Kerber, R. *Chimerge: Discretization of numeric attributes*. in *Proceedings of the tenth national conference on Artificial intelligence*. 1992. Aaai Press.

[17] Liu, H. and R. Setiono, *Feature selection via discretization.* IEEE Transactions on knowledge and Data Engineering, 1997. **9**(4): p. 642-645.

[18] Alcalá, J., et al., *Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework.* Journal of Multiple-Valued Logic and Soft Computing, 2010. **17**(2-3): p. 255-287.

[19] Ma, B.L.W.H.Y. and B. Liu. *Integrating classification and association rule mining*. in *Proceedings of the 4th*. 1998.

[20] Agrawal, R. and R. Srikant. *Fast algorithms for mining association rules*. in *Proc. 20th int. conf. very large data bases, VLDB*. 1994.

[21] Liu, B., Y. Ma, and C.-K. Wong, *Classification using association rules: weaknesses and enhancements.* Data mining for scientific applications, 2001. **591**.

[22] Salzberg, S.L., *C4. 5: Programs for machine learning by j. ross quinlan. morgan kaufmann publishers, inc., 1993.* Machine Learning, 1994. **16**(3): p. 235-240.

[23] Kurgan, L.A. and K.J. Cios, *CAIM discretization algorithm.* IEEE transactions on Knowledge and Data Engineering, 2004. **16**(2): p. 145-153.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 855-864 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

A novel model for Time-Series Data Clustering Based on piecewise SVD and BIRCH for Stock Data Analysis on Hadoop Platform

Ibgtc Bowala[1], Mgnas Fernando[*, 2]

[1]*Undergraduate, University of Colombo school of Computing, 00700, Sri Lanka*

[2]*Senior Lecturer, University of Colombo school of Computing, 00700, Sri Lanka*

ARTICLE INFO

Article history:
Received: 21 April, 2017
Accepted: 31 May, 2017
Online: 20 June, 2017

Keywords:
Clustering
Time series analysis
SVD
BIRCH
Hadoop
MapReduce

ABSTRACT

With the rapid growth of financial markets, analyzers are paying more attention on predictions. Stock data are time series data, with huge amounts. Feasible solution for handling the increasing amount of data is to use a cluster for parallel processing, and Hadoop parallel computing platform is a typical representative. There are various statistical models for forecasting time series data, but accurate clusters are a pre-requirement. Clustering analysis for time series data is one of the main methods for mining time series data for many other analysis processes. However, general clustering algorithms cannot perform clustering for time series data because series data has a special structure and a high dimensionality has highly co-related values due to high noise level. A novel model for time series clustering is presented using BIRCH, based on piecewise SVD, leading to a novel dimension reduction approach. Highly co-related features are handled using SVD with a novel approach for dimensionality reduction in order to keep co-related behavior optimal and then use BIRCH for clustering. The algorithm is a novel model that can handle massive time series data. Finally, this new model is successfully applied to real stock time series data of Yahoo finance with satisfactory results.

1. Introduction

A stock market, equity market or share market is the aggregation of buyers and sellers of stocks, which enables the trading of company stocks collective shares [1]. With the rapid development of trades in the world, people are paying more attention on investing in Stock markets [1]. By reviewing the stock's financial curves, an educated decision can make whether the company is stable, growing and has an improving future. Thus, it is necessary to find a way to identify stocks with similar trend curves, which is very difficult due to uncertainty [2]. Financial institutes such as stock markets produce massive datasets [1]. Large amount of data is a barrier to analyze and summarize stock market data.

To visualize stock market behavior, researchers have used data mining techniques such as decision tree [1,3], neural network [1,4], association rules [5], factor analysis [6], etc. The decision trees are a powerful beginning step, but very costly. Neural networks (NN) such as Self-Organizing Feature Maps (SOFM) have been effectively applied in a many previous approaches. However, using decision trees or NNs for cluster large data sets cause performance degradation. Association rule is a popular and well-studied method for discovering interesting relations among variables in huge databases. But, researchers have also shown that it can produce better index return only with fewer trades. Factor analysis is an important step towards effectual clustering. But it can use only a limited number of stocks and, can only find fewer relations like the best stock. The overall analysis for large number of stocks cannot be achieved using factor analysis.

In statistics, signal processing and many other fields, a time series is a sequence of data points measured typically at successive [7] uniform times [1]. Stock data, being time series data, shares a common set of analysis problems with other time-series data. Time-series analysis includes methods that attempt to understand time series, to make forecasts [1]. Time-series data are often large

[*] Mgnas Fernando, Senior Lecturer, University of Colombo school of Computing, 00700, Sri Lanka | Email: nas@ucsc.cmb.ac.lk

and might contain outliers. In addition, time series are a special type of data set where elements have a temporal order [8] and time-series data are essentially high-dimensional data [7]. Mining high-dimensional data involves handling a range of challenges such as curse of dimensionality, the meaningfulness of the similarity measure in the high-dimensional space, and handling outliers. Furthermore, time series analysis requires multiple arrays to represent the series, which make it computationally very expensive. Therefore, an effective dimensionality-reduction method is needed to reduce memory consumption and to fit matrix into memory and it should capture the temporal order and the highly-correlated features of time series [7]. Today, dimension reduction [7,8,9] is a highly-attractive research area and researchers focus on new techniques for dimension reduction [10,11,12] because it affects both the accuracy and the efficiency. The stock data only differs from other time-series data in terms of the data distribution. Therefore, our primary focus of research is time-series analysis.

Clustering, which is another tool for data analysis [3,13], provides the basis for most of the data analysis, decision making, designing, and forecasting problems. Thus, it is very important to achieve accurate clusters. But, due to the special structure of time-series data with high level of noise, building an effective model for clustering needs a huge effort and time. Furthermore, building an effective model for time series clustering cannot be achieved without a comprehensive study of theories and previous approaches.

Moreover, due to the volatility of stock data, which is not directly recognizable [2,14], and due to the type of distribution of stock data, the clustering becomes harder. Unlike financial return series, price series is harder to handle, because it has more attractive statistical properties [2]. Even for a stock return series, the volatility is not directly recognizable and it becomes harder as far as the price curves are concerned. Thus, the financial analysis sector has a thirsty of identifying relationships between original price curves, rather than trend and other curves.

Many stock analysis researchers and companies are searching for the right methods to cluster stock data in order to perform their analysis. This paper introduces a novel model for time-series clustering, including a new dimensionality-reduction approach. This can be taken as a case study for time-series clustering big-data fields. Comparative decisions made in the noise removal stage enhanced the clustering quality. The novel dimensionality-reduction approach is suitable for time-series data, for stock data, and for other huge data sets. The novel model is applied successfully on real stock data of Yahoo finance to evaluate the accuracy and performances. The cluster evaluation shows that, this can cluster stock price curves very effectively and efficiently. The next section will discuss the related work.

2. Related Work

A time series is "a sequence $X = (x_1, x_2, \dots, x_m)$ of observed data over time", where $i = 1, 2, \dots, m$ are time units and m is the number of observations [7]. It has a temporal order [8], often large and might contain outliers and essentially high-dimensional data [7]. The time series X can be considered as a point in n-dimensional space. This suggests that time series could be clustered using clustering methods. Time series tend to contain highly-correlated features. Thus, time-series data are usually good applicants for dimensionality reduction [9].

Let P_t be the price of an asset at time index$t$. Thus, the One-Period Simple Return is holding the asset for one period from date $t-1$ to date t would result in a simple gross return of $R_t = \frac{P_t - P_{t-1}}{P_{t-1}}$ [2]. Stock return series data follows a normal distribution of $N(0, \sigma^2)$ [14], where mean $\mu = 0$ and σ^2 is the variance.

In order to reduce defects in the clustering stage, data standardizing is needed [13]. Calculating the mean absolute deviation and calculating the standardized measurement (z-score) are some methods. Z-score is defined as $Z = \frac{x - \mu}{\sigma}$, where μ is the mean and σ is the standard deviation of the sequence.

Some of the dimensionality reduction techniques researchers which have been using for time-series data are Singular Value Decomposition (SVD) [7,9,14,15], the Discrete Fourier transforms (DFT) [7,9,16], the Discrete Wavelets Transform (DWT) [7,9,17], Piecewise Aggregate Approximation (PAA) [7,9], Principal component analysis (PCA) [12,18], Factor analysis (FA) [12,19], Adaptive Piecewise Constant Approximation (APCA) [7], Piecewise Linear Approximation (PLA) [7], Independent component analysis (ICA), Chebyshev Polynomials (CHEB) [7], etc. Researchers have been using wavelets for dimension reduction [7,9,17] but, its only defined for sequences with length which are an integral power of two [9]. Therefore, this method cannot be used for time series processing with various lengths, which is a very huge limitation. PAA ignores the co-related behavior of time-series data. Thus, PAA is not a good solution to use for dimension reduction of time-series data.

Singular Value Decomposition (SVD) had successfully been used [20], for time-series indexing [21]. Singular value is a good feature of a matrix and is suitable when data follow a normal distribution [14]. SVD is a global transformation, which is a weakness from the point of large data sets and strength from an indexing point of view. Additionally, the insertions to the clusters already have required re-computing SVD for the entire dataset. In order to eliminate these drawbacks, this research introduces a new extension of SVD to perform dimension reduction without these drawbacks.

There are two categories of sequence matching methods, named, Whole Matching and subsequence Matching [7,9]. Whole matching needs comparing the query sequence to every candidate sequence. This can be reached by evaluating the distance function [9].

When handling time series, the similarity between two time-series sequences of the same length can be calculated by summing the ordered point-to-point distance among them [7]. One of the widely used distance function is the Euclidean Distance [7] which is a good "gold standard" used to compare different approaches [9].

Previous probability based clustering approaches do not adequately consider the case that the dataset can be too large to fit in main memory [22]. They do not recognize that the problem must be viewed in terms of limited resources such as keeping the I/O costs low [22]. Using probability based approaches for time-series data clustering is problematic due to co-related features

[8,23]. Distance based approaches like k-means assume that all the data points are given in advance and can be scanned frequently [22], which is not true in the case of large data sets.

Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) has some good characteristics compared to the requirements. It is suitable for very large data sets, because it makes the time and memory constraints explicit [22]. The clustering decision is made without scanning all the data points [22], and in clustering, outliers can be removed optionally [22]. It ensures accuracy by fully utilizing the available memory to derive the finest possible sub clusters and ensures efficiency by minimizing the I/O costs which cause a linear running time [22]. It only scans the dataset once [22].

BIRCH uses a clustering feature (CF) which is a triple, summarizing the information that we maintain about a cluster. Then, CF vector of the cluster that is formed by merging the two disjoint clusters can be calculated using CF additive theorem. The CF vectors can be stored easily and can be calculated easily and accurately when merging. Storing the CF vector is sufficient for making the clustering decisions. Thus, BIRCH only stores this CF vector [22]. Thus, BIRCH is suitable for a data set with lot of entries, such as huge number of stocks. CF Tree is a height balance tree, which stores CF entries of clusters [22]. The CF tree can build dynamically, while inserting data. Thus, BIRCH is very suitable for real-time, time-series big data clustering like in stock markets. Thus, future researchers can extend this study easily to real-time clustering, because, an updating is just like B+- tree [22]. The CF tree is a very compact representation of the dataset [22]. Thus, BIRCH is suitable for huge number of time-series sequences.

In order to achieve a considerable efficiency level, parallel processing is needed. Since stock data falls into the big data category, parallel processing technique for big data is needed. According to K. Gutfreund [24], functional programming ideas and message passing techniques are intrinsic to MapReduce. Thus, we can achieve parallel processing using an implementation of MapReduce.

Time series consists of four components [25]: Trend component, Seasonal variation, Cyclic component, and Irregular fluctuations. An upward or downward movement is known as a 'trend'. A trend is the price that is continuing to move towards a certain direction. Moving-average lines are used to help a trader to identify the direction of the trend more easily. The simple moving average is formed by computing the average price of a security over a specific number of periods. This time period can be selected as 10, 20, 50, 100 or 200. The next section will discuss the design stage.

3. Design

The main objective of this research is to study a novel model for time-series data clustering, which is suitable for stock market data. It is assumed that the stock data follows a normal distribution as explored in the literature.

Most of the financial studies involve in returns. When P_t indicates the closing price of day$t$, and P_{t-1} indicates the closing price of day$t-1$, one can calculate the stock-returns series using:

$$R_t = \frac{P_t - P_{t-1}}{P_{t-1}} \text{ for } t > 1 \text{ and } 0 \text{ when } t = 1 \quad (1)$$

Sometimes, time-series data usually contains missed values. Let's assume that day t is missing. Then, in order to handle missed values, we use:

$$R_t = \frac{R_{t-1} + R_{t+1}}{2} \quad (2)$$

Stock-returns data series follows a normal distribution. Standardization or z-scores converts all indicators in to a common scale with an average of zero and standard deviation of one. Therefore, the z-score normalization is used.

$$Z = \frac{R_t - \mu}{\sigma} \quad (3)$$

$$\mu = {}^{1}/_{n} \sum_{t=1}^{n} R_t \quad (4)$$

$$\sigma = \sqrt{\frac{\sum_{t=1}^{n}(R_t - \mu)^2}{n}} \quad (5)$$

4. Piece wise Singular value decomposition (SVD)

The time series are usually good candidates for dimensionality reduction because they contain highly-correlated features [9]. To represent sequences, matrix is required. Singular value is a good feature of a matrix. It is feasible to use the SVD, when data follows a normal distribution [14] and also suitable for representing neighborhood. Therefore, the SVD is selected in this design. But, in order to reduce drawbacks of SVD, this paper introduces a novel approach to accomplish dimensionality reduction. This method is motivated by the simple observation that most of the time-series datasets can be approximated by segmenting the sequences into equal length sections and then recording the SVD of these sections, similar to the vectoring process used in PAA with mean value [9]. We can efficiently represent a "neighborhood" of data points with the reduced SVD value. These SVD values can be indexed efficiently in a lower dimensional space. This method can calculate SVD locally, and this new piecewise SVD removes the necessary of re-computing SVD for the entire dataset, when inserting to the clusters that we already have. Let us denote a time series query as $X = x_1, \dots, x_n$ and the set of time series of the dataset as $Y = \{Y_1, \dots, Y_K\}$. Let us assume that each sequence in Y is n units long (Previously, used filling value techniques for achieving this). Let N be the dimensionality of the transformed space that we wish to index($1 \leq N \leq n$). N may or may not be a factor of n. N being a factor of n is not a requirement of this approach. A time series X of length n can be represented in the N space by a vector $\bar{X} = \overline{x_1}, \dots, \overline{x_N}$. Let us take a nonnegative real number σ_i as the singular value for the matrix, generated by the vector$\bar{x}_i$. Then, the new reduced time series can be represented as $W = \sigma_1, \dots, \sigma_N$. The ith element of W is calculated by the following equation:

$$\sigma_i = SVD_{\left\lfloor j=\frac{n}{N}(i-1)+1 \right\rfloor}^{\left\lceil \frac{n}{N} i \right\rceil} x_j \quad (6)$$

In simple terms, to reduce the time series from n dimensions to N dimensions, the data is divided into N equal sized "frames". The SVD of the data falling within a frame is calculated and that is a nonnegative real number. The sequence of these σ_i values becomes the reduced time series representation.

5. Time series matching technique

In order to use with SVD, whole matching is selected. Here, since we are using the whole matching technique, as the previous original series, time series $X = x_1, \dots, x_n$ needs to be compared with all the other series in $Y = \{Y_1, \dots, Y_K\}$. After dimension reduction, the time series X has been reduced to W which has N dimensions. Furthermore, all the other time series in the data set $Y = \{Y_1, \dots, Y_K\}$ has been reduced to set of series with N dimensions. Let us take all the other series in the reduced data set as $Y_N = \{Y_{1N}, \dots, Y_{KN}\}$ and $Y_{iN} = \sigma_1', \dots, \sigma_N'$. Thus, to perform whole matching, one has to compare the reduced time series $W = \sigma_1, \dots, \sigma_N$ with the other time series in the data set $Y_N = \{Y_{1N}, \dots, Y_{KN}\}$. This comparison accomplished with measuring the distance as stated in the next section (III).

6. Distance Measure

For certain applications the Euclidean distance measure can produce notions of similarity which are very unintuitive. Thus, as a prototype, we use Euclidean distance measure. The Euclidean distance is a good "gold standard" [9] and BIRCH also supports Euclidean distance.

$$D(X,Y) = D(Z,Y_N) \equiv \sqrt{\sum_{i=1}^{N} (\sigma_i - \sigma_i')^2} \qquad (7)$$

7. Time series clustering

In this research study, we propose clustering to identify dense and sparse regions in set of stock time series and, to discover overall distribution patterns and interesting correlations among stock time series [13]. Clustering can also be used for outlier detection. Alternatively, the proposed clustering method may serve as a preprocessing step for other algorithms, such as characterization; attribute subset selection, and classification, which would then operate on the detected clusters. Clustering can reduce a lot of time in selection of stocks as stocks of similar categories [26].

A. Requirements of clustering:

First, it is required to identify the requirements of the proposed clustering algorithm. Since stock data are time-series data, be able to cluster time series data. Number of stocks is huge and volume of stocks is also large. Therefore, we need to find a method for efficient and effective cluster analysis of large datasets. Ability to handle outliers, minimal requirements for domain knowledge to determine input parameters, and ability to incorporate with newly inserted data into existing clusters are the requirements.

B. Clustering algorithm – BIRCH

As described, BIRCH only stores CF vectors. CF vectors can be stored easily and can be calculated easily and accurately when merging. Storing CF vector is sufficient for clustering decisions [22]. Thus, BIRCH is good for a data set with lot of entries, such as a huge number of stocks. Furthermore, each entry in a leaf node is a sub-cluster, and not a single data point. Thus, the CF tree is a very compact representation of the dataset [22]. Therefore, BIRCH is suitable for a dataset with huge number of time series sequences.

Furthermore, the CF tree can be built dynamically, while inserting data. Therefore, BIRCH is very suitable for real-time, time-series big data clustering like in stock markets. Thus, future researchers can extend this research study to real time clustering [22]. Therefore, BIRCH which is used in this research is to give an efficient and accurate solution.

8. Map Reduce

In order to give an efficient solution, parallel data processing technique is used. Therefore, Java Application Program Interface (API) of Hadoop MapReduce is selected for the implementation. Next section will discuss the implementation stage.

9. Implementation

The implementation is divided into three main phases. The first phase involves in processing the time series, generating returns and data normalization. The second phase is to generate the reduced series. Clustering using BIRCH is the last phase. This phase generates clusters with the same trend, which can be used for further clustering using another method or can be used for further analysis process. Furthermore, clusters with single stock can be used for stock fraud detection.

Stock data from yahoo finance, during 1st of January 2000 and 30th of November 2016 were selected. Thus, Number of records per company is 6179 and Number of companies used is 9000. The data is in 'metastock' format.

Apache Hadoop MapReduce is used to achieve the parallelism of processing. Apache Hadoop Java API, Hadoop libraries, and Apache Mahout Libraries were used in the implementation. Next sections will describe the steps of the implementation.

First, filled missing values and then normalization process were implemented using java. Then as described in earlier sections, 'piecewise SVD' was implemented. Due to the local behavior of new piecewise SVD, it can effectively be used for parallel processing. This piecewise SVD removes the necessity of re-computing SVD for the entire dataset, when inserting the clusters that we already have. Thus, we only need to calculate the reduced series for the current insertion series. Then the clustering process was implemented. In order to accomplish the implementation of BIRCH, we modified the 'jbirch' implementation of 'Roberto Perdisci'[1] according to our requirements. Next section will focus on evaluation and results.

10. Evaluation And Results

This section evaluates the difference phases of the proposed approach based on comparisons, time complexity and visual analysis of results.

First, we focus on the novel dimensionality-reduction approach and time complexity of it. Let us take an example time series: closing prices of stock Aalberts Industries (AALB.AS). This series contains 4401 values and is shown in Figure 1. (Using the chart from Yahoo Finance to visualize easily)

[1] https://github.com/perdisci/jbirch

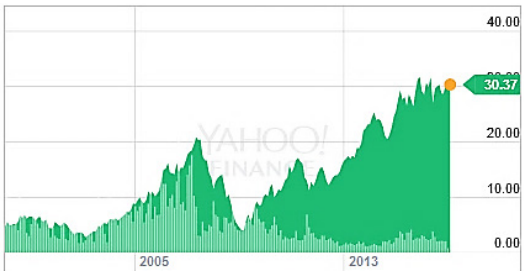


Figure 1- Closing prices of stock Aalberts Industr (AALB.AS)

When using piecewise SVD and number of dimensions as 8, this time series is reduced to:

[532.6446625096315, 416.8873393376198, 930.5871525547732, 1102.8832536583386, 206.39692810698506, 325.16997582033906, 472.1037659773537, 663.7071469029098] and is shown in Figure 2.

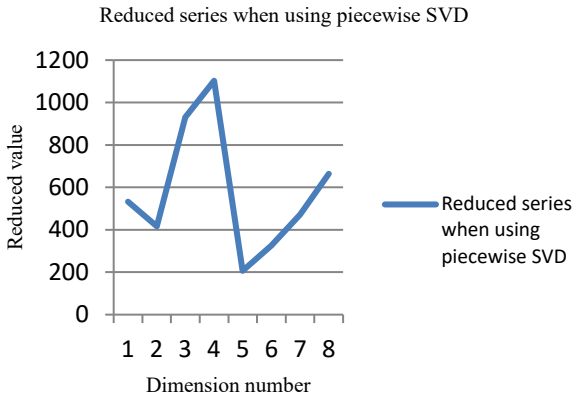


Figure 2- Closing prices of AALB.AS, when using piecewise SVD and 8 dimensions

In order to do a comparison, particular original series processed using piecewise aggregate approximation (PAA) with 8 dimensions. Results are:

[22.53275164735288, 17.36341740513517, 38.1274619404681, 39.638409452397134, 8.413749602363096, 13.765098841172463, 19.827730061349694, 28.209734151329258] and shown in Figure 3.

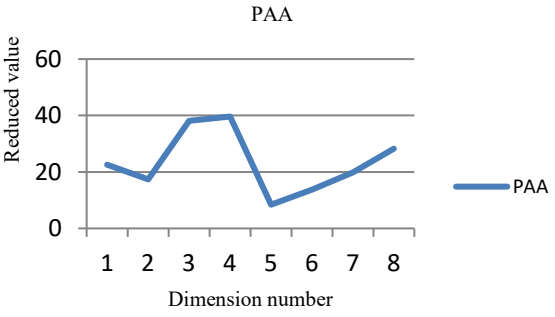


Figure 3- Closing prices of AALB.AS, when using PAA and 8 dimensions

When comparing two graphs (Figure 2 and Figure 3), both have much similar shapes.

Let us take the same time series: closing prices of stock AALB.AS. When using piecewise SVD and number of dimensions as 20, this time series has reduced to:

[332.3925127014745, 364.7341703213453, 290.44596055032366, 203.1714561644916, 297.2408049040378, 417.95588810303883, 591.8952354935798, 846.2187567644681, 891.780199993249, 214.3798407033649, 111.96008132365769, 125.80219373683428, 182.04229624732812, 216.06621785924793, 200.37451241363, 241.1271972942913, 331.1352621663843, 349.7118797667588, 426.1323476339242, 437.8459535327464] and shown in Figure 4.

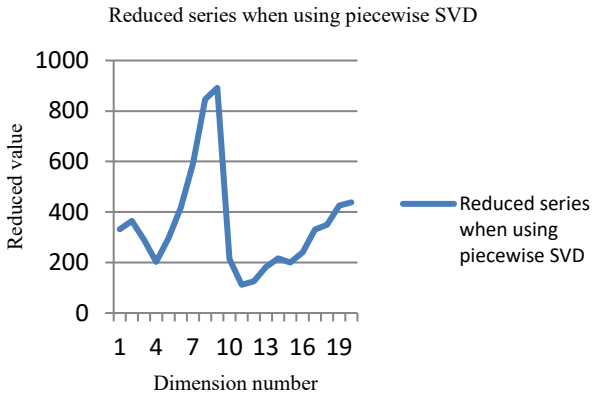


Figure 4- Closing prices of AALB.AS, when using piecewise SVD and 20 dimensions

In order to do a comparison, particular original series processed using PAA with 20 dimensions. Results are:

[22.284253578732113, 24.53078845716883, 19.513928652578947, 13.629266075891842, 19.782185866848458, 27.921654169506937, 39.799772778913876, 56.72833446943877, 55.645944103612784, 14.31897296069074, 7.061522381276979, 8.299890933878666, 12.191888207225627, 14.44708020904339, 13.446739377414218, 16.193206089525106, 22.225585094296747, 23.433946830265832, 28.684389911383786, 29.556532606225836] and shown in Figure 5.

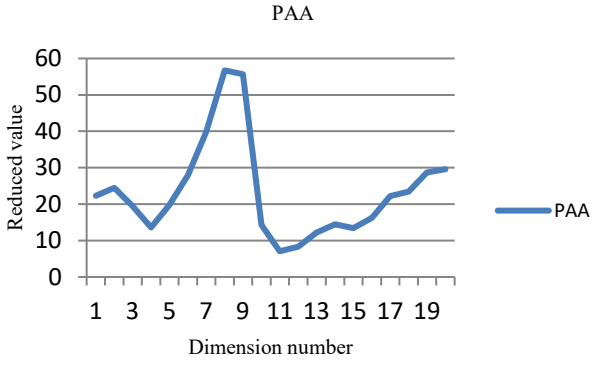


Figure 5- Closing prices of AALB.AS, when using PAA and 20 dimensions

When comparing two graphs (Figure 4 and Figure 5), both have much similar shapes but, piecewise SVD has a better gap between values with dimension number. Thus, this helps to give more accurate clusters in the clustering process. Further, when comparing Figure 3 and Figure 5, PAA has changed the shape of the graph, because PAA is ignoring the co-related behavior of values of time series. But, piecewise SVD remains the same shape, when comparing Figure 2 and 4. Thus, when using piecewise SVD, number of dimensions has reduced accurately. Therefore, this technique can be used to reduce stock time-series data. Furthermore, because of the new local behavior of SVD, can use this new piecewise SVD in the clustering process effectively. Effectiveness will be evaluated in the next section.

11. Evaluation of the time complexity

For full SVD, on an $m \times n$ matrix A, $[U,S,V] = SVD(A)$, the time complexity is in $O(\min(mn^2, m^2n))$ [27]. Thus, when K is the number of stocks and n is the length of a time series, as the previous old method, time complexity is:

$$O(\min(Kn^2, K^2n)) \tag{8}$$

For a dataset with a higher number of series (higher K), or for a dataset with long time series (higher n), above value is high. In the piecewise SVD in this research, when computing SVD for one vector in one sequence, $m = 1$ and $n = n/N$. Thus,

$$O(\min(mn^2, m^2n)) = O\big(\min\big((n/N)^2, (n/N)\big)\big) = O(n/N) \tag{9}$$

The complexity of one time series is given by:

$$N \times O(n/N) = O(n) \tag{10}$$

Computing reduced time series for different sequences are processed in parallel. Thus, there is no impact of K in the piecewise SVD. Therefore, the piecewise SVD has a lower time complexity than the original SVD. Another major advantage of the piecewise SVD is the ability to perform SVD locally. As a result of using one time series at a time, we can perform SVD locally and thus, can add a new series to existing clusters without computing the SVD for the whole data set. Therefore, the local behavior removes the necessity of re-computing the SVD for the entire dataset, when inserting the clusters that we already have by adding a parallelization strategy.

12. Evaluation of the clustering

In most of the previous research approaches, the quality of the stock clustering and time-series clustering have been validated via visualizing the resulting clusters instead of statistical measurements [5,14,28,29,30]. The reason is that even though the statistical measurements give a numerical number, they cannot give accurate measurements. Thus, to assess how well this methodology clusters stock-market data, we performed a visual analysis of sets of stocks in the same groups and those in different groups. Yahoo Finance line charts were used to plot stock price movements of these stocks over the maximum period of available time, and the actual time period used in this research (i.e., Jan 1, 2000 to Nov 30, 2016) is visualized using a vertical line when needed.

```
Cluster 1: [ABL]
Cluster 2: [ACU, BTI, BTN, BVX, CAW, CDY, CKX, CMT, CVU, EVI, FAA, GLJ, GVP, ITI, IYW]
Cluster 3: [ACY, CCF, CIX, DLA, EML, GHM, GRC, IMO, IOT, ISL]
Cluster 4: [AE, CEF, CH, DXR, FCL]
Cluster 5: [AGX, DSS, EPM, GBR]
Cluster 6: [AIM, IYR]
Cluster 7: [AIP, CEV, CTO, DHY, EMI, EVP, EVY, EWA, FAX, FCO, IAF, IEV, IJT, IWD, IYE, IYF, IYG]
Cluster 8: [ALT]
Cluster 9: [AMS, AWX, BCV, BDL, BDR, BHB, CCA, CET, CIK, CRVP, CVR, DGT, DIA, ECF, EVJ, EVO, EWC,
EWD, EWG, EWH, EWK, EWL, EWN, EWO, EWQ, EWW, EWY, EWZ, EZU, HWG, IOO, IVE, IVV, IVW, IWF, IWO,
IWV, IYC, IYH, IYJ, IYK, IYM, IYY]
Cluster 10: [APT, DPW, IPT]
Cluster 11: [BBH, CNR, GV, INS]
Cluster 12: [BKR]
Cluster 13: [BRN, ESP]
Cluster 14: [BSC]
Cluster 15: [BTC, FCD, IMH]
Cluster 16: [BTX, EGAS, FRD, HRT, IJK, IJR, IJS, IWM, IWN]
Cluster 17: [CONM]
Cluster 18: [CRF, EWJ, EWM, HEB, HH, JOB]
Cluster 19: [CUO, ENA, HMG]
Cluster 20: [CVM, IDN]
Cluster 21: [DGSE, FCQ, IDU, IYZ]
Cluster 22: [DIT]
Cluster 23: [ELR]
Cluster 24: [EMAN]
Cluster 25: [EMM]
Cluster 26: [ESA]
Cluster 27: [EWI, EWS, EWT, EWU]
Cluster 28: [FLL]
Cluster 29: [FSE]
Cluster 30: [FSG]
Cluster 31: [FSI]
Cluster 32: [GIT]
Cluster 33: [IAH]
Cluster 34: [IFO]
Cluster 35: [INO]
Cluster 36: [INUV]
Cluster 37: [KGN]
```

Figure 4- Final clustering results of a sample with 156 stocks from AMEX sector

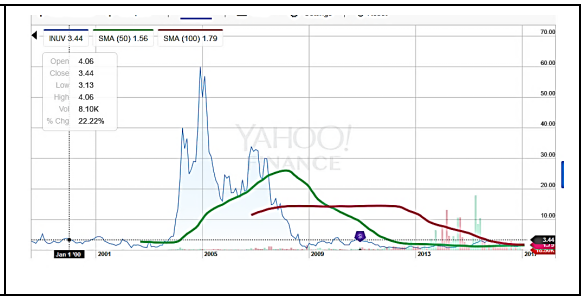
Let us take closing prices and trends of some stocks of cluster 16, as an example. Graphs are taken from the current yahoo exchange pages of particular stock[2]. Blue line indicates the original closing prices of stocks. Also, in order to compare only the trend part, let us take simple moving averages of IJR, IWN, IJS and IWM stocks of cluster 16. The simple moving average is used to get the trend out of the stock-price time series. Thus, now, seasonal and irregular parts have been removed. Here, the green line is the simple moving average with 50 periods and the red line is the simple moving average with 100 periods, which indicates trends.

[2] https://finance.yahoo.com

IJR	
IWN	
IJS	
IWM	

Table 1- Price curves and trend curves of cluster 16

Thus, according to the graphs of cluster 16, price curves of stocks of cluster 16 are much similar. Also, it shows that, the trend curves of same cluster stocks are same. Therefore, financial analyzers can use trend curves of stocks of a cluster to build a model of trend curves. Furthermore, it can be used to forecast stocks of that cluster, using models like Autoregressive Integrated Moving Average (ARIMA) models. Let us take some outliers, which are clusters with only one stock. Let us take simple moving averages of those outliers in order to compare only the trend part of outliers. Starting point of our data set is "Jan 1 '00", which is indicated by a crossed horizontal and vertical line.

cluster 36: INUV	

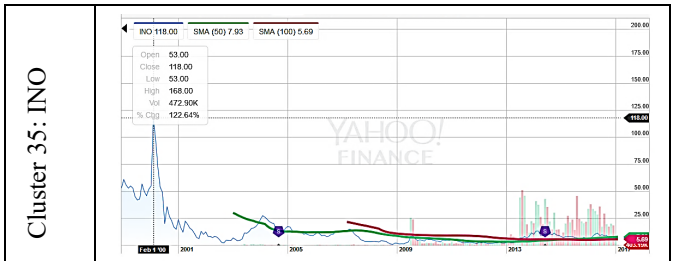
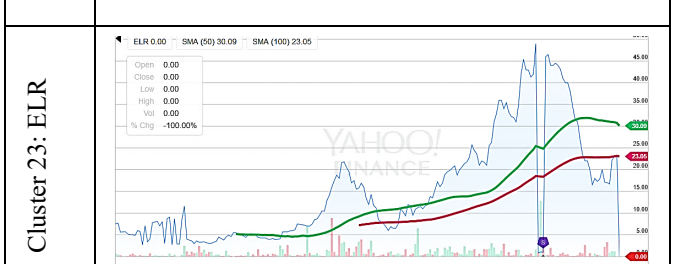
Cluster 35: INO	
Cluster 22: DIT	
Cluster 23: ELR	

Table 2- Price curves and Trend curves of outliers

Thus, there exist a big different between shapes of price curves between outliers. For example, there is an abnormal behavior due to sudden downtrend in ELR of cluster 23. Thus, financial analyzers can further process outliers for fraud detection using price curves or trend curves. Furthermore, price curves of stocks of cluster 16 and price curves of outliers have a dissimilarity. Thus, from these stock price plots, it is apparent that the shapes of the same cluster price curves are similar, but there is a great difference in different clusters. The trend curves of outliers are different from other stock trends and can detect sudden decreases of trends like in ELR of cluster 23.

When the distance threshold increases, it can be got outliers far away from other clusters, which is good for further processing of fraud detection. When the distance threshold decreases, more refined clusters can be gained to receive accurate forecasting results.

13. Stress testing results

In order to give some indication about the running time of the program, running times with different number of stocks are listed below. Here, one stock is 6179 dimensions lengthier.

- Total time: Total running time in milliseconds
- CommittedVirtualMemorySize: The amount of virtual memory that is guaranteed to be available to the running process in bytes, or -1 if this operation is not supported.
- FreePhysicalMemorySize: The amount of free physical memory in bytes.

- FreeSwapSpaceSize: The amount of free swap space in bytes.
- ProcessCpuLoad: The "recent CPU usage" for the Java Virtual Machine process.
- ProcessCpuTime: The CPU time used by the process on which the Java virtual machine is running in nanoseconds.
- SystemCpuLoad: The "recent CPU usage" for the whole system.
- TotalPhysicalMemorySize: The total amount of physical memory in bytes.
- TotalSwapSpaceSize: The total amount of swap space in bytes.

Number of stocks processed	Size of stocks	Running time, CPU and Memory usage
156	29.1 MB (30,525,923 bytes)	Total time: 13530 milliseconds CommittedVirtualMemorySize = 3466719232 FreePhysicalMemorySize = 728338432 FreeSwapSpaceSize = 8589930496 MaxFileDescriptorCount = 4096 OpenFileDescriptorCount = 73 ProcessCpuLoad = 0.006662867826430245 ProcessCpuTime = 17900000000 SystemCpuLoad = 0.7195349045079402 TotalPhysicalMemorySize = 8252899328 TotalSwapSpaceSize = 8589930496
500	61.3 MB (64,380,826 bytes)	Total time: 35659 milliseconds CommittedVirtualMemorySize = 3466719232 FreePhysicalMemorySize = 211812352 FreeSwapSpaceSize = 8589930496 MaxFileDescriptorCount = 4096 OpenFileDescriptorCount = 73 ProcessCpuLoad = 0.0012549780525146542 ProcessCpuTime = 38600000000 SystemCpuLoad = 0.3898377989635702 TotalPhysicalMemorySize = 8252899328 TotalSwapSpaceSize = 8589930496
1000	124 MB (130,960,596 bytes)	Total time: 32280 milliseconds CommittedVirtualMemorySize = 3466719232 FreePhysicalMemorySize = 130936832 FreeSwapSpaceSize = 8589930496 MaxFileDescriptorCount = 4096 OpenFileDescriptorCount = 73 ProcessCpuLoad = 8.664680788157126E-4 ProcessCpuTime = 30850000000 SystemCpuLoad = 0.40124565536868134 TotalPhysicalMemorySize = 8252899328 TotalSwapSpaceSize = 8589930496
2000	254 MB (267,264,965 bytes)	Total time: 73661 milliseconds CommittedVirtualMemorySize = 3466719232 FreePhysicalMemorySize = 204488704 FreeSwapSpaceSize = 8589930496 MaxFileDescriptorCount = 4096 OpenFileDescriptorCount = 74 ProcessCpuLoad = 0.0011113291481388054 ProcessCpuTime = 73020000000 SystemCpuLoad = 0.3958333999369435 TotalPhysicalMemorySize = 8252899328 TotalSwapSpaceSize = 8589930496
4000	525 MB (550,963,670 bytes)	Total time: 220727 milliseconds CommittedVirtualMemorySize = 3466719232 FreePhysicalMemorySize = 122253312 FreeSwapSpaceSize = 8589930496 MaxFileDescriptorCount = 4096 OpenFileDescriptorCount = 73 ProcessCpuLoad = 0.0022682967493810216 ProcessCpuTime = 150880000000 SystemCpuLoad = 0.3980579663474845 TotalPhysicalMemorySize = 8252899328 TotalSwapSpaceSize = 8589930496
6000	789 MB (828,032,464 bytes)	Total time: 315866 milliseconds CommittedVirtualMemorySize = 3466719232 FreePhysicalMemorySize = 179867648 FreeSwapSpaceSize = 8589930496 MaxFileDescriptorCount = 4096 OpenFileDescriptorCount = 73 ProcessCpuLoad = 0.003132518603259838 ProcessCpuTime = 211690000000 SystemCpuLoad = 0.4013075514566676 TotalPhysicalMemorySize = 8252899328 TotalSwapSpaceSize = 8589930496
8000	1.07 GB (1,149,946,977 bytes)	Total time: 443453 milliseconds CommittedVirtualMemorySize = 3466719232 FreePhysicalMemorySize = 122851328 FreeSwapSpaceSize = 8589930496 MaxFileDescriptorCount = 4096 OpenFileDescriptorCount = 73 ProcessCpuLoad = 0.004165920567014453 ProcessCpuTime = 287090000000 SystemCpuLoad = 0.40548878789980664 TotalPhysicalMemorySize = 8252899328 TotalSwapSpaceSize = 8589930496
9000	1.23 GB (1,327,082,337 bytes)	Total time: 410532 milliseconds CommittedVirtualMemorySize = 3466719232 FreePhysicalMemorySize = 134823936 FreeSwapSpaceSize = 8589930496 MaxFileDescriptorCount = 4096 OpenFileDescriptorCount = 73 ProcessCpuLoad = 0.0031908212468572204 ProcessCpuTime = 226650000000 SystemCpuLoad = 0.4075407278451208 TotalPhysicalMemorySize = 8252899328 TotalSwapSpaceSize = 8589930496

Table 3- Running time, CPU and Memory usage with different loads of stocks

Thus, according to these comparisons, it shows that this method can process massive data sets. Therefore, this method is suitable for process time series big data such as stock time series.

14. Conclusions and Future Work

In this paper, clustering method for time-series data based on SVD is proposed. The method takes advantages of the piecewise SVD and clustering to visualize efficiently the time-series data, which improves the processing efficiency and analysis of time-series data. We successfully implemented this method on Hadoop platform which provides the capabilities to process massive data. Furthermore, the real historical stocks are analyzed by this method on Hadoop platform: first, the data normalization, then, the feature extraction by piecewise SVD, and finally, the clustering by BIRCH with good results.

15. Conclusions

This research added a value to time series, data mining, and big data fields. First of all, this research explores a theoretical background related to time-series data. This can be seen as a case study to time-series mining research area. Very special decisions made in the stock data normalization phase can apply to time series data sets which follow a normal distribution. Furthermore, the novel piecewise SVD approach has numerous advantages than previous approaches. It is simple to understand and implement and can be applied to dimensionality reduction very accurately and efficiently. This will solve most of the problems with high-dimensionality and high-noisy in time-series data. Furthermore, the novel piecewise SVD has a local behavior, which is an opposite characteristic when considering the original SVD which has a global behavior. Therefore, this new piecewise SVD is suitable for big data. Furthermore, because of this local behavior, piecewise SVD can be used for parallel processing, and this piecewise SVD removes the necessity of re-computing the SVD for entire dataset, when inserting to the clusters we already have.

Previous stock and other time-series clustering methods have only succeeded with traditional clustering methods like k-means, canopy, AprioriAll, SOM, etc. This research concludes that, BIRCH is suitable for data distributions which have time-series behavior with high noise and high dimensions such as stocks and can be used to cluster massive time series data very accurately and efficiently.

Therefore, the overall methodology can be applied to process big time-series data with a normal distribution such as stocks in an efficient manner.

As stated in literature, the previous stock data clustering methods focused only on clustering the trend curve, because, it is very difficult to cluster the actual prices. But, in this research, the price curves of the stocks have clustered efficiently in an accurate manner. The first reason for this is that the correct methods used in the preprocessing stage, which was selected after a good study about time series data behaviors and stock data behaviors. The second reason is the novel approach of using SVD accurately with a local behavior.

Therefore, the method of this research can support exploratory search and facilitate serendipitous discovery. Besides revealing groups of similar stocks, this visualization can be used to find stocks that show different trading patterns. Such discovery can help traders to identify, for example, a set of hot stocks, stocks with a same trend and outliers and can further analysis of fraud detection or for forecasting. Instead of chasing the hot stocks, some investors may prefer to buy undervalued or oversold stocks. To find such stocks, they can explore clusters of stocks that have a sudden downtrend. This strategy may help financial analyzers to reduce the volatility of a portfolio and assist in decision-making process of buying and selling stocks. Therefore, this research enhances the prediction procedure of time-series data via an optimum clustering method.

Future directions

In order to identify subtle patterns with very short life times, Piecewise SVD will be extended for subsequent matching, instead of whole matching. Furthermore, future work should be continued to cluster data with different dimensions and should be continued to time series forecasting, with respect to particular data domains separately. Furthermore researchers can extend the methodology for different data distribution types, instead of normal distributions. Moreover, the researchers can extend this to real time stream processing using big data engine such as Spark because of BIRCH.

References

[1] Hamed Davari Ardakani, Jamal Shahrabi Ehsan Hajizadeh, "Application of data mining techniques in stock markets: A survey," *Journal of Economics and International Financex*, vol. Vol. 2(7), pp. pp. 109-118, July 2010.

[2] Ruey S. Tsay, *Analysis of financial time series*, Second Edition ed., Noel A. C. Cressie, Nicholas I. Fisher David J. Balding, Ed. Hoboken, New Jersey and Canada: A John Wiley & Sons, Inc., Publication, 2005.

[3] Dr. Samidha D. Sharma Abhishek Gupta, "Clustering-Classification Based Prediction of Stock Market Future Prediction," *(IJCSIT) International Journal of Computer Science and Information Technologies*, vol. Vol. 5 , no. 3, pp. 2806-2809, 2014.

[4] Farhat Roohi, "Artificial Neural Network Approach to Clustering," *The International Journal Of Engineering And Science (Ijes)*, vol. 2, no. 3, pp. 33-38, 2013.

[5] Yung-Piao Wu and Hahn-Ming Lee Kuo-Ping Wu, "Stock Trend Prediction by Using K-Means and AprioriAll Algorithm for Sequential Chart Pattern Mining," *Journal of Information Science and Engineering*, vol. 30, pp. 653-667, 2014.

[6] Sun Xin, "The application of Factor Analysis in Chinese Stock Market," Collage of Science Tianjin Polytechnic University, Tianjin (300160), Research paper.

[7] Marco Aliotta, Andrea Cannata, Carmelo Cassisi, Alfredo Pulvirenti Placido Montalto, "Similarity Measures and Dimensionality Reduction Techniques for Time Series Data Mining," in *Advances in Data Mining Knowledge Discovery and Applications*.: InTech, ch. 3, pp. 71-96.

[8] M.Punithavalli V.Kavitha, "Clustering Time Series Data Stream – A Literature Survey," *(IJCSIS) International Journal of Computer Science and Information Security*, vol. 8, no. 1, pp. 289-294, April 2010.

[9] Kaushik Chakrabarti, Michael Pazzani, Sharad Mehrotra Eamonn Keogh, "Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases," Department of Information and Computer Science, University of California, Irvine, California 92697 USA, Research Paper 2000.

[10] J. Vargas, A. Pascual-Montano C.O.S. Sorzano, "A survey of dimensionality reduction techniques," National Centre for Biotechnology (CSIC), C/Darwin, 3. Campus Univ. Autonoma, 28049 Cantoblanco, Madrid, Spain, Survey.

[11] Eric Postma, Jaap van den Herik Laurens van der Maaten, "Dimensionality Reduction: A Comparative Review," Tilburg centre for Creative Computing, Tilburg University, 5000 LE Tilburg, The Netherlands, Review 2009.

[12] Imola K. Fodor, "A survey of dimension reduction techniques," Center for Applied Scientific Computing, Lawrence Livermore National Laboratory, Livermore, survey 2002.

[13] Micheline Kamber Jiawei Han, "Cluster Analysis," in *Data Mining: Concepts and Techniques*, Microsoft Research Jim Gray, Ed. San Francisco, CA 94111: Diane Cerra, 2006, ch. Chapter 7, pp. 383-466.

[14] Aziguli Wulamu, Yantao Wang, Zheng Liu Yonghong Xie, "Implementation of Time Series Data Clustering Based on SVD for Stock Data Analysis on Hadoop Platform," *2014 9th IEEE Conference on Industrial Electronics and Applications* , pp. 2007-2010, June 2014.

[15] Alexander Thomasian and Chung-Sheng Li Vittorio Castelli, "CSVD: Clustering and Singular Value Decomposition for Approximate Similarity Searches in High Dimensional Spaces," IBM Research Division, Almaden, Austin, Beijing, Haifa, T.J. Watson, Tokyo, Zurich, IBM Research Report RC 21755 (98001), 2000.

[16] Aliaksei Sandryhaila, Jonathan Gross, Markus Doru Balcan, "Alternatives to The Discrete Fourier Transform," Carnegie Mellon University, Pittsburgh, PA 1521, Survey 0634967,.

[17] Jon Wickmann, "A wavelet approach to dimension reduction and classification of hyperspectral data," Faculty of Mathematics and Natural Sciences, University of Oslo, Oslo, Thesis 2007.

[18] J.R.G. & Barrios, E.B. Lansangan, "Principal components analysis of nonstationary time series data," *Statistics and Computing*, vol. 19, no. 173, June 2009.

[19] Paul D. Gilbert and Erik Meijer, "Time Series Factor Analysis with an Application to Measuring Money," University of Groningen, Research School SOM, Groningen, Research report 05F10, 2005.

[20] Jiri Dvorsky, Vaclav Snasel Pavel Praks, "Latent Semantic Indexing for Image Retrieval Systems," Dept. of Applied Mathematics and Department of Computer Science, Technical University of Ostrava, listopadu, Ostrava - Poruba, Czech Republic, Research Paper.

[21] Pavel V. Senin, "Literature Review on Time Series Indexing," Collaborative Software Development Lab, Department of Information and Computer Sciences, University of Hawaii, Honolulu, HI, CSDL Technical Report 09-08, 2009.

[22] Raghu Ramakrishnan, Miron Livny Tian Zhang, "BIRCH: An Efficient Data Clustering Method for Very Large Databases," in *SIGMOD '96 6/96*, Montreal, Canada, June 1996, pp. 103-114.

[23] Geeta Sikka Sangeeta Rani, "Recent Techniques of Clustering of Time Series Data: A Survey," *International Journal of Computer Applications (0975 – 8887)*, vol. Volume 52, no. No.15, pp. 1-9, August 2012.

[24] Keith Gutfreund, "Big Data Techniques for Predictive Business Intelligence," in - *(Held to be in future)*, -, 2015, pp. 64-70.

[25] Assoc. Prof. Dr. Sevtap Kestel, "Time Series Analysis - Classical Time Series," 2013.

[26] B. Mahanty, M.K. Tiwari S.R. Nanda, "Clustering Indian stock market data for portfolio management," *Expert Systems with Applications*, vol. 37, pp. 8793-8798, 2010.

[27] Alexander G. Gray, Charles Lee Isbell, Jr Michael P. Holmes, "Fast SVD for Large-Scale Matrices," College of Computing, Georgia Institute of Technology, Atlanta, Research report.

[28] Joel Joseph and Indratmo, "Visualizing Stock Market Data with Self-Organizing Map," in *Proceedings of the Twenty-Sixth International Florida Artificial Intelligence Research Society Conference*, Florida , 2013, pp. 488-491.

[29] Kate A. Smith, Rob Hyndman and Damminda Alahakoon Xiaozhe Wang, "A Scalable Method for Time Series Clustering," School of Business Systems and Department of Econometrics and Business Statistics, Monash University, Victoria, Australia, Research paper.

[30] Vamsidhar Thummala, Kamalakar Karlapalem Vipul Kedia, "Time series forecasting through clustering - A case study," International Institute of Information Technology, Hyderabad, India, Research paper.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 865-871 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Analysis of Wireless Traffic Data through Machine Learning

Muhammad Ahsan Latif*[1], Muhammad Adnan[2]

[1]*Department of Computer Science, University of Agriculture Faisalabad, 38000, Pakistan*

[2]*Research assistant, Institute of Manufacturing Information and Systems, Department of Computer Science, National Cheng Kung University(NCKU), 701, Tainan, Taiwan*

ARTICLE INFO

Article history:
Received: 23 April, 2017
Accepted: 02 June, 2017
Online: 20 June, 2017

Keywords:
NextGen wireless networks
Resource management
Neural networks
Principal component analysis

ABSTRACT

The paper presents an analytical study on a wireless traffic dataset carried out under the different approaches of machine learning including the backpropagation feedforward neural network, the time-series NARX network, the self-organizing map and the principal component analyses. These approaches are well-known for their usefulness in the modeling and in transforming a high dimensional data into a more convenient form to make the understanding and the analysis of the trends, the patterns within the data easy. We witness to an exponential rise in the volume of the wireless traffic data in the recent decade and it is increasingly becoming a problem for the service providers to ensure the QoS for the end-users given the limited resources as the demand for a larger bandwidth almost always exist. The inception of the next generation wireless networks (3G/4G) somehow provide such services to meet the amplified capacity, higher data rates, seamless mobile connectivity as well as the dynamic ability of reconfiguration and the self-organization. Nevertheless, having an intelligent base-station able to perceive the demand well before the actual need may assist in the management of the traffic data. The outcome of the analysis conducted in this paper may be considered in designing an efficient and an intelligent base-station for better resource management for wireless network traffic.

1. Introduction

Today, the communication has become an integral part of the daily activities for everyone. For example, the services such as video on demand (VoD), downloading music, stream video data, voice over IP (VoIP) and live video conference calls have become a part of the daily activities of the users. Such applications need a high quality of service, particularly for voice in real time. Resultantly, a lot of wireless technologies such as 3rd Generation (3G), wireless fidelity (Wi-Fi), Worldwide Interoperability Microwave Access (WiMAX), Global System for Mmobile Communications (UMTS) and the Evolution of the Long-Term (LTE) are growing to meet the accelerated needs of the users to provide them anywhere, anytime access to the Internet. In the recent past, the main point in the telecommunications sector is the next generation wireless networks (NGWNs). These networks consist of different techniques to gain access to the network to provide unlimited access to the internet, global roaming, high-speed and an increased user satisfaction.

The NextGen Wireless Networks are a combination of circuit-switched networks, various services, and packet switching. It is covering a wider area in which the user can take advantage of different wireless networks to improve the call quality. NextGen comprises of the services independent of the core technologies related to transport [1, 2]. It can be defined as a complex process, easily accessible, convenient and converging economic, efficient and flexible, personal and real, reliable and secure [3].

The fundamental technical issues for the NextGen Wireless Networks include end-to-end QoS, mobility management, energy efficiency, call admission control, resource allocation, and security. This has encouraged many researchers to develop the QoS models for the NextGen Wireless Networks [4]. In mobility the connected base station cannot support the ongoing session, and it has to handover the session to the new base station [5]. Under the wireless communication, most of the network equipment

* Muhammad Ahsan Latif, Asst. Prof., Department of computer science, University of Agriculture, Faisalabad, Pakistan | Email: mahsanlatif@uaf.edu.pk

operate on battery power and with the limited resources just not capable to provide the energy to keep the equipment alive for extended time period [6]. Controlling calls in wireless communication is a process of administering the traffic volume. The decision of whether a new call has to be admitted, delayed, dropped or forwarded to a neighboring network is decided by this sub-system. The incoming traffic can further be split into real time i.e. video and voice, and non-real time i.e. images, text, etc. [7]. As the NextGen Wireless Networks are distributed and open architectures so it offers easy access to the services, information and the resources together with constant abuse from the fraudsters, hackers and crime units [8]. User identification based on the IP layer can be easily tampered within such networks. The packets sent over the network can be easily identified with a "borrowed" IP address, which grants the unauthorized users to mimic the valid ones. These intruders abuse services and gets benefit with the expenditures of legitimate users, who do not know the actual situation [9]. Such intruders can get a valid electronic serial number and mobile identification number during the registration process of the call. They can duplicate the same number on the other handset and utilize the services in the name of real user. A mobile node utilizes Mobile IP and Session Initiation Protocol (SIP) under subscriber mobility [10]. The installation of small cell base stations on the existing macro cellular systems is presented as a solution for an optimized coverage, offloading traffic and to boost the capacity of the next generation wireless networks in [11]. A comprehensive overview of the machine learning approaches used to address different issues in the wireless sensor network is provided in [12]. The researchers supported the use of machine learning techniques to better monitor the dynamic behavior that emerges out of the sensor network over time. For enhanced traffic management, the researchers performed classification of the traffic data through supervised machine learning approaches [13]. For mobile big data, the challenges of computing capabilities, spectrum efficiency and backhaul / fronthaul link capacity are discussed in [14]. The study presented different opportunities and the bottlenecks in the design of the scalable wireless systems to get adapt to the big data.

The foremost issue in the NextGen Wireless Networks is guaranteeing the QoS and provision of requisite resources to the connected subscribers from the service provider at any time. This claims the network reliability, bandwidth, timeliness, jitter, fault tolerance and seamless mobility among heterogeneous access networks. These needs are the driving factors for the NextGen Wireless Networks. Particularly, VoIP will be the most popular application in the future. It demands high QoS for the moving user. In this regard, many problems exist to make the NextGen Wireless Networks fully resourceful and well-matched with the existing services and technologies. For the moving subscribers providing the required service, by sustaining the adequate QoS while considering the various networks as well as the global roaming is a tricky job. The appearance of the irregular needs in the NextGen Wireless Networks require an effective assurance to satisfy the subscribers.

To ensure the QoS for the subscribers, an understanding of the behavior of the network traffic is necessary. This would assist in following an adaptive policy for using the available network resources to guarantee QoS at an optimal level. To analyze the network traffic behavior, we used the self-organizing map and further employed the principal component analyses. An intelligent module within a base-station could be embedded with the outcomes of the analysis to take adaptive decisions intelligently. The module would estimate the future requirements of the bandwidth for a base-station as per the previous history and commit in advance the requisite resources from the backhaul network for the guaranteed bandwidth provision to the end-users. The real time data of a 3G wireless network established by the Pakistan Telecommunication Co. LTD i.e. CDMA2000 1xEV-DO Rev is used in this study.

2. Intelligent Base Station

A schematic diagram of an intelligent base station is shown in Figure 1.

Following are the components of this intelligent model:

- Monitoring Unit: This component monitors the data transferred from the Base Station to the connected subscribers periodically.
- Storage Unit: This component stores the observed data in the memory to use for another component of this model i.e. Intelligent Forecasting Unit.
- Intelligent Forecasting Unit: This component is the heart of the model, which facilities in forecasting of upcoming demands on the Base Station in a specific hour. It uses the previous stored data from the storage unit for predicting the upcoming resources requirement and subsequently forwards the request to the backhaul network in advance to allocate the forecasted bandwidth allocation.

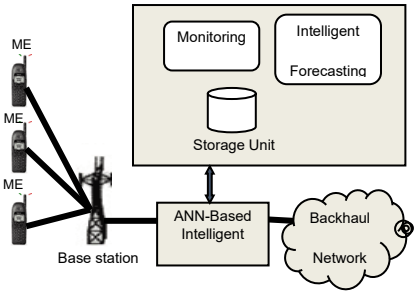


Figure 1: Intelligent base station model

3. Neural Network Based Analysis

The ANNs, owning a proven learning paradigm can be used effectively to develop such intelligent agent [15-27]. Three different NNs are tested to understand the patterns and the behavior lying within the network traffic data. The backpropagation feedforward neural network and the nonlinear autoregressive with external input neural network are used to better learn the demand time-wise. Whereas the self-organizing map is used to learn the key parameters to focus on. The training data comprises of the six-month wireless traffic data acquired from the Pakistan Telecommunication Company Ltd. The training data comes in periodic shape i.e. hourly data with reference to the week-days and an average accumulative throughput in a particular hour. The mathematical models of the three NNs are given as under.

$$Y = F^2(LW^{2,1}F^1(IW^{1,1}P + b^1) + b^2) \quad 1a$$

$$X_{k+1} = X_k - a_k g_k \quad 1b$$

$$E = MSE = 1/N \sum_{i=1}^{N} (t_i - a_i)^2 \quad 1c$$

$$g_k = \frac{\partial E_k}{\partial w_{i,j}} \quad 1d$$

$$Y(t) = F(Y(t-1), ..., Y(t-n), X(t-1),..., X(t-n)) \quad 2$$

$${}_iW(q) = {}_iW(q-1) + \alpha(P(q) - {}_iW(q-1)) \quad 3a$$

$$N_i(d) = \{j, d_{ij} \leq d\} \quad 3b$$

The (1) define the feedforward neural network completely. The (1a) defines a three layered architecture with LW as the layer weights, IW as the input weight matrices and P as the input vector The sub-equation (1b) gives the learning rule to update the network parameters wherein X_k represents the vector of the current weights and the biases, g_k is the current gradient and a_k is the learning rate. The (1c) defines the mean square error (mse) which is also one of the stopping criteria used during the training phase. The (1d) defines the relation how to determine the gradient. The system of nonlinear autoregressive with external input (NARX) is defined in (2). In our case, the previous two terms (of both the input & output) are taken into account to calculate the current term. The model uses the same learning rule as given in (1b). The system of (3) provides the rule to update the weights of the winning neuron as well as of the neighboring neurons in a neighborhood d' defined in (3b). All of the three models are programmed and tested in Matlab. The used parameters are given in the following tables Table-1, Table-2 and Table-3 against each model.

The results of all the three models provide a very clear understanding of the relations among the variables embedded in the data. The function approximation performed by the feedforward model (1) provides a strong evidence of the relation between the output variable (throughput) and the input-vector (comprising of eight input elements). The training-performance for the feedforward model went perfectly well as shown in Figure-2 as all the three segments of the training data sets, i.e., the training, the testing and the validation got to converge optimally. The value of 'R' achieved in this case is R = 0.9867 (regression value).

The NARX model (2) is also found capable of learning the underlying behavior of the network traffic data. The function Bayesian Regularization is used for training purpose as it produces the better generalization for a complex and noisy data. The principle of adaptive weight minimization is followed to stop the training. In this model, the previous two states of the input-time-series and the target-time-series are taken into account to predict the next value in the output-time-series. The regression value is found to an acceptable level as R=0.983. An important graph of the NARX model, i.e., the time-series response is shown in Figure-3. A close similarity can be seen between the outputs and the targets. The lower part of the graph shows the error graph which is found well within the confidence interval. The objective of testing the SOM model over the traffic data is to reduce the data-dimensionality to see which variables are more intrinsic to the resource-requirement. A network of 100 neurons fully mapped the underlying data space. The impact of the variables on the network is fully conveyed in the input weight-planes. All the input weight-planes corresponding to the input variables are given in the Figure-4. The weight-planes for the input-1 to the input-7 contain most of the part as dark regions indicating zero or lower values for the weights between the network-neurons and that particular input. This would suggest the days of the week are not to play much role in deciding the network resource management as the traffic varies. The weight-plane for the 8th input variable, i.e., the hours, comprised of more bright spots and fewer dark ones presenting a strong connection strength with the network-neurons. The 9th weight-plane owns much of the bright spots as compared to the rest of the weight-planes showing the resilient connections with the neurons of the network.

(Dataset Distribution-%) Training/Testing/Validation	(70)(15)15)
ANN Architecture	8-25-1
No. of input neuron	8
No. of hidden layer	1
No. of output neuron	1
Training Algorithm	Levenberg-Marquardt
Error performance function	MSE
Maximum Epochs	1000

Table. 1. Parameters used for Feedforward ANN Model

(Dataset Distribution-%) Training/Testing/Validation	(70)(15)15)
ANN Architecture	A 3-layered time delay network
No. of input neuron	8
No. of hidden layer	1
No. of output neuron	1
Training Algorithm	Bayesian Regulation
Error performance function	MSE
Maximum Epochs	1000

Table. 2. Parameters used for NARX Model

SOM Architecture	10 x 10
Neighborhood size	3 x 3
No. of input neurons	9
No. of output neurons	100
Training Algorithm	Batch weight / Bias
Error performance function	MSE
Maximum Epochs	200

Table. 3. Parameters used for SOM Model

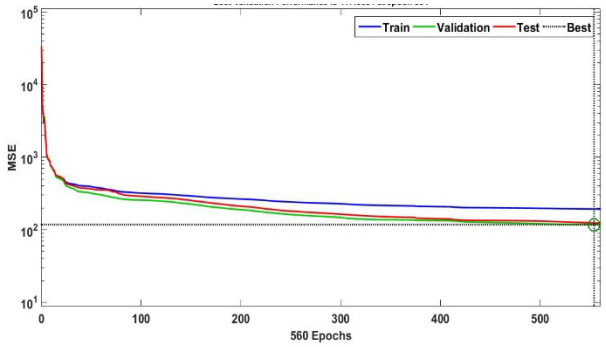


Figure 2. Feedforward Training Performance

4. Principal Component Analysis

The principal component analysis is regarded as one of the fundamental approach to simplify the data. A wide use of this approach has been reported in the relevant literature [28-29-30]. The approach basically transforms a redundant data comprising of linearly correlated variables into a non-redundant dataset of orthogonal (uncorrelated) variables. The transformation results into a projection of the existing data to a new data-space where each axis corresponds to an orthogonal variable, i.e., principal component. The principal components are arranged in terms of their variance, i.e., the first component (axis) holds the maximum variance, the second component, being orthogonal to the first component, holds the second-highest variance along its axis and so on. This projection highlights the real trends within the data which drive the overall behavior of the data. In the following, the results of the approach are presented conducted on wireless traffic dataset. The dataset is comprised of nine variables.

Principal Components	Latent	Explained (%)
1	1.60	17.85
2	1.16	12.96
3	1.16	12.96
4	1.16	12.96
5	1.16	12.96
6	1.16	12.96
7	1.14	12.73
8	0.41	4.590
9	1.53e-28	1.70e-27

Table 4. Variances of Principal Components

The variance and the variance-percentage for each principal component is given in the Table 4. This shows only the very first principal component contributes more in terms of variance as its share is 17.85 %. Most of the principal components, i.e., from 2 to 7 have almost a constant share of variance whereas the last two components present negligible share. Figure-5 provides a graphical representation of the share of each principal component variance-wise. In general cases, the first three or four principal components are considered enough to represent much of the variance in the data but in our case except the very last two principal components, all of the first seven principal components contribute significantly in the variance. In Figure-6 all the nine variables are represented as vectors wherein the direction and the length of each vector shows the contribution of each vector to the first two principal axis. For example, the first principal component along the horizontal axis, has the positive coefficients for the variables Hr. (hour), Throughput, Sunday, Saturday, Friday and Thursday. This gives an idea that more variance occur in these days. The maximum variance are for the variables Hr. and the Throughput as both are also in complete parallel to the component' axis. This also shows that the variables Mon, Wed and Tue have no influence on the first principal component as these fall on the negative side of the horizontal axis. The second principal component, on the vertical axis, has positive coefficients for the variables Mon, Wed, Sun, Sat, Hr. and Throughput whereas the negative coefficient for Tue, Fri and Thu. The PCA analysis shows the main driving forces behind the variances in the data come from the variable Hr., Throughput and the days like Thu, Fri, Sat and Sun. These results are in complete compliance with the findings of self-organization map presented in Figure-4. The weights of the network out of the input-9 (Throughput) and the input-8 (Hr.) are much stronger as compared to the weights from the remaining inputs. Also the weights from the input-4, input-5, input-6 and input-7 (correspond to Thu, Fri, Sat and Sun) contain slightly more grey areas as compared to the weights from the remaining counterparts, i.e., input-1, input-2 and input-3 which correspond to Mon, Tue and Wed.

5. Results and Discussion

The results of all the three models provide a very clear understanding of the relations among the variables embedded in the data. The function approximation performed by the feedforward model (eq. 1) provides a strong evidence of the relation between the output variable (throughput) and the input-vector (comprising of eight input elements). The training-performance for the feedforward model went perfectly well as shown in Figure-2 as all the three segments of the training data sets, i.e., the training, the testing and the validation got to converge optimally. The value of 'R' achieved in this case is R The NARX model (eq. 2) is also found capable of learning the underlying behavior of the network traffic data. The function Bayesian Regularizationis used for training purpose as it produces the better generalization for a complex and noisy data. The principle of adaptive weight minimization is followed to stop the training. In this model, the previous two states of the input-time-series and the target-time-series are taken into account to predict the next value in the output-time-series. The regression value is found to an acceptable level as R=0.983. An important graph of the NARX model, i.e., the time-series response is shown in Figure-3. A close similarity can be seen between the outputs and the targets. The lower part of the graph shows the error graph which is found well within the confidence interval. The objective of testing the SOM model over the traffic data is to reduce the data-dimensionality to see which variables are more intrinsic to the resource-requirement. A network of 100 neurons fully mapped the underlying data space. The impact of the variables on the network is fully conveyed the input weight-planes. All the input weight-planes corresponding to the input variables are given in the Figure-4. The weight-planes for the input-1 to the input-7 contain most of the part as dark regions indicating zero or lower values for the weights between the network-neurons and that particular input. This would suggest the days of the week are not to play much role in deciding the network resource management as the traffic varies. The weight-plane for the 8th input variable, i.e., the hours, comprised of more bright spots and fewer dark ones presenting a strong connection strength with the network-neurons. The 9th weight-plane owns much of the bright spots as compared to the rest of the weight-planes showing the resilient connections with the neurons of the network. The results from PCA analysis confirm the neural-network based modeling. The analysis presented here as a case study for the network traffic data shows a strong potential to employ the machine learning techniques to manage the network traffic. With the surging rate of network traffic across different platforms and networks, the provision of QoS has become an issue for the service providers especially for the telecom industry. The communication

has taken multi-dimensional shape, i.e., voice, video, text, etc., all being delivered over the same infrastructure. Though, new paradigms for the increased band-width, improved communication protocols, etc., are being worked on in parallel yet the efforts should also be made to better plan the network traffic within the same framework to ensure the maximum optimum use of the existing resources. In this regard, the machine learning approaches may play a significant role to assist the service providers in dealing with the traffic management. More research is required in this context to develop robust traffic planning.

6. Conclusion

The behavior of the wireless traffic data is studied under neural networks and the principal component analysis. The three ANN models used include backpropagation feedforward, the time-series NARX and the self-organizing map. The FF model and the NARX model showed the potential to closely model the relation between the throughput and the other variables. The results from the SOM model showed the dimension reduction in the original data by presenting strong weight diagrams for only the two inputs while representing the remaining inputs with average weight strengths. The results from the principal component analysis further verify the results from the SOM. The results of all this study could be incorporated in designing an intelligent module at a network base station to provide adaptive services for the better resource management, hence improving the efficiency of the network.

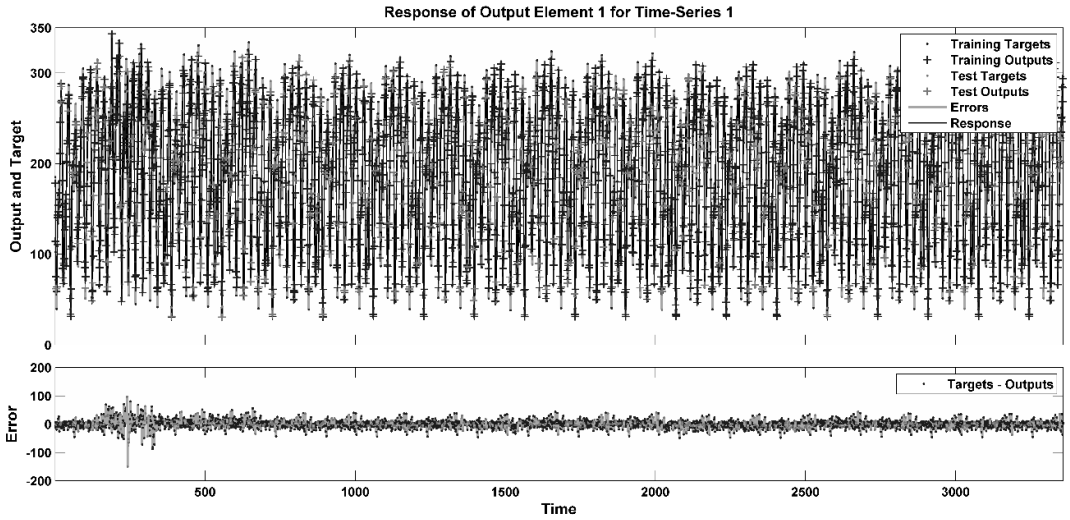


Figure. 3. Time series response of NARX model

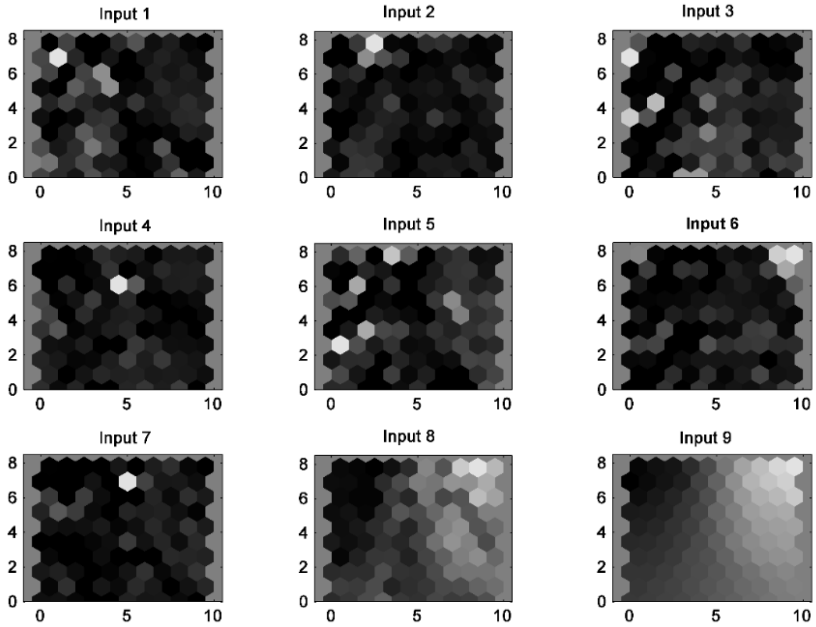


Figure. 4. Some weight planes for the inputs

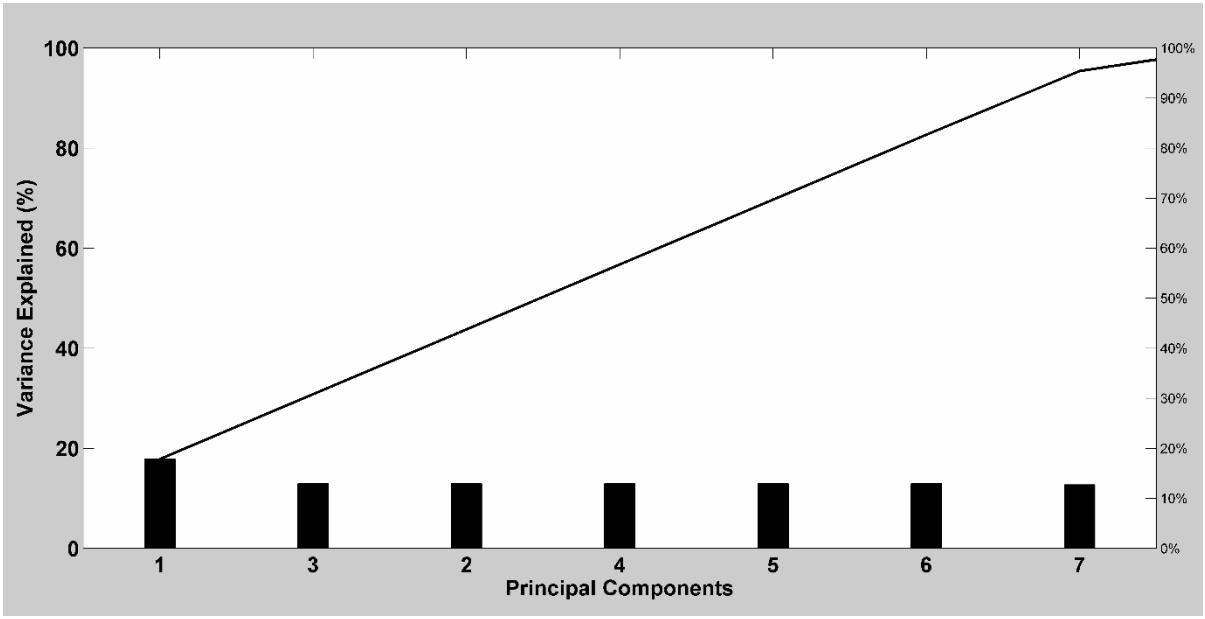


Figure. 5 Variance-Percentage of principal components

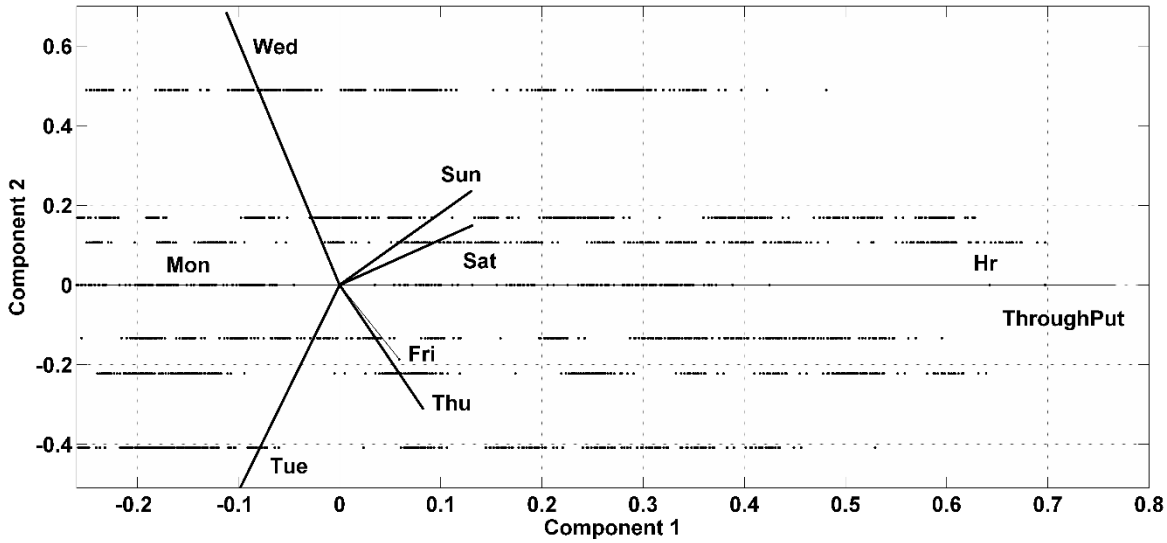


Figure. 6 Projection of all the variables onto the first two orthogonal principal components

References

[1] Latif, M. Ahsan, and M. Adnan. "ANN-Based Data Mining for Better Resource Management in the Next Generation Wireless Networks." Frontiers of Information Technology (FIT), 2016 International Conference on. IEEE, 2016.

[2] Lee, Chae-Sub, and N. Morita. "Next Generation Network Standards in ITU-T." 2006 1st IEEE International Workshop on Broadband Convergence Networks.

[3] Korotky, Steven K., and Thomas Pfeiffer. "Continuing advances in next-generation communication technologies, services, and networks." Bell Labs Technical Journal 14.1 (2009): 1-5.

[4] Li, Bo, et al. "QoS enabled voice support in the next generation Internet: issues, existing approaches and challenges." IEEE Communications Magazine 38.4 (2000): 54-61.

[5] Akyildiz, Ian F., Jiang Xie, and Shantidev Mohanty. "A survey of mobility management in next-generation all-IP-based wireless systems." IEEE Wireless communications 11.4 (2004): 16-28.

[6] Berardi, Victor L., and Guoqiang Peter Zhang. "An empirical investigation of bias and variance in time series forecasting: modeling considerations and error evaluation." IEEE Transactions on Neural Networks 14.3 (2003): 668-679.

[7] Wu, Chen-Feng, et al. "A novel call admission control policy using mobility prediction and throttle mechanism for supporting QoS in wireless cellular networks." Journal of Control Science and Engineering 2011 (2011): 21.

[8] Bella, MA Bihina, M. S. Olivier, and J. H. P. Eloff. "A fraud detection model for Next-Generation Networks." Proceedings of the 8th Southern African Telecommunications Networks and Applications Conference (SATNAC 2005), Central Drakensberg, KwaZulu-Natal, South Africa. 2005.

[9] Ericsson 2004. Categorizing telecommunications fraud, an introduction for those new to the subject.

[10] Lee, Hyejeong, Sung Won Lee, and Dong-Ho Cho. "Mobility management based on the integration of mobile IP and session initiation protocol in next generation mobile data networks." Vehicular Technology Conference, 2003. VTC 2003-Fall. 2003 IEEE 58th. Vol. 3. IEEE, 2003.

[11] Bennis, Mehdi, et al. "When cellular meets WiFi in wireless small cell networks." IEEE Communications Magazine 51.6 (2013): 44-50.

[12] Alsheikh, Mohammad Abu, et al. "Machine learning in wireless sensor networks: Algorithms, strategies, and applications." IEEE Communications Surveys & Tutorials 16.4 (2014): 1996-2018.
[13] Ertam, Fatih, and Engin Avci. "Classification with intelligent systems for internet traffic in enterprise networks." Int. J. Comput. Commun. Instrum. Eng 3 (2013): 1469-2349.
[14] Bi, Suzhi, et al. "Wireless communications in the era of big data." IEEE Communications Magazine 53.10 (2015): 190-199.
[15] Atiya, Amir F., et al. "A comparison between neural-network forecasting techniques-case study: river flow forecasting." IEEE Transactions on neural networks 10.2 (1999): 402-409.
[16] Balkin, Sandy D., and J. Keith Ord. "Automatic neural network modeling for univariate time series." International Journal of Forecasting 16.4 (2000): 509-515.
[17] Chen, Yuehui, et al. "Time-series forecasting using flexible neural tree model." Information sciences 174.3 (2005): 219-235.
[18] Cottrell, Marie, et al. "Neural modeling for time series: a statistical stepwise method for weight elimination." IEEE Transactions on Neural Networks 6.6 (1995): 1355-1364.
[19] Giordano, Francesco, Michele La Rocca, and Cira Perna. "Forecasting nonlinear time series with neural network sieve bootstrap." Computational Statistics & Data Analysis 51.8 (2007): 3871-3884.
[20] Günther, Frauke, and Stefan Fritsch. "neuralnet: Training of neural networks." The R journal 2.1 (2010): 30-38.
[21] Jain, Ashu, and Avadhnam Madhav Kumar. "Hybrid neural network models for hydrologic time series forecasting." Applied Soft Computing 7.2 (2007): 585-592.
[22] Kaastra, Iebeling, and Milton Boyd. "Designing a neural network for forecasting financial and economic time series." Neurocomputing 10.3 (1996): 215-236.
[23] Lapedes, Alan, and Robert Farber. Nonlinear signal processing using neural networks: Prediction and system modelling. No. LA-UR-87-2662; CONF-8706130-4. 1987.
[24] Zhang, Guoqiang, B. Eddy Patuwo, and Michael Y. Hu. "Forecasting with artificial neural networks:: The state of the art." International journal of forecasting 14.1 (1998): 35-62.
[25] Poli, I., and R. D. Jones. "A neural net model for prediction." Journal of the American Statistical Association 89.425 (1994): 117-121.
[26] Weigend, Andreas S., Bernardo A. Huberman, and David E. Rumelhart. "Predicting the future: A connectionist approach." International journal of neural systems 1.03 (1990): 193-209.
[27] Zhang, Guoqiang, B. Eddy Patuwo, and Michael Y. Hu. "Forecasting with artificial neural networks:: The state of the art." International journal of forecasting 14.1 (1998): 35-62.
[28] Le Borgne, Yann-Aël, Sylvain Raybaud, and Gianluca Bontempi. "Distributed principal component analysis for wireless sensor networks." Sensors 8.8 (2008): 4821-4850.
[29] Le Borgne, Y., and Gianluca Bontempi. "Unsupervised and supervised compression with principal component analysis in wireless sensor networks." Proceedings of the Workshop on Knowledge Discovery from Data, 13th ACM International Conference on Knowledge Discovery and Data Mining. 2007.
[30] Li, J., and Y. Zhang. "Interactive sensor network data retrieval and management using principal components analysis transform." *Smart Materials and Structures* 15.6 (2006): 1747.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 872-877 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Video-frame Quality Improvement before Shot boundary Detection by using Logarithm, Wavelet and Contoulet Transform

Songpon Nakharacruangsak*

Faculty of Science and Technology, Southeast Bangkok College, 10260, Thailand

ARTICLE INFO

Article history:
Received: 30 April, 2017
Accepted: 06 June, 2017
Online: 20 June, 2017

Keywords:
Shot Boundary Detection
Adaptive Threshold
Logarithm Transform
Wavelet Transform
Contourlet Transform

ABSTRACT

Video shot boundary detection is important step for the research in the content analysis and retrieval fields. In this paper, firstly we presented an efficient method for Video-frame quality improvement to suppress flash occurred within video frame using logarithm, wavelet and contourlet transform. In addition, wavelet and contourlet transform also performed denoising. Secondly, for shot boundary detection, we used gray-scale histogram differences with edge change ratio. Furthermore, the adaptive threshold algorithm for shot transition detection was proposed. The experiment results showed that using logarithm with contourlet transform gain the precision and recall higher than using logarithm with wavelet transform.

1. Introduction

Currently, digital video is widely produced and used due to advanced multimedia technology. Also, the rapid growth of Internet and computer technology results in a tremendous number of digital videos. Therefore, many researches were interested in the development of how to manage, search, retrieve and conduct the data mining of digital video as efficiently as possible for many years ago [1].

Video structure analysis is the first important step of managing video before automatic video content analysis. The analysis results are then utilized in indexing for digital video retrieving later. In various video structure (frame, shot, scene, etc.), the first investigation is video shots. Therefore, a sequence of frame within video shot is suitable for searching and retrieving video data [2].

The frames taken by the only one camera are arranged to be shots with the continuous action in the same area and time. An activity between one shot to the next shot is the shot transitions which is the main problem because the location of video shot boundary has to be identified. Generally, there are two types of shot transition: 1) abrupt transition: CUT, shot transitions in one frame and 2) gradual transition: GT, There are many kinds of the editing technique used in the GT such as dissolve, wipe and fade in/out that could appear between the last frames of the previous shots and the beginning frames of the next shots. Therefore, the detection is harder than that of the abrupt transition because it depends on these techniques of editing [3]. Hence, the point of the difference of two consecutive frames would show shot transitions that need decision function to evaluate threshold. If the result of function is more than threshold, it means that the two frames are different shots. There are a lot of techniques to automatically detect the shot boundary by calculating the difference between frames such as the difference of pixels [3, 4], the difference of statistic [5, 6], edge change ratio [7, 8] and comparison of histogram [1, 9]. However, if flash or abrupt change illumination active in a frame of shots, it may introduce the wrong difference of two consecutive frames that causes the wrong detection.

Therefore, in this paper, we propose the improvement of video frames quality before video shot boundary detection to solve the flash and abrupt change illumination within frames. Then we used the comparison of gray-scale histogram difference, edge change ratio and adaptive threshold algorithm in order to identify the position of shot transition within the video.

The rest of this paper is organized into 4 sections. Section 2 describes about method and material. The experimental results and analysis are presented in section 3, and section 4 provides conclusions and future work.

* Songpon Nakharacruangsak, Southeast Bangkok College, Thailand
Email: songpon_n@hotmail.com

2. Method and Material

We proposed an algorithm for shot boundary detection which composed of four stages, as shown in the Figure 1. Firstly, frame quality preparation for reduction of the light or flash effect in video frame though Logarithm Transform and noise reduction by contourlet transform are proposed. Next, the gray-scale histogram difference method and edge change ratio are used to find the frame differences. Both methods are utilized to detect both the abrupt transition and gradual transition. Finally, the detection results of such methods are merged to specify position of shot boundary.

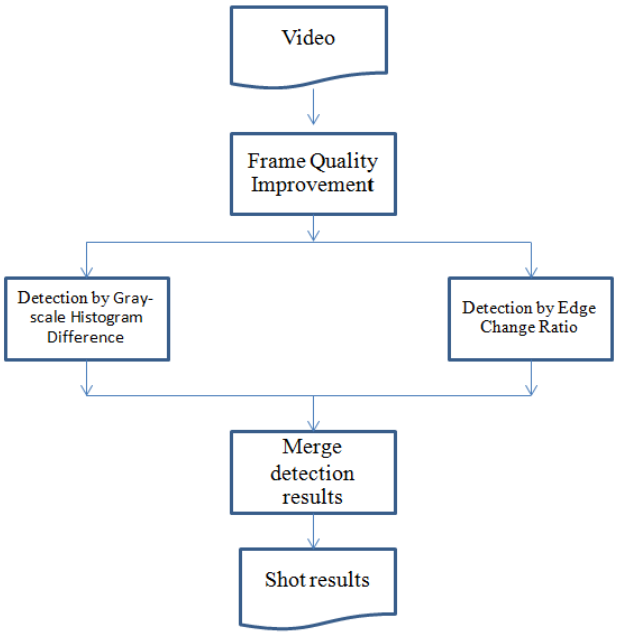


Figure 1: Shot Boundary Detection Framework

2.1. Frame Quality Improvement

Zhang et al. [1] stated that the flash and illumination change in the two consecutive frames have strong effect on gray-scale histogram, i.e., given the high differences. It might make the detection of the abrupt transition in the wrong position .The sample of gray-scale histogram differences of video sequence with flashes is shown in Figure 2.

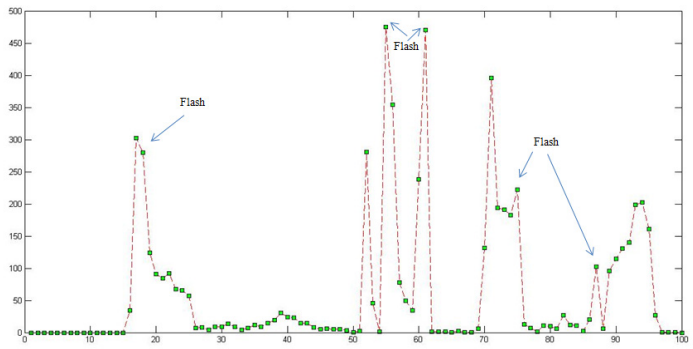


Figure 2: The sample of gray-scale histogram differences of video sequence with flashes

As illustrated in Figure 2, two consecutive peaks appear in the position which a flash occurs.

Logarithm Transform

We applied logarithm transform to improve the frame resolution or contrast of frame in dark area or in low intensity area to have higher brightness. The calculation could be done by using;

$$g=c \times \log(1+\text{double}(f)) \quad (1)$$

Where c is constant value and f is the required frames that need to be increased the brightness. Thus, if c is high, the frames would be more brightness, shown in Figure 3.

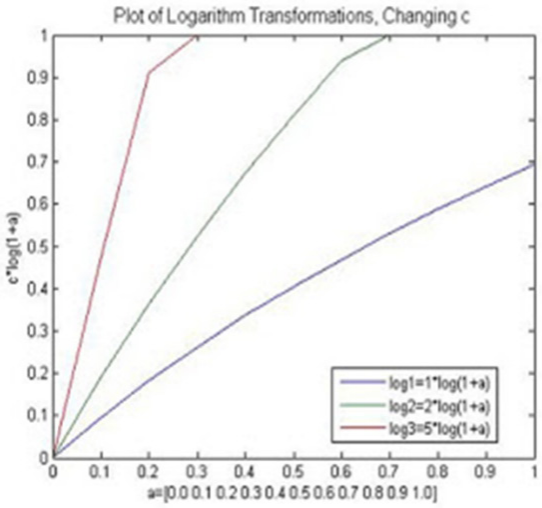


Figure 3: Shot Boundary Detection Framework

According to the proposed method above, under condition of the brightness of the light within a different frame, brightness adjustment may increase noise within the frame. Consequently, we will reduce the noise inside the frame with wavelet transform and contourlet transform and compare the results of the two methods.

Wavelet Transform

We used wavelet transform to reduce noise, proposed by Li et al. [10] Firstly, this method decomposes the noisy image in order to get different sub-band image. Secondly, the low-frequency wavelet coefficients remainly unchanged. After taking into account the relationship of horizontal, vertical and diagonal high-frequency wavelet coefficients and comparing them with Donoho threshold, we make them enlarge and narrow relatively. Thirdly, we use soft-threshold denoising method to achieve image denoising. Finally, the denoising image was achieved by performed inverse wavelet transform. The result shown in Figure 6 c)

Then, we used this method for video-frame quality improvement. The result shown in Figure 4.

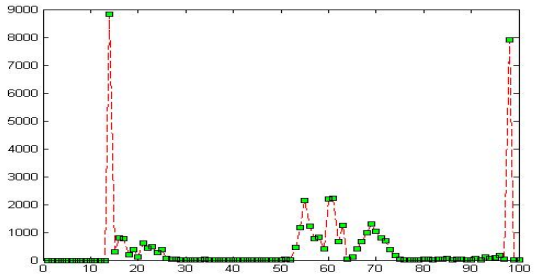


Figure 4: The gray-scale histogram differences of video sequence after suppress the flash problem by wavelet transform.

Contourlet Transform

Contourlet transform, proposed by Do and Vetterli [11], combined Laplacian pyramid (LP) and Directional filter bank (DFB), to efficiently produce more details of sparse curve because of its direction and anisotropy. First of all, the result images after applying the logarithm transform are then divided into subbands by the Laplacian pyramid (LP). Detail of each image is analyzing by Directional filter bank (DFB), shown in Figure 5.

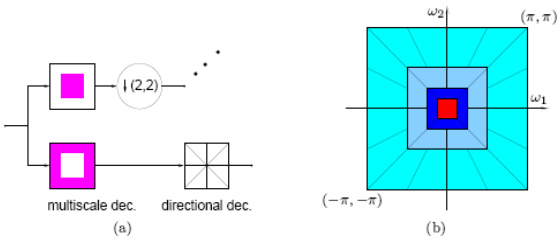


Figure 5: Contourlet Transform Operation and example of DFB classified by frequency

Then, we tested contourlet transform to denoising images. The results shown in Figure 5 d)

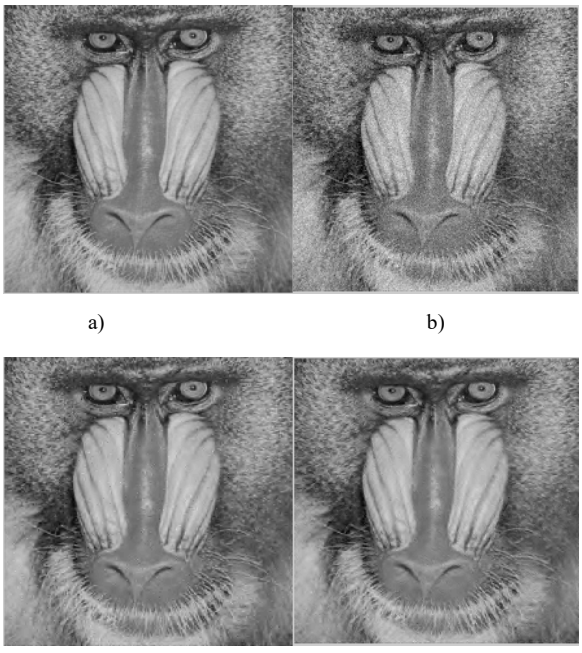


Figure 5: a) Original Images, b) Noisy Image (SNR = 9.52 dB c) Denoise using Wevelet Transform (SNR = 12.38 dB), d) Denoise using Countourlet Transform (SNR = 13.45 dB)

Then, we tested contourlet transform method with the video sequence with flashes. The results shown that, the flash problem and abrupt change illumination can be suppressed, shown in Figure 6.

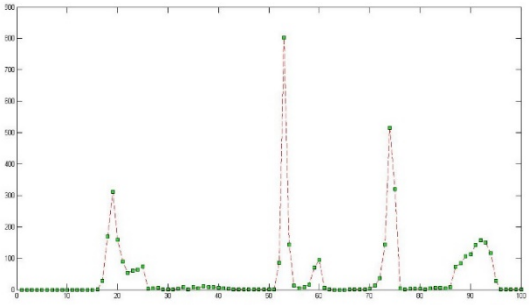


Figure 6: The gray-scale histogram differences of video sequence after suppress the flash problem by our method

2.2. *Video Shot Boundary Features*

Evaluation of efficiency in video shot boundary detection depends on function selection to be used to evaluate the two difference frames. We choose multiple features that combine the difference value of gray-scale histogram, calculated by using Euclidean distance and results from edge change ratio to identify the position of the shot transition. The features of the differences have been shown in the following section.

Gray-scale Histogram Difference

For research in image and video processing, most researchers usually take images in gray scale more than RGB color images because RGB color images composed of color, light and brightness values. It is difficult to compare them because of it is complicated processing. The transformation of RGB color images to gray scale images can be done by using the following equation;

$$Y=0.3 \times R+0.59 \times G+0.11 \times B \quad (2)$$

Y is the gray-scale level at a specific pixel. R, G and B are the red, the green and the blue level at that pixel, respectively.

According to digital images, histogram gives number of pixels that have the same intensity value. In the histogram, the abscissa is range of intensity and the ordinate is the number of pixels. The calculation can be done using the equation;

$$H(i)=\frac{n_i}{N} \quad i=1,2,3,4,\ldots\ldots,k \quad (3)$$

N is the number of all pixels in the image, n is the number of pixel of the ith intensity, k is number of intensity value in the histogram .Therefore, histogram of image M is a vector $H(M) = (h_1,h_2,\ldots..h_k)$

Euclidean distance is the function used for evaluating the similarity of frames. It considers the distance of gray histogram between 2 consecutive frames. The calculation could be conducted by using the equation;

$$Dis(x,y)=\left[\sum_{i=0}^{n}\left|x_i-y_i\right|^2\right]^{\frac{1}{2}} \quad (4)$$

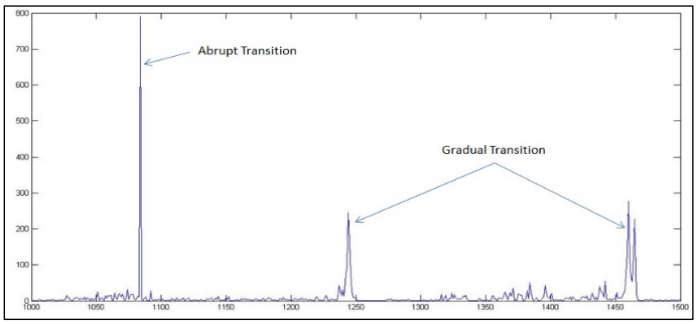
Where x and y are the feature vector. n is the dimension of the vector.

To identify the position of shot transition, we focus on the difference of gray-scale histogram between 2 frames. The calculation could be conducted by using the equation;

$$Diff[i]= \left[\sum_{j=0}^{n}\left|h_{i,j}-h_{i-1,j}\right|^{2}\right]^{\frac{1}{2}} \quad (5)$$

where Diff[i] is the difference between frame i and frame i-1, h_i is gray-scale histogram, n is the dimension of frame i.

The difference of gray-scale histogram shows the frame positions of the shot transition .It is calculated by using equation 5 and is shown in Figure 7.



Edge Change Ratio

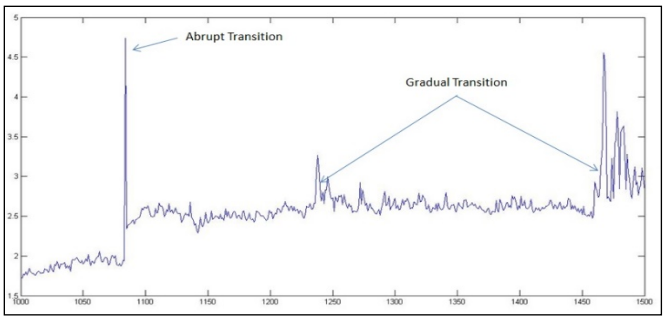
Zabih et al. [12] presented edge change ratio: ECR. The main idea is that the continuous frames come along with the continuous structure. Therefore, the edge of object in the last frame before the hard cut usually cannot be found in the first frame of the next shot. The edge of object in the first frame after the hard cut usually cannot be found in the last shot before the hard cut. The ECR can be described in the following steps:

1. Detect edge of object in frame f_n and f_{n+1}, respectively.
2. Count the number of edge pixel δ_n and $\delta_{n\ 1+}$in the f_n and f_{n+1} frame.
3. Define the entering and exiting edge pixels $En+1^{in}$ and En^{out}.

To assume that two image frames, $lm_{(n)}$ and $lm_{(n+1)}$, the $En+1^{in}$ are fractions of edge pixel in the $lm_{(n+1)}$ frame which farther than fixed distance r away from the closed edge pixel in $lm_{(n)}$. Similarly, the En^{out} are fractions of edge pixel in the $lm_{(n+1)}$ frame which farther than fixed distance r away from the closed edge pixel in the $lm_{(n+1)}$ frame. Therefore, ECR_n between f_n and f_{n+1} can calculated by:

$$ECR_{(n,n+1)}= \left(\frac{En+1^{in}}{\delta n+1}, \frac{En^{out}}{\delta n}\right) \quad (6)$$

If ECR is larger than a predefined threshold, the shot transition occurs at that frame. The process how to be performed on every frame in the video. The ECR sample is shown in the Figure 8.



2.3. Shot Boundary Detection

In shot boundary detection, gray-scale histogram difference and edge change ratio need to be functioned for defining the correct positioning of shot transition which are very important . Therefore, it is more appropriate to set the threshold to suitable to that area .In this paper, adaptive threshold is used in principle of sliding window 2W+1, to compare the similarity of two consecutive frames .The calculation could be done by using to the following equation;

$$T_{(f_n)}= \sum \left(\frac{S(f_{n-1}+f_{n+1})+c}{2w}\right) \quad (7)$$

In this paper, we defined w to 1, c is constant value and $S(f_{n-1},f_{n+1})$ as the function of similarity of two consecutive frames such as f_n and f_{n+1}. Then we compares the results with $T(f_n)$. If the results are greater than $T(f_n)$, then the position of f_n is shot transition. However, it is possible that $S(f_{n-1},f_{n+1})$ is 0. Thus, it is necessary to specify c to 3-5 and apply to shot boundary detection with gray-scale histogram difference and edge change ratio respectively to detect abrupt transition and gradual transition shown in Figure 9 and Figure 10.

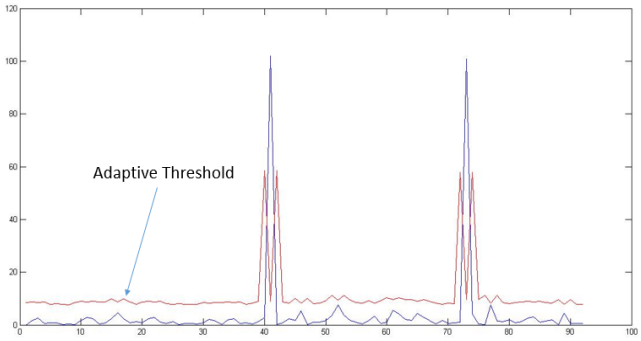


Figure 9: The difference of ECR

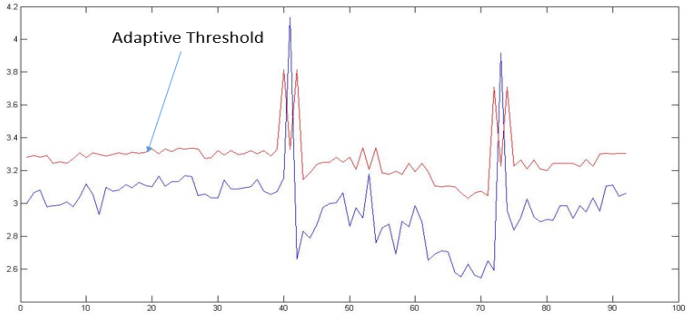


Figure 10: The difference of ECR

After, we obtained the results of shot boundary detection from Gray-Scale Histogram Difference and Edge Chang Ratio shown in Table 1.

Table 1: Results of Shot Boundary detection from two methods

Frame#	Shot1	Shot2	Shot3	Shot4	Shot5	Shot6	Shot7
Histogram Difference	**998**	**1167**	**1292**	1359	**2081**	**2184**	
ECR	86	**998**	**1167**	**1292**	**2081**	**2184**	2312

We will merged the results of two methods and selected the duplicate number of frames to define position of shot transition. The results are shown in Table 2.

Table 2: The Position of Shot Transition

Shot No.	Shot 1	Shot 2	Shot 3	Shot 4	Shot 5
Frame Number	**998**	**1167**	**1292**	**2081**	**2184**

2.4. Video Data and Evaluation

Video clips having inside operation such as object motion, camera panning and zooming were selected in order to evaluate the efficiency of the our proposed method. All data are downloaded from the Open Video Project (http//:www.open-video.org). The evaluation of shot boundary detection is divided into abrupt transitions and gradual transitions their measurement methods are different. In addition, recall and precision are used as metrics which can be calculated as follows:

$$\text{recall}= \frac{N_c}{N_c+N_m} \text{ x } 100\% \quad (8)$$

$$\text{precision}= \frac{N_c}{N_c+N_f} \text{ x } 100\% \quad (9)$$

Where N_c is number of correct detection. N_m is number of missed and N_f is number of incorrect detection.

3. Results and Discussion

The results revealed that the proposed method about a frame quality improvement before shot boundary detection could suppress the flash problem and abrupt change illumination. The experimental results of shot boundary detection by denoising use wavelet transform and contourlet transform shown in Table 3 and are shown performance evaluation of each method in Table 4, that the abrupt transition and gradual transition can be obtained with high accuracy due to using the advantages of the detected object motion and the number of pixels of edge in comparison with edge change ratio to be replaced disadvantages of gray-scale histogram difference which caused by comparison between two consecutive frame with a similar color scheme. Those affected shot boundary detection to have more accuracy and precision.

Table 3: Shot Boundary detection Results

Video	**Number of Frames**	**Number of Cut**	**Number of Gradual**	**Using Wavelet Transform**			**Using Contourlet Transform**		
				Nc	***Nm***	***Nf***	***Nc***	***Nm***	***Nf***
Vipscenc	93	2	-	1	1	-	2	-	-
FlashPreset	398	7	-	5	2	-	6	1	-
anni006	1590	5	5	8	2	-	8	2	
anni007	2775	2	12	11	3	1	13	1	1
Indi114	1813	5	2	7	-	-	7	-	-

Table 4: Performance Evaluation

Wavelet Transform		**Contourlet Transform**	
Recall	***Precision***	***Recall***	***Precision***
80.00%	96.96%	90.00%	97.29%

4. Conclusions

In this paper, we proposed a new method for a frame quality improvement by using logarithm transform and denoising by wavelet transform compared contourlet transform, which can to suppress the flash problem and abrupt change illumination within video. Then, frame differences were calculated by gray-scale histogram difference with edge change ratio. We proposed adaptive threshold algorithm to detect abrupt transition and gradual transition of the boundary. After performing some experimenting to evaluate the performance, we found that the precision and recall value of the boundary position in using wavelet transform less than contourlet transform. However, both methods are not still absolutely successful because some shots are missing and false positive occurs. As a result of this, in the future, we will improve this method by combining multiple features such as motion feature within frame, and statistics so as to reduce missed shot detection and false positive.

Conflict of Interest

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This study was supported by Department of Information Technology, Faculty of Science and Technology, Southeast Bangkok College. We are grateful to website www.open-video.com for supported video database.

References

[1] H. Zhang, R. Hu, L. Song, "A Shot Boundary Detection Method Based on Color Feature", International Conference on Computer Science and Network Technology, Computer Science and Network Technology (ICCSNT), 2011, pp.2541-2544. https://doi.org/10.1109/ICCSNT.2011.6182487

[2] J.Yuan, H. Wang, L. Xiao, W. Zheng, J. Li, F. Li and B. Zhang, "A Formal Study of Shot Boundary Detection", IEEE Transection on Circuits and Systems for Video Technology, Vol.17, No.2, February 2007, pp.168-185. https://doi.org/10.1109/TCSVT.2006.888023

[3] J. S. Borecsky, L. A. Rowe, "Comparison of video shot boundary detection techniques", In Proceedings of SPIE, vol. 2670, pp. 170-179, 1996. https://doi.org/10.1117/12.238675

[4] N. Hirzalla, "Media processing and retrieval model for multimedia documents", PhD thesis, Ottawa University, Jan. 1997

[5] R. Kasturi, R. Jain, Dynamic vision, in Computer Vision: Principles (R. Kasturi and R. Jain, eds.), IEEE Computer Society Press, 1991, pp. 469-480.

[6] T. Kikukawa, S. Kawafuchi, "Development of an automatic summary editing system for the audiovisual resources", IEEE Trans. on Electronics and Information, J75-A, pp. 204-212, 1992.

[7] A.Nasreen, S. G., "Key Frame Extraction Using Edge Change Ratio for Shot Segmentation", International Journal of Advanced Research in Computer and Communication Engineering Vol.2 Issue 11, November 2013, pp.4421-4423.

[8] J. Mann, N. Kaur, "Key Frame Extraction from a Video using Edge Change Ratio", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 5, Issue 5, May 2015, pp.1228-1233.

[9] Z. Wu, P. Xu, "Shot Boundary Detection in Video Retrieval", Electronics Information and Emergency Communication (ICEIEC), pp.86-89. https://doi.org/10.1109/ICEIEC.2013.6835460

[10] L. Hongqiao, W. Shengqian, "A New Image Denoising Method Using Wavelet Transform", Information Technology and Applications, 2009, pp.111-114. https://doi.org/10.1109/IFITA.2009.47

[11] Do, M. N. and Vetterli, M., "The Contourlet Transform: an efficient directional multiresolution image representation", IEEE Transaction on Image Processing 14(12), 2005, pp.2091-2106. https://doi.org/ 10.1109/TIP.2005.859376

[12] R. Zabih, J. Miller, and K. Mai, "A feature-based algorithm for detecting and classifying scene breaks", Proc. ACM Multimedia '95, San Francisco, CA, pp. 189-200, Nov. 1995. https://doi.org/ 10.1145/217279.215266

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 878-883 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Electromagnetic Field and Lumped Circuit Analysis of Planar Transformer at Medium Frequency for Converter Application

Chengjian LIAN, Daming Zhang*

School of Electrical Engineering and Telecommunications, University of New South Wales, 2052, Australia

ARTICLE INFO

Article history:
Received: 16 May, 2017
Accepted: 04 June, 2017
Online: 20 June, 2017

Keywords:
SST
Planar Transformer
High Frequency
Litz Structure
PSFB DC-DC converter
DC-AC inverter

ABSTRACT

High frequency planar transformer is a new type of transformer that operates at high frequency. Its significance is proved through its miniaturised features, high power density and other performances. The design of planar transformer needs to consider several fundamental factors, including core material, dimension of layers and winding pattern. As the volume decreases, extra attentions need to be paid in power losses and safety requirements. Different converter topologies, with planar transformer integrated, are also investigated and simulated for maximising the overall efficiency of conversion. Overall, the use of planar transformer in converters can bring several benefits, and, hence, there is a strong interest of comprehending its structure and high frequency behaviours.

1. Introduction

The increasing demand of miniaturisation of electronic devices limits the use of traditional wire-wound type of transformers[1]. DC converter is a desired technology which has relatively smaller dimension and higher efficiency. Previous works were focusing on the development of Very High Frequency (VFH) used in aerospace and telecommunications, and less effort was put in power electronics at medium frequency. With the increasing popularity of using DC devices, planar structure and converter topologies should be re-investigated.

This paper contains two types of analysis of high frequency planar transformer: electromagnetic field and lumped circuit analysis. The field analysis is performed using Maxwell 3D, which applies finite element analysis on a three-dimensional structure to solve functions iteratively. Electrical parameters, such as resistance and inductance, are calculated and plotted. The circuit analysis is performed in Matlab Simulink using derived mathematic model. The integrated system, containing a full bridge phase-shifting (PSFB) DC-DC converter and a hysteresis current control (HCC) DC-AC inverter, has been also constructed and analysed.

Compared to the conventional winding of planar transformer, a new type of winding structure, litz wire, is analysed. This configuration can effectively reduce the power loss under high operating frequency by providing a fairly uniform current distribution across the conductors.

*Daming Zhang. 650, Material Science Building, UNSW Sydney, NSW 2042. Contact No: +61 2 9385 4070. Email: daming.zhang@unsw.edu.au

2. Design of Litz Wire Planar Transformer

2.1. Design Procedure

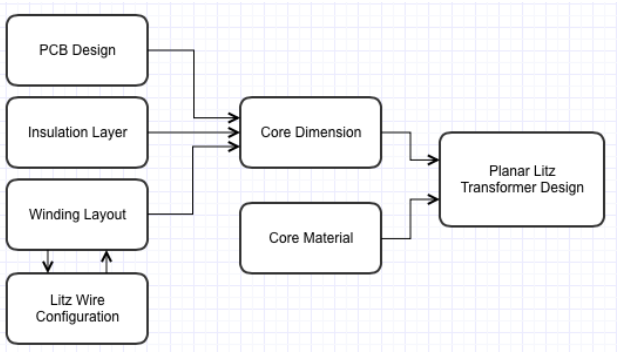


Figure 1. Design procedure of planar transformer

Figure 1 demonstrates the design procedure of planar transformer. Wire configuration depends on winding layout, while its double-sided nature alters conductor arrangement. Core dimension is determined by the design of windings, insulation layers and PCBs. Selection of commonly used materials for core is combined with the core dimension to produce the overall design of planar transformer.

2.2. Winding layout

Current flowing through conductors would cause power dissipation in the form of heat and temperature rise in PCBs. A symmetrical winding structure is used to minimise the effect of thermal expansion. The standard thickness of conductor is selected, which is 70 μm.

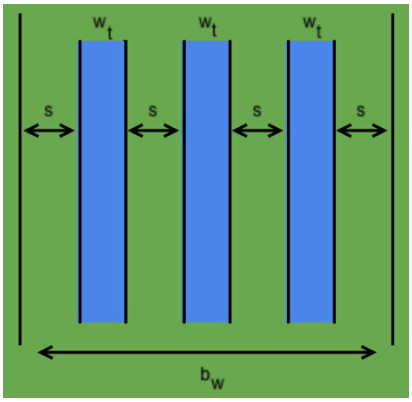


Figure 2. Winding structure

Figure 2 illustrates the fundamental design of winding structure. Based on the international safety standards, the minimum width between two adjacent windings is 150 μm. The labelled parameters in Figure 2 follow the equation:

$$w_t = \frac{b_w-(N_1+1)\cdot s}{N_1} \quad (1)$$

2.3. Strand layout

The length of strand varies as the strand angle changes. The relationship between average length of strands, l_{strand}, width of PCB, b_w, and strand angle, θ_{strand}, is given by the following inequality [2]:

$$l_{strand} \cdot \cos(\theta_{strand}) \leq b_w - 2 \times s \quad (2)$$

The electrical isolation requirements of strands in litz wire configuration is relatively low, because of the minimised potential difference between adjacent strands. Therefore, the space between strands is designed to be small in order to maximise the surface utilisation factor. Reduced gap would, however, make proximity effect significant. Therefore, there is a trade-off between surface utilisation factor and AC resistance in litz wire configuration.

2.4. Insulation layout

Plexiglass is one of the common materials used for insulating conductors. It has widespread utilisation when toughness and thermal conductivity are required.

2.5. PCB design

The standard thickness of a PCB is 1.6 mm, and it is applied in this design. The dimension of PCB is adjusted based on the industrial availability of core, particularly the window size of the core section.

2.6. Core material

The selection of core material is based on three critical characteristics: saturation flux density, permeability and resistivity. High saturation flux density is required to prevent core saturation under high operating frequency. High permeability alleviates the effect of leakage flux and fringing, so that magnetic flux is more concentrated in the core. High resistivity would reduce the power loss caused by eddy current.

There are two commonly used ferromagnetic materials in high frequency transformer, 3F3 and 3F4[3,4]. Both are suitable for medium frequency applications ranging from 10kHz to 2MHz [1]. Moderate permeability and saturation flux density also make them to be optimal in planar transformer.

The optimal operating frequency of two materials are different. 3F3 operates between 10kHz and 500kHz, and the saturation flux density is 370 mT at 100℃ [2]. 3F4 can operate at 2MHz, and its saturation flux is 350 mT at 100 ℃ [4]. Therefore, both materials are used in the design for a wide range of result.

2.7. Design summary

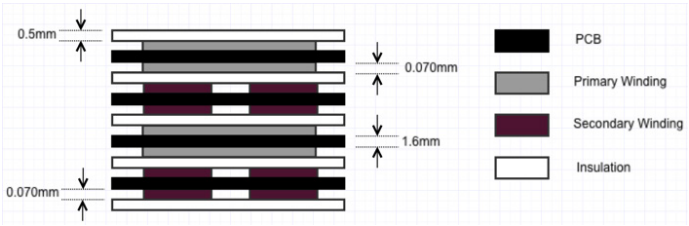


Figure 3. Cross-section of designed planar transformer

Figure 3 shows the cross section of the designed structure. The thicknesses of different layers are labelled based on the design. The total thickness is calculated as:

$$0.5 \times 5 + 1.6 \times 4 + 0.070 \times 8 = 9.46\ mm \quad (3)$$

The core dimension is selected based on the total thickness and the industrial availability, and E43/10/28 is selected. The winding configuration is set to be interleaved for the purpose of minimising the leakage inductance. The overall structure of the planar transformer is shown in Figure 4.

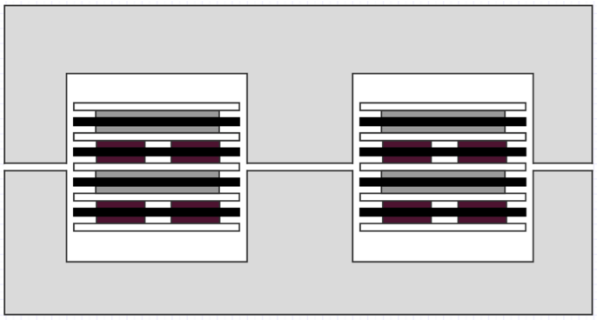


Figure 4. Cross section of overall design

3. Simulation of planar transformer.

3.1. Modelling of planar transformer in Maxwell 3D

The 3D structure of planar transformer is separated into two main parts: windings and core. Figure 5 shows the 3D structure of

windings, PCBs and insulation layers. All layers are assigned to different materials based on the theoretical design.

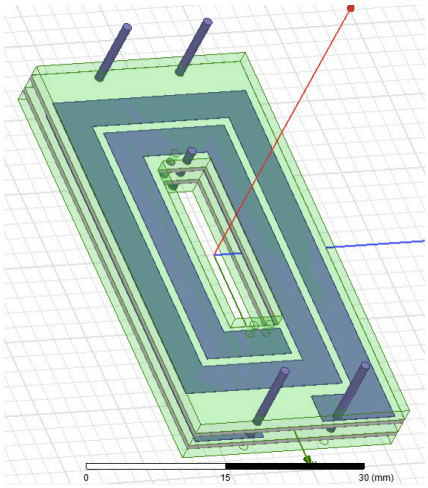


Figure 5. 3D model of windings and PCBs

Figure 6 shows the 3D modelling of core section. Materials are used accordingly to the operating frequency range. Two EE segments are modelled with small air gaps included for simulating real applications.

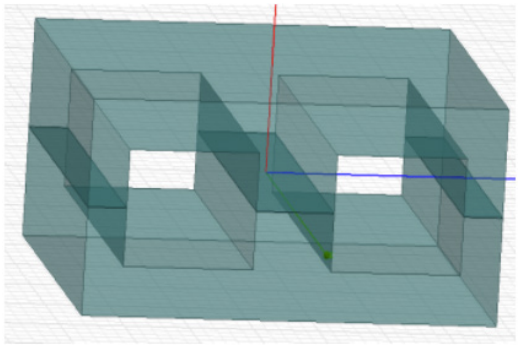


Figure 6. 3D model of core section

Figure 7 shows the overall model of planar transformer, and it is used for field analysis. Current density distribution and electrical parameters under different operating frequencies are evaluated.

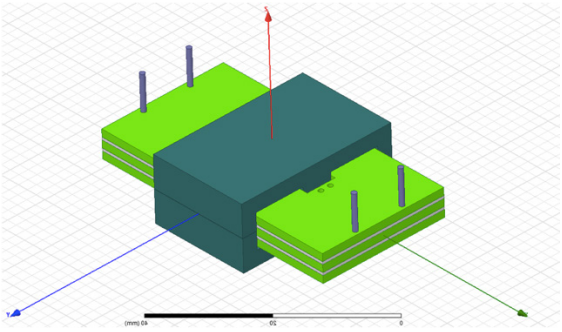


Figure 7. 3D model of planar transformer

3F3 and 3F4 are not standard materials in Maxwell 3D. Assumptions have been made to reduce the time of simulation. Relative permeability is assumed to be constant, and the operating temperature is fixed at 70 degrees for constant conductivity and permeability.

Maxwell 3D applies iteration method to solve mesh equations. The iteration convergence is set to be 5% for accurate result.

3.2. Litz wire 3D modelling

Compared to planar transformer using solid conductor, the liz wire winding structure is also modelled for comparing the overall performance.

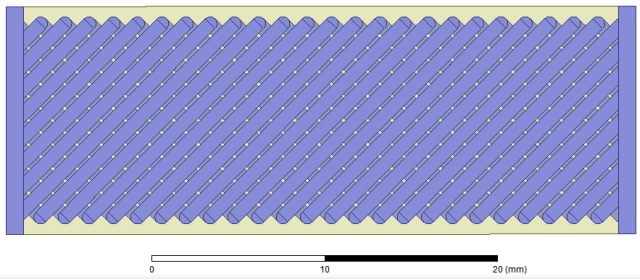


Figure 8. 3D model of litz wire configuration

Figure 8 shows the design of litz wire configuration. This winding structure could provide a relatively uniform current distribution across the conductor, which alleviates the skin effect.

According to the theory, with uniform current distributed across the cross-section of conductors, the AC resistance of litz wire is much smaller than that of transformer using solid conductor. This is, however, not always the case which is proved in the following simulation.

4. Field analysis

4.1. Simulation results

Figure 9 shows the electric field distribution in secondary winding. The copper loss of the winding can be calculated from the electric field strength and the conductivity of copper, and this loss is used as an intermediate result for iterative calculations.

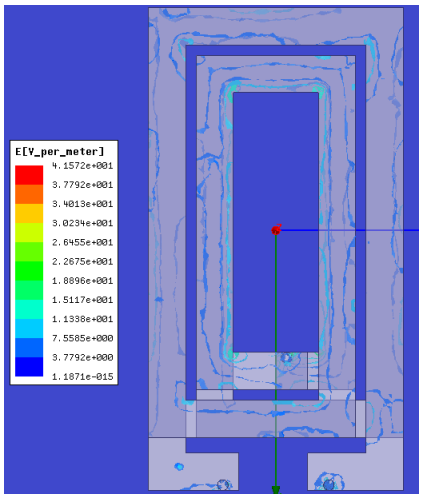


Figure 9. Electric field distribution

Figure 10 shows the fringing effect of air gap. This air gap is placed intentionally to simulate the non-perfect assembly of two EE ferrite segments. Large air gap is avoided for a concentrated magnetic flux density.

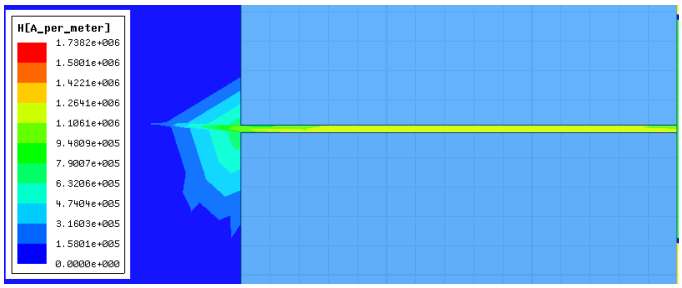


Figure 10. Magnetic field in air gap

Further simulation results, including current density distribution and AC resistance of both solid conductor and litz wire structure, can be found in [5].

4.2. Discussion

The AC resistance of litz wire is relatively smaller than that of solid conductor under medium frequency range. However, the proximity effect becomes significant under very high operating frequency compensating the benefits from the uniform current density distribution.

This proves that the application of litz wire structure is appropriate in a certain frequency range. Operating at frequency outside this range would result in low efficiency, and the solid conductor should be applied instead of litz wire.

5. Design of DC-DC converter

5.1. DC-DC schematic

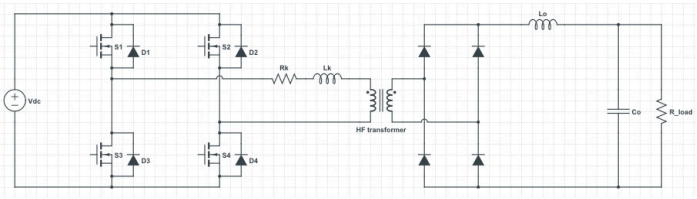


Figure 11. Circuit schematic of PSFB DC-DC converter

Figure 11 shows a PSFB DC-DC converter using a high frequency transformer installed between two DC stages. Possible switching on-states are listed as follow: S1 and S4, S1 and S2, S2 and S3, S3 and S4. The gating signals are shown in Figure 12.

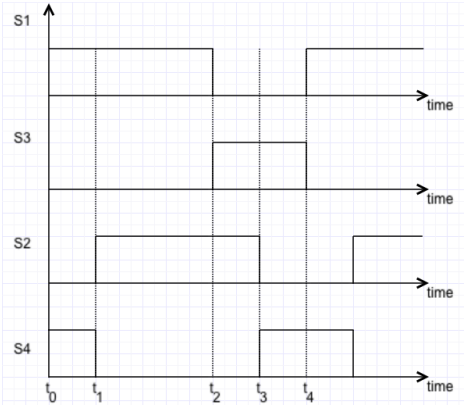


Figure 12. Gating signals

5.2. Derivation of transfer functions

The transfer functions are derived from a unified approach of modelling switching converter, the state space averaging method, which requires possible state functions of corresponding switching state [6].

Let $n = N_1/N_2$ be the turns ratio, where N_1 and N_2 are the number of turns on primary and secondary side; V_1, V_2, i_1 and i_2 be the voltages and currents of two sides respectively. Assume that electrical parameters follow transformer dot convention.

State 1: $t \in (t_0, t_1)$. Switches S1 and S4 are turned on, and current at primary side is drawn from voltage source and increases linearly. The circuit is shown in Figure 13.

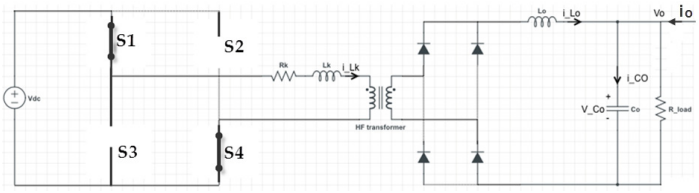


Figure 13. Circuit schematic of state one

$$i_{L_k} = \frac{1}{n} \cdot i_{L_0}, V_1 = n \cdot V_2 \tag{4}$$

$$V_1 = V_{dc} - R_k \cdot i_{L_k} - L_k \cdot \frac{di_{L_k}(t)}{dt} \tag{5}$$

$$V_0 = V_{C_0}(t) = V_2 - L_0 \cdot \frac{di_{L_0}(t)}{dt} \tag{6}$$

Therefore, the following state space matrices are derived:

$$\dot{x} = \begin{bmatrix} i_{L_0} \\ V_{C_0} \end{bmatrix} = \begin{bmatrix} -\frac{R_k}{n^2 \cdot L_0 + L_k} & -\frac{R_k}{n^2 \cdot L_0 + L_k} \\ \frac{1}{C_0} & -\frac{1}{RC_0} \end{bmatrix} x + \begin{bmatrix} \frac{R_k}{n^2 \cdot L_0 + L_k} \\ 0 \end{bmatrix} V_{dc} \tag{7}$$

$$y = V_{C_0}(t) = \begin{bmatrix} 0 \\ 1 \end{bmatrix} x, \tag{8}$$

where state-space parameters are:

$$A_1 = \begin{bmatrix} -\frac{R_k}{n^2 \cdot L_0 + L_k} & -\frac{R_k}{n^2 \cdot L_0 + L_k} \\ \frac{1}{C_0} & -\frac{1}{RC_0} \end{bmatrix}, b_1 = \begin{bmatrix} \frac{R_k}{n^2 \cdot L_0 + L_k} \\ 0 \end{bmatrix}, \ c_1^T = [0\ 1]$$

Applying this method to the other three states, with the assumption that $L_k \ll n^2 \cdot L_0$ and R_k is negligible, the average of state equations is derived for the corresponding time duration.

Let t_i *(i = 1, 2, 3, 4)* be the time duration for corresponding state with duty cycle $D = \frac{t_1 + t_3}{T}$ and $D' = \frac{t_2 + t_4}{T}$, where T is the switching period.

$$L_0 \cdot \frac{di_{L_0}(t)}{dt} = D \cdot \left[-V_{C_0}(t) + \frac{1}{n} \cdot V_{dc} \right] + D' \cdot [-V_{C_0}(t)] \tag{9}$$

$$C_0 \cdot \frac{dV_{C_0}(t)}{dt} = i_{L_0}(t) - \frac{1}{R} \cdot V_{C_0}(t) \tag{10}$$

$$y = V_{C_0}(t) \tag{11}$$

For small signal analysis, time-variant parameters are separated into two parts: static and small signal. Parameters with bar and hat are used to represent small-signal and static parts respectively.

By omitting DC and second-order small terms, small signal equations can be derived:

$$L_0 \cdot \frac{d\widehat{i_{L_0}(t)}}{dt} = -\widehat{v_{C_0}(t)} + \frac{1}{n}(\overline{D} \cdot \widehat{v_{dc}} + \hat{d} \cdot \overline{V_{dc}}) \quad (12)$$

$$C_0 \cdot \frac{d\widehat{v_{C_0}(t)}}{dt} = \widehat{i_{L_0}(t)} - \frac{1}{R} \cdot \widehat{v_{C_0}(t)} \quad (13)$$

$$\hat{y} = \widehat{v_{C_0}(t)} \quad (14)$$

By applying Laplace transformation to these time domain equations, three transfer functions are obtained:

$$G_{vd}(s) = \frac{\widehat{v_0(s)}}{\widehat{d(s)}}\Big|_{\widehat{v_{dc}(s)}=0,\widehat{i_0(s)}=0} = \frac{\overline{V_{dc}}}{n} \cdot \frac{R}{s^2RL_0C_0+sL_0+R} \quad (15)$$

$$G_{vi}(s) = \frac{\widehat{v_0(s)}}{\widehat{v_{dc}(s)}}\Big|_{\widehat{d(s)}=0,\widehat{i_0(s)}=0} = \frac{\overline{D}}{n} \cdot \frac{R}{s^2RL_0C_0+sL_0+R} \quad (16)$$

$$Z_{out}(s) = \frac{\widehat{v_0(s)}}{\widehat{i_0(s)}}\Big|_{\widehat{d(s)}=0,\widehat{v_{dc}(s)}=0} = \frac{sRL_0}{s^2RL_0C_0+sL_0+R} \quad (17)$$

With the assumption that output current and input DC voltage are constant, a AC small-signal model of DC-DC PSFB converter can be obtained using these transfer functions. The system schematic is shown in Figure 14.

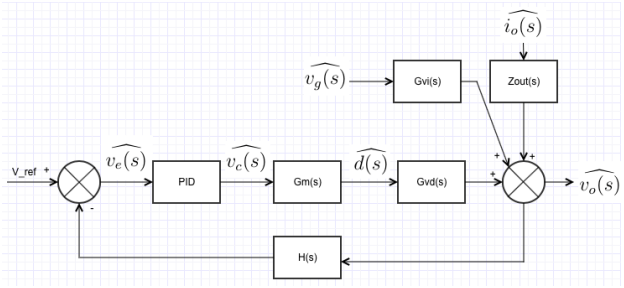


Figure 14. Small-signal model of PSFB DC-DC converter

5.3. Simulation in Simulink

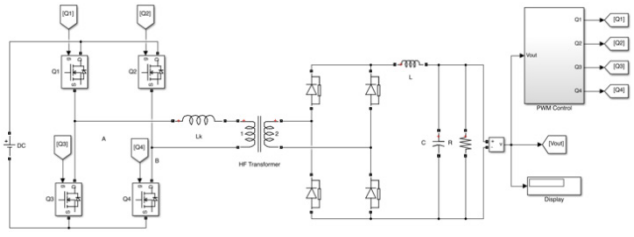


Figure 15. Circuit schematic of PSFB DC-DC converter

Figure 15 shows the circuit schematic of designed PSFB DC-DC converter. Assume that input voltage and output current are constant, the open loop transfer function is simplified into:

$$T(s) = G_{vd}(s) = \frac{\overline{V_{dc}}}{n} \cdot \frac{R}{s^2RL_0C_0+sL_0+R} \quad (18)$$

The design of feedback loop is shown in Figure 16, including inputs, function blocks and outputs:

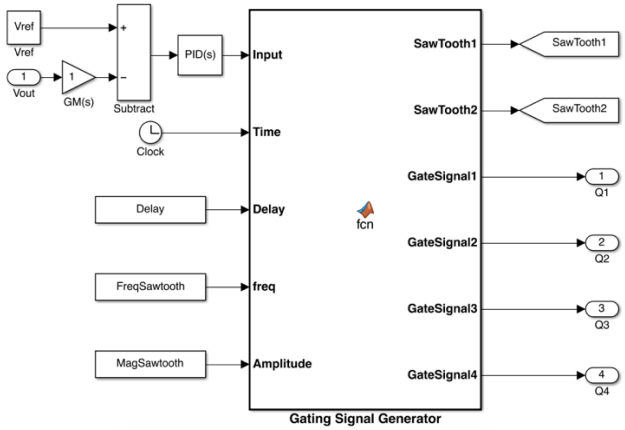


Figure 16. Feedback loop of DC-DC converter

Figure 17 shows the step response of the closed-loop transfer function, which is calculated from the open loop transfer function.

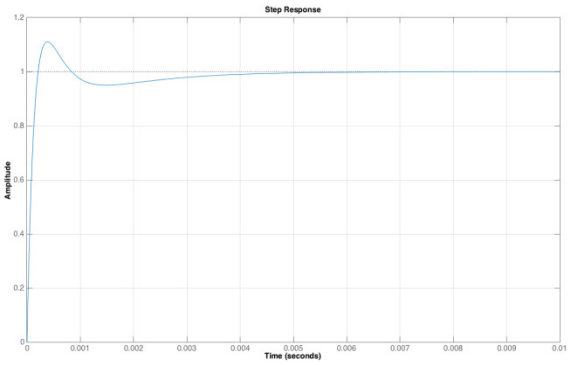


Figure 17. Step response of closed-loop transfer function

Further simulation results, including output voltage waveforms and total harmonic distortion result, are presented in [4]. The design and simulation results of HCC DC-AC inverter can also be found in [4].

5.4. Discussion

With the proper application of state space average method and design of feedback loop, both DC and AC voltages could stabilise at target values with small transient time. The total harmonic distortion of AC voltage is small illustrating a smooth AC waveform.

The phase shift between the converted AC signal and the reference signal from power grid is insignificant and negligible, which proves that this AC output signal can be connected to the power grid.

6. Conclusion

At high frequency range, the operating performance of planar transformer is primarily dependent on the frequency. Litz wire winding configuration is preferred under high frequency range, but its manufacturing complexity and cost need to be considered during the process of design.

In addition, by applying the state space averaging method with reasonable assumptions combined, an accurate open loop transfer function of PSFB DC-DC converter is derived. The DC output voltage of this converter can stabilise at required reference voltage within a short period of time. Additional DC-AC inverter can be easily installed to convert DC power into AC power at target amplitude and frequency for AC applications.

Further study could be done in the electromagnetic field analysis, including current density and magnetic flux density distribution. With more sophisticated devices used, a more comprehensive analysis could be performed and a more reasonable simulation result can be obtained.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

I would like to express my gratitude to the School of Electrical Engineering and Telecommunications UNSW for providing necessary resources to conduct this research.

I would also like to thank my family and friends who have supported and encouraged me throughout the time.

References

[1] Z. Ouyang and M.A. Andersen, "Overview of planar magnetic technology – fundamental properties", IEEE Trans. Power Electronics, vol. 29, No. 9, 2014.

[2] Shen Wang, Michael Andrew de Rooij, Willem Gerhardus Odendaal, Jacobus Daniel van Wyk, and Dushan Boroyevich, "Reduction of High-Frequency Conduction Losses Using a Planar Litz Structure", IEEE Transactions on Power Electronics, Vol. 20, No. 2, 2005, pp. 261-267.

[3] Ferroxcube, "3F3 material specification". Datasheet, 2004

[4] Ferroxcube, "3F4 materoal specification". Datasheet, 2004.

[5] C. Lian and D. Zhang, "Field-circuit analysis of planar transformer at medium frequency for converter application", IEEE Annual Southern, Power Electronics (SPEC), 2016.

[6] R.D. Niddlebrook and S. Cuk, "A general unified approach to modelling switching-converter power stages", Power Electronics Specialists Conference, pp. 18-34, June 1976.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 884-890 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Indoor monitoring in Public Buildings: workplace wellbeing and energy consumptions. An example of IoT for smart cities application

Stefania Nanni[1,*], Elisa Benetti[1], Gianluca Mazzini[2]

[1]*LepidaSpa, Viale Aldo Moro, 64, 40127 Bologna, Italy*

[2]*University of Ferrara, Via Savonarola, 9, 44121 Ferrara FE, Italy*

ARTICLE INFO

Article history:
Received: 05 April, 2017
Accepted: 12 May, 2017
Online: 20 June, 2017

Keywords:
Indoor monitoring
consumptions monitoring
Sensor integration
IoT
Smart Cities
environmental awareness
Workplaces wellbeing

ABSTRACT

In the recent years LepidaSpa created Sensornet platform for the control of the Emilia-Romagna territory managing heterogeneous environmental sensor networks distributed all over the region. LepidaSpa decided to integrate into Sensornet a brand-new family of sensors measuring indoor environmental parameters inside public buildings, in order to monitor energy consumption on one hand and workers' wellbeing on the other. The innovative aspects of the adopted solution (the extreme simplicity of the sensor network architecture, its ease of installation, the real time detection of parameters directly related to power consumption and the availability of environmental parameters measurements (temperature, humidity and light) simultaneously providing context information on the power consumption and correlation to working conditions) made it a perfect candidate for a seamless integration within the platform and for an intensive monitoring campaign. This paper is presenting the results of such monitoring activity, where an IoT infrastructure acts as an enabler of a Smart City solution, focused on a sustainable wellbeing of the working places.

1. Introduction

The concept of "smart building" is focused on scientific and technological development and social advancement of automated and integrated management of intelligent buildings systems. While the automation of new buildings, being able to be provided in the process of realization, has a negligible impact on the overall costs, the automation of already existing buildings often requires an expensive investment both on the infrastructure point of view and on the economic one. To these ends, in the latest case, which is clearly the most frequent for the Public Administration, the support of a system for monitoring power consumptions, being able to identify wastes and allow their reduction, is necessary, not necessarily through a control system, but through a more careful and aware use of the conditioning system by the users themselves.

The development of new architecture for radio links based on low bit-rate, long range and low power consumption, (LPWAN) and the presence of a sensor platform available to all the PA of the Emilia Romagna region for the collection of data (SensorNet), constitute the right prerequisites for the development of an indoor monitoring system for energy consumption in public buildings, without the need for infrastructure works on already existing buildings.

The above assumptions were the basis of the experiment conducted by LepidaSpa to demonstrate the generality and repeatability of the system and to make available to all PAs of the Emilia-Romagna region a tool for monitoring energy consumption, as well as economic and environmental benefits, derived from it.

This paper will explain and demonstrate the main innovative aspects of the proposed solution, such as the extreme simplicity of the network architecture used, the ease of installation of the indoor monitoring system, the simplification of the energy monitoring through the measurement of indirect parameters (temperature, humidity and light), the detection in real time of parameters related to the consumption, the possibility of immediate intervention on their reduction by a direct, conscious and spontaneous action of the users and its repeatability on a large scale.

Section II of this paper shows the evidence of the big interest on the theme of consumptions and of the limit of the traditional wireless technologies used until today for indoor monitoring systems. Section III explains the conceptual, architectural and functional model of the proposed solution. Section IV shows the technical and logistic features and constraints of the building used for the test. The analysis of the experimental data is presented in section V, and their correlation with wellbeing parameters is

*Stefania Nanni, LepidaSpa, Viale Aldo Moro, 64, 40127 Bologna, Italy
Email: stefania.nanni@lepida.it

provided in section VI. In section VII the consequences in Energy saving are provided. Conclusions are provided in Section VIII.

2. The State of The Art

Reducing energy consumptions, especially those related to electric energy, in buildings is a topic of big interest [1] as it relates to the problem of greenhouse gas emissions: the CO2 emitted by the latter each year is equal to 15% of emissions total worldwide (source World Research Institute). The legislation at EU and national level is increasingly stringent on energy containment in buildings, as evidenced by the past European Directive 2010/31 / EC and the more recent 2012/27 / EU. Until the beginning of 2012 every year the new buildings with high energy performance and existing ones improved, represented a share of 2,5% of the total housing stock, while on the remaining part of the buildings was not carried out any type of improvement [2].

The availability of new technologies is making a significant contribution to the creation of "smart buildings" and the achievement of the European objectives.

Both wired and wireless environmental monitoring systems have their own benefits over the other, as well as their own drawbacks. For the purpose and constraints of this research, dealing with buildings already existing, only wireless technologies will be treated.

The paradigm of sensor and actuator networks is based until now on short-range transmission technologies, often in the frequency band around 2.4 GHz, with multi-hop architecture, such as the ZigBee system [3]. In metering application another standard “de facto” protocol is the wireless M-BUS (169MHz / 868MHz), which requires, in general, the use, on top of a concentrator, also of one or more repeaters to extend the signal range radio peripheral nodes [8][9][12].

Currently however networks with wide geographic coverage operating in the band around the 868 MHz in Europe and 915 MHz in the USA are emerging. An example of this type of system is the one used by SIGFOX, an operator dedicated to the internet of things (IoT - Internet of Things), which, from France, is extending its coverage to Russia and other European countries, also including, from 2015, Italy. [10]

A technology that fits into this trend, but that can be used in private applications (that is, without necessarily requiring the presence of an operator) is the Lo-Ra™ of Semtech technology [4]. The Lo-Ra™ (Long Range) technology can be considered the actual innovation in the IoT world as it extends by a factor of ten the coverage range compared to the de facto standards, such as ZigBee, constituting an infrastructure enabler for new and different application areas, starting from that of monitoring of consumption.

3. The Proposed Solution

In an article that appeared in the chronicle of Bologna in early February 2016 [11] the virtuous and rather isolated example of the Sala Borsa was reported, highlighting its compliance of the legislation limits regarding the maximum temperature allowed inside public buildings. To be fair, the same article also reported the use of coats and scarves by users who were reading in the library. This example shows how the imposition of limits on the permissible temperature in public offices is not always easy to apply, both for technical and technological issues, but also for contour conditions such the time of permanence or the activity carried out in the offices, which can determine a different degree of perception of the temperature. Having said that, however, is crucial to have tools that can detect a possible misuse of energy resources in buildings, starting with those public, enabling energy savings just from the elimination of unnecessary waste, and that constitutes a concrete solution, effective and replicable on a large scale.

The proposed solution is a monitoring system, and not a control system, which aims at providing feedback to the buildings manager on the consumption. The conceptual model of the proposed system is based on a centralized control system of a distributed building stock that could bring significant energy savings with very low investment costs being primarily based on control, intelligence and awareness [5].

The technology Lo-Ra™ of Semtech allows monitoring very extensive building via wireless. It uses a star topology (in which the various nodes-sensors are connected directly to a hub-gateway node and is suitable for low traffic applications (such as monitoring). The advantages of the technology Lo-Ra™ [6] compared to existing technologies for sensor networks are:

- a range 10 times more extended, compared to the short-range technologies, thanks to a 198dB link budget (obtained via a modulation of the spread spectrum type) that allows for connections of up to 15 km in the sub-urban area, 2-5 km in densely populated urban areas and 2 km inside buildings;
- simplified network architecture that avoids the multi-hop and all related problems;
- low consumption of nodes (up to 10 years on battery power);
- very high network capacity: a single-concentrator gateway can connect thousands of nodes (up to 50,000)
- very low cost of the transmitters nodes.

All the above features make the Lo-Ra™ system an optimal solution, in the current scenario, for the implementation of energy monitoring systems of buildings because a single concentrator can cover one or more neighbouring buildings with a star architecture, which eliminates the problems associated with multi-hop (routing problems, latency, etc.) and because is characterized by high energy efficiency which can enable the duration of the batteries of nodes even for several years.

The monitoring system is based on a network of sensors installed in every building that transmit the indoor parameters to a central system in real time. These sensors are low-power and easily replaceable because of radio data transmission and because of being battery powered. The network of sensors, widely distributed within each building, allows the monitoring of the parameters, such as temperature, humidity and brightness, providing useful indications about the consumption resulting from the conditioning system. The transmission system, integrated on each sensor node, transfers the measured data to a single centralized point in real time. In the "control room" the parameter information is stored and processed on a server (remote intelligence). The data analysis module is able to detect in real time any conditions of energy waste, such as the exceeding of the reference temperature thresholds, different for summer and winter [7]. An automatic "warning" system, allows to report any abnormal conditions detected, by sending email or text message to the involved user, which will have to spontaneously provide to

an immediate regulation of local conditioning system and pay more attention on unnecessary consumptions. System web interface allows all users to view at any time and in real time the value of the parameters measured in their own office, favouring the acquisition of a greater awareness of consumption and use of energy resources.

4. The Test Bed

As a test bed has been chosen a building known and verifiable, the headquarters of LepidaSpa, located in a tower of the Bologna exhibition, divided on three floors with a total of thirty offices (Fig. 1, point A).

Figure 1: The Test Bed

In this building, as in all those of the same area, the conditioning system is centralized, with forced convection and independently adjustable in every office. The central conditioning system is switched on and off automatically on the working days from 6.00 to 19.00. The windows of the building are locked. The test bed consists of the installation of a sensor node in each office of all three floors. The lack of constraint on the presence of an electrical power point has allowed to select the most suitable installation point, as close as possible to the workers to measure their real comfort, but so it does not interfere with normal business activities, and also far away from sources of heat or exposure to direct sunlight. Each sensor node is equipped with a temperature, humidity and brightness sensor. The sampling period is 15 minutes, while the data transmission period one hour. In each transmission, therefore, the sensor nodes send 4 samples for each of the three measured parameters, for a total of 12 measures.

An extension of the experimentation has provided, in its terminal stage, the displacement of some sensors in some offices of the adjacent towers to those of LepidaSpa (points G, F and H of Fig. 1), which has allowed to verify the actual coverage of the data transmission system, as well as the monitoring of environmental parameters relating to different entities and buildings.

5. Experimental Data

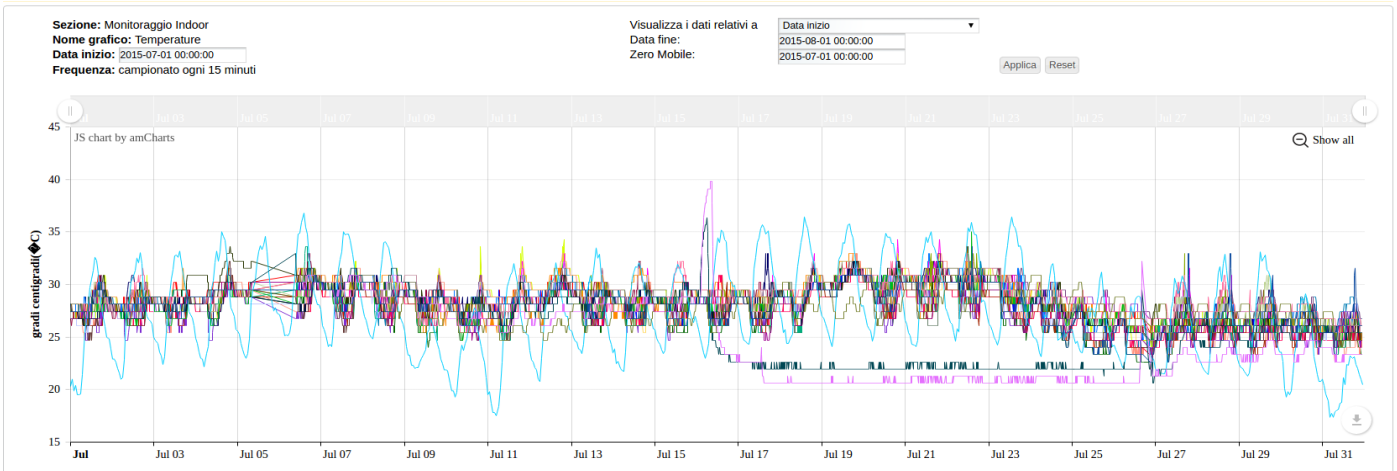
The monitoring has been focused on the detection of three environmental parameters inside the offices: temperature, humidity and brightness. The parameters were calculated automatically and continuously 24 hours a day, 7 days a week.

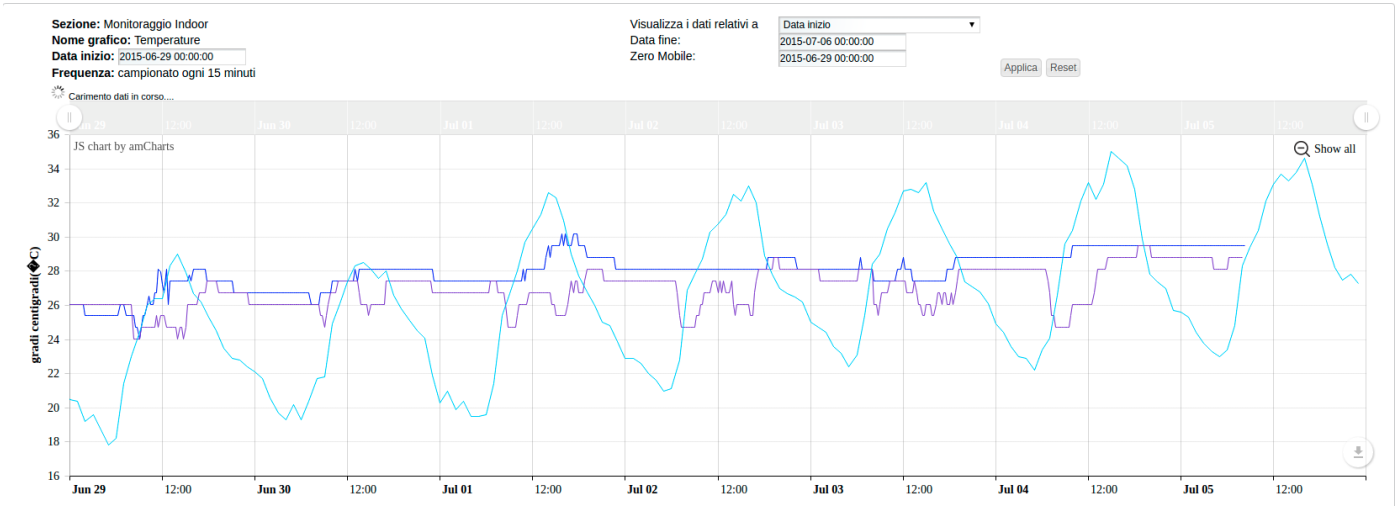
The monitoring period lasted a year, from May 2015 to April 2016, and it made possible, therefore, to monitor the use of the air conditioning system both in summer and in winter. Temperature measurements data, coming from the indoor monitoring system, were correlated with outside temperature measurements, coming from ARPA hydro-pluviometric monitoring system installed in the city of Bologna, to allow a deeper understanding in the reading of the data.

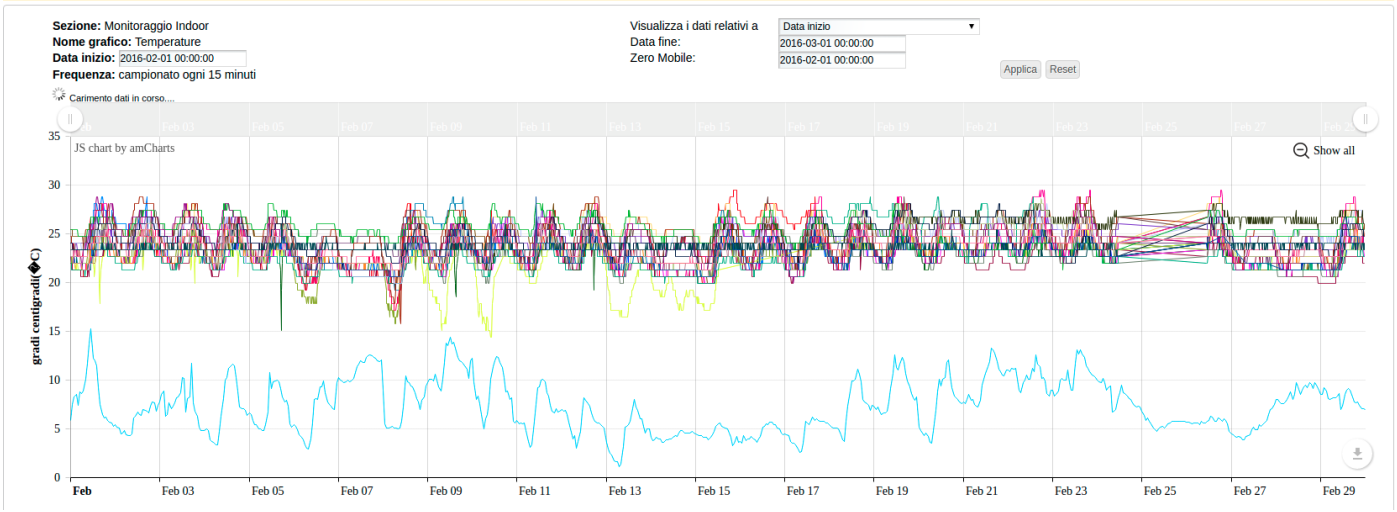
5.1. Temperature monitoring

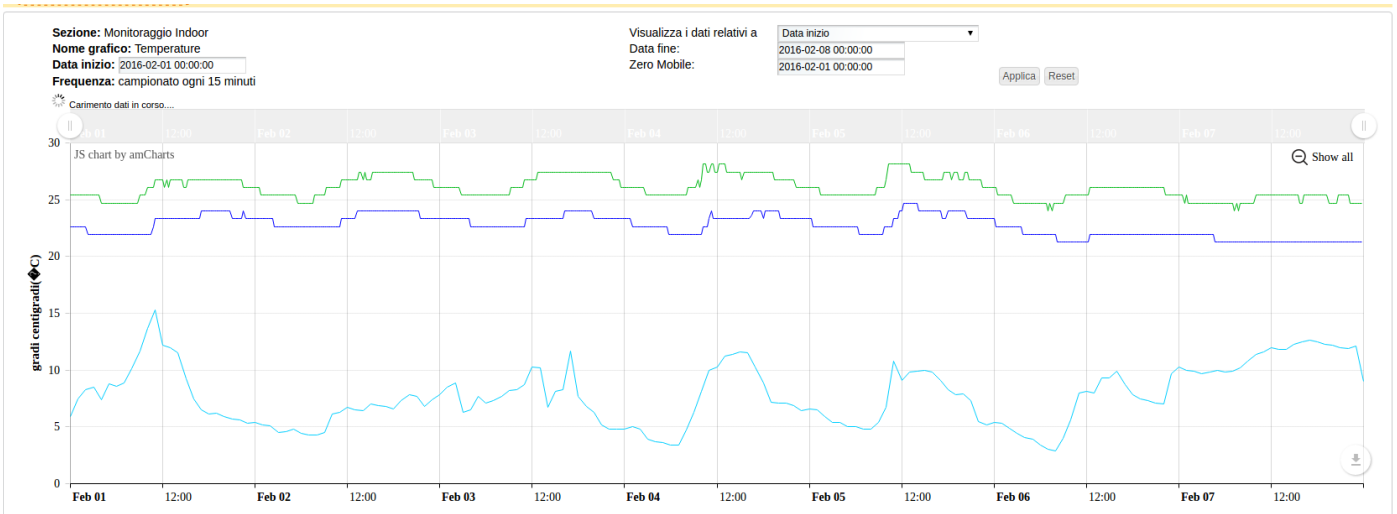
Fig. 2 shows the graphs of the temperatures in Celsius degrees measured by the sensors installed in the offices in July 2015. A correlation between the internal temperatures and the outside temperature (blue sinusoid) can be noticed, but also a dispersion of several degrees in the indoor temperature measurements during the on-time of conditioning systems.

In particular, Fig. 3 shows the comparison of the temperature trend detected in two offices in the period from 29/06/2015 to 05/07/2015. It's evident a rather different use of the air conditioning system on every working day, which it brings a difference of temperature up to a couple of degrees in the hottest hours between the two offices. Office users corresponding to the blue graph appear to have adopted the most efficient use of the conditioning system, since the temperature never drops below 26 degrees. From the chart it may also be noted that the air conditioning system in the office corresponding to the purple graph is always active regardless of the presence or absence of office staff, because it evident that it works when the offices are certainly closed.




Sezione: Monitoraggio Indoor
Nome grafico: Temperature
Data inizio: 2015-06-29 00:00:00
Frequenza: campionato ogni 15 minuti
Visualizza i dati relativi a
Data fine:
Zero Mobile:
Data inizio
2015-07-06 00:00:00
2015-06-29 00:00:00
Applica
Reset
Caricamento dati in corso....
JS chart by amCharts
Show all
gradi centigradi(°C)
Jun 29
Jun 30
Jul 01
Jul 02
Jul 03
Jul 04
Jul 05
12:00


Sezione: Monitoraggio Indoor
Nome grafico: Temperature
Data inizio: 2016-02-01 00:00:00
Frequenza: campionato ogni 15 minuti
Visualizza i dati relativi a
Data fine:
Zero Mobile:
Data inizio
2016-03-01 00:00:00
2016-02-01 00:00:00
Applica
Reset
Caricamento dati in corso....
JS chart by amCharts
Show all
gradi centigradi(°C)
Feb
Feb 03
Feb 05
Feb 07
Feb 09
Feb 11
Feb 13
Feb 15
Feb 17
Feb 19
Feb 21
Feb 23
Feb 25
Feb 27
Feb 29


Sezione: Monitoraggio Indoor
Nome grafico: Temperature
Data inizio: 2016-02-01 00:00:00
Frequenza: campionato ogni 15 minuti
Visualizza i dati relativi a
Data fine:
Zero Mobile:
Data inizio
2016-02-08 00:00:00
2016-02-01 00:00:00
Applica
Reset
Caricamento dati in corso....
JS chart by amCharts
Show all
gradi centigradi(°C)
Feb 01
Feb 02
Feb 03
Feb 04
Feb 05
Feb 06
Feb 07
12:00

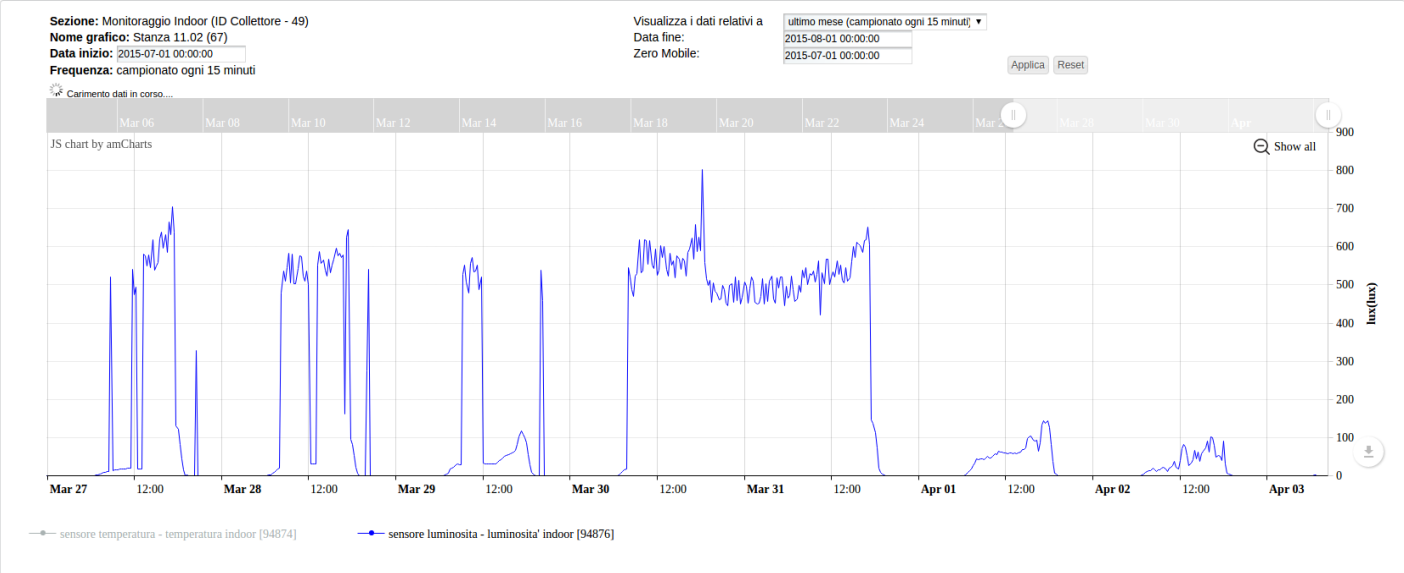


Fig. 4 corresponds to the graphs of the temperatures measured in the offices in February 2016. The variance of the measured values is very high, up to five degrees. The average temperature during working hours is about 25 degrees, with peaks up to over 28.

Fig. 5 shows the comparison between the temperature trend detected in two offices during the first week of February 2016 in which a quite different use of the air conditioning system, even more than in the summer, is evident in all the days of working week. In particular, an average difference of more than two degrees between the temperature of the two offices can be noticed: while the temperature represented by the blue line is always below 24 degrees, the one in the green line is always above 24 degrees, with a typical value of 27 degrees.

5.2. Lighting monitoring

The brightness sensor installed on each sensor node enabled a very interesting lighting monitoring. The graphs corresponding to the data detected by the brightness sensor (expressed in lux) easily highlight the alternation between day and night, as well as the use of artificial light in place of the natural one.

Fig.6 shows a chart related to the lighting monitoring in an office during a week, in which it is evident a generally correct use of the light during the working hours (light-on), the night and the weekend (light-off), except for an occurrence of light left on during the night.

The analysis carried out during one year of testing has shown that the oversights were few. As a matter of fact, only five offices resulted having left the lights on for a total of only 11 occurrences in a year.

5.3. More data analysis

The graphs of the light sensors, as well as main information analysed in the previous paragraph, can provide some additional one, useful in the control and management of buildings and personnel occupying them. In Fig.6, for example, the entry and exit time of the employee sitting in the office can be easily distinguished as well as the proper use of the lights even in the lunch break, during which are kept switched off. The same graph also shows a typical reactivation of lights around 8PM, linked to the presence of the cleaning personnel, authorized to operate only after the working time

6. Conditions of Health and Wellness within the Offices

Environmental parameters monitoring inside the offices are also providing useful information about health and wellness conditions inside workplaces.

As a matter of fact, among the aspects considered in the EHS guidelines for the workplaces and environmental protection in working activities [13], on top of indoor air quality, passive smoking and noise, it is also possible to find indicators on microclimate and lighting matching the data analysis on the environmental monitoring.

6.1. Microclimate

The range of physical and climatic parameters (temperature, relative humidity, air speed) in a confined environment is called microclimate. Thermal wellbeing depends on the performed activity, on the clothing and on individual sensations that can be influenced by environmental parameters such as temperature, relative humidity and air speed. Besides, the spread of microbiological contaminants in a confined area is influenced and enabled by other factors such as temperatures above 26°C and Relative Humidity above 65%.

PERIOD	AIR TEMP.	recommended value	R.H.
Summer	19-24	22	40-60
Winter	17.5 – 21.5	19.5	40-60

Table 1: thermal comfort zone in sedentary working conditions and seasonal clothing

From the data analysis performed, it is possible to identify a significant overpass of the recommended temperature during the months when study was conducted, but it is also possible to note an overall respect of the recommended values for relative humidity, both in winter and in summer, with some isolated cases in the summertime (Fig. 7, Fig. 8).

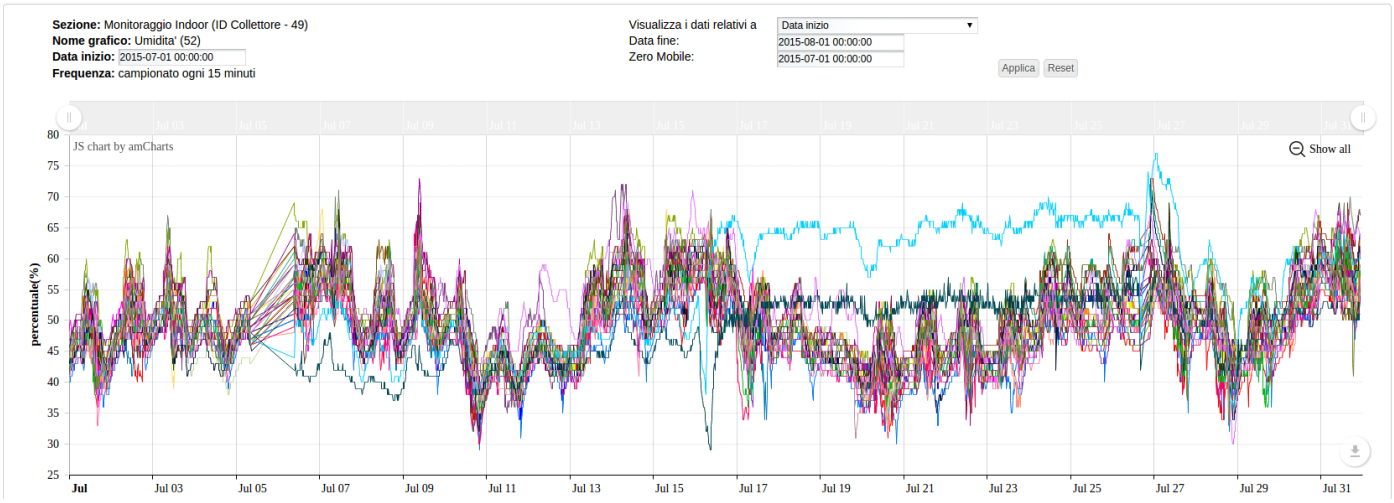


Figure 7: Humidity in the offices in July 2015

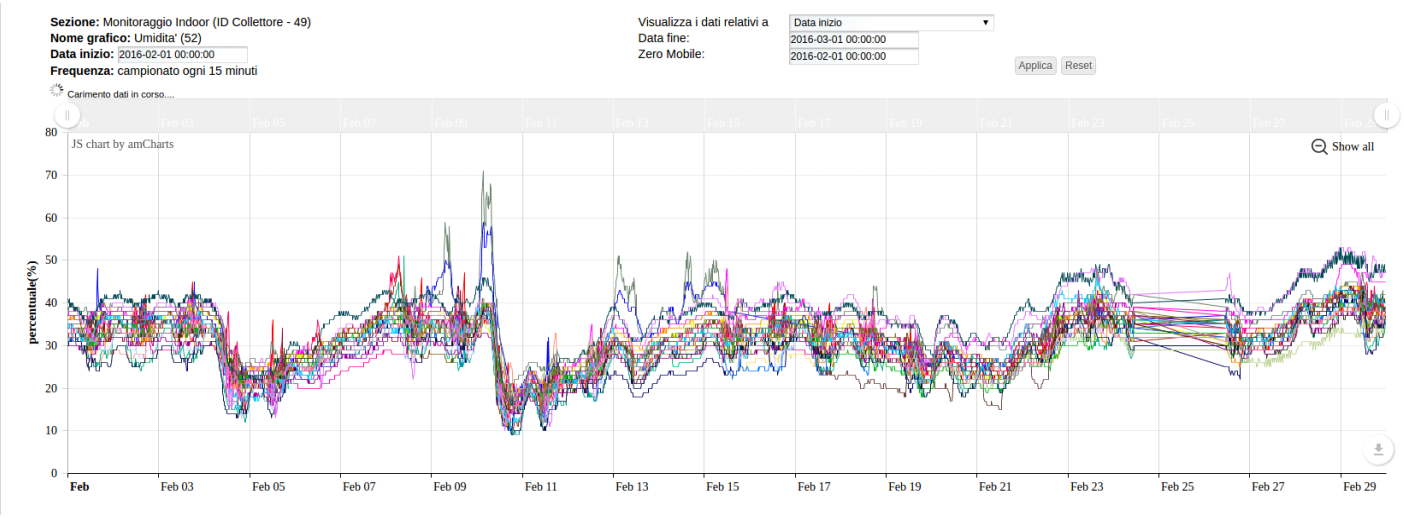


Figure 8: Humidity in the offices in February 2016

6.2. Lighting

The vast majority of the sensorial stimuli in the offices are visual, and eyes are evidently one of the most stressed organs. Natural light, although fundamental, is not sufficient alone to grant an adequate lighting in a confined area, since it is subject to a number of factors, among which the time of the day, the orientation of the room, the weather conditions, etc. In case of insufficient natural lighting, it is therefore necessary to provide a correct artificial lighting. There are several parameters defining a bright room, such as distribution and direction of the light sources, dazzling, etc… as an example, comparing in a table the measured parameters with minimum indoor requirements, it is possible to conclude that offices are correctly lightened (Fig.6).

UFFICI	(ILLUMINAMENTO) LUX
Copy room	300
Structure	500
Data centre	500
Technical Drawings	750

Table 2: Fonte normative: UNI EN 1264-1/2001

7. Energy Saving

Once the power of the light sources is known, the saving that can be achieved if lights are not turned on when nobody's around can be easily calculated

The saving that can be achieved if the room temperatures is lowered by 1°C is much less straightforward to calculate, and even less known. It is reported that each degree Celsius less in the room temperatures is allowing a 7% annual saving on heating costs. Under this assumption, the annual saving can be calculated as

Taking into account that 1 KWh of electricity equals 0,72 Kg of CO2, it is possible to evaluate the quantity of carbon dioxide that is not emitted

These simple formulas tell us that a significant saving is possible even with small actions and good practices that can also

be learnt by means of an awareness system such the one depicted highlighting the waste, misuse or dangerous use of resources.

8. Conclusions

The results of the experiment have shown the effectiveness of the energy monitoring system developed by LepidaSpa, highlighting that there are wide opportunities for savings coming from a more rational use of the heating and air conditioning system inside of LepidaSpa offices. In particular, while it has overall demonstrated a correct use of the lights, it has shown an incorrect use of the heating system in the winter period, detecting a much higher temperature than the allowed limits and much more than any reasonable margin of discretion.

Another important result was the verification of the radio coverage provided by the transmission system. Fig. 1 shows the positioning of the gateway (point A) and the different points in which the coverage has been verified (points B, C, D, E, F, G, H). The only area not covered consists of the E tower, inside which the received signal has not been sufficient to ensure reliable communications. The cause is probably to be identified in the C tower and in the interference caused by its height and its position between the extreme points A and E.

The results have, indeed, demonstrated the adequacy of the new technologies adopted, Lo-Ra™, which allowed the realization of unprecedented solutions and constitutes an enabling element for the IoT paradigm.

The solution illustrated is a real example of "Smart city" application not only because it provides a value-added service for the Public Administration through the use of new technology [13], but also because it has been made by the reuse of the already existing Sensornet platform [14].

The results of the experiment have finally demonstrated the generality and the possibility of large-scale implementation of the system, which can be adopted since now by all PAs of the Emilia-Romagna region interested in the implementation of energy savings policies through the detection of waste within their buildings, as well as of wellbeing conditions in their offices.

References

[1] Chen, Han, et al. "The design and implementation of a smart building control system. "e-Business Engineering, 2009. ICEBE'09. IEEE International Conference on. IEEE, 2009.

[2] "Monitoraggio e modellazione innovativa di consumi elettrici in smart building", M. Annuziato, C. Meloni, S. Pizzuti (ENEA), F. Lauro (Università degli sutdi della Calabria), Novembre 2011.

[3] Akyildiz, Ian F., Xudong Wang, and Weilin Wang. "Wireless mesh networks: a survey. "Computer networks 47.4 (2005): 445-487.

[4] http://www.semtech.com/wireless-rf/iot.html.

[5] Tuan Anh Nguyen, Marco Aiello, Energy intelligent buildings based on user activity: A survey, Energy and Buildings, Volume 56, January 2013, Pages 244-257, ISSN 0378-7788

[6] http://www.semtech.com/images/mediacenter/collateral/ism_sg.pdf

[7] E. Benetti, S. Nanni, G. Mazzini, "Indoor Monitoring: SensorNet Integration and Actuations for Energy Saving". Submitted to Software, Telecommunications and Computer Networks (SoftCOM), 2016 24th International Conference on. IEEE.

[8] http://www.telecomitalia.com/content/dam/telecomitalia/it/archivio/documenti/Innovazione/NotiziarioTecnico/2013/n3-2013/07.pdf

[9] Spinsante, Susanna, et al. "Wireless m-bus sensor networks for smart water grids: analysis and results. "International Journal of Distributed Sensor Networks 2014 (2014).

[10] Bhagwat, Pravin. "Bluetooth: technology for short-range wireless apps. "IEEE Internet Computing 5.3 (2001): 96-103.

[11] http://bologna.repubblica.it/cronaca/2016/02/04/news/termosifoni_bassi_contro_lo_smog_ma_gli_uffici_pubblici_non_si_adeguano-132657870/

[12] Squartini, Stefano, et al. "Wireless M-Bus sensor nodes in smart water grids: the energy issue. "*Intelligent Control and Information Processing (ICICIP), 2013 Fourth International Conference on*. IEEE, 2013.

[13] Carlo Alberto Rosini, Davide Barbieri, Tiziana Chiara, Dante Milani, "Sicurezza sul lavoro e tutela dell'ambiente nelle attività di ufficio", Area Ambiente e Sicurezza.

[14] NANNI, Stefania; MAZZINI, Gianluca. From the Smart City to the Smart Community, Model and Architecture of a Real Project: SensorNet. Journal of Communications Software & Systems, 2014, 10.3.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 891-899 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

EEG Mind Controlled Smart Prosthetic Arm – A Comprehensive Study

Taha Beyrouthy*, Samer Al Kork, Joe Akl Korbane, Mohamed Abouelela

American University of the Middle East, Egaila, Kuwait

ARTICLE INFO

Article history:
Received: 05 April, 2017
Accepted: 11 June, 2017
Online: 26 June, 2017

Keywords:
EEG
Prosthetic arm
Sensors
Mind-controlled
Robotics
Emotiv
Signal processing

ABSTRACT

Recently, the field of prosthetics has seen many accomplishments especially with the integration of technological advancements. In this paper, different arm types (robotic, surgical, bionic, prosthetic and static) are analyzed in terms of resistance, usage, flexibility, cost and potential. Most of these techniques have some problems; they are extremely expensive, hard to install and maintain and may require surgery. Therefore, our work introduces the initial design of an EEG mind controlled smart prosthetic arm. The arm is controlled by the brain commands, obtained from an electroencephalography (EEG) headset, and equipped with a network of smart sensors and actuators that give the patient intelligent feedback about the surrounding environment and the object in contact. This network provides the arm with normal hand functionality, smart reflexes and smooth movements. Various types of sensors are used including temperature, pressure, ultrasonic proximity sensors, accelerometers, potentiometers, strain gauges and gyroscopes. The arm is completely 3D printed built from various lightweight and high strength materials that can handle high impacts and fragile elements as well. Our project requires the use of nine servomotors installed at different places in the arm. Therefore, the static and dynamic modes of servomotors are analyzed. The total cost of the project is estimated to be relatively cheap compared to other previously built arms. Many scenarios are analyzed corresponding to the actions that the prosthetic arm can perform, and an algorithm is created to match these scenarios. Experimental results show that the proposed EEG Mind-controlled Arm is a promising alternative for current solutions that require invasive and expensive surgical procedures.

1. Introduction

In recent studies, the World Health Organization (WHO) reported that about 15% of the world's population suffers from a form of disability, half of which cannot afford health care [1]. Due to various political, economic, scientific, and demographical reasons, the overall rates of amputees and limb dysfunction patients are increasing [2]. There are over 10 million amputees worldwide, out of which 30% are arm amputees [3]. Although prosthetic limbs exist since decades, they are not very natural in terms of operation and interaction with the environment. They require undergoing an invasive surgical procedure [4]. The main goal of such complex procedures is to reassign nerves and allow amputees to control their prosthetic devices by merely thinking about the action they want to perform as stated by the John Hopkins Applied Physics Laboratory.

A team at the John Hopkins University developed a robotic arm in their physics lab that is controlled by brain signals [5]. The arm has 26 joints, and can lift up to 45 pounds. It is based on an idea, which can be altered to fit any need, from someone missing just a hand to an entire limb. Engineers from Ossur, one of the biggest bionic arms markets in the world, designed an artificial bionic advanced leg and it was tested on a farmer who lost a leg [6]. A comparison was made between the bionic and the mechanical artificial legs, in which the engineers noticed some flaws in the mechanical leg, and thus causing the user to fall while walking if wrong amount of pressure was applied to the toe where the leg is

*Taha Beyrouthy, American University of the Middle East, +965 2225 1400 Ext. 1134, Email: taha.beyrouthy@aum.edu.kw

connected. The bionic leg, on the other hand, was constructed out of more than one piece. The knee, the foot and the leg were assembled to create a fully functional bionic limb [7]. This limb has a brain of its own, and can sense what surrounds it by the processors that analyze the inputs. According to the head of this group of engineers, the leg costs more than a few sedan cars.

Prosthetic limbs need to be measured and fitted to the patient for his needs [8]. To apply prosthetics on a patient, intense medical observation and a training course for the patient are needed so that he can use the limb comfortably. There are several techniques used as a means of controlling robotics arms, and the top three methods are highlighted hereafter. The first method is to use an electroencephalogram (EEG) device [9], which will record the person's brain waves when he is thinking of a certain action or implementing a facial expression. These readings are then converted to commands for the arm. The author in [10] states that the mind regulates its activities by electric waves registered in the brain that emits electrochemical impulses having different frequencies, which can be registered by an electroencephalogram. For instance, beta waves are emitted when a person feels nervous or afraid with frequencies ranging from 13 to 60 Hertz. Alpha waves are emitted when a person feels relaxed mentally and physically with frequencies from 7 to 13 Hertz. On the other hand, delta waves are emitted when a person is in a state of unconsciousness. The advancement in technology made it possible to process these EEG frequencies and data directly in real time by the use of a brain-computer interface which is a combination of hardware and software.

The second method is the surgical implantation. The arm is surgically connected to the person's torso. Connections are also made to the nerves to allow the reading of electrical signals so that these signals can then be filtered and converted to commands. The last control method consists in using sensors, which will be connected to the robotic arm in order to take specific readings. Some of the most common sensors used in this case are EMG, gyroscope, and accelerometer sensors. This will allow the user to be aware of the position his arm can be in as well as expand and enclose it. All the arm types used are summarized and compared with each other in Table 1 hereafter in terms of usage, flexibility, cost potential.

Table 1. Arm Types comparison

Arm Type	Usage	Flexibility	Potential
Robotic	Using mounts	Depends on the motors, material and design	Big potential with the rise of 3D printers
Surgical	Surgically Implanted	Depends on the training and the physiotherapy	Very high especially for the war victims
Bionic	Very effective for everyday usage	Depends on how well can the processor analyze the surroundings	Good potential, especially with the microprocessor revolution
Prosthetic	Similar to the surgical limbs, but less medical attention	Very high flexibility (custom made for each person using exact measurements)	Overshadowed by the surgical and the robotic limbs
Static	No mechanical use	Not flexible	No potential

The control techniques used are summarized and compared with each other in Table 2 hereafter in terms of cost, installation, degree of control and accuracy.

Table 2. Control techniques comparison

Type	Approximate Cost (U.S. Dollars)	Installation	Degree of Control	Accuracy
EEG	100 – 400	Detachable	Complete control	Accurate
Surgical	10,000 – 120,000	Permanent	Complete control	Very accurate
Sensors	Below 100	Detachable	Limited	Accurate

The EEG method is not only cost effective, but it is also accurate and gives the patient complete control of the arm. It also gives the user the luxury of taking it off when feeling discomfort. EEG is a noninvasive method of monitoring brain activity. Typically, it uses electrodes placed on the outside of the head, and measures voltage oscillations in the neurons of the brain caused by ionic current. It has been used in medical applications for a very long time. The Emotiv EPOC is an example of an EEG headset with 14 sensors and having an internal sampling rate of 2048 Hz. After filtering the signals, it sends the data to the computer at approximately 128 Hz. The signals are transferred from the headset to the computer through wireless technology. This offers much greater mobility, and instead of requiring a special gel, the electrodes of the EPOC simply need to be dampened using a saline solution that in disinfectant and common.

The project presented in this paper aims to develop a low-cost and versatile human-like prosthetic arm controllable via brain activity using EEG neuro-feedback technology. The arm is equipped with a network of smart sensors and actuators that give the patient intelligent feedback about the surrounding environment and the object in contact. It also allows the arm to react and execute pre-programmed series of actions in critical cases (extremely hot or fragile objects, etc.) A first prototype has been developed to test the prosthetic arm with the embedded electro-mechanical system. This prototype is controlled using flex sensors integrated within a wearable glove. A microcontroller is added to the system, thus allowing to perform programmed actions and tasks. The arm is created using a 3-D printer in order for it to be cost efficient. All parts are printed separately then assembled together. This prototype focuses on the arm-environment interaction. A second prototype based on the EEG

control has also been developed and still under test. Preliminary experimental results show that the EEG technique is a promising and good alternative to other existing techniques.

2. System Architecture

The proposed system is divided into 4 major units [11] as shown in Figure 1:

1- The Input Unit – EEG sensors
2- The Processing Unit – Pattern recognition
3- The Electro-Mechanical Unit - The arm
4- The Interface Unit – Smart sensor network

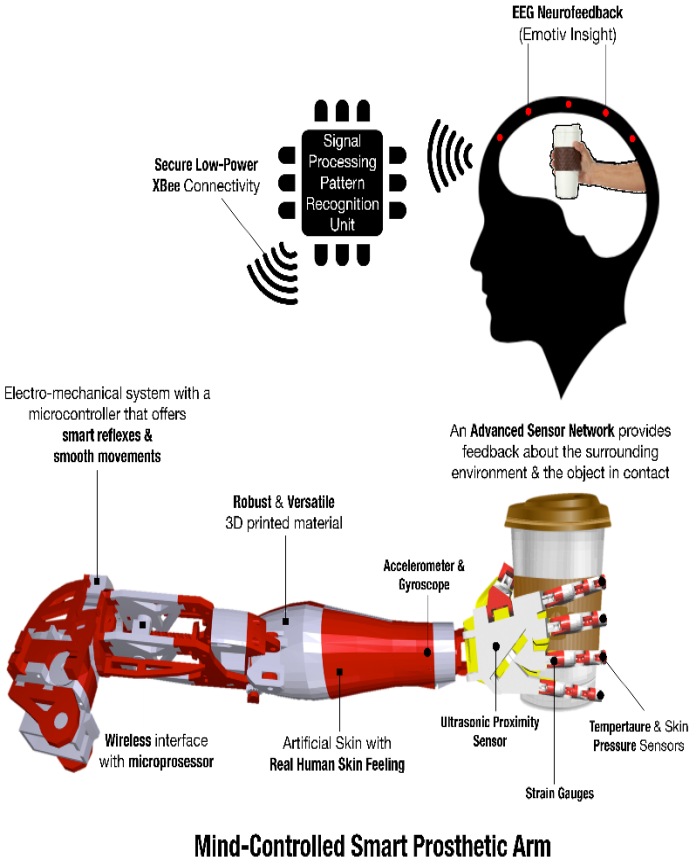


Figure 1. Mind-controlled smart prosthetic arm architecture

2.1. Input Unit

In this unit, brain signals were captured by an array of advanced EEG sensors communicating with a Signal Processing Unit via low-power and secure connectivity using Bluetooth technology. This device has an internal sampling rate of 2048Hz and 14 sensors arranged according to the international 10-20 System as shown in Figure 2 in order to cover the most relevant area over the brain. This allows a maximum and efficient coverage of the brain activity. EEG signals are acquired using the Emotiv EPOC wireless recording headset bearing 14 channels (AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8), referenced to the common mode sense [12].

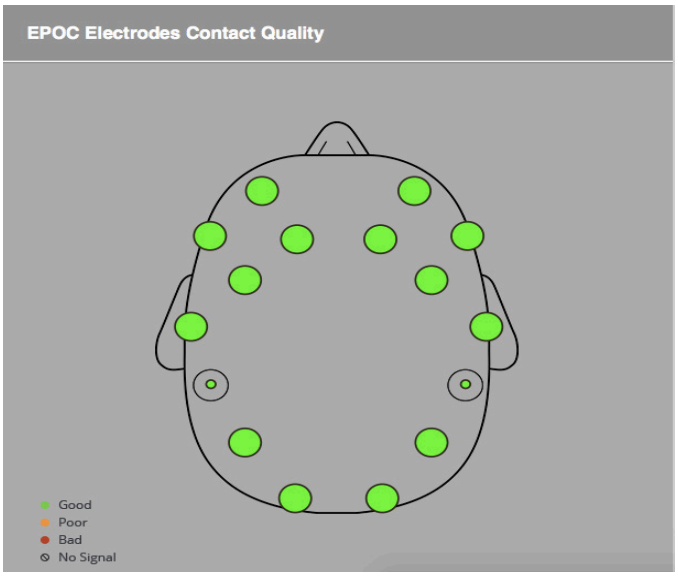


Figure 2. EEG 14-sensor arrangement

2.2. Processing Unit

The EEG signals provided by the input unit were sampled and processed on a lightweight wearable device – the Processing Unit. The processing activity consists of two main parts: a pattern recognition part that identifies different brain behavior captured by the input unit, and a command part that generates a series of commands to be sent to the mechatronics system of the arm.

This unit was programmed to distinguish between several states of the mind representing different levels of "meditation" and "focus". Every mind state was captured and encoded to represent a set of desired tasks to be performed by the arm. Due to the diversity and the complexity of brain wave activities among different humans, machine-learning techniques were required to train patients to specific arm movements according to a set of mind states.

2.3. Electro-mechanical Unit

This unit was designed and built from various lightweight high-strength materials that can handle high impacts and fragile elements as well. It integrated servos capable of handling 800 oz.-in. of stall torque. These servos were strategically placed to minimize hardware and facilitate complex moves. A microcontroller was also integrated to this setup to provide the interface between the Mechanical Unit and the Processing Unit. It can also be programmed to perform a series of predefined movements, allowing the arm to have a sophisticated and realistic real hand behavior.

2.4. Interface Unit – Smart Sensor Network

This unit is composed of a network of smart sensors, including temperature, skin pressure and ultrasonic proximity sensors, accelerometers, potentiometers, strain gauges and gyroscopes. The main features of this unit allow the arm to interact with and adapt to the surrounding environment. Moreover, a bi-directional communication was required to give commands to the arm and provides feedback to the patient. This integrated network of sensors and actuators required custom communication protocols and networking techniques that allow seamless interaction and control between the arm and patient. By default, controlling the

arm was handled by the brain (patient); however, it can be transferred to the arm to proactively protect itself against damage.

Due to its unique features, the proposed Mind-controlled Smart Prosthetic Arm should be able to improve the quality of life for millions of patients and their families around the world. Its low cost design makes it accessible for a wide range of beneficiaries, especially those with limited or no access to advanced health care.

3. Technical Specifications

3.1. Hardware Overview

The Smart Prosthetic Arm architecture was based on an advanced electro-mechanical system controlled by EEG neuro-feedback technology. This architecture is divided into 4 main units in terms of hardware (see Figure 3 and Figure 4) [13]:

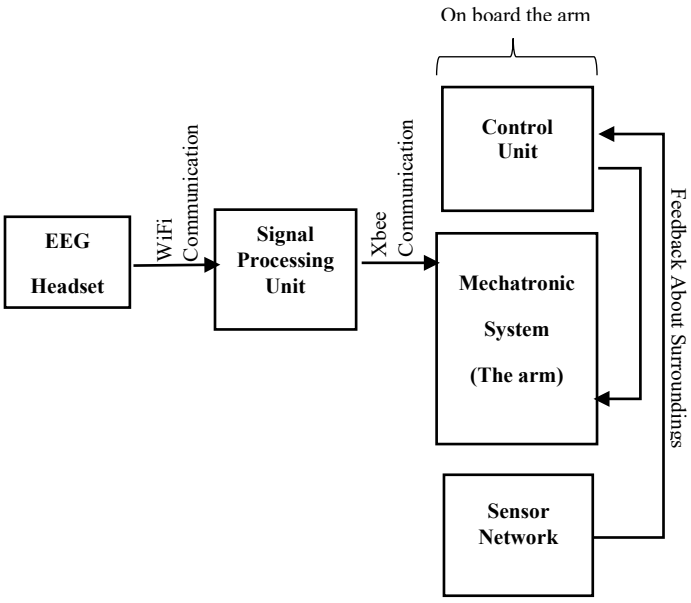


Figure 3. System overview – block diagram

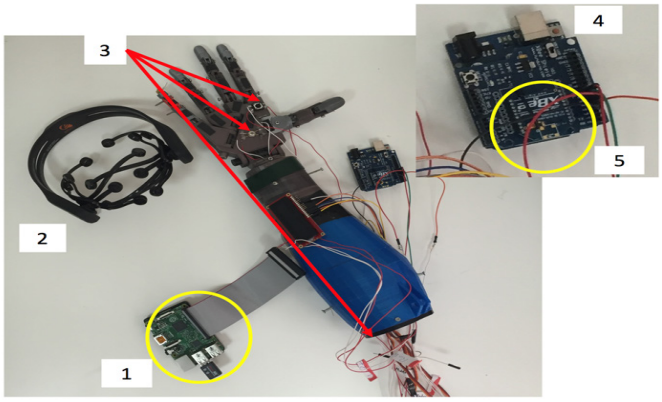


Figure 4. System overview – photo of the actual components

As shown in Figure 4, component number 1 is the processing unit that receives the EEG waves from the headset and communicates with the control unit; component number 2 is the Emotiv EEG headset with 14 sensors; component number 3 corresponds to the embedded sensors (temperature, proximity and touch); component number 4 is the control unit, which is Arduino based; component number 5 is the XBee driver (wireless communication with the processing unit).

The Data Sampling Unit includes 14 EEG sensors implemented in the EEG-based headset. This unit converts EEG signals to digital signals and send them to the processing unit through Bluetooth communication channel.

The Wireless Communication Unit includes low power connection Bluetooth module interfaced with the Raspberry III microcomputer. This part implements a communication protocol between the user (EEG sensors installed on the head) and the embedded microcontroller in the prosthetic arm.

The Computing and Processing Unit: Sampled EEG signals are channeled through the wireless communication unit to reach the Computing and Processing Unit, which is embedded in the arm and includes a Raspberry Pi III microcomputer, interfaced with an Arduino Mega microcontroller that handle the mechanical servo units installed in the arm. The main function of the processing unit is to treat the digitized EEG signals. It is programmed to compare between the headset reading and a set of premeasured patterns related to different states of the mind. Different mind states are captured and encoded to represent a set of desired tasks to be performed by the arm. This Unit is also programmed with multiple hand reflexes, triggered by the smart sensor network embedded in the arm. It gives the arm a human-like behavior via smooth movements and smart reflexes.

Finally, the Electro-Mechanical Unit includes nine servo motors installed on the 3D designed model. The 3D hand model is built from various lightweight high-strength materials that can handle high impacts and fragile elements as well. This unit integrates all servos, which are capable of handling 800 oz.-in. of stall torque. It is embedded in the arm and links servos to joints to perform different motions. The servos are strategically placed to minimize hardware and facilitate complex moves. A microcontroller is also integrated to this setup to provide the interface between the Mechanical Unit and the Processing Unit.

One of the main features of this arm is its affordability, which is accomplished mainly due to the fact that the arm is entirely 3D printed, which is one of the most affordable, yet durable ways of manufacturing such a product in this day.

For the process of building, different materials were used to create a full arm. The usage of different materials provided a more practical build.

The hand was created using a material called EcoPLA, which is PLA mixed with certain chemicals to provide a sense for heat recognition. This material can change color when exposed to heat. The added sensors provide visual aid to the user of the arm so that the user can estimate the heat of the held item. EcoPLA is considered to be a clean material, mainly extruded from starch, corn, and sugar cane, which makes it more environmental friendly. For the mechanical moving parts that need better endurance and will be exposed to more pressure, PLA is also used as the material for the build because of its strength, flexibility, and heat resistance, which makes it more reliable; therefore, it is the best material to

execute the job with no flaws. It is also easily sanded and machined.

For the forearm, shoulder, and elbow, PLA is used to complete the build. PLA is a printing material that is easily machined under normal conditions, on the other hand ABS needs special heating bed during the printing. PLA is also an environment friendly plastic which can be created using a mixture of different plant substances, such as potatoes and corn, yet, it does not bio-degrade easily. PLA is also considered strong and rigid material, yet it has a lower melting temperature which makes it unreliable under pressure.

The project required the use of nine servomotors in different places in the arm. To make sure the hand is able to carry a load varying from 3-5 kg, first we had to know how much torque each servo motor needed. To get that, the following formula shown in (1) for the static mode is used:

$$+M = (L_1 \times \cos(\theta_1) + L_2 \times \cos(\theta_2))x9.8 \quad (1)$$

+M is the momentum; L_1 is the length between the force and the first angle θ_1; L_2 is the length between the force and the second angle θ_2 as shown in Figure 5.

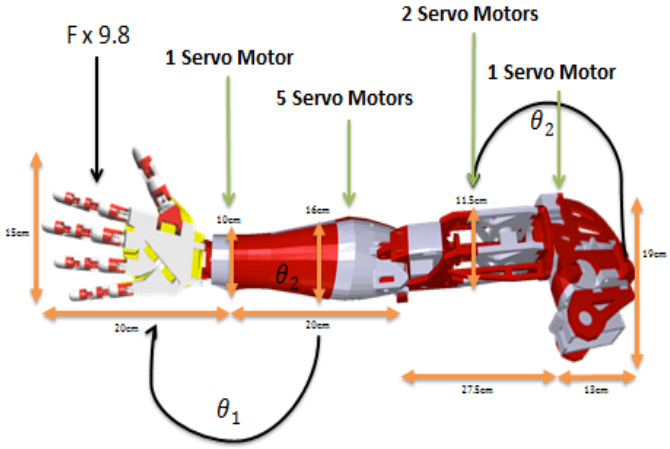


Figure 5. Prosthetic Arm Technical Specifications

For the dynamic mode of the servo motors, friction of the arm parts along with some other parameters were neglected. The following formula shown in (2) is used:

$$M = F \times l + J \times \alpha \quad (2)$$

F is the force applied on the hand; l is the length of the arm; J is the polar movement of inertia; α is the angular acceleration or polar acceleration.

The equation for the dynamic mode shown in (2) changes from one part to another (shoulder, wrist, hand or elbow).

The next calculations are obtained while the hand is in the static mode, not the dynamic mode. While the angles are equal to zero, the servos will be outputting their highest torque in the static mode as shown in Table 3.

Table 3. Load calculation for a servo motor

Mass (Kg)	Force (N)	Wrist S. Torque	Hand S. Torque	Biceps S. Torque	Shoulder S. Torque
1 Kg	9.8	1.96	3.43	6.615	7.889
2 Kg	19.6	3.92	6.86	13.23	15.778
3 Kg	29.4	5.88	10.29	19.845	23.667
5 Kg	49	9.8	17.15	33.075	39.445
7 Kg	68.6	13.72	24.01	46.305	55.223
10 Kg	98	19.6	34.3	66.15	78.89

3.2. Dynamic Modeling of Prosthetic Arm

In the time domain, the mechanical model of the Prosthetic Arm can be represented as a generalized second order differential equation form for the translation motions as shown in (3):

$$f(t) = m\,\frac{d^2x}{dt^2} + C_1\frac{dx}{dt} + K_1.x(t). \quad (3)$$

where: f(t) represents the applied tension forces in the two sides of the tendons to create displacement between fingers at different configurations, m is the mass of the movable part, x(t) is the displacement between fingers at different time, C1 is the friction damping coefficient for the translation motion , K1 is the spring constant . the same model can be restated for the torque $T(t)$ at the elbow and shoulder joints as shown in (4):

$$T(t) = J\,\frac{d^2\theta}{dt^2} + C_2\frac{d\theta}{dt} + K_2.\theta(t). \quad (4)$$

where: J is the mass polar moment of inertia (kg.m2) of the rotated part, θ(t) is the angular displacement (rad.), C2 is the friction damping coefficient for the rotational motion , K2 is the rotational spring constant .

From the multibody dynamic model of the different parts of the Prosthetic Arm, we can get the required toque for each of the 8 motors at all possible movement configurations, the selection design value for the motors specifications was based on the maximum calculated required torque & Power.

3.3. Power Consumption

The power study of the proposed Smart Prosthetic Arm is divided into 2 major parts.

The EEG headset uses a lithium (Li) battery that provides up to 12 hours of continuous use when fully charged. If fully discharged (after 12 hours of usage), it can take up to 4 hours to be recharged. However, the charging process takes between 30 minutes to 2 hours when not fully discharged. Therefore, it is recommended to recharge it before 12 hours of continuous use. For safety reasons, the headset does not work while charging it.

The electro-mechanical system is composed of eight servo motors of rated power 0.5 Watts and 3.5 Watts, which are placed on different parts of the prosthetic arm (wrist, elbow, shoulder and fingers). Three servomotors with a rated power of 3.5 Watts are placed on the wrist, elbow and shoulder area to provide three degrees of freedom. In addition, five servomotors with a rated power 0.5 Watts each are used to control the five fingers. Thus, all the servomotors consume a total power of 13 Watts. A 5V battery with an output current of 2.6 Ampere is required to provide

power to all servomotors embedded in the arm. The wireless communication unit operates when a 50 mA is supplied at 3.3 V, which requires a power of 0.165 Watts to operate.

In addition, a network of smart sensors is used, including temperature, skin pressure and ultrasonic proximity sensors, accelerometers, potentiometers, strain gauges and gyroscopes. These sensors consume a total current of 100 mA. The power consumption of the low-power single board computer (Computing and Processing Unit) is 0.1 Watts corresponding to a current of 30 mA. To power all the units listed above, the system requires a power source with an output current not less than 2.8 A and 5V as output voltage.

Two 10,000 mAh lithium ion batteries are chosen with an output current of 2A each. They include a charging circuit (via USB Cable), and a boost converter that provides 5V DC. These batteries have an 80% efficiency loss on both ends, meaning that it is not recommended to operate the arm while the battery is being charged. With these two batteries, the arm will be operational for 7 continuous hours knowing that the average hand movements per person during a day are equivalent to 1 to 3 hours of continuous movements depending on the daily activity performed. In conclusion, the whole system can be operational for 2 full days.

3.4. Smart Sensors

The sensors included in the proposed solution can be classified in two categories: user-end sensors and environment-end sensors. The first category consists of the 14 EEG sensors presented previously which are installed on the user headset.

As for sensors of the second category, their main feature is to allow the arm to interact and adapt to the surrounding environment, by providing intelligent feedback about critical condition, such as high temperature or pressure, etc. When interfaced with the embedded microprocessor installed on the arm, this network will give the prosthetic arm a human-like behavior with smart reflexes and smooth movements. Note that the feedback coming from some of these sensors will not only be used to operate some servos of the arm, but will also be displayed on small LCD screen mounted on the forearm.

3.5. Control

The proposed system was based on both fully autonomous and semi-autonomous control.

A bi-directional communication channel was implemented between the smart sensor network and the embedded microprocessor in such a way to autonomously control the electro-mechanical unit and provide feedback to the user by displaying it on an LCD mounted on the arm. This setup offered the arm the ability to have smart reflexes when it counters delicate, dangerous and critical situations such as protecting the arm from very hot surface contact or over squeezing fragile objects (glass, human hand, etc…). This integrated network of sensors and actuators requires custom communication protocols and control mechanisms techniques that allow seamless interaction and control hand over between the arm and the user.

By default, the brain signals control the arm movements semi-autonomously via a wireless connection. The EEG headset has a proprietary wireless USB dongle that can be connected to the processing unit via a Bluetooth module. It reads the neuro-electrical signals and interprets them as a set of predefined outputs that reflect facial expressions, mood and conscious intentions. These predefined outputs are received by the processing unit, compared with user-dependent library of pattern and then converted into functions. These functions are then labeled using variables and sent to the Arduino microcontroller through via UART channel. Based on these variables, a certain movement of the arm occurs according to the mapping that is done between the variables and the readings.

4. Project Cost

The cost of the project was divided into components and equipment cost (servo motors, sensors, Bluetooth, Raspberry PI II and EEG-based headset, two USB battery packs), labor costs and the 3D printing cost of the prosthetic arm. All the required components are relatively cheap off-shelf products that can be purchased from multiple providers. The total cost of the system is estimated to be around 1140 USD as shown in Table 4.

Table 4. Cost of project components

Components needed	Approximate Cost (U.S. Dollars)
EEG-based headset (including battery and software)	400
Servo motors	135
Bluetooth module	95
Smart network sensors	150
Raspberry PI II & Arduino Mega	60
Batteries	100
3D model arm	200
Total cost	**1140**

The total cost of the proposed solution is estimated to be relatively affordable and more economical when compared to the existing solutions (surgical, bionic and static prosthetic arms) manufactured by different companies around the world.

5. Testing and Results

During the training session performed, the signals corresponding to valence, engagement, frustration, meditation, short and long-term excitement detected by the Emotiv EPOC headset for 30 and 300 seconds are shown in Figure 6.

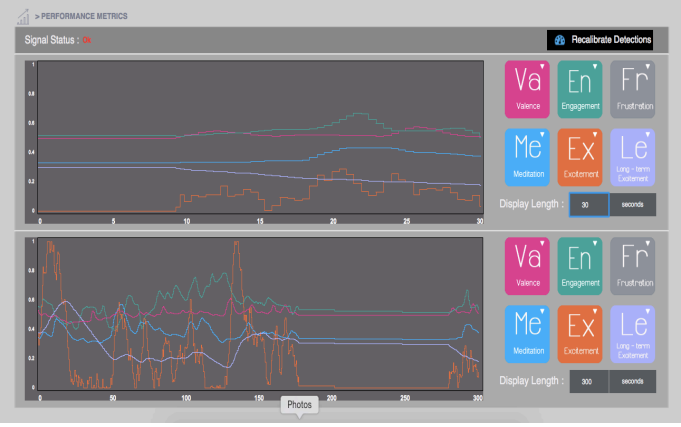


Figure 6. Different signals detected by the Emotiv headset

A 3D cube was displayed, and a training session was available in order to guide the user to move the object either to the left or to the right with his thoughts as shown in Figure *7*. The process started by recording the user's mind when he was not thinking of anything. Then, the user could select an action he prefered to move the object. His thoughts were recorded for the time interval when he was thinking of the object.

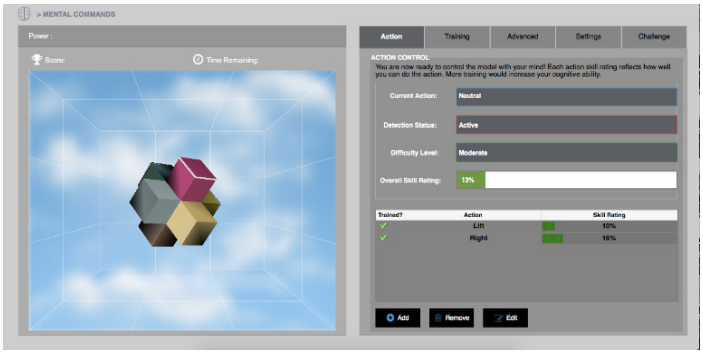


Figure 7. Training session with the 3D cube

Thinking of moving the cube to the left stimulated the hand of the prosthetic arm to close (see Figure 8). However, thinking of moving the cube to the right stimulated the hand to open (see Figure 9).

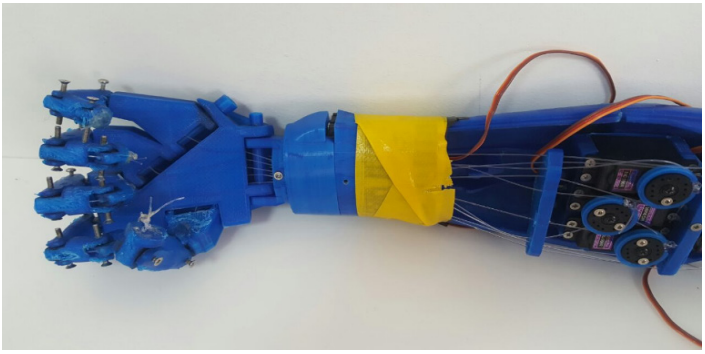


Figure 8. Prosthetic arm – closing hand

Coming up for the scenarios and the algorithms of the arm are one of the things that makes this arm considered as a smart arm. The scenarios simply define the smart reflexes of the arm which can protect the user from the surrounding environment or the people who are communicating with the user. Initially, the arm automatically checks the three sensors (pressure, IR temperature, and proximity). If all of the sensors are in a safe state, the following scenarios can be initiated:

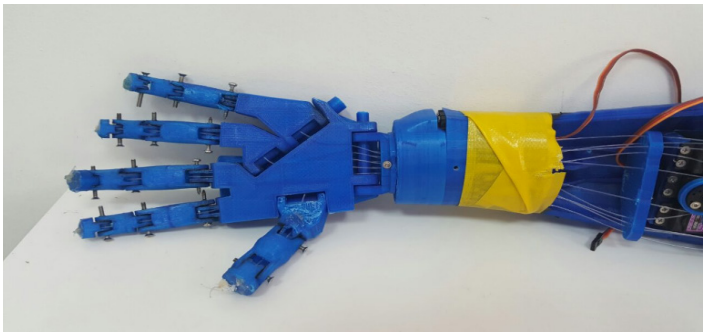


Figure 9. Prosthetic arm – opening hand

Scenario 1: User will be able to initiate brain command and control the arm if all the sensors are in a safe state.

Scenario 2: If the user initiates "Action 1" command, this will enable the handshake. The servomotors in the hand will increment the angle and check if it has the optimum pressure for the handshake. If it reaches it, it will stop automatically.

Scenario 3: If the user imitates "Action 2" command, this will enable the arm to grab a soda can and pour it in a cup. The servo motor in the bicep moves up and stops at a certain degree, then the servo motor in the wrist is rotated so that the soda will be poured in the cup.

On the other hand, if one of the three sensors is not in a safe state, the following scenarios will be taken:

Scenario 4: If the hand holds an object, the temperature sensor, which is embedded within the hand, will measure the temperature. If the temperature is too hot for human skin and reaches the pain threshold, "Action 3" will be automatically initiated. Once Action 3 is initiated a red LED lights up.

Scenario 5: It is the same as the previous scenario, but in this case if the temperature is too hot for the material of the hand itself, "Action 4" is automatically initiated. The servomotors in the hand open the fingers and release the object.

Scenario 6: Here if the proximity sensor which is placed in the back side of the arm reads a distance which is not in the safe range (too close to an object), the servo motors within the shoulder will move the arm forward to keep the arm within a safe range. This is defined as "Action 5".

Finally the code on the Arduino was updated so that all the individual codes were combined together. The codes for each sensor, the codes for testing the servomotors and the codes for the

different gestures based on the input were all put together in a way for the arm to function effortlessly. Within the code, the first thing that was taken into consideration and checked is the sensor network. If the sensors are not in a safe state, they will follow the scenarios as shown previously. Otherwise, the serial port will be checked if it is receiving any data. The serial port is checked because the Arduino communicates with the XBee through a serial connection. If data is being received, it will go through different if statements, each one checking if the input is representing a certain letter. Each letter corresponds either to a raw action (move arm up) or a certain gesture.

The Emotiv EPOC headset has 16 electrodes, which it uses to measure the electromagnetic pulses from the brain. These electromagnetic pulses are created by intentional thoughts, facial expressions and even the mental state of the user. What we will be using for the inputs are the intentional thoughts and the facial expressions.

Before wearing the headset, Saline solution must first be applied to the electrodes. Once the electrodes are damp, they can be connected to the headset. Saline solution is applied in order to increase contact and the quality of the signal read; however, if too much is applied, it could lead to affecting the electrical equipment. Once the user slides the headset on to his head, he/she must insure that the electrodes are placed in the correct position and are in contact with the scalp; this ensures a good signal (green). To get the most accurate output, it is best that all electrodes are green.

The flowchart of the algorithm is shown in Figure 10.

Figure 10. Flowchart of the algorithm

6. Conclusion

The proposed arm hosts state-of-the art technological advancement, communication protocols, control systems, and human interfacing. This gives it great potential in many applications related to the health care field as well as other fields. As long as health care is considered, the idea could be expanded to other body parts as well as to patients having other dysfunctions as nerve damage. On the other hand, many industrial and commercial applications can utilize many features of the proposed arm. Within the health care field, there exists a class of patients who need extra help with their daily lives. This includes elderly people, people under rehabilitation, and people with limited mobility, etc. The proposed arm may be interfaced to a robotic-structure and function as a helper or caregiver to this group of people. It can be programmed to do various functions according to specific patient needs. This may vary from cooking to assistance with bathing or dressing. Another example in the medical field is remote high precision surgical procedures, where surgeons can undergo operations remotely with the aid of the robotic arm. Many industries employ robots in the manufacturing process, many of which can make use of a modified version of the proposed arm. Based on a specific application, this smart arm can be programmed to execute a series of predefined actions, and customized with dedicated sensors, actuators and customized algorithms (such as image and signal processing, gesture and voice recognition etc…). In addition, connecting the arm to the Internet, and making it part of an Internet of Things network (IOT) will increase the performance and productivity of many industry applications. A first prototype is designed, built and was under test. The testing required long training sessions in order first to build a user-dependent library of brain activity patterns, and second to make the user more familiar and comfortable using this hand.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] S. W. Hawking, "World report on disability," World Health Organization, Geneva, Switzerland, 2011.

[2] NBC News. (2010, March 20). *Limb loss a grim, growing global crisis* [Online].Available: http://haitiamputees.nbcnews.com/_news/2010/03/19/4040341-limb-loss-a-grim-growing-global-crisis

[3] M. LeBlanc. (2011, January 14). *Give Hope – Give a Hand* [Online]. Available: https://web.stanford.edu/class/engr110/Newsletter/lecture03a-2011.html

[4] C. Moreton. (2012, August 4). *London 2012 Olympics: Oscar Pistorius finally runs in Games after five year battle* [Online]. Available: http://www.telegraph.co.uk/sport/olympics/athletics/9452280/London-2012-Olympics-Oscar-Pistorius-finally-runs-in-Games-after-five-year-battle.html

[5] Y. Jeong, D. Lee, K. Kim and J. Park, "A wearable robotic arm with high force-reflection capability," in *9th IEEE International Workshop on Robot and Human Interactive Communication*, Osaka, 2000, pp. 411-416.

[6] A. Bennett Wilson Jr., B. (n.d.). Retrieved Octobor 17, 2015, from oandplibrary: http://www.oandplibrary.org/alp/chap01-01.asp

[7] E. Sofge (2012, May 28). *Smart Bionic Limbs are Reengineering the Human*. [Online]. Available: http://www.popularmechanics.com/science/health/a7764/smart-bionic-limbs-are-reengineering-the-human-9160299

[8] R. M. Coupland, *War Wounds of Limbs: surgical management*. Geneve, Switzerland, ICRC, 1993.

[9] Jerkey. *Brain-Controlled Wheelchair* [Online]. Available: http://www.instructables.com/id/Brain-Controlled-Wheelchair

[10] H. Heyrman. *Brainwaves* [Online]. Available: http://www.doctorhugo.org/brainwaves/brainwaves.html

[11] S. Sequeira, C. Diogo and F.J.T.E. Ferreira, "EEG-signals based control strategy for prosthetic drive systems," in *IEEE 3rd Portuguese Meeting in Bioengineering*, Braga, 2013, pp. 1-4.

[12] V. Charisis, S. Hadjidimitriou, L. Hadjileontiadis, D. Ugurca and E. Yilmaz, "EmoActivity – An EEG-based gamified emotion HCI for augmented artistic expression: The i-Treasures paradigm," in *Springler-Verlag Berlin Heidelberg,* Berlin, 2011.

[13] T. Beyrouthy, S. K. AlKork and J. A. Korbane, "EEG mind controlled smart prosthetic arm," in *IEEE International Conference on Emerging Technologies and Innovative Business Practices for the Transformation of Societies*, August 2016.

[14] S. K. Al Kork, "Development of 3D finite element model of human elbow to study elbow dislocation and instability," in *ASME 2009 Summer Bioengineering Conference*, American Society of Mechanical Engineers, 2009.

[15] T. Beyrouthy and L. Fesquet, "An event-driven FIR filter: design and implementation," in *22nd IEEE International Symposium on Rapid System Prototyping (RSP),* May 2011.

[16] C. Heckathorne, "Upper-limb prosthetics: Components for adult externally powered systems," in Atlas of Limb Prosthetics, H. K. Bowker and J. W. Michael, Eds. Rosemont, IL: Am. Acad. Orthoped. Surg., 2002

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 900-905 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Features based approach for indexation and representation of unstructured Arabic documents

Mohamed Salim El Bazzi[*1], Driss Mammass[1], Abdelatif Ennaji[2], Taher Zaki[1]

[1]*IRF-SIC Laboratory, Ibn Zohr University, Agadir, Morocco.*

[2]*LITIS Laboratory, University of Rouen, France.*

ARTICLE INFO

Article history:
Received: 05 April, 2017
Accepted: 03 June, 2017
Online: 28 June, 2017

Keywords:
Unstructured documents
Arabic text mining
Keyphrases
Indexation
Classification

ABSTRACT

The increase of textual information published in Arabic language on the internet, public libraries and administrations requires implementing effective techniques for the extraction of relevant information contained in large corpus of texts. The purpose of indexing is to create a document representation that easily find and identify the relevant information in a set of documents. However, mining textual data is becoming a complicated task, especially when taking semantic into consideration. In this paper, we will present an indexation system based on contextual representation that will take the advantage of semantic links given in a document. Our approach is based on the extraction of keyphrases. Then, each document is represented by its relevant keyphrases instead of its simple keywords. The experimental results confirms the effectiveness of our approach.

1. Introduction

Document indexing is an essential step in the text mining process because it determines how the knowledge contained in the documents is represented. It takes place each time a document is added to the corpus. Indexation must then deal with two main problems, the choice of the most representative terms of each document and the evaluation of their power of representation.

Several approaches are proposed in the literature, particularly in English, but they are unusable for documents in Arabic, given its specific characteristics and the complexity of its morphology, grammar and vocabulary. For instance, association rules and semantic approaches based on external references are more complicated to use on Arabic documents.

The indexation, under all its forms, aims to extract the most relevant descriptors from a given document. The more sophisticated the selection is, the more accurate are the subsequent tasks of text mining using indexation systems, like classification and information retrieval.

Document indexing involves extracting keywords that best represent a document. In spite of the essential role of this phase in the next step of the natural language processing process, few are the works identified at this level [1][2][3].

While preparing this paper, we stopped particularly on descriptor's type extracted from a document in a text mining process. We noticed that stemming is the most used method to represent a document.

This paper is an extension of work originally presented in the 11th International Conference on Intelligent Systems: Theories and Applications (SITA 2016) [4]. It presents a comparison of the two text representation methods: stem-based representation and keyphrase-based representation, as well as their applications and compatibility with the Arabic language. However, our proposed approach can be applied to other languages.

The rest of this paper is organized as following. The second part introduces the related works. The third part is dedicated to the presentation of a conventional text mining process. The fourth part describes the proposed indexing systems and the used methods. We will present and discuss the experimental results in

* Mohamed Salim El Bazzi, IRF-SIC Laboratory, Ibn Zohr University, Agadir, Morocco, Email: elbazzi@yahoo.fr

the fifth section and we conclude by highlighting our contribution and any possible improvements.

2. Related Works

Before starting the keywords extraction, a very important task must be accomplished, which is cleaning and normalizing the text. Tokenization, stop words removal and normalization are frequently used as a preprocessing step. While stemming is the most complicated and discussed issue.

The stemming method consists of extracting the root of a word. In other words, associating the words linked morphologically to the same root. The number of terms is then reduced, which makes the system more effective. This technique reveals a major drawback which is ambiguity.

A root can be defined as a word that cannot be created from other word. For example, the root of the Arabic word ("التصفيات", *the qualifications*) is ("صف", *row*). Although stemming methods cause a big loss of meaning from the original words, these techniques are competitive and provide good classification results. Hence, several works on Arabic text indexing opt for roots as document's descriptors.

Mansour et al. in their work [5], they perform a morphological analysis of the words in a document to extract indexes. First, the authors propose a process of extracting stems. Second, they set up a recognition system of names and verbs, based on rhymes (grammatical patterns) and grammatical rules. A weight is then assigned to each stem taking into account its occurrence and introducing a function indicating how the word is spread in the document

El-Khoribi [6] applied the stemming as a representation of features. Features are represented as vectors that consist of a number of elements equal to the number of classes which the probability of belonging to a class where stems occur. Then a stems lookup table is built from roots and labels of the classes to which they belong. Then, a Hidden Markov Model is used to evaluate the membership of a new document to a class.

Al Molijy et al. [7] used a linguistic method for the extraction of indexes from Arabic documents. Their method is based on the parsing of words document. The proposed algorithm consists of breaking words into N-grams (where N equals 3, 4 or 5), calculate their frequencies, and then retain the first 100 most frequent words, constituting thereby an N-gram document profile. The advantage of this method is the reduction of the number of words representing a document.

Bawaneh et al. [8] compared between the two classifiers K-Nearest Neighbor (KNN) and Naïve Bayesian (NB). The light stemmer was used as feature and TFIDF (Term Frequency – Inverse Document Frequency) measurement as a method of weighting features. The K-NN classifier was considered more efficient.

Gharib et al. [2] applied four classifiers on Arabic corpus which are: SVM (Support Vector Machine), NB, K-NN and Rocchio method, using stemming as features representation technique and TFIDF as weighting method. The Rocchio classifier works better when the feature space is small but the SVM performs better when space becomes larger.

Raheel et al. [3] have shown in a comparative study the influence of the choice of entities representing a document, on manipulating the performance of classifiers. They selected as descriptors, words in their original form, lemmas, roots, and the n-grams. Two classifiers were used, the SVM and Naive Bayesian Networks. SVM based on the 3-grams gave better classification results.

Harrag et al. [9] conducted a comparative study of three pretreatment techniques: light stemming, root-based stemming and dictionary lookup stemming in order to reduce the feature space. Then two classifiers were tested, Artificial Neural Networks and SVM.

Since the stemming is the method of representation which is the most common in text mining preprocessing, several studies have contributed to its improvement [10][11][12][13][14] [15]. However, the extraction of keyphrases has not experienced the same progress [16].

3. Arabic Text Mining Process

المغرب يتجاوز تداعيات أزمة المال هذه السنة
أبدى بنك المغرب ارتياحاً لأداء الاقتصاد في المغرب، خلال 2009، وتوقع أن يتراوح معدل النمو بين 5 إلى 6 في المائة للعام الثاني على التوالي، رغم تداعيات أزمة الاقتصاد العالمية وتأثيرها على إيرادات المغرب من العملات الصعبة والاستثمارات الخارجية.
وتوقع المركزي أن يحقق النمو الاقتصادي 3 إلى 4 في المائة في 2010، ارتباطاً بالإنتاج الزراعي وتحسن أسعار القطاعات الصناعية.

Figure 1: Original text from the used corpus.

3.1. Preprocessing

Before starting the document indexing phase, a very important task is to clean up and standardize the text. The following are the most commonly used steps:

- *Tokenization*

Tokenization is the production of a sequence of segments separated by spaces or punctuation marks. The output is a list of words without neither punctuation, nor special characters.

المغرب يتجاوز تداعيات أزمة المال هذه السنة أبدى بنك المغرب ارتياحاً لأداء الاقتصاد في المغرب خلال 2009 وتوقع أن يتراوح معدل النمو بين 5 إلى 6 في المائة للعام الثاني على التوالي رغم تداعيات أزمة الاقتصاد العالمية وتأثيرها على إيرادات المغرب من العملات الصعبة والاستثمارات الخارجية وتوقع المركزي أن يحقق النمو الاقتصادي 3 إلى 4 في المائة في 2010 ارتباطاً بالإنتاج الزراعي وتحسن أسعار القطاعات الصناعية

Figure 2: text after tokenization.

- *Stop words removal*

Stop words are non-significant terms in a text. It is a about eliminating words whose occurrence is very frequent and do not bring any added value to the process of indexation. These words are usually pronouns, articles and conjunctions.

المغرب يتجاوز تداعيات أزمة المال السنة أبدى بنك المغرب ارتياحاً لأداء الاقتصاد المغرب خلال توقع يتراوح معدل النمو المائة للعام الثاني التوالي تداعيات أزمة الاقتصاد العالمية تأثيرها إيرادات المغرب العملات الصعبة الاستثمارات الخارجية توقع المركزي يحقق النمو الاقتصادي المائة ارتباطاً بالإنتاج الزراعي تحسن أسعار القطاعات الصناعية

Figure 3: text after stop words removal.

- *Stemming*

Stemming can be defined as the process of removing prefixes, infixes and suffixes from words to reduce these words to their stems (roots). To remedy the problem of big loss of meaningful information, the notion of light stemming has been evoked in several works and consists in eliminating just the prefixes and suffixes of a given word without having to go back to its root.

- *Conversions*

The first conversion that can be applied to a document is the elimination of diacritical marks. Diacritical marks are signs added to the top or bottom of the Arabic letters to specify the pronunciation of the word. This phonological role also influences the meaning of this word. Indeed, two words can be written in the same way but differentiated by the addition of different diacritical signs. For example, if the word "عالم" is pronounced (عالم, âalim) it means "scientist", and if pronounced (عالم, âalam) it means "world". This procedure aims to standardize the documents because it is rare to find a fully accented corpus.

مغرب جوز دعا ازم مول سنة ابدى بني مغرب ارتياحا لاداء قصد مغرب خلل وقع روح عدل نمي عوم ثنى ولى رغم دعا ازم قصد علم اثر ردا مغرب عمل صعب ثمر خرج وقع مركز حقق نمي قصد ارتباطا توج زرع حسن سعر قطع صنع

Figure 4: text after stemming and conversions.

The second conversion is that of characters which has for purpose to normalize the letters which can be written in several forms. Thus the characters "إ" ,"أ" and "آ" are replaced by "ا", similarly "ة" is converted into "ه" and "ئ" ,"ي" into "ى". In our system, we did not consider this conversion.

3.2. Term weighting

The weighting of an indexing term is the association of numerical values called weight to that term to represent its power of discrimination for each document in the collection. This characterization is linked to the informativeness of the term for the given document. This notion refers to the amount of meaning that a word carries.

There are several ways to determine the weight of a term. In our case, we have used the TF-IDF (Term-Frequency Inverse Document Frequency) method.

$$Weight_i(j) = TF_{ij} \times IDF_j \quad (1)$$

Where TF_{ij} is the occurrence frequency of term the **j** in a document **i** and IDF_j is the inverse absolute frequency of term **j** in the collection. Thus the weight of a term increases if it is frequent in the document and decreases if it is frequent in the collection.

3.3. Document representation

Document representation is one of the techniques used to reduce the complexity of documents and makes them easier to manipulate, and the document is then transformed from its textual version to a matrix [Document × Term] (Figure 5). The representation of the most used document is the Vector Space Model. The documents are represented by vectors of words. This representation has its own drawbaks as the large dimension of representation and loss of correlation between adjacent words, the thing that leads to the loss of the semantic relationship that exists between the terms in the original document. To overcome these problems, weighting methods are used to assign appropriate weights to the term as shown in Figure 5:

	T_1	T_2	…	T_m	
D_1	w_{11}	w_{12}	…	w_{1m}	C_a
D_2	w_{21}	w_{22}	…	w_{2m}	C_b
…	…	…	…	…	…
D_n	w_{n1}	w_{n2}	…	w_{nm}	C_k

Figure 5 : Document × Terme Matrix.

Each input represents a vector of terms, where w_{nm} is the weight of the term T_m in the document D_n, and C_i is the class assigned to the document D_i.

3.4. Classification

The text classification is an important part of the text mining. It consists in providing a set of learning data (labeled documents) to the classification system. The task is then to determine a classification model that is capable of affecting the correct class for a new document.

Lately, the task of automatic text classification has been widely studied and the progress seems to be efficient in this field [2]. Several classification methods have been compared and proven to be effective, including the Bayesian classifier, decision trees, K-nearest neighbor, Support Vector Machines, neural networks.

The classification of textual documents presents many challenges and difficulties. First, it is difficult to grasp high level semantics and abstract concepts of natural language only a few keywords. This confirms that the efficiency of the indexing step is paramount and entirely decisive.

3.5. Evaluation

Experimental evaluation of classifiers is the last step in the indexing process, and generally attempts to evaluate the effectiveness of a classifier, namely, its ability to make categorization decisions. To this end, there are many measures, each emphasizing one or other property of the system.

We used the most commonly used measures: recall, precision and the f-measure which synthesizes the first two formulas. Let's consider the following nominations:

TP (True positive), FN (False Negative), FP (False positive), TN (True Negative).

$$\textbf{Recall} = \frac{TP}{TP + FN} \quad (2) \qquad \textbf{Precision} = \frac{TP}{TP + FP} \quad (3)$$

$$\textbf{F-measure} = \frac{2 \times recall \times Precision}{(recall + Precision)} \quad (4)$$

4. Proposed Approach: Keyphrases based indexing

4.1. Text features

Descriptors are the units of text that represent its content. There are several categories of descriptors, which attempt, on the one hand, to reduce the size of the document, and on the other hand to preserve its semantic aspect. We quote:

Word: the text is simply segmented into single words after removing stop words.

Lemma: the words of the text are reduced to their canonical forms. The process of lemmatization consists on using grammatical rules to replace verbs by their infinitives and names by their singular forms.

Root (or stem): the stemming process extracts the roots of each word from the text. The words are then associated with their common root.

Concepts: they take the form of a word or group of words, coming from a list of controlled lexicon, generally dictionaries or ontologies.

Keyphrase: it is a sequence of words that describes more perfectly a document than a single word. They maintain better context and semantics. For example, the term "middle east" is more descriptive than "middle" and "east" taken distinctly.

4.2. Keyphrases Extraction Methotds

In this section, we try to give an overview of different used methods for unsupervised keyphrases extraction (table 1). Consequently, our proposed system is completely automatic. The unsupervised methods have the particularity to abstract the specific nature of the processed data.

Table 1: Overview of unsupervised methods for keyphrases extraction

Reference	Approach
Nabil et al. [26]	Combining linguistic methods with statistical methods
Ali et al. [27]	enrich statistical and linguistic information
Mihalcea et al. [28]	A Graph based ranking algorithm
El-Beltagy et al. [29]	Consider the position of the first occurrence of any given phrase
Elshishtawy et al. [30]	Combining the linguistic knowledge and the machine learning techniques
Najadat et al. [31]	Phrase frequency (PF), summation of phrase terms frequencies (Tf), PFIDF (Phrase Frequency - Inverse Document Frequecy), Phrase Position, Title Threshold and phrase distribution
Liu et al. [32]	The best keyphrase must be: understandable, semantically relevant with the document and have high coverage of the whole document
Sarkar et al. [33]	A candidate keyphrase is considered as a sequence of words that does not contain neither punctuations nor stop words, and then this sequence is broken into smaller phrases

4.3. The Used Keyphrases Extraction Method

Our implemented method for keyphrases extraction is inspired from Najadat's [31] approach which uses Phrases frequency. Hence, to extract keyphrases we follow those steps:

1. Segmentation of the text into simple words;
2. Calculating the number of co-occurrences of two (or more) consecutive words. To get a candidate keyphrase, the consecutive words should co-occur at least two times in the given document.
3. Calculate the TFIDF of obtained keyphrases to get the final list of indexes.

5. Results and Discussion

In this work, we have conducted a study about the influence of the choice of features' type on the efficiency of documents indexing. Consequently, experimental results should clarify our hypothesis which considers that keyphrases describe better the content of a document than simple keyword.

For validation, we tested our system on a corpus of 1000 documents of electronic press, extracted from Aljazeera [1] and Alarabiya [2]. This corpus is classified into 3 categories. These categories are: Economics, Politics and Sport.

In this section, a series of experiments was conducted. For each experiment, we used TFIDF as weighting method and the optimal and same Thresholds' values of K-Nearest Neighbor and features selection. Tables 2 and 3 show different classification results obtained for each feature type. These results are expressed through recall, precision and F-measure criteria.

Table 2: classification results using stems

	Recall	Precision	F-measure
Economics	0.164	0.392	0.231
Politics	0.477	0.326	0.387
Sport	0.393	0.351	0.371
Average	**0.345**	**0.356**	**0.350**

Table 3: classification results using keyphrases.

	Recall	Precision	F-measure
Economics	**0.447**	**0.461**	**0.454**
Politics	**0.537**	**0.327**	**0.406**
Sport	0.136	**0.36**	0.197
Average	**0.373**	**0.382**	**0.378**

[1]http://www.aljazeera.net

[2]http://www.alarabiya.net

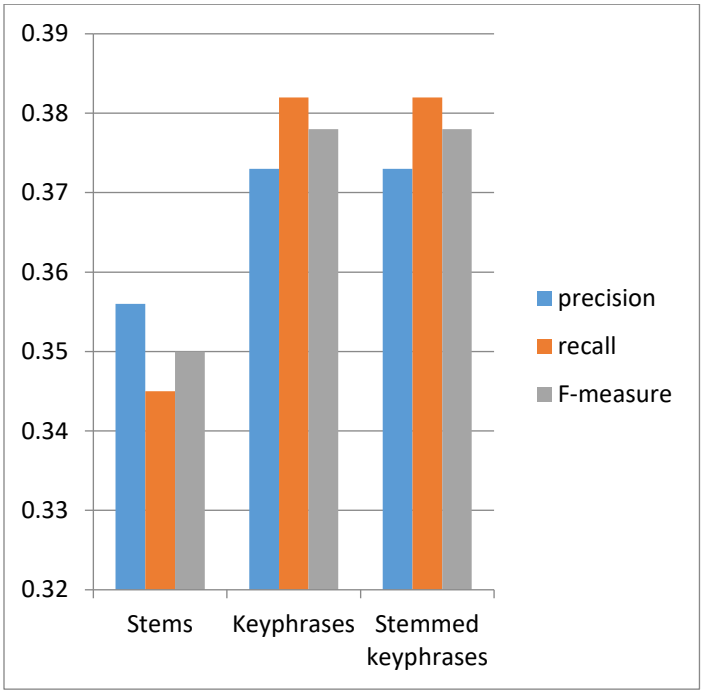
We conducted another test by performing the stemming firstly and then proceeding to extract keyphrases. After this combination, we obtained a list of stemmed-keyphrases. Classification results (table 4) do not show a significant impact of stemming, which explains that the system can stand alone.

Table 4: classification results using stemmed keyphrases.

	Recall	Precision	F-measure
Economics	0.149	0.370	0.212
Politics	**0.731**	0.365	**0.487**
Sport	0.242	**0.410**	0.304
Average	**0.373**	**0.382**	**0.378**

The obtained results show that our approach has an impact on the behavior of the indexes in the classification phase.

Figure 6: Graphic representing the obtained results.



Despite the fact that we were not particularly selective on the keyphrases extraction method, the results are encouraging.

Figure 7: Recall curves depending on each class.

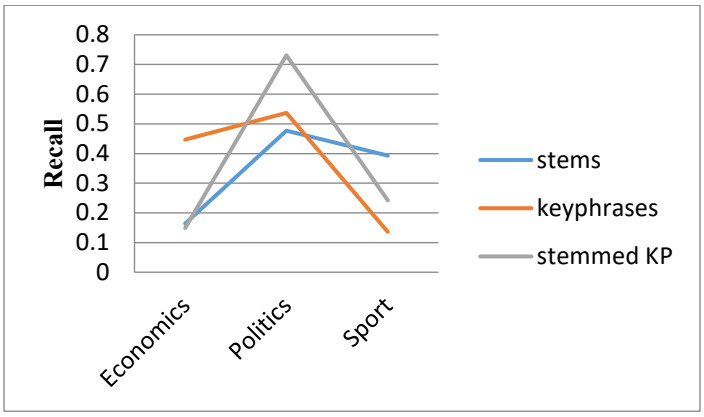
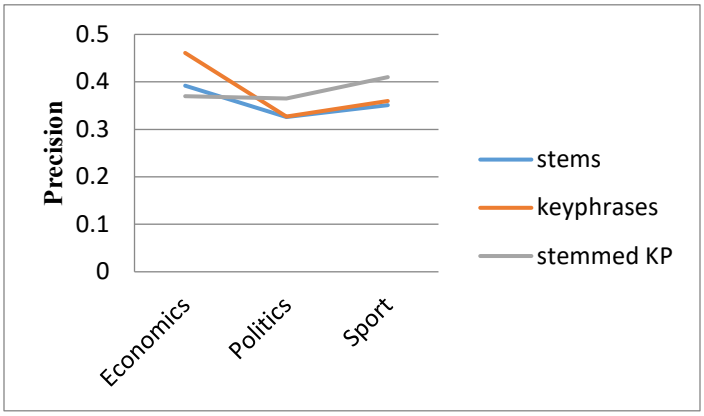


Figure 8: Precision curves depending on each class.



Despite the fact that the results are close for the two approaches (keyphrases and stemmed keyphrases) on average recall and precision (table 3and 4), we clearly distinguish the changes in values according to each domain (figure 7 and 8). This is probably due to the homogeneity of the classes in the corpus. However, our approach reaches up to 73% of recall and 46% of precision in some cases. This leads us to test our approach on other bigger corpus with various classes, and considering adding other improvements.

The experimental results confirms our assumption that keyphrases express better documents than simple keywords. However, this comparison does not, in any way, thwart the approaches presented in related works. Nevertheless, our work could be taken as an advance on what is proposed in the state of the art of Arabic documents indexing.

6. Conclusion

. In this paper, we introduced keyphrases based indexing for Arabic documents. After segmenting texts into simple words, we select candidate terms using the co-occurrence method, then we used the TFIDF formula to obtain the final list of indexes. Our objective is to assess the ability of keyphrases to represent a given document.

On one hand, stemming techniques are used in Arabic text preprocessing to reduce multiple forms of the word to one root. However, the results that we have obtained show that the stem's power of discrimination is relatively low. This may be due to the loss of semantic linking between words during the stemming process. On the other hand, we have proposed a method that extract keyphrases in order to represent document. Although the keyphrases extraction method we used is basic, and based on calculation of frequencies, the results of indexation outperform those obtained by the stem based indexing.

In our future works, we will continue to reveal the semantic information by developing new indexing methods. Nonetheless, we consider developing an efficient method in order to extract Arabic keyphrases. Many other improvements are planned for Arabic document indexation.

References

[1] H. Al-Mahmoud, M. Al-Razgan. "Arabic Text Mining: A Systematic Review of the Published Literature 2002-2014", International Conference on Cloud Computing (ICCC), 2015.

[2] T. F. Gharib, Habib M. B,.and Z. T. Fayed. "Arabic Text Classification Using Support Vector Machines". International Journal of Computers and Their Applications ISCA, vol. 16, no. 4, pages 192–199, 2009.

[3] S. Raheel and J. Dichy. "An empirical study on the feature qs type effect on the automatic classification of arabic documents". In Proceedings of the 11th international conference on Computational Linguistics and Intelligent Text Processing, CICLing'10, pages 673–686, Berlin, Heidelberg, 2010.

[4] M.S. EL Bazzi, T. Zaki, D. Mammass, A. Ennaji. "Stemming Versus Multi-words Indexing For Arabic Documents Classification". 11th International Conference on Intelligent Systems: Theories and Applications, SITA, 2016.

[5] N. Mansour, R.A. Haraty, W. Daher, M. Houri. "An auto-indexing method for Arabic text". Information Processing and Management, volume: 44 issue: 4, pages: 1538-154, 2008.

[6] R. El-Khoribi, A. and M. A. Ismael. "An Intelligent System Based on Statistical Learning For Searching in Arabic Tex"t. ICGST International Journal on Artificial Intelligence and Machine Learning, AIML, vol. 6, pages 41–47, 2006.

[7] A. Al Molijy, I. Hmeidi et I. Alsmadi. "Indexing of Arabic documents automatically based on lexical analysis". International Journal on Natural Language Computing (IJNLC) Vol. 1, No.1, April 2012.

[8] M.J. Bawaneh, M. S. Alkoffash and A. I. Al Rabea. "Arabic Text Classification using K-NN and Naive Bayes". Journal of Computer Science, vol. 4, pages 600–605, 2008.

[9] F Harrag,, E. El-Qawasmah and A. Al-Salman. "Stemming as a Feature Reduction Technique for Arabic Text Categorization". In Proceedings of The 10th International Symposium on Programming and Systems, ISPS 2011, pages 128–133, April 2011.

[10] S.Khoja and R.Garside (1999). "Stemming Arabic Text". Computing Department, Lancaster University, Lancaster, http://www.comp.lancs.ac.uk/computing/users/khoja/stemmer.ps.

[11] M. Y. Al-Nashashibi, D. Neagu, and A. A. Yaghi, "Stemming techniques for Arabic words: A comparative study", 2nd International Conference on Computer Technology and Development, 2010, pp. 270–276.

[12] M. Y. Al-Nashashibi, D. Neagu, and A. A. Yaghi, "An improved root extraction technique for Arabic words", 2nd International Conference on Computer Technology and Development, 2010, pp. 264–269.

[13] M. N. Al-Kabi, "Towards improving Khoja rule-based Arabic stemmer", IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT), 2013, pp. 1–6.

[14] A. Zitouni, A. Damankesh, F. Barakati, M. Atari, M. Watfa, and F. Oroumchian, "Corpus-Based Arabic Stemming Using N-Grams," in Information Retrieval Technology, vol. 6458, P.-J. Cheng, M.-Y. Kan, W. Lam, and P. Nakov, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 280–289.

[15] S. Ghwanmeh, G. Kanaan, R. Al-Shalabi, and S. Rabab'ah, "Enhanced Algorithm for Extracting the Root of Arabic Words," 2009, pp. 388–391.

[16] S. R. El-Beltagy and A. Rafea, "KP-Miner: A keyphrase extraction system for English and Arabic documents," Inf. Syst., vol. 34, no. 1, pp. 132–144, Mar. 2009.

[17] G. Wei, X. Gao and S. Wu. "Study of Text Classification Methods for Data Sets With Huge Features". In Proceedings of the 2nd International Conference on Industrial and Information Systems, volume 1, pages 433–436, 2010.

[18] I. Mallak. "De nouveaux facteurs pour l'exploitation de la sémantique d'un texte en Recherche d'Information". Human-Computer Interaction. Université Paul Sabatier – Toulouse III, 2011.

[19] A. Alajmi, E.M. Saad, R.R. Darwish. "Toward an ARABIC Stop-Words List Generation". International Journal of Computer Applications (0975-8887) Volume 46- No.8, May 2012.

[20] I. Hmeidi, B.Hawashin, E.El-Qawasmeh. "Performance of KNN and SVM classifiers on full word Arabic articles". Advanced Engineering Informatics 22 (2008) 106–111.

[21] B. Trstenjak, S. Mikac, D. Donko. "KNN with TF-IDF Based Framework for Text Categorization" . 24th DAAAM International Symposium on Intelligent Manufacturing and Automation, 2013.

[22] T. Zaki, D. Mammass, A. Ennaji. "A semantic proximity based system of Arabic text indexation". International Conference on Image and Signal Processing (ICISP) 2010.

[23] A. Mountassir. "Sentiment Analysis: Classification supervisée de documents arabes". Proceedings of 7th International Conference on Intelligent Systems : Theories and Applications. May 16-17, 2012, Mohammedia, Morocco.

[24] S. Alsaleem. "Automated Arabic Text Categorization Using SVM and NB". International Arab Journal of e-Technology, vol. 2, no. 2, 2011.

[25] B. Al-Salemi. and M. J. A. Aziz. "Statistical Bayesian Learning For Automatic Arabic Text Categorization". Journal of Computer Science, vol. 7, no. 1, pages 39–45, 2011

[26] M. Nabil, M.Aly, Amir F. Atiya. "New Approaches for Extracting Arabic Keyphrases". 2015 First International Conference on Arabic Computational Linguistics.

[27] N. G. Ali, N. Omar. "ARABIC KEYPHRASES EXTRACTION Using a Hybrid of STATISTICAL AND MACHINE" Learning Methods. International Conference on Information Technology and Multimedia (ICIMU), November 18 – 20, 2014, Putrajaya, Malaysia

[28] Mihalcea R. et P. Tarau. Textrank: Bring- ing order into texts. In Proceedings of EMNLP, pages 404–41, 2004.

[29] S. R El-Beltagy and Ahmed Rafea. Kp-miner: A keyphrase extraction system for english and arabic documents. Information Systems, 34(1):132–144, 2009.

[30] El-shishtawy T.A. & Al-sammak A.K. "Arabic Keyphrase Extraction using Linguistic knowledge and Machine Learning Techniques". Proceedings of the Second International Conference on Arabic Language Resources and Tools, 2009

[31] Hassan M. Najadat, Mohammed N. Al-Kabi, Ismail I. Hmeidi, Maysa Mahmoud Bany Issa. "Automatic Keyphrase Extractor from Arabic Documents". (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 2, 2016

[32] Liu, Z. , Peng, L. , Yabin, Z. et S. Maosong. 2009. "Clustering to find exemplar terms for keyphrase extraction". In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 257–266.

[33] Kamal Sarkar. "A Hybrid Approach to Extract Keyphrases from Medical Documents". International Journal of Computer Applications (0975 – 8887) Volume 63– No.18, February 2013

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 906-912 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

The Analysis and Comparison of Multiport Converter used for Renewable Energy Sources

Sarab Al-Chlaihawi[1,2*] , Ammar Al-Gizi[1,3], Aurelian Craciunescu[1]

[1]*Electrical Engineering Faculty, University Politehnica of Bucharest, 060042, Romania*

[2]*Al-Furat Al-Awsat Technical University, Najaf Technical Institute, 31001, Najaf, Iraq*

[3]*Al-Mustansiriyah University, Faculty of Engineering, 10001, Baghdad, Iraq*

ARTICLE INFO

Article history:
Received: 05 April, 2017
Accepted: 07 June, 2017
Online: 30 June, 2017

Keywords:
Dual Input Single Output
DC-DC converter
PV panel
Half bridge converter
Full bridge converter

ABSTRACT

The concept of renewable energy is increasing day by day because the electrical applications are moving towards utilization of renewable sources and due to this factor electric power is also gaining ground and new type of converters are modelled in order to get high utilization of renewable sources. In this paper, two types of multiport converters are presented which includes a DC-DC converter with a transformer attached and a DC-DC converter without a transformer. The converter with transformer is further subdivided into a full bridge and a half bridge DC-DC converter. The simulation results for all converters are presented and in the end, a case is taken for comparison which involves fixed resistance. The converters are compared with a constant load resistance of 70 ohms in terms of cost, area, complexity, control, isolation, weight and efficiency etc.

1. Introduction

There is an increased trend in the use of multiport converters because of the demand for renewable energy. Electrical applications are moving towards the utilization of renewable energy due to which research in electrical power electronics has increased. With the advancement in renewable energy usage, the DC-DC converter research has also increased and hence multiport converters are being proposed. The multiport converter helps in interfacing multiple sources at a time and delivers power from sources to load simultaneously as well as separately.

In [1] a review of multiport converter is presented. The multiport converters presented in [1] comprises of converters with the transformer attached and detached. The converter with the transformer attached is further subdivided in to Half Bridge Three Port (HBTP) DC-DC Converter and Full Bridge Three Port (FBTP) DC-DC Converter. The advantages and disadvantages of each converter type for each DC-DC converters are presented.

This proposed paper is an extension of [1] where the multiport converters were analyzed at constant load resistance of 70 ohms for each case and they are compared in terms of losses within the converters, area of the converters, cost of converters, controlling of converters and complexity of converters etc.

In [2] an overview is presented regarding multiport converters. The first section contains the information of multiport converter introduction. In the second section, a comparison between the multiport converter and conventional converter is presented.

The conclusion of research involves; every dc-dc converter has its own advantages and disadvantages so it depends upon the applications for which the converter is to be used. In the case of isolation, converters with transformers are to be used while if the isolation is not required then DC-DC converters without transformers are to be used.

In [3] a dc-dc converter topology is presented which has the following advantages:

- Simplification of power conditioning
- Improved utilization of source
- The ports of converter supports bidirectional power flow
- Each port can be controlled individually
- The connected sources of current can be decoupled

*Sarab Al-Chlaihawi, Ph.D. student, University Politehnica of Bucharest, 060042, Romania, +40723062504, sarab.haedar@yahoo.com

With this converter, we can control the level of every individual delivery of power and decouple the associated source of the current in series sources.

In [4] different types of DC to DC converters are used and their computational performances are analyzed along with their efficiencies. The Buck converter gives better computational performance and a high efficiency with the resistance lesser than Resistance at the maximum power point (R_{MPP}) but a low efficiency with the Resistance greater than the R_{MPP}. The Boost converter gave a low computational performance and a low efficiency with the Resistance less than R_{MPP} but a high efficiency with the Resistance greater than R_{MPP}. Similarly, the Buck-Boost converter gave a better computational performance and a high efficiency with the Resistance less than R_{MPP} along with a high efficiency with the Resistance greater than R_{MPP}.

The Cuk converter gave an average computational performance with a medium efficiency with the Resistance less than R_{MPP} but a high efficiency with the Resistance greater than R_{MPP}. The SEPIC converter also gave an average computational performance with a medium efficiency with the Resistance less than R_{MPP} but a high efficiency with the Resistance greater than R_{MPP} while the Zeta Converter gave a better computational performance with a medium efficiency with the Resistance less than R_{MPP} but a high efficiency with the Resistance greater than R_{MPP}.

From the above discussion it can be concluded easily that the buck-boost converter is suitable for Maximum Power Point Tracker (MPPT) because its efficiency is higher in both cases whether the resistance is greater than R_{MPP} or less.

In [5] multiport converters are discussed. The converters are compared in terms of soft switching and isolation. In the conclusion of [5] the advantages and disadvantages of a multi-port dc-dc converter are presented. These are:

- utilization of two or more simultaneous power sources
- lower cost due to semi-conductor elements and fewer passive components
- the power to load can be in higher productions because of multiple sources
- the management of energy can be monitored in a better way

In [6] the comparative study of MPPT algorithm is presented. The MPPT algorithm is used by the controller to get maximum power from the PV panel.

The algorithms presented for comparison are based on Fuzzy Logic for MPPT controller, Incremental Conductance and Perturb & Observe (P&O). The algorithms are tested in standard conditions with atmospheric condition variance. The results obtained are:

- The response of Incremental Conductance algorithm is better at a steady state but it has some errors when compared to the Perturb & Observe algorithm.
- In Perturb & Observe algorithm there is some oscillation at the operating point which results in loss in available power thus PV module efficiency is decreased.
- In Perturb & Observe algorithm the MPP tracking is directly proportional to the step size of perturbation. By higher steps the Maximum Power Point (MPP) is achieved faster but then there are oscillations at steady state.
- The Fuzzy Logic controller provides a better steady state and transient state performance as compared to classical methods. Fuzzy Logic controller tracks MPP with high speed and it has more stability at the steady-state in comparison with Perturb & Observe algorithm and Incremental Conductance algorithm.

For different types and capacities of renewable energy sources, an isolated multiport DC-DC converter is proposed in [7] for its corresponding simultaneous management of power. Only one controllable switch is utilized for each port in which a source is connected. As a result, it has a very limited number of power switches and a very simple topology. The proposed converter's operation and characteristics are explained and analyzed and its applications are discussed.

In order to replace the regular transformer distribution, a solid state transformer (SST) is proposed in [8] which helps enables the formation of power quality features along with providing ports for the integration of the distributed generation (DG) (e.g. photovoltaic (PV)) as well as storage. A quad-active-bridge (QAB) converter based on SST topology is modelled in [8] which facilitates the load isolation along with the isolation of the distributed generation as well as the storage. An average model formulated on the gyrator is developed which helps facilitates a general multi-active-bridge (MAB) converter. Using these expressions, the power ratings of the MAB ports are derived. These results are then applied onto the QAB converter and later on it is analyzed.

For a stand-alone renewable power system application, a systematic strategy is proposed in [9] in order to interface a renewable source, a load and a storage battery with the generated Three-Port Half-Bridge Converters (TPHBCs). If the DC bias current is allowed in the transformer then the half-bridge converter's primary circuit operates as a rectified synchronous buck converter which can help configure a power flow path between the battery and the renewable source which are, as a result, connected in parallel with one of the dividing capacitors. A post and a synchronous regulation has been proposed for various implementations in order to make the voltage on any two of the ports individually controlled.

A three-port converter (TPC) derived from a full-bridge converter (FBC) is proposed in [10] and its systematic approach has been discussed. The methodology helps strips a pair of the switching legs of the corresponding FBCs into two different switching cells with distinct sources attached to it and along with this a DC bias current is also allowed in the transformer. By utilizing the proposed strategy, a full-bridge TPC (FB-TPC) is developed for the applications of the renewable power systems which composes of multiple characteristics which includes: a single-stage power conversion between any two of the three ports, reduced number of devices, an easy and understandable control method and topology.

A three-port DC-DC converter which integrates the photovoltaic (PV) and the power of the battery for a high step-up applications is proposed in [11] in which the topology comprises of two coupled inductors (they help to acquire high step-up voltage

gain and also help reduces the voltage stress produced by the two input switches), two active-clamp circuits (which helps recycles the energy which is stored in the leakage inductors and also improves the efficiency of the system) and five power switches. The operation mode does not require changing in case a transition between the discharging and charging occurs.

The major contribution of the proposed paper is to analyze multiport converters at constant load resistance of 70 ohms for each case and after that, they are compared in terms of losses within the converters, area of the converters, cost of converters, controlling of converters and complexity of converters etc. So this paper is organized as follows:

- Section 2 explains in detail about the Half bridge and Full bridge converters.
- Section 3 discusses the Dual Input Single Output converter (DISO).
- Section 4 is detailed with the comparison between a converter with a transformer and a converter without a transformer.
- Section 5 contains the in depth analyses of the results.
- And at the end in Section 6, the conclusion has been given.

2. Half Bridge Converter and Full Bridge Converter

In the above section, the literature about Multiport converters has been presented. According to the literature the multiport converters have resolved the problems of single input converters and improved the renewable energy interfacing.

The interfaced sources in the multiport converter can be unidirectional or bidirectional and the sources are able to deliver power to load individually and simultaneously. By the use of multiport converters, components are reduced hence cost is reduced, while reliability and power are increased.

The characteristics of multiport converter in comparison with traditional converter are:

- It has a high efficiency.
- It is more reliable.
- It has less number of components but complexity increases.
- Insulation can be attained by adding transformer.
- It supports unidirectional as well as bidirectional power flow.
- It has more control area.
- It has flexibility.

2.1. Operation of converter

Figure 1 shows the half bridge single phase converter. It is composed of two switches (out of which the mostly used switch is MOSFET) and two capacitors, one with source and one with load.

The flow of power inside the inverter varies after every half cycle.

Let the source voltage be 'Vs' for half bridge converter and switching frequency be 'f' and 'T' be the total time of the whole cycle.

$$f = \frac{1}{T} \tag{1}$$

For time $0 < t \leq \frac{T}{2}$ the power flow path is through MOSFET 'S1' and capacitor 'C2'. The voltage obtained at load 'V_L' is

$$V_L = \frac{V_s}{2} \tag{2}$$

For time $\frac{T}{2} < t \leq T$ the power flow path is through the MOSFET 'S2' and capacitor 'C1'. The voltage obtained at load 'V_L' is

$$V_L = -\frac{V_s}{2} \tag{3}$$

It can be seen that the magnitude of 'V_L' is half as half of the voltage is dropped across the capacitor.

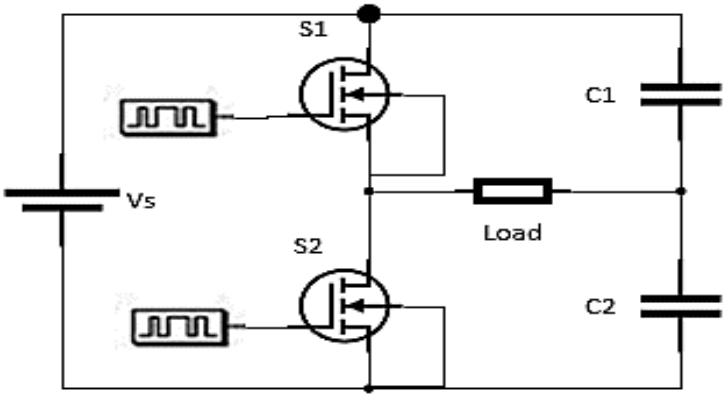


Figure 1: Half Bridge Converter

Figure 2 shows the full bridge single phase converter. It is composed of four switches (of which the mostly used switch is MOSFET), with source and load. The flow of power in the inverter varies after every half cycle.

Let the source voltage for the full bridge converter be 'Vs' and switching frequency is 'f' and 'T' be the total time of the whole cycle.

$$f = \frac{1}{T} \tag{4}$$

For time $0 < t \leq \frac{T}{2}$ the power flow path is through MOSFET 'S1' and MOSFET 'S4'. The voltage obtained at load 'V_L' is

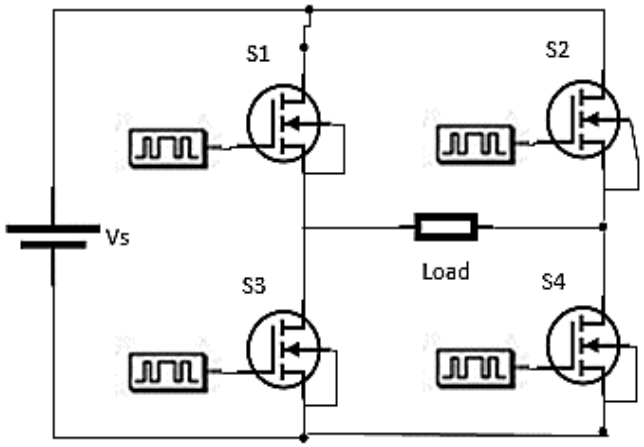
$$V_L = V_s \tag{5}$$

For time $\frac{T}{2} < t \leq T$ the power flow path is via MOSFET 'S2' and MOSFET 'S3'. The voltage obtained at load is 'V_L'

$$V_L = -V_s \tag{6}$$

It can be seen that the magnitude of the voltage at output 'V_L' is increased in comparison to the half bridge converter because half

of the voltage which dropped across the capacitor is now retrieved by using switches instead of capacitors.



The half bridge converter limitation is overcome by using the full bridge converter. So with the help of the full bridge converter, the output voltage is increased.

3. Dual Input Single Output

The dual input single output (DISO) DC-DC converter shown in Figure 3 is presented in [1]. It is a DC-DC converter without a transformer. The converter is composed of one inductor, two capacitors and four switches.

The advantage of this converter is the controlled flow of power between the input and the output. The two input sources are, the unidirectional battery and the PV panel. The use of the battery is to provide power to the load side.

The PV panel is used as a primary source for utilization of the renewable energy while at the secondary source; the battery delivers the power when there is a power reduction from the solar system.

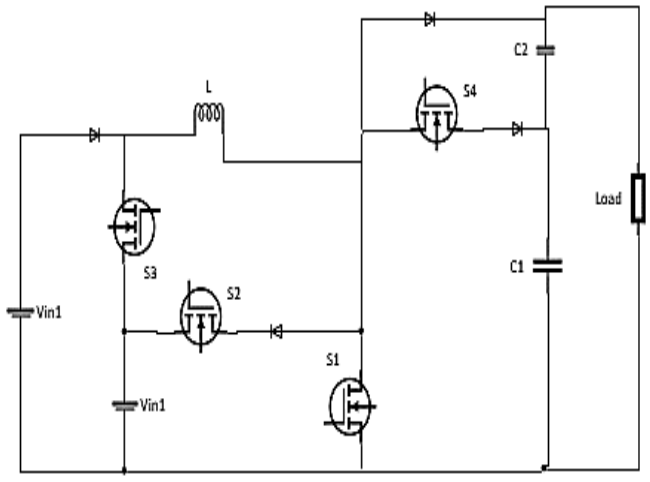


Figure 3: Dual Input Single Output Converter

4. Comparison between Transformer and Transformer-less Multiport Converter

Both the systems, with a transformer and without a transformer, delivers power but with some specifications whether a transformer is required or not in an application, depends on the particular application.

4.1. With transformer

A transformer is used in any of the applications listed below:

- For step up of voltage.
- For step down of voltage.
- For creation of either magnetic or galvanic isolation.
- For reduction of transformer current, Delta-wye transformers should be used.
- Fault current can be restricted.
- For obtaining multiple output voltages, taps in the transformer can be used.

There are also some disadvantages in using transformers:

- Cost increment
- Area increment
- Weight increment
- Reduction in Efficiency

4.2. Without transformer

In comparison to a system with a transformer, the system without a transformer also works, and it has some unique characteristics as well:

- Reduction in Cost
- No heavy circuitry
- Reduction in Weight
- Reduction in number of components
- Efficiency increment
- Fast Response

Some disadvantages of system without transformer are:

- Isolation is not present
- The level of voltage remains the same
- No taps availability as in the case of the transformer

5. Results

5.1. Half Bridge converter

Figure 4 displays the output voltage of the half bridge converter. It can clearly be seen from the graph that the output voltage reaches a maximum magnitude of 19 V after which it stabilizes steadily.

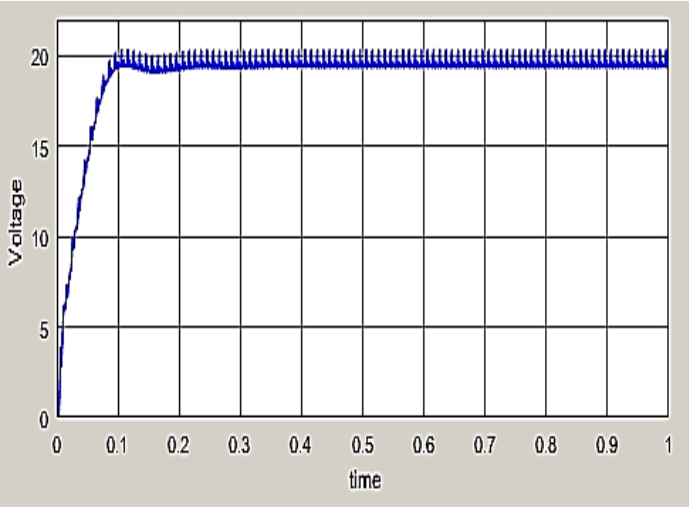


Figure 4: Half Bridge Converter Output Voltage

There is an overshoot produced at 0.1seconds but till 0.2 seconds the voltage stabilizes completely. The ripples produced in the output are due to the capacitor of a smaller magnitude being used. If the value of capacitor is increased the output stabilization time will be high.

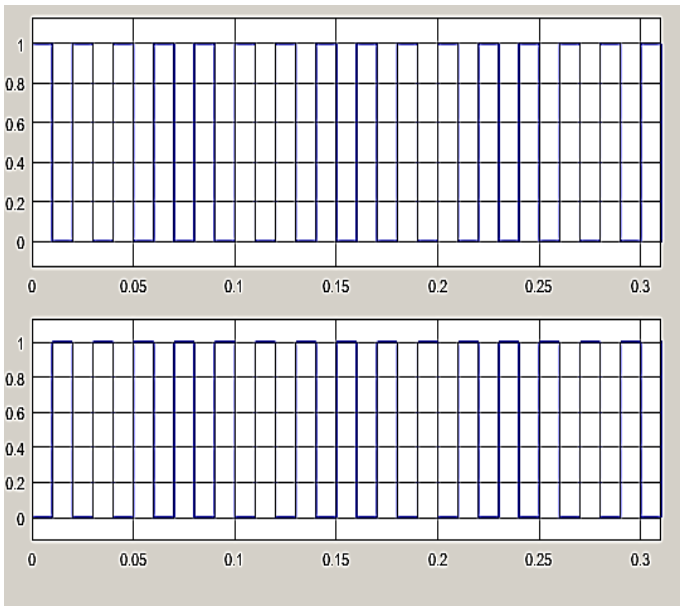


Figure 5: Half bridge Converter Switches Signals

The states of the switches has been displayed in Figure 5, where the first signal displayed has been provided to switch 'S1' while its inverted wave that is displayed in the second signal in Figure 5 is for the switch 'S2'.

This means that while the switch 'S1' is ON, 'S2' will remain OFF and vice versa. Similarly, the same case is observed for the remaining bridges in the converter.

5.2. Full Bridge Converter

In Figure 6, the voltage at the output of the full bridge converter is displayed. It is observed that the voltage can be stabilized at around 190 V till 0.02s with the help of high capacitances.

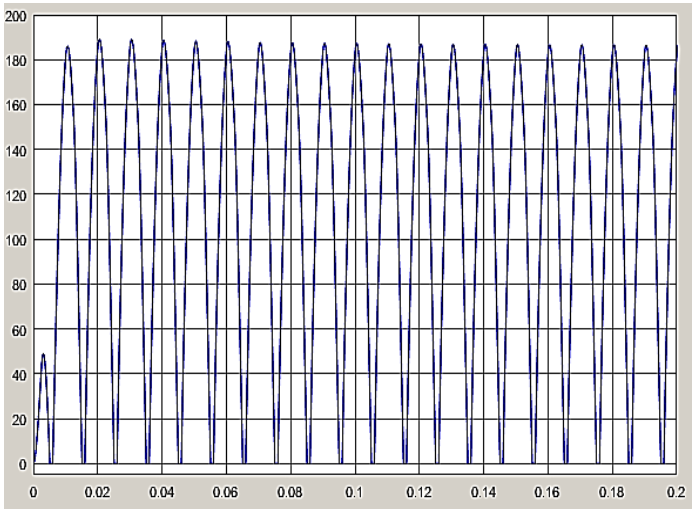


Figure 6: Full Bridge Converter Output Voltage

The produced ripples in the output voltage are a result of using capacitors with small magnitudes. If the magnitude of the capacitors is increased, the output time taken for the converter to stabilize would increase.

The switching signals provided to the H-bridge have been represented in Figure 7. The four signals (from top to bottom) are being provided to the four switches from 'S1', 'S2', 'S3' and 'S4'. For controlling on side of the H-bridge 'S1' and 'S3' are used while 'S2' and 'S4' are used for controlling the other end of the H-bridge. It can be viewed that while 'S1' (signal waveform 1) is switched ON, 'S3' (signal waveform 3) remains OFF and vice versa while when 'S2' (signal waveform 2) is switched ON, 'S4' (signal waveform 4) remains OFF during that interval and vice versa.

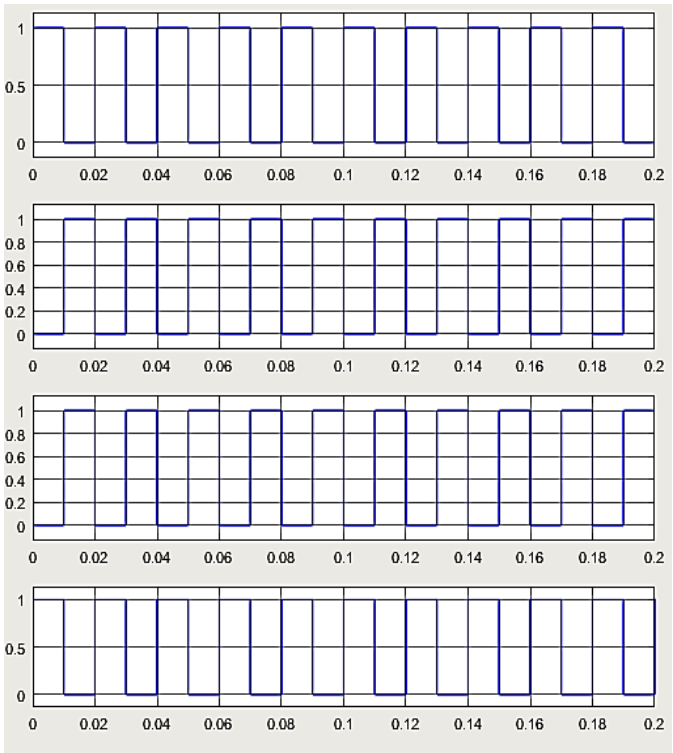


Figure 7: Full Bridge Converter Switches Signals

5.3. Dual Input Single Output

The result obtained at fixed resistance of 70 ohm for all three converters are shown in the table below.

Table: Voltage at inputs and output of multiport DC-DC converter

Type of Converter	Source 1 (V)	Source 2 (V)	Output Voltage (V)
DISO converter (without transformer)	20	20	39
FBTP DC-DC converter (with transformer)	34.5	50	400
HBTP DC-DC converter (with transformer)	34.5	50	152

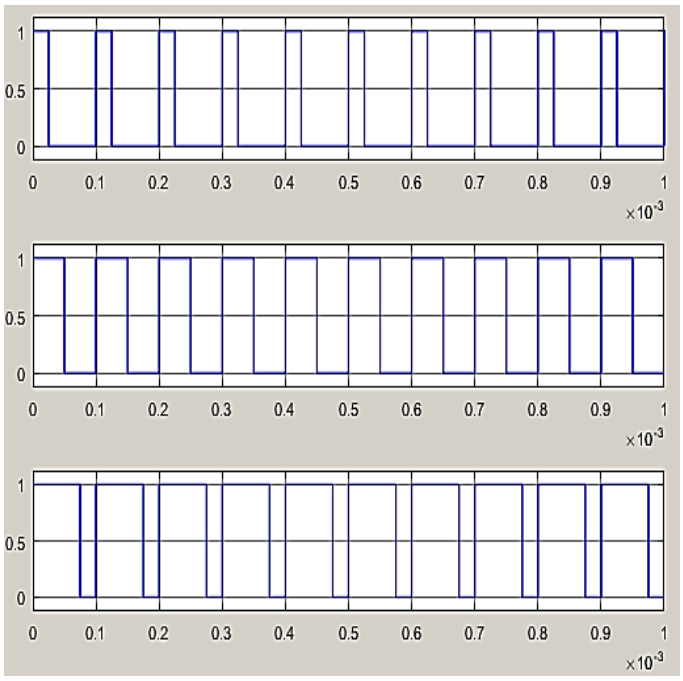


Figure 8: Dual Input Single Output Converter Switches Signals

Figure 8 represents the switching pulses for Dual Input Single Output converter while Figure 9 displays the output voltage for the same DISO converter.

It can be seen from Figure 8 that the switching frequency has been altered a bit for all the three switching signals to a ratio of 0.25%.

The first switch remains ON for 0.25% while the second switch remains ON for 50% and the third one remains ON for 75% for the same time interval (that is from 0 to 0.1ms).

The final output of this converter is displayed in Figure 9, which reaches a maximum of 57V in the overshoot and later on stabilizes at 39V. The time taken for the converter to stabilize in 0.3s.

The efficiency of this converter is quite high compared to the other remaining converters but its major drawback is that one user cannot alter the voltage level in this converter.

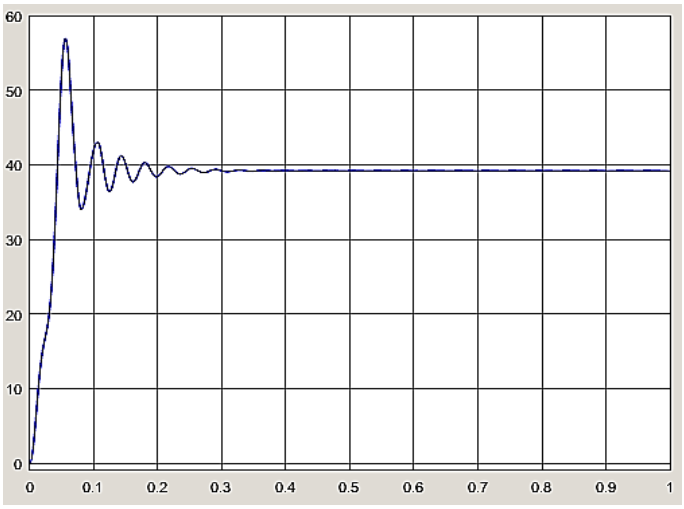


Figure 9: Dual Input Single Output Converter Output Voltage

It can be seen from the above Table that the voltage level of Dual input single output converter remains the same as it is nearly equal to the input source's voltage. The result obtained shows that it is a highly efficient system as losses are very low in this converter. It has a low cost because less number of components are involved in it.

The ripples in the output are very low and the harmonics are not present in this converter. The structure of this converter is quite simple and easy to implement but its control strategy is a complex task.It can be seen from the Table that the input voltage level of the three port full bridge converter is increased. The output result is almost 12 times higher than the input voltage level.

This converter is not efficient as compared to the DISO converter and its cost is higher. The advantage of this converter over the DISO converter is that the voltage level can be increased or decreased as per requirement. The harmonics are less compared to the half bridge converter but the ripples produced in the output are high.

The structure of this converter is complex but its controlling strategy is simple. A little complexity is to cater the switching time of the transistors or switches, so that all the four switches can work properly. It can be seen from the Table that the voltage level of the three port half bridge converter is increased.

The output result is 4.4 times higher than input voltage level. This converter is less efficient compared to the above two converters and its cost is high. In this converter, the efficiency is very low because of high losses in the converter.

The harmonics produced in the output is high and the ripples are also high compared to the full bridge converter. The structure of this converter is not simple but control is quite easy, as only two switches are involved. When one switch is open, the other switch is closed and vice versa.

6. Conclusion

In this paper, multiport converters are discussed. The converters are of different types and a comparison between the converters is presented on the basis of the multiport converter with a transformer and without a transformer. The multiport converters presented in research are divided into two groups. One with a transformer converter and the other without a transformer

converter. The converters with transformer are further subdivided into two parts.

- Full Bridge Three Port DC-DC Converter
- Half Bridge Three Port DC-DC Converter

The advantages of the converter with a transformer are: It steps-up or steps-down the voltage level, provides isolation between the input and output sides, the transformer current can be reduced by Delta-wye transformers, fault current is reduced by impedance, magnetic isolation can be attained and by the taps in transformer we can attain multiple voltages in the output.

The disadvantages of the converter with a transformer are: efficiency reduction, in some cases weight is increased; area increment of converter is also present while the cost also increases. The harmonics components are greater due to the transformer and it has a complex structure (whereas its control is quite easy). Now moving towards the converters without a transformer. The transformer-less converter presented was the Dual Input Single Output converter.

The advantages of the transformer-less converter in comparison to the converter with transformer are: it has a reduced cost factor, less components are used so the area of circuit is reduced. The harmonics are very low and its efficiency is high due to low losses in the converter since the output voltage (39V) is almost equal to the input source voltages (20V each). Its structure is easy to implement.

The disadvantages of transformer-less converter in comparison with the converter with transformer are: only a single level of voltage is obtained, there is no magnetic isolation or galvanic isolation between input and output side, the fault current in not affected as there is no transformer present. The control of this converter is difficult because we have to control more than two pulses at a single time (Source 1 and Source 2).

The complete research work concludes that; every DC-DC converter has its own advantages and disadvantages, so it depends upon the application for which the converter is to be used. For example if we need to change the voltage level, we have to use a converter which contains a transformer while in the case when voltage level change is not required we can use a transformer-less converter.

Acknowledgment

I, personally, thank the Iraqi government represented by the Ministry of higher education and scientific research, Al-Furat Al-Awsat Technical University, Najaf, Iraq for its financial support.

References

[1] Al-Chlaihawi, S. J. (2016, October). Comparative study of the multiport converter used in renewable energy systems. In Applied and Theoretical Electricity (ICATE), 2016 International Conference on (pp. 1-6). IEEE.

[2] AL-Chlaihawi, S. J., & Al-GIZI, A. G. (2016, June). A survey of multiport converters used in renewable energy. In Fundamentals of Electrical Engineering (ISFEE), 2016 International Symposium on (pp. 1-4). IEEE.

[3] Hawke, J. T., Krishnamoorthy, H. S., & Enjeti, P. N. (2014, September). A multiport power sharing converter topology for renewable-to-grid interface. In Energy Conversion Congress and Exposition (ECCE), 2014 IEEE (pp. 4992-4999). IEEE.

[4] Ramki, T., & Tripathy, L. N. (2015, January). Comparison of different DC-DC converters for MPPT application of photovoltaic system. In International Conference on Electrical, Electronics, Signals, Communication and Optimization (EESCO) (pp. 1-6).

[5] Khosrogorji, S., Torkaman, H., & Karimi, F. (2015, February). A short review on multi-input DC/DC converters topologies. In Power Electronics, Drives Systems & Technologies Conference (PEDSTC), 2015 6th (pp. 650-654). IEEE.

[6] Al-Gizi, A. G., & Al-Chlaihawi, S. J. (2016, June). Study of FLC based MPPT in comparison with P&O and InC for PV systems. In Fundamentals of Electrical Engineering (ISFEE), 2016 International Symposium on (pp. 1-6). IEEE.

[7] Zeng, J., Qiao, W., Qu, L., & Jiao, Y. (2014). An isolated multiport DC–DC converter for simultaneous power management of multiple different renewable energy sources. IEEE journal of emerging and selected topics in power electronics, 2(1), 70-78.

[8] Falcones, S., Ayyanar, R., & Mao, X. (2013). A DC–DC multiport-converter-based solid-state transformer integrating distributed generation and storage. IEEE Transactions on Power Electronics, 28(5), 2192-2203.

[9] Wu, H., Chen, R., Zhang, J., Xing, Y., Hu, H., & Ge, H. (2011). A family of three-port half-bridge converters for a stand-alone renewable power system. IEEE Transactions on Power Electronics, 26(9), 2697-2706.

[10] Wu, H., Sun, K., Chen, R., Hu, H., & Xing, Y. (2012). Full-bridge three-port converters with wide input voltage range for renewable power systems. IEEE Transactions on Power Electronics, 27(9), 3965-3974.

[11] Chen, Y. M., Huang, A. Q., & Yu, X. (2013). A high step-up three-port dc–dc converter for stand-alone PV/battery power systems. IEEE Transactions on Power Electronics, 28(11), 5049-5062.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 913-917 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Structural Optimization of Wavy FinFET for Leakage Reduction and Performance Enhancement

Anju Chakkikavil[*, 1], Nisha Kuruvilla[1], Ayoob Khan[1], Shahul Hameed[2]

[1]*Department of Electronics Engineering, College of Engineering Chengannur, 689121, India*

[2]*Department of Electronics and Communication Engineering,* TKM College of Engineering*, 691005, India*

A R T I C L E I N F O

Article history:
Received: 05 April, 2017
Accepted: 24 May, 2017
Online: 30 June, 2017

Keywords:
FinFET
Ultrathin Body
Wavy
SOI
Leakage
Spacer

A B S T R A C T

Wavy FinFET is a hybrid device that integrates FinFET and Ultrathin Body FET technologies on SOI platform to provide high density and drivability without causing area penalty. The problem associated with this device is higher leakage and lower threshold voltage. This problem can be solved by structural modification of the device. This work analyses the performance of Wavy FinFET under various optimization methods such as device/channel engineering, work function engineering, spacer engineering and variation of isolation oxide thickness. The obtained results indicate that optimized Wavy FinFET can act as a solution for low power, highly reliable device topology. Leakage power is found to be reduced by 40.39%, 30.39% and 43.75% with channel engineering techniques such as substrate doping., halo doping and retrograde doping, respectively. Leakage power is lowered by 35.48% and 32.25% with increase in gate work function and isolation oxide thickness respectively. By using high k spacer material 54.77% reduction in leakage is further obtained without compromising drive current. In ADSE(Asymmetric Drain Side Extention) wavy FinFET,leakage is reduced by 61.35%. Whereas in symmetric and asymmetric Dual k wavy FinFET, leakage reduction is found to be 44.19% and 28.25% respectively.

1. Introduction

Gordon Moore predicted the scaling trends in integrated circuits in 1965 [1]. The technology driving factors are reduced cost, improved performance and greater circuit functionality. Scaling causes reduction in gate control over current flow in channel and which will in turn results in short channel effects. In order to mitigate these effects, engineers had to innovate advanced MOSFET structures which can be used in smaller technology nodes. Gate oxide scaling combined with gate length scaling was the traditional MOSFET scaling which ran out in the early 2000's due to leakage limitations. Engineers had to innovate new structures since market was demanding low leakage. Strained silicon is an example of a revolutionary technology that provided improved performance when traditional MOSFET scaling methods were no longer effective [2]. It provided valuable performance enhancements at 90 nm and 65 nm generations. Intel's 45 nm logic technology was the first to introduce high-k + metal gate transistors for improved performance and reduced leakage [3]. Further scaling was achieved using structural change in transistor architecture. A layered silicon-insulator-silicon substrate replaced conventional silicon substrates to form SOI technology[4]. The ultrathin body FET eliminates the leakage paths between source and drain , and thereby control short-channel effects with thin body [5]. Multiple gate devices such as double gate, trigate, pi-gate, GAA devices could achieve better gate control over channel [6]. ITFET (Inverted T channel FET) combines double gate FinFET and single gate SOI to achieve layout efficiency[7].

Wavy FinFET is a hybrid device comprising of FinFET and Ultrathin body FETs [8]. Since UTB (Ultrathin Body) is inserted in the otherwise wasted fin-fin spacing, drivability increases (because of the increased path for current flow) without causing area penalty. The major issues related to wavy FinFET architecture are increase in leakage and reduction in threshold voltage. Optimization are required to mitigate these problems and thereby to obtain better device performance.

* Anju Chakkikavil, Department of Electronics Engineering, College of Engineering Chengannur, 689121, India | Email: anjucraman@gmail.com

In the following session, comparative analysis of trigate SOI FinFET and wavy FinFET are done. After that optimization methods such as channel engineering, work function engineering, spacer engineering and isolation oxide thickness variation are done. Finally the last session concludes this work.

2. Trigate v/s Wavy FinFET

3D structure of Trigate SOI FinFET, segmented trigate SOI FinFET with 3 fins and wavy FinFET having dimensions as per 2013 Overall Roadmap Technology Characteristics Fin-FET parameters specified by the International Technology Roadmap for Semiconductors (ITRS) at the 7-nm technology node [9] were drawn using COGENDA TCAD [10]. In all structures fin height width and fin termination edge are 30nm, 6.8nm and 25nm respectively. Gate length is 18nm and EOT is 0.5nm. The channel is intrinsic and UTB thickness in wavy FinFET is 4.5nm. Fin spacing is 19nm in the case of multifin architecture. The structures are given in Figure 1. It can be seen that Ultrathin body is inserted in the otherwise wasted fin spacing to achieve advantages of two competing technologies. Figure 2 shows the gate cross section of trigate FinFET with single fin, with 3 fins and that of wavy FinFET. We can see that channel area under gate in wavy FinFET is higher compared to that of conventional FinFET.

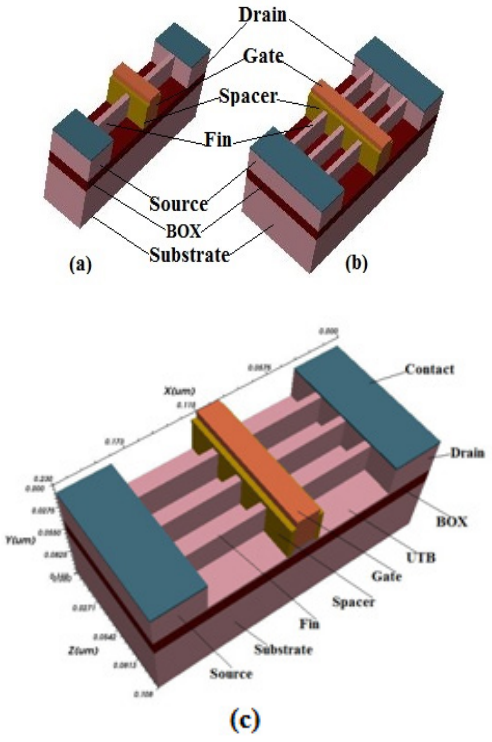


Figure 1.(a) Trigate FinFET with single fin, (b) Trigate FinFET with 3 fins, (c) Wavy FinFET

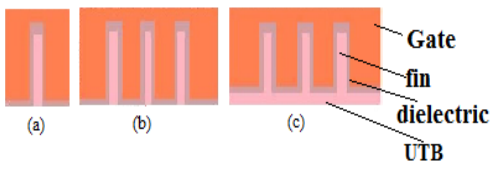


Figure 2. Gate cross section of (a) Trigate FinFET with single fin, (b) Trigate FinFET with 3 fins and (c) Wavy FinFET

Simulations are carried out to analyase the performance of normal trigate FinFET with single fin, trigate FinFET with multiple fins and wavy FinFET. Also the Ultrathin body thickness is varied from 2.5nm to 4.5nm to see the variation of current with ultrathin body thickness. In Table 1, W2.5, W3.5, W4.5 represents wavy FinFET with UTB thickness 2.5nm, 3.5nm, 4.5nm respectively.

Table 1. Comparison of conventional FinFET and wavy FinFET (UTB thickness=2.5nm, 3.5nm, 4.5nm)

Stru Cture	I_{off} (A)	I_{on} (A)	V_t (m V)	SS (m V/d ec)	DIB L
trigate 1 fin	2.05e-13	1.03e-05	372	62	40
trigate 3 fins	3.21e-13	2.43e-05	371	62.8	40.2
W 2.5	1.09e-12	2.79e-05	367	65.7	43.6
W 3.5	3.35e-12	2.95e-05	365	70.2	47.3
W 4.5	1.55e-11	3.82e-05	355	73.9	53

It can be seen that the drive performance of segmented FinFET is better than that of conventional FinFET. In order to increase the drive current further, it is required to array more number of fins which will result in area penalty. At the same time wavy FinFET overcomes this difficulty and current is increased keeping the same footprint area. Also it is found that drive capability and short channel characteristics varies with the thickness of UTB in wavy FinFET. It can be seen that, higher UTB thickness results in I_{ON}. But the problem is along with ON current, OFF current also increases. Threshold voltage of wavy FinFET is lesser compared to that of conventional FinFET which further reduces with increase in UTB thickness.

Due to threshold voltage mismatch between UTB and fins, subthreshold swing and DIBL becomes higher in wavy channel transistor. Hence wavy FinFET has to be optimized to achieve higher I_{ON} with lower leakage power.

3. Optimized Low Power Wavy FinFET

Substrate back biasing is one of the methods for threshold voltage tuning [8]. Reverse back biasing helps to increase the V_t and reduce I_{OFF}. The problem here is that the reduction in leakage is obtained at the expense of ON current. Hence another methods are needed to optimize wavy FinFET, which is discussed in the following sessions of this work.

3.1. Substrate Doping

Doping concentration of the silicon is varied from undoped to 1e17 cm^{-3}. Fig. 3 represents the variation in V_t with doping concentration. V_t increases as doping increases. Table 2 represents the effect of doping on I_{ON}, I_{OFF} and leakage power. Both linear and logarithmic plot of V_{gs}- I_d characteristics with different doping concentration are shown in Figure 4. It can be observed that along with increase in doping concentration, leakage is reduced. Hence leakage power is also reduced. 40.39% reduction in leakage current is obtained by choosing the doping concentration to be 1e17/cm^3. Here a 4.71% reduction in ON current is observed.

Table 2. Effect of doping

Doping (cm^{-3})	I_{off} (A)	I_{on} (A)	P_{leak} (W)
Undoped	1.55e-11	3.82e-05	1.24e-11
1e16	9.79e-12	3.67e-05	7.83e-12
1e17	9.24e-12	3.64e-05	7.39e-12

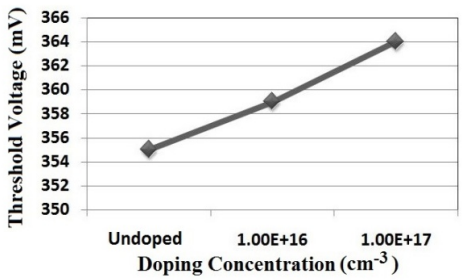


Figure 3. V_t variation with doping concentration

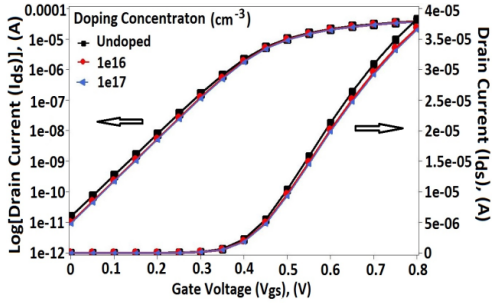


Figure 4. Transfer characteristics with different doping concentration

3.2. Halo and Retrograde Doping

Halo doping can be performed by incorporating highly doped p type region near source and drain ends. Here the drain and source junction depletion region width reduces. It can help in the reduction of DIBL (Drain Induced Barrier Lowering) and punch through. For retrograde doping surface concentration is kept low to increase surface mobility and the subsurface doping is kept high to act as a barrier against punch through [11]. Effect of both halo doping and retrograde doping are shown in Figure 5 and Table 3. By using these methods leakage can be reduced.

Table 3. effect of halo and retrograde doping

	undoped	Halo doping	Retrograde Doping
I_{off}	1.55e-11	1.079e-11	8.718e-12
I_{on}	3.82e-05	3.72e-05	3.696e-05
P_{leak}	1.24e-11	8.632e-12	6.968e-12
DIBL	53	48	45.8

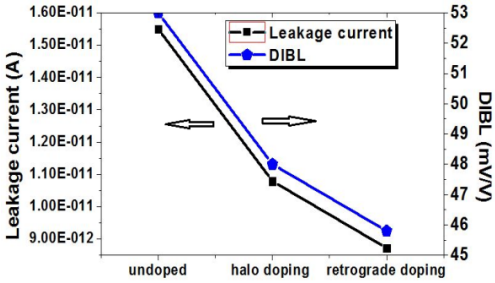


Figure 5. Effect of Halo and retrograde doping

Here the advantage is that we get reduction in leakage power without much sacrificing drive performance. Ion/Ioff ratio also will be higher.

3.3. Work function Engineering

Threshold voltage is dependent work function. Hence by varying gate work function of wavy FinFET, better performance can be achieved. Gate work function is varied from 4.5eV to 4.6 eV. Effect of work function is shown in Figure 6 and Table 4. As the work function increases, threshold voltage increases and leakage reduces. 35.48% reduction in leakage power is obtained by increasing work function to 4.6 eV.

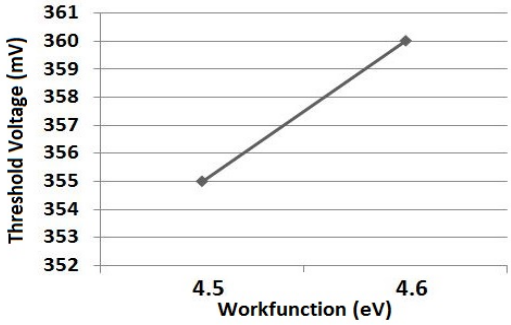


Figure 6. Effect of work function on V_t

Table 4. Effect of work function on performance of wavy FinFET

Work Function(eV)	I_{off} (A)	I_{on}(A)	V_t (mV)	P_{leak}(W)
4.5	1.55e-11	3.82e-05	355	1.24e-11
4.6	1.00e-11	3.73e-05	360	8.00e-12

3.4. Isolation oxide thickness variation

This is a method for leakage suppression. Threshold voltage is directly proportional to isolation oxide thickness. Increasing isolation oxide thickness leads to higher threshold voltage which further leads to reduced leakage. Isolation oxide thickness is varied from 20nm to 40nm. Effect of BOX thickness variation is given in Figure 7. Leakage power is found to be lower in the case of higher oxide thickness.

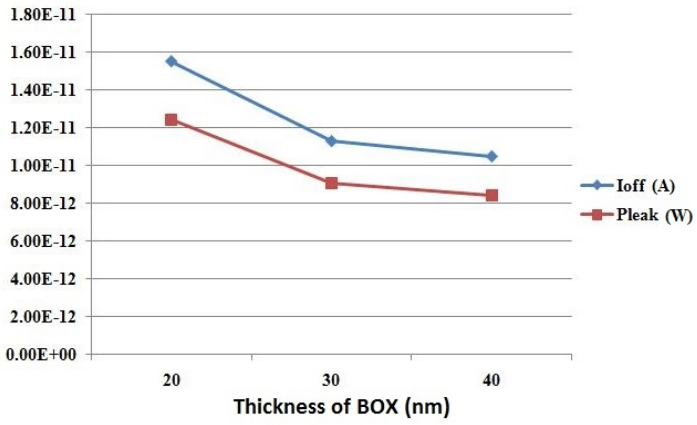


Figure 7. Effect of isolation oxide thickness on I_{off} and leakage power

3.5. High-k Spacer

In sub-20 nm region, short channel effects are reduced using underlap regions. But here on current reduces. Even at high Vds, gate-source/drain barrier restricts source to drain carrier flow. Introduction of high-k spacers can provide strong field coupling from the gate. Effect change in spacer material is checked by using SiO_2 (k=3.9), Si_3N_4 (k=7.5) and Hfo_2 (k=22) as the spacers. Simulation result is shown in Figure 8. As the dielectric constant of the spacer material increases, an increase in I_{on} and reduction in I_{off} can be seen. And thus device can be optimized for lower leakage power and higher I_{on}/I_{off} ratio. Effect of change in spacer material on I_{off}, I_{on}, V_t and leakage power is shown in Table 5.

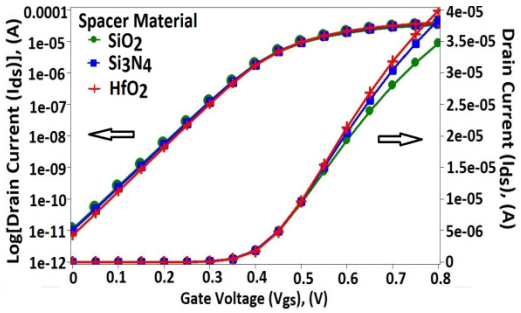


Figure 8. Transfer characteristics of wavy FinFET with different spacer materials

Table 5. Effect of change in spacer material

Material	I_{off}(A)	I_{on}(A)	V_t(mV)	P_{leak}
SiO_2	1.09e-11	3.46e-05	356	8.72e-12
Si_3N_4	1.01e-11	3.84e-05	362	8.08e-12
HfO_2	7.01e-12	3.98e-05	396	5.61e-12

Higher the dielectric strength of spacer, lesser the leakage, and higher the threshold voltage can be obtained which leads to a better device performance. By using HfO2 as the spacer material on both sides of gate, 54.77% reduction in off current is achieved.

3.6. Asymmetric Drain Spacer Extention (ADSE) Wavy FinFET

In asymmetric drain spacer extention FinFET as shown in Figure 9, an additional spacer is introduced only on the drain side to introduce drain-side underlap in FinFET. Thus, bidirectional currents in these transistors are not the same. The spacer thickness at drain side is higher which will result in asymmetry in current and reduction in DIBL. DIBL and OFF current is found to be reduced by using ADSE structure.

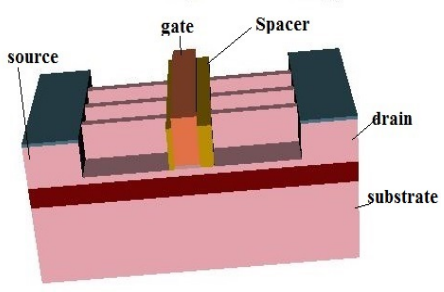


Figure 9: ADSE Wavy FinFET

Effect of ADSE wavy FinFET on drain current is shown in Figure 10. Drain current in logarithmic scale is shown on left Y axis and from which it can be understood that the OFF current is lower in ADSE Wavy FinFET compared to normal wavy FinFET. Here the leakage power is reduced by 61.35% compared to that of normal wavy FinFET.

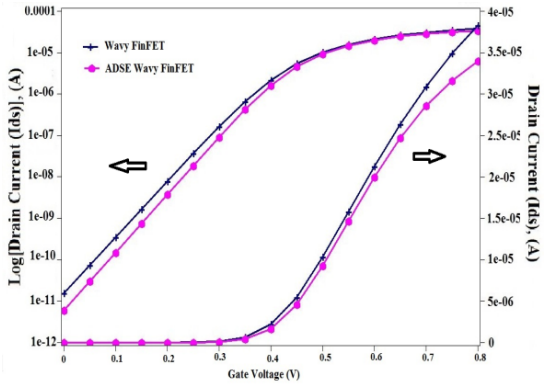


Figure 10: Effect of ADSE Wavy FinFET

3.7. Symmetric and Asymmetric Dual-k Spacer Architecture

The problem associated with high-k spacer is that it increases the fringe capacitance (C_{fr}) that worsens the circuit delay in digital applications [12]. By introducing dual k spacer, instead of single spacer material, inner high-k and outer low-k spacer material is introduced. Interface region between high-k and low k region can make significant changes in device properties. Ion/Ioff can be increased.

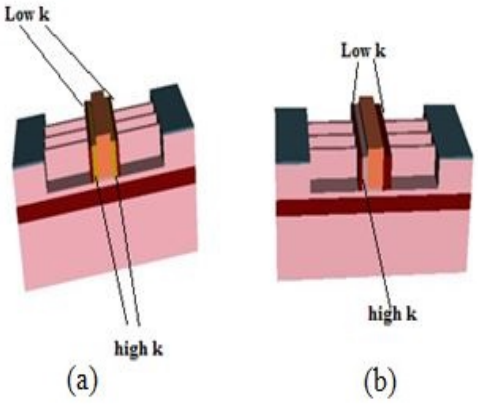


Figure 11: (a)Symmetric dual-k and(b)asymmetrc dual-k wavy FinFET.

In symmetric dual-k as shown in Figure 11(a), there will be dual k spacer on both sides of gate. Where as in asymmetric dual k structure as shown in Figure 11(b), dual k spacer is present only on source side. Effect of introducing symmetric dual-k spacer on leakage power and Ion/Ioff ratio are shown in Figure 12. For dual-k, Si3N4 and HfO2 are used as high k spacers along with SiO2 spacer. 44.948% increase in I_{ON}/I_{OFF} is obtained with symmetric dual-k structure (HfO2- inner high-k spacer) compared to conventional with SiO2 spacer.

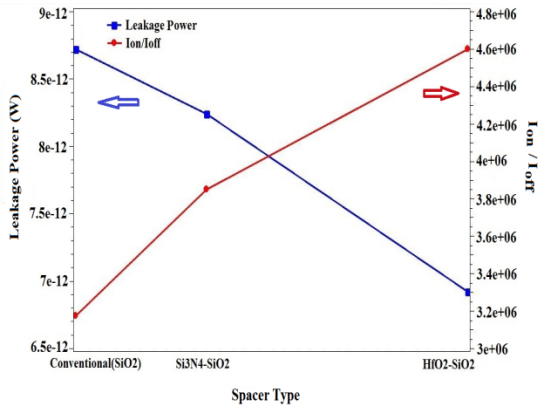


Figure 12: Effect of symmetric dual-k structure

Table 6 shows the effect of introducing asymmetric dual k spacer in wavy FinFET architecture. 28.25% reduction in leakage power is achieved by introducing asymmetric double dielectric spacer in wavy FinFET architecture

Table 6: Effect of asymmetric dual k spacer

Structure	I_{OFF}(A)	I_{ON}(A)	I_{ON}/I_{OFF}	$P_{leakage}$
Wavy FinFET	1.55e-11	3.82e-05	2.46e06	1.24e-11
Asymmetric dual k	1.11e-11	3.76e-05	3.38e06	8.89e-12

3. Conclusion

Since wavy FinFET is a hybrid structure of UTBFET(Ultrathin Body FET) and FinFET, it is found to be better than conventional FinFET in terms of drivability, integration density and area efficiency. The problems associated with wavy FinFET are increased leakage power and degraded SCE. This work could propose structural modifications of wavy FinFET using methods such as channel engineering, work function engineering, variation of isolation oxide thickness and spacer engineering to optimize the device for better performance. I_{ON}/I_{OFF} ratios of optimized structures are given in Figure 13 and Percentage change in leakage power is shown in Table 7. It can be observed that these optimization techniques leads to reduction in leakage and improvement in performance of wavy FinFET.

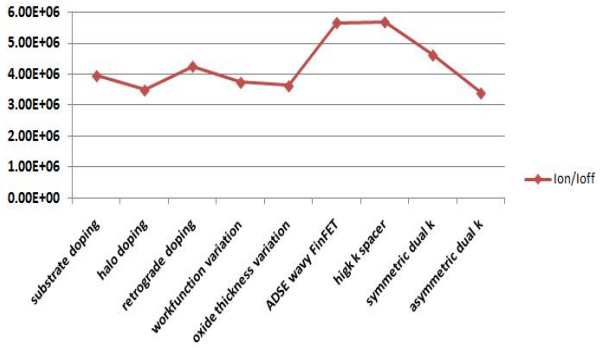


Figure 13: I_{ON}/I_{OFF} in various optimization methods.

Table 7: Percentage change in leakage power and on current with optimization methods

Methods	Value/Percentage reduction in $P_{leakage}$
Unoptimized Wavy FinFET	1.24e-11
Substrate Doping	40.39%
Halo Doping	30.39%
Retrograde Doping	43.75%
Work function Engineering	35.48%
BOX thickness Variation	32..25%
Higk k spacer	54.77%
ADSE wavy FinFET	61.35%
Symmetric dual-k spacer	44.19%
Asymmetric dual-k spacer	28.25%

Acknowledgment

This work was carried out by utilizing resources acquired by College of Engineering, Chengannur, through funded project granted by Department of Science Technology –India (DST-FIST Type-0).

References

This paper is an extension of work originally presented in 2016 IEEE International Symposium on Nanoelectronic and Information Systems [13]

[1] Gordon E. Moore, "Cramming more components onto integrated circuits", https://doi.org/10.1109/N-SSC.2006.4785860
[2] Tatjana Pesic-Brdanin, Branko L. Dokic, "Strained Silicon Layer in CMOS Technology", Electronics, vol. 18, no. 2, , 63-69, 2014. doi:10.7251/ELS1418063P
[3] M. Bohr, "The evolution of scaling from the homogeneous era to the heterogeneous era," in Proc. IEEE Int. Electron Devices Meeting, 2011, pp. 1.1.1–1.1.6. https://doi.org/10.1109/IEDM.2011.6131469
[4] Yuan Taur , "Fundamentals of Modern VLSI Devices"
[5] Y.-K. Choi et al., "Ultrathin-body SOI MOSFET for deep-sub-tenth micron era,"IEEE Electron Device Lett., vol. 21, no. 5, 2000, pages. 254–255. https://doi.org/10.1109/55.841313
[6] Jong-Tae Park et.al, "Multiple-Gate SOI MOSFETs:Device Design Guidelines", IEEE transactions on electron devices, vol. 49, no. 12, 2222-2229, 2002,. https://doi.org/10.1109/TED.2002.805634
[7] W. Zhang et.al, "The ITFET: A novel FinFET-based hybrid device,"IEEE Trans. Electron Devices, vol. 53, no. 9, 2006, pages. 2335–2343.
[8] Hossain M. Fahad et.al, "Simulation Study of a 3-D Device Integrating FinFET and UTBFET", in press. https://doi.org/10.1109/TED.2014.2372695
[9] Overall Roadmap Technology Characteristics, ITRS,2013[Online].Available:http://www.itrs.net/links/2013ITRS/2013Tables/2013ORTC_Detailed Table.pdf, accessed Nov. 25, 2014
[10] VisualTCAD Semiconductor Device Simulator Version 1.7.2 VisualTCAD User's Guide:Available :http://www.cogenda.com
[11] Kausik Roy, "Low voltage, low power VLSI subsystems". http://dx.doi.org/10.1036/007143786X
[12] Pankaj Pal et.al, "Investigation of Symmetric Dual-k Spacer Trigate FinFETs From Delay Perspective", IEEE transactions on electron Sdevices, vol. 61, 3579-3585, 2014. https://doi.org/10.1109/TED.2014.2351616
[13] Anju.C, Nisha Kuruvilla, AyoobKhan T E, Shahul Hameed T A, "Performance Analysis of Wavy FinFET And Optimization for Leakage Reduction", IEEE Inis 2016. https://doi.org/10.1109/iNIS.2016.030

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 918-921 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Program with Ixquic

Carmen Ramos[1], Tania Patiño[2,*]

[1]Department of Visualization and Virtual Reality, Ixtli, UNAM. National Autonomous University of Mexico, UNAM

[2]Circuito Exterior, Nº 3000, Ciudad Universitaria –Coyoacán, CDMX. cramos@unam.mx

ARTICLE INFO

Article history:
Received: 17 April, 2017
Accepted: 19 May, 2017
Online: 30 June, 2017

Keywords:
Story Telling
Serious Games
Game Design
Augmented Reality
Educative Games

ABSTRACT

The goal of this paper is to explain the story telling, game design, game play and how is the interaction with the game Program with Ixquic. Here is the description about the new approach to this video game I called Program with Ixquic. Initially focused on an educational video game but now is a hybrid, now has more levels and provides moments of great fun, where each player can have some much fun in each different level and environments of interaction.

1. Introduction

In this article we consider pertinent to ask the following question from which we set out to create this video game. How can you learn to program in object-oriented philosophy with languages such as Java?. To answer this question is that we created a video game that allows players to learn languages such as Java. This video game we call *Program with Ixquic* and we started to create it from the year 2014, conceptualizing and generating ideas, that would allow us to create what we now have a complete and very fun game. *Program with Ixquic.*

Program with *Ixquic*, is a game that started as an educational game to learn object oriented programming in Java language. After some changes now it's an entertainment game with a bit of the topic of education. Program with *Ixquic*, contains three game levels, consisting of a card game, hidden objects and a 3D environment, in Augmented Reality like a fairy tale, where the player can interact and play in different ways to learn some programming.

Fortunately this game has evolved and with this improved remarkably. In this article we show an advance of this development that has been changing over time from being a simple 2D card game to a small *pandora* box full of incredible surprises that each player will enjoy to the maximum.

Program with Ixquic, is a game (in the future a complete *Serious Game*) that contains some ingredients related to training applications, application that improving cognitive functions, memory training, development of mental skills and improvement of strategic skills.

However, there are many applications that are directed to control sensors to calculate heart rate, moods and is applied in either sensory, psychological or motor rehabilitation of the users [8].

2. Serious Games

2.1 How to define a Serious Games

"Serious Games are a mental contest, played with a computer in accordance with specific rules, that uses entertainment to further government or corporate training, education, health, public policy, and strategic communication objectives. [9]"

2.2 What are the objectives of SG?

SGs seek to improve a particular skill of the user interacting with the game. The idea is that while the player interacts more and more with the game will improve some cognitive, sensory or physical ability [9].

3. Story Telling Program with Ixquic

This is the story that starts this video game *Program with Ixquic*, described in a comic that tells the story of a girl named *Cucita* who is one of the characters of *Program with Ixquic*. The story begins in the CDMX in the traffic of a sunny day and *Cucita*

* Tania Patiño, Madagascar Studio, Coyoacán, CDMX, (52)5539813442 & taniapc@live.com.mx

in a moment of thinking about the immortality of the crab, in traffic imagine that it would be a good idea to create a programming application to learn to program in some language programming in basic form. This is the intro of *Program with Ixquic* and from there starts this video game [2], [7],[8].

Figure.1. Story Telling *Program with Ixquic*

4. Game play of Program with Ixquic

Program with Ixquic, consists of a card game, hidden objects and a 3D environment in Augmented Reality with expanded object display, where the player can interact and play in different ways by learning some programming.

To enter the game appears this interface of *Program with Ixquic*:

Figure.2. Interface *Program with Ixquic* [2].

4.1 User-Centered Design in Program with Ixquic

This is an example of how we were creating each level of play, for example, is the card game level, we first generated it with paper and pencil. We tested user to see if it worked and how it worked. Once we have finished, we begin with the design of each letter and each symbol representing an element of nature, a number or a Mayan symbol. This is an application where is involved the user in each phase of the development process which ensures that the final product responds to their needs and characteristics [10]. Then this is the result of the level 1 of the video game *Program with Ixquic.*

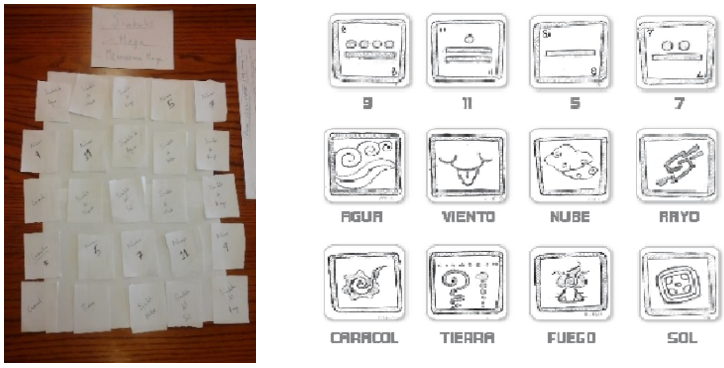


Figure.3.User-Centered Design in *Program with Ixquic*

4.2. Level 1. Card Game

A Card Game, was where we applying the design philosophy and a process related to User-Centered Design (DCU) [10], we understand the requirements and limitations of the end user then we create this type of card game that had magical symbols where we expected that the users can learn and practice in an original way topics related to programming.

This card games comes from ancient times and distinct places in the world: Egypt, India and China. In this case each card has a description of a data structure or a programming concept, thus giving a description of its function[1].

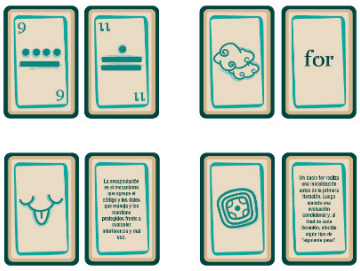


Figure.4. Basic programming concepts

4.3 Level 2 Hidden Figures

In this level called Hidden Figures, we show a landscape where each user have to find (photographs, drawings, graphs, etc.), that are hidden and have to be found. Those things are related to programming concepts or structures, maybe a question is done and then find the thing.

Here is an example of a landscape Yucatán, México, where an investigator have to find many things related to programming topics of Java [6].

Figure.5. Hidden Figures

4.4 Level 4 Augmented Reality

The third level in *Program with Ixquic* is a 3D environment in Augmented Reality with expanded object display. The concept of Augmented Reality brings together a wide variety of user experiences. There are three main categories of RA tools [4],[5].

- Augmented reality games generate immersion experiences.

- Augmented reality 3D visualizer, as it is the case of Augment, allows to place 3D models in real size in the environment, using a 'tracker' to move the model in space.

- Augmented reality browsers enrich the camera viewer with context information. For example, one can point his or her smartphone at a building to visualize a story.

In *Program with Ixquic*, we create a story tale and some interactions with characters in augmented reality technology (RA), and a book with scenarios come to life, allowing gamers to interact with them and experience stories that are told in a more engaging way. This books only need an Internet connection, a camera (it can be a smartphone, a tablet or laptop) to display the tale [2],[3],[4].

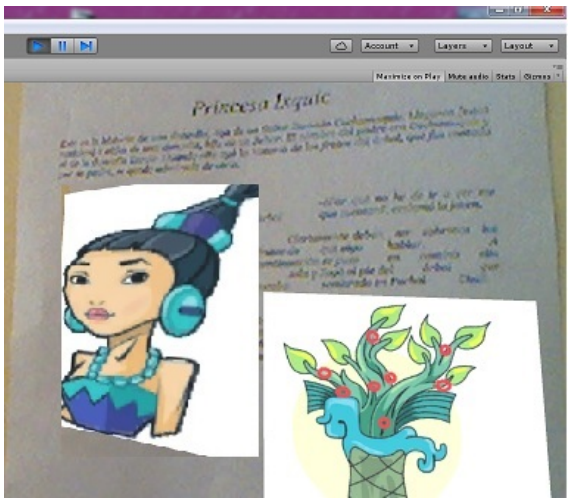


Figure.6. AR *Ixquic Princess Tale*

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

We thank all the team and friends at the National Autonomous University of Mexico, UNAM, specially the Department of Visualization and Virtual Reality, Ixtli, UNAM. Where the development of this Serious Game: Program with Ixquic, started.

Continuing with the creation of *Program with Ixquic,* all the Game Design, Programming levels and Game tests were made at my studio called Madagascar Computer Games and Systems S.A. de C.V. This paper is dedicated to my dear friend *Israel Figueroa Angulo*.

References

Journal articles: Journal titles should be abbreviated according to ISI Web of Science abbreviations.

[1] Patino, T., and Ramos, C., "Program with Ixquic: How to Learn Object-Oriented Programming with a Game," 2015 7th International Conference on Games and Virtual Worlds for Serious Applications (VS-Games), Skövde, 2015, pp. 1-2. doi: 10.1109/VS-GAMES.2015.7295781

[2] Pelletier, C. "Producing difference in studying and making computer games: How students construct games as gendered in order to construct themselves as gendered," in Y. B. Kafai, C. Heeter, J. Denner, & J.Sun (Eds.). Beyond Barbie and Mortal Kombat: New Perspectives on Gender and Games (Cambridge MA, 2006), MIT Press.

[3] McGonigal J. All Game Play is Performance: The State of the Art Game University of California @Berkeley/42 Entertainment. (www.avantgame.com). May 13, 2005.

[4] D.Wagner and D. Schmalstieg, "Making Augmented Reality Practical on Mobile Phones, Part 1," in IEEE Computer Graphics and Applications, vol. 29, no. 3, pp. 12-15, May-June 2009.doi: 10.1109/MCG.2009.46

Conference Papers:

[5] Ramos C., and Patino, T., "Program with Ixquic: Educative Games and Learning in Augmented and Virtual Environments," 2016 8th International Conference on Games and Virtual Worlds for Serious Applications (VS-GAMES), Barcelona, 2016, pp. 1-2. doi: 10.1109/VS-GAMES.2016.7590359

[6] Solís A., Patino T., and Radillo J., Improving Cognitive Abilities with Gaming Concepts in Computer Graphics and Videogames courses. Proceedings of the 2007 DiGRA International Conference: Situated Play, DiGRA 2007, Tokyo, Japan, September 24-28, 2007. Digital Games Research Association 2007.

Books:

[7] Kushner, D. Masters of Doom: How Two Guys Created an Empire and Transformed Pop Culture. May 11, 2004.

[8] McGonigal J. SuperBetter: A Revolutionary Approach to Getting Stronger, Happier, Braver and More Resilient Powered by the Science of Games. September 15, 2015.

[9] Michael D., Chen S.,"Serious Games: Games That Educate, Train, and Inform". Muska & Lipman/Premier-Trade ©2005. ISBN:1592006221

[10] Shneiderman, B. (1998). Designing the user interface: Strategies for effective human-computer interaction(3rd ed.). Reading, MA: Addison-Wesley.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 922-931 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

A Design of SDN Based IP Mobility Management Considering Inter-Domain Handovers and Its Evaluation

Misumi Hata*, [1], Mustafa Soylu[2], Satoru Izumi[1], Toru Abe[1, 3], Takuo Suganuma[1, 3]

[1]*Graduate School of Information Sciences, Tohoku University, 980-8577, Japan*

[2]*Fuculty of Electrical and Electronics Engineering, Istanbul Technical University, 34398, Turkey*

[3]*Cyberscience Center, Tohoku University, 980-8577, Japan*

ARTICLE INFO

Article history:
Received: 18 April, 2017
Accepted: 03 June, 2017
Online: 30 June, 2017

Keywords:
IP mobility
mobility management
Software Defined Network
inter-domain

ABSTRACT

To use IP services continually while users are moving with their mobile devices, IP mobility management that enables them to keep IP communication is necessary. Although Mobile IP is a popular method to manage IP mobility, Mobile IP has a problem that there is a possibility of a mobile node (MN) communicating over a redundant route. Then, Software Defined Network (SDN) based IP mobility management has emerged to solve this problem. Most of these solutions focus on intra-domain handovers and routing. However, since wireless network environment has spread, mobile devices move across domains during communication. To deal with this, we need to look upon inter-domain handovers and routing. In this paper, we propose a SDN based IP mobility management scheme considering situation of mobile devices moving across domains. This proposed scheme focuses on inter-domain handovers and introduces efficient functions of management information sharing and inter-domain routing. Experimental results show the desired effects of the proposed scheme: optimizing communication route and keeping the management information exchange traffic low.

1. Introduction

1.1. Background and Overview

This paper is an extension of work originally presented in 2016 18th Asia-Pacific Network Operations and Management Symposium (APNOMS) [1]. Due to the widespread of wireless network environment, it is now possible to use Internet services in various places and scenes, and IP services like file transfer service, remote terminal service, and virtual desktop service are popularly used [2,3]. These services can be used with small-sized devices capable of wireless communication like laptop PCs, smart phones, tablet PCs. We call them mobile node (MN) from now on. MNs are portable, thus people tend to move around while using IP services with MNs. Especially with session continuous IP services, continuity of service during movement may be a problem. When a MN moves and connects to an access point in a different network domain, MN's IP address may change. If MN's IP address changes during communication and communication session is disconnected, this causes some problems that would affect the use of services may occur such as lack of data and logout from services. Thus, IP mobility management is essential for enabling use of IP services while moving.

Mobile IP is one of the IP mobility management standardized by Internet Engineering Task Force (IETF) [4-6]. Mobile IP is a technology to enable continuous communication when a MN moves during a communication. Even when a MN moves to a different domain, it enables the MN to continue its communication with the IP address that it was using before the movement. However, there are cases that a MN's communication route falls into a triangular routing. This might lead to a degrade in quality of service caused by communication delay. Moreover, it is necessary for a MN to be equipped with Mobile IP function, and this leaves a challenge of application to all MNs.

There are some researches that apply Software Defined Network (SDN) to IP mobility management to solve the problems

*Misumi Hata, 2-1-1 Katahira, Aoba-ku, Sendai 980-8577, Japan, +81222175080, Email: m-hata@ci.cc.tohoku.ac.jp

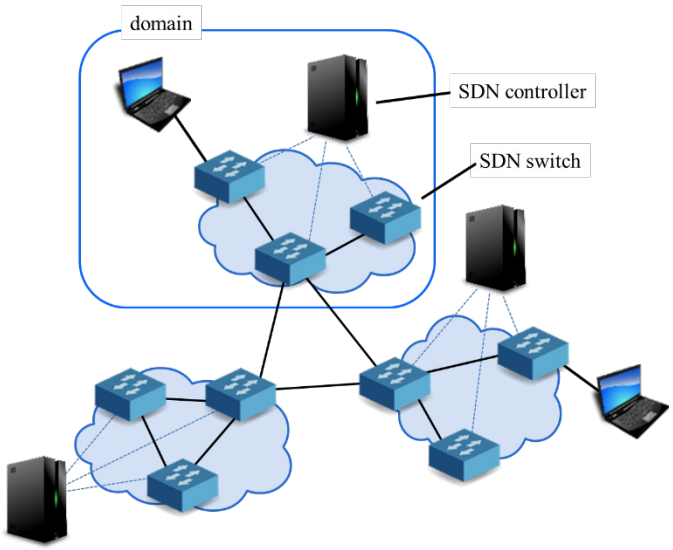


Figure 1 Overview of Software Defined Network (SDN).

mentioned above [7,8]. SDN is a technology to control a network dynamically with a software [9]. We show an overview of SDN in Figure 1. By using SDN, a SDN controller centrally manages SDN switches in their domain and this enables nodes to communicate with suitable route. However, existing researches focus on intra-domain move. To cope with inter-domain movement with SDN, SDN controllers in different domains have to exchange information about movement of a MN. On this occasion, communication cost between SDN controllers may increase and it might become a problem in practical use. Therefore, it is difficult to maintain communication efficiently when a MN makes an inter-domain move.

There were two major trends in ways to handle inter-domain handover with SDN. One approach is to use Mobile IP when inter-domain handovers occurs [8]. By using Mobile IP, SDN controllers doesn't need to search for domains to exchange information of MNs. However, Mobile IP has difficulty in communication route optimization. Another is to announce information of MNs to every controllers in the network [7]. This approach allows the communication routes to be optimized. However, with a network consisted of many domains, communication for the announcement might compress the communication bandwidth. These conventional approaches could only realize efficient data exchange or route optimization.

To overcome both of the problems, we propose an SDN based IP mobility management scheme considering inter-domain handovers. Our proposed scheme consists of two functions. The Management Information Sharing Function (MISF) searches a destination domain of a moved MN effectively and chooses minimum SDN controllers to share minimum information related to MN's communication to manage information efficiently. The Inter-domain Routing Function (IDRF) calculates an appropriate route to allow the MN to continue effective communication.

We also conducted some simulation experiments to confirm the effectiveness of our proposed scheme. The simulation results show that our proposal achieved to optimize communication route and keep traffic between SDN controllers low.

1.2. Novelty and Contribution

To realize efficient inter-domain handovers, we propose a novel scheme to efficiently exchange MNs' information between SDN controllers in each domains and optimize communication route between a MN and a CN. Existing approaches don't realize these two requirements simultaneously because these approaches differ in target, purpose and assumed environment therefore they do not match the purpose of what we concentrate on this time. At the time Mobile IP emerged, route optimization had little meaning because there were not so many candidates of route to be selected. Moreover, it aimed for easy implementation without any modifications in network infrastructure, as it focused on practical use so Mobile IP doesn't have full support to route optimization. The SDN based methods don't aim at covering wide area network and they are approaches for environment with some limitation on scale. Hence they can handle inter-domain handovers but not efficiently. Our proposed scheme realizes 1) efficient information exchange between SDN controllers, and 2) route optimization simultaneously to achieve the inter-domain handovers in large-scaled wide-area network efficiently with functional support of network infrastructure.

With this novelty, mobile communication will be more seamless and delayless. This leads to an improvement of communication quality for users moving around with their communicating MNs. For example, when users with mobile PCs move across domains while downloading data from a cloud server, the disconnection of communication would occur and lead to packet loss. Moreover, when the users move across domains while using interactive-type applications such as remote login terminal or remote virtual desktop environment (VDI), a session would be interrupted. In addition, redundant communication route leads to serious communication delay especially when users are handling large data. By using the proposed scheme, MN users will be able to download or upload their data smoothly and perfectly while moving around. As for the interactive-type applications, MN users will be able to use the services without any termination of the process, even without the applications having the function of switching connections when the IP address changes.

1.3. Paper Organization

In the following sections, we first summarize the problems of Mobile IP and SDN based IP mobility management in section 2. Then we propose the SDN based IP mobility management scheme in section 3, followed by details of functions. In section 4, we evaluate our proposed scheme and show its effectiveness. Finally in section 5, we conclude the paper and discuss future works.

2. Related Work

2.1. Mobile IP

In this section, we introduce related works of IP mobility management and summarize their problems. Internet Engineering Task Force (IETF) standardized Mobile IPv4 to realize IP mobility management [4]. Mobile IPv4 is a protocol to maintain communication between a MN and a Correspondent Node (CN), which communicates with the MN, after the MN moves to other network domain. Figure 2 shows its overview. Mobile IPv4 uses a Home Address (HoA) and a Care-of Address (CoA) to continue communication after MN's move. HoA is an address that MN was using in the source domain and CoA is an address MN is using in the destination domain. The MN registers its binding-cache to its

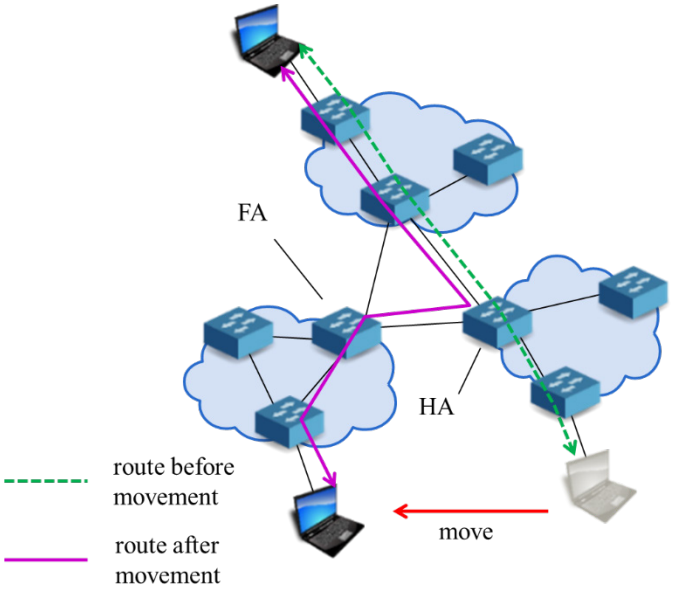


Figure 2 Overview of Mobile IPv4.

Home Agent (HA). Binding-cache is a set of HoA and CoA of a MN. The HA transfers the packets sent to HoA, to CoA. However, there is a possibility that a MN communicate with a CN over a redundant route, which we call triangle routing problem, depending on the location of the HA.

Mobile IPv6 is also standardized by IETF [5]. In addition to functions of Mobile IPv4, Mobile IPv6 has a route optimization function which enables MNs to communicate with the most suitable route. However, this is an optional function so only the MNs with support for route optimization can use this function.

In order to realize IP mobility management, we need to share a variety of information such as HoA, CoA, lifetime of Binding Cache, etc. among a MN, a HA, and a Foreign Agent (FA). IETF standardized Mobile IPv6 Management Information Base [10]. It defines large amount of information needed for Mobile IPv6. It is essential to select where to share plentiful information efficiently to realize IP mobility management. In Mobile IPv4 and Mobile IPv6, the MN registers its binding cache to its HoA. Thus, there is no need of searching for where to share information.

In Mobile IPv4 and Mobile IPv6, the MN needs to conduct IP mobility management. On the other hand, Proxy Mobile IPv6 (PMIPv6) executes IP mobility management on network equipment. Figure 3 shows its overview. The network equipment execute process for IP mobility management in PMIP. Therefore, there is no need of MN to involve in IP mobility management. Local Mobility Anchor keeps bindings of HoA and Proxy-CoA. Mobile Access Gateway (MAG) detects MNs and registers bindings to LMA. After a MN moves, MAG registers the MN's HoA and Proxy-CoA binding to LMA. Packets sent to the MN's HoA are transferred to the MN via LMA. Like Mobile IPv4 and Mobile IPv6, PMIP also has some possibility of triangular routing.

2.2. SDN based IP mobility management

SDN enables us to control networks intensively and flexibly. Therefore, SDN has attracted attentions as a means to solve the route optimization problem. Using SDN, network devices execute

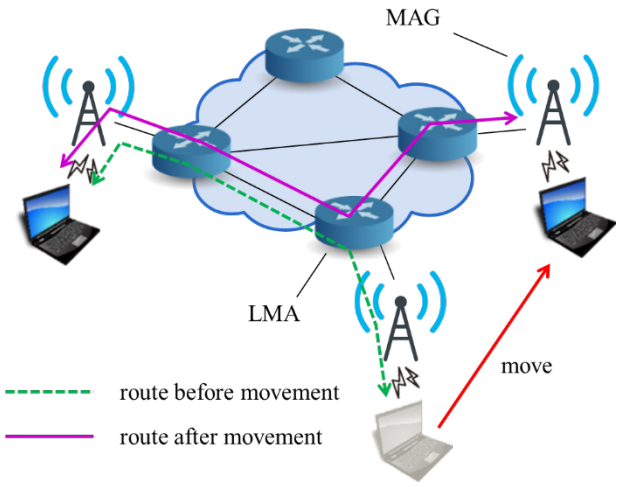


Figure 3 Overview of Proxy Mobile IPv6.

all the IP mobility management processes. Thus, unlike Mobile IPv4 and Mobile IPv6, the MN does not need to be equipped with IP mobility management functions. For this reason, there are some IP mobility management approaches using SDN. Papers [7] and [8] show route optimization mechanisms in the case when intra-domain handover occurs. However, in the case of inter-domain handover, the mechanism in [7] uses Mobile IP, therefore MNs need to be equipped with route optimization function to optimize route after handover. The mechanism in [8] regards inter-domain handover as infrequent and broadcasts this event to all other controllers. Therefore, amount of traffic increases when inter-domain handover occurs frequently in a network consisted by many domains.

2.3. Problems of related works

We summarized the problems of the related works in Table. We confirm that existing works cannot realize intra-domain route optimization, inter-domain route optimization, and efficient selection of domain to share information all together.

The papers [7] and [8] applied SDN to solve the problem of Mobile IP, route optimization, but SDN controllers can only control the network under their managed domain. This means that these approaches focus on intra-domain route. However, in recent network environment, users move freely to various domain networks while using IP services. Therefore, we have to consider the IP mobility management in multiple domain networks.

To realize IP mobility management in multiple domain networks, we need SDN controller in each domain to share information of MNs. However, if a SDN controller broadcasts the information to all other SDN controllers, it might cause large amount of traffic in a situation that inter-domain handover occurs frequently. Hence, we need a mechanism to share information efficiently in multiple domain networks.

3. SDN Based IP Mobility Management Scheme Considering Inter-Domain Handovers

3.1. Overview

In this section, we propose a SDN based IP mobility management scheme considering inter-domain handovers to solve the problems explained in section 2. We first show the

Table Summary of related works

approaches	Mobile IPv4 [4]	Mobile IPv6 [5]	PMIPv6 [6]	Paper [7]	Paper [8]
Intra-domain routing	×	△	△	○	○
Inter-domain routing	×	△	×	○	△
Domain selection to share information	○	○	○	×	○

overview of our proposed scheme and discuss the details of composing functions later. We show the overview of our scheme in Figure 4. This approach realizes inter-domain route optimization while sharing information efficiently with two functions; Management Information Sharing Function and Inter-domain Routing Function.

The Management Information Sharing Function (MISF) searches for the domain which MN was in before movement (source domain) and modifies the domain to exchange MN information to reduce traffic load between SDN controllers. The Inter-domain Routing Function (IDRF) calculates end-to-end route between MN and CN in the domain MN moved into (destination domain) and announces this route only to the domains in the path to optimize inter-domain routes.

We assume that inter-domain topology changes do not occur frequently. Therefore, the SDN controllers have an inter-domain topology information. Inter-domain topology information consists of SDN controller's IP address of each domain, network address of each domain, and inter-domain link information. The functions we mentioned above use this inter-domain topology information to search domain and calculate routes.

We show the components of our proposal in Figure 5. The components include SDN controllers, SDN switches, a MN, and a CN. Our scheme supports IPv4 network with two layers of network: a SDN controller network and a SDN switch network. SDN controller network is consisted of each domain's SDN controller and is used for SDN controllers to exchange information to optimize MN-CN route. SDN switch network is consisted of SDN switches. We call the SDN switch network managed by one SDN controller "domain".

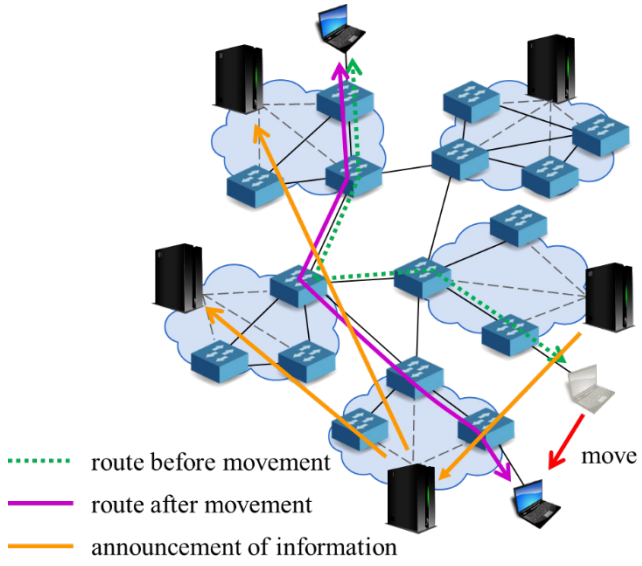


Figure 4 Overview of SDN based IP mobility management considering inter-domain handovers.

3.2. Management Information Sharing Function

We describe the detailed design of MISF with an example. The network we use for explanation is shown in Figure 4. We assume that a MN moves to a destination domain from a source domain during a communication with a CN and the SDN controller of the destination domain detects the connection of the MN. The MN's IP address changes after move. We show below the MAC address, IP address which the MN was using in the source domain (former IP address) and IP address which the MN is using currently in the destination domain (current IP address).

- MAC address: 00:00:00:00:00:11
- former IP address: 10.0.1.1
- current IP address: 10.0.2.1

We define the names of domains and SDN controllers that involve in MN handover as below:

- Domains

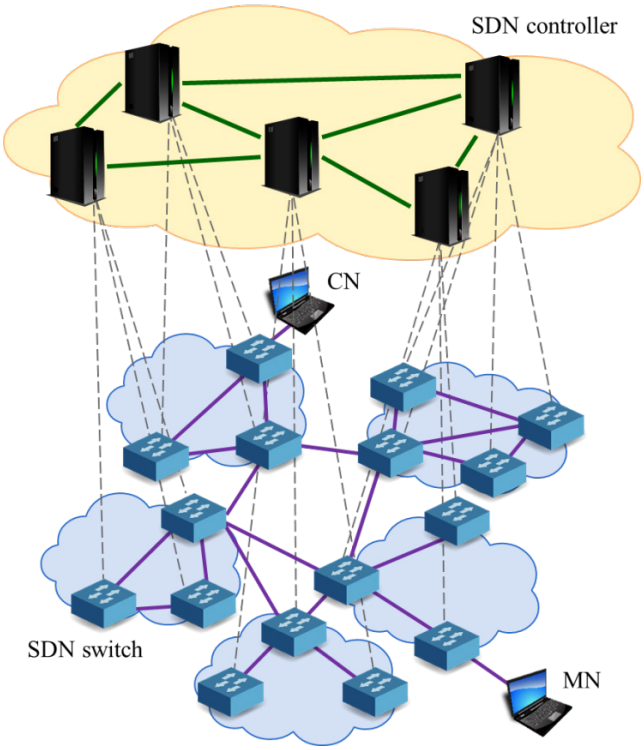


Figure 5 Network configuration.

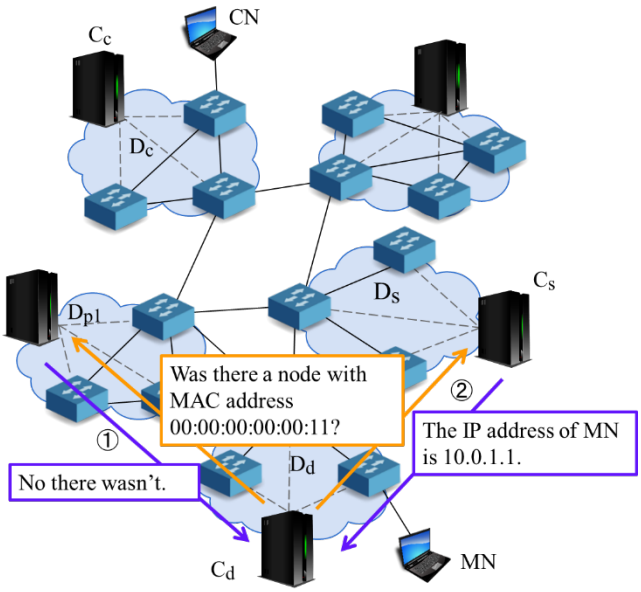


Figure 6 Search for D_s.

- D_d: the domain that MN belongs to after move
- D_s: the domain that MN belongs to before move
- D_c: the domain that CN belongs to

- SDN controllers
  - C_d: SDN controller that manages D_d
  - C_s: SDN controller that manages D_s
  - C_c: SDN controller that manages D_c

We describe the procedure of MISF step by step.

1. C_d searches for D_s based on MAC address of MN.

   C_d sends queries in order from neighboring domains like in Figure 6. SDN controllers of the domains that received the query search their domain to check if MN belonged to, and sends the result to C_d for a reply. When MN was not in their domain, they reply the absence of MN. When MN was in their domain, they reply the IP address MN was using.

2. C_d generates node connection information from the IP address sent from C_s.

   Node connection information consists of following information.

   - MAC address of MN
   - former IP address
   - current IP address

3. C_d sends node connection information to C_c.

   C_d sends the information as shown in Figure 7. On this occasion, C_d gets the IP address of CN from header of the packet sent from MN to CN and derive the information of C_c.

3.3. Inter-domain routing function

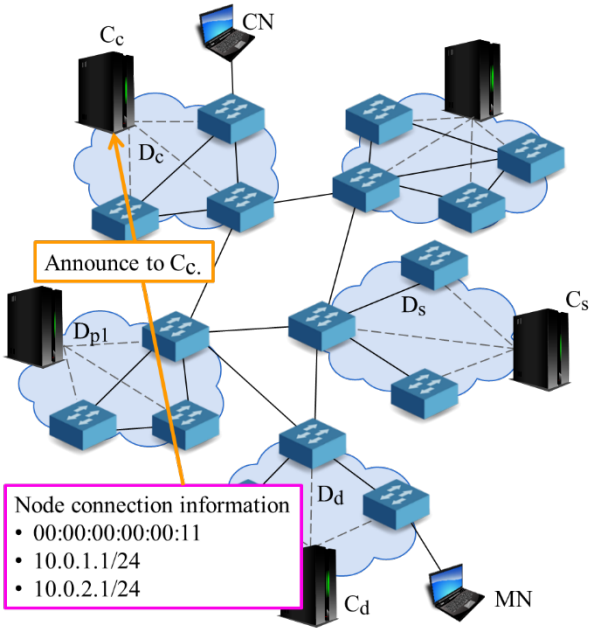


Figure 7 Announcement of node connection information.

Next, we explain the detailed design of IDRF following the example we used in the explanation of MISF in section 3.2.

1. C_d calculates the MN-CN inter-domain route.

   We regard each domain as a node and calculate inter-domain route. We take the number of inter-domain hops (N_{hop}) and the total number of flow entries that are needed to be installed in SDN switches (N_{flow}) into account to select a route that is the shortest and able to suppress the resource consumption of SDN controllers.

2. C_d announces the route information to the SDN controllers of the domains to go via the selected route.

   In the example, C_d announces the route to the domains shown in Figure 8.

3. Each SDN controller generates flow entries based on the received route information and installs them to their corresponding SDN switches.

   Especially, the SDN switch that the CN connects to check flow entries and rewrites the packets sent between the MN and the CN based on the node connection information. This keeps the IP address of the MN seen from the CN the same as before and it enables the communication to continue between the MN and the CN as before movement. Concretely, we set the flow entries as shown below.

   - Rewrite the destination IP address of the packets sent from the CN to the MN from former IP address to current IP address.
   - Rewrite the source IP address of the packets sent from the MN to the CN from current IP address to former IP address.

We would explain the detail of Step 1 mentioned above. Concretely, we choose the route with minimum N_{flow} in the routes with minimum N_{hop} in the following steps.

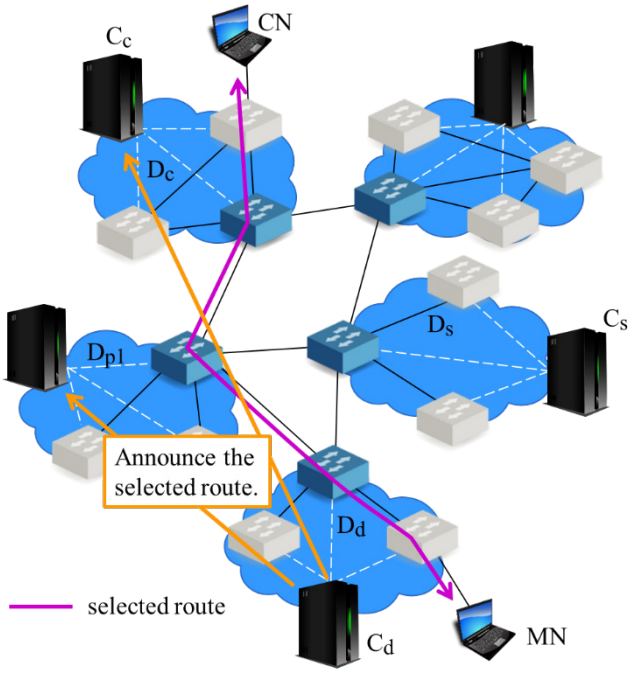


Figure 8 Announcement of inter-domain route.

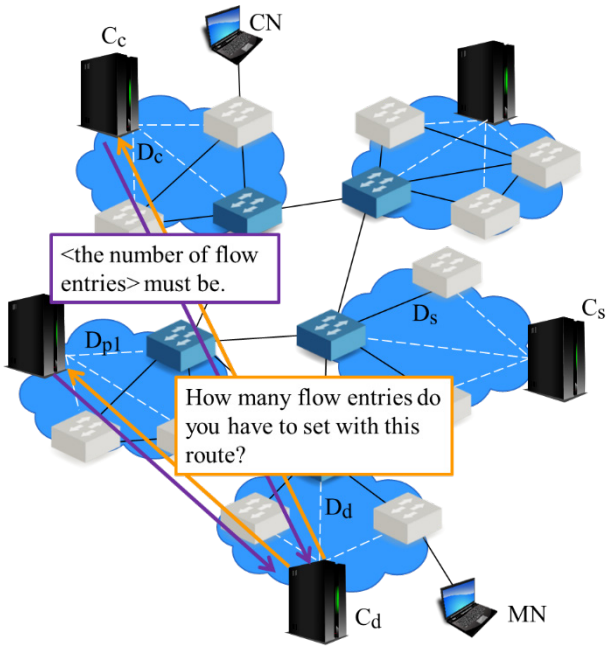


Figure 9 Inquiring N_{flow}.

1. C_d checks for routes with minimum N_{hop} based on inter-domain topology information.

   We use breadth first search algorithm for this search.

2. If there is only one route with minimum N_{hop}, we choose this route as the shortest route. If there are two or more routes with minimum N_{hop}, we then compare N_{flow} of each route.C_d asks SDN controllers of each domain via listed route the number of flow entries they need to install for each route like in Figure 9. The SDN controllers reply the number of flow entries to C_d and C_d calculates N_{flow} of each route.

3. C_d chooses the route with minimum N_{flow} as the finally decided route.

4. Performance Evaluation

4.1. Overview

We carried out simulation experiments to validate an effectiveness of our proposed scheme. Especially, we confirmed the reduction on both communication load between SDN controllers by MISF and reduction effect on communication delay between a MN and a CN by IDRF.

We used OpenDaylight Lithium [11] as SDN controllers, Open vSwitch [12] as SDN switches, and OpenFlow1.3 [13] as the communication protocol between the SDN controllers and the SDN switches. We implemented the two functions we introduced in section 3 into SDN controllers.

We equipped two approaches for comparison with our proposed scheme.

- approach 1: announces move of MN to all domains

  We replaced MISF to a function to announce the movement of MN to all the SDN controllers [8].

- approach 2: transfers packets sent to MN from source domain

  We replaced IDRF to a function to set route via the domain where the MN belongs to before moving [4-7].

We constructed a virtual network as an experimental network with Mininet2.2.1 [14]. The bandwidth capacity of the links between SDN switches are set to 1Mbps and delay of each links is set to 10ms.

The experiments are conducted in the following scenario:

1. A MN moves from D_s to D_d 5 seconds after the MN and a CN started communication.
2. C_d searches for source domain and retrieve former IP address of the MN.
3. C_d generates node connection information and announces it to C_c.
4. C_d calculates new communication route between the MN and the CN.
5. C_d announces the new route and SDN controllers of the domains involved in the new route install flow entries.

In this paper, we carried out three types of experiments.

- Experiment 1: evaluation of MISF
- Experiment 2: evaluation of IDRF
- Experiment 3: comprehensive evaluation of the proposed scheme

4.2. Experiment 1: Evaluation of MISF

In Experiment 1, we focused on traffic loads of communication between SDN controllers. We arranged three networks: each network consists 5, 10, and 15 domains, respectively. The number of inter-domain hops between source domain and destination

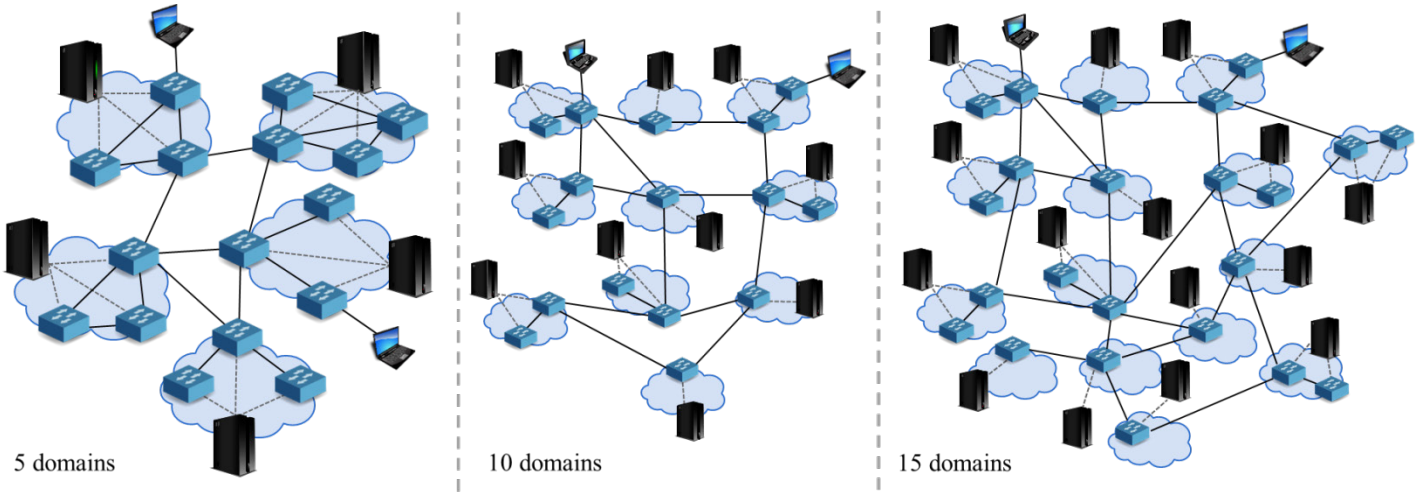


Figure 10 Experimental network topology in Experiment 1.

domain is 1. The domains increase concentrically like shown in Figure 10.

Figure 11 shows how the traffic load changed as the number of domains changed. In case of approach 1, traffic load increased according to the increase of number of domains. In contrast, traffic load in the case of proposed scheme did not raise even though the number of domains increased.

4.3. Experiment 2: Evaluation of IDRF

In Experiment 2, we focused on the delay of communication (Round Trip Time: RTT) between the MN and the CN after MN's move. We arranged a network with 10 domains and set the number of inter-domain hops between D_d and D_s to 1, 2, and 3, respectively, as shown in Figure 12.

Figure 13 shows the result of Experiment 2. In the case of approach 2, RTT increased according to the increase of number of inter-domain hops between D_d and D_s. In contrast, the proposed approach succeeded to keep RTT less than that of approach 2 in any case.

4.4. Experiment 3: Comprehensive evaluation of the proposed scheme

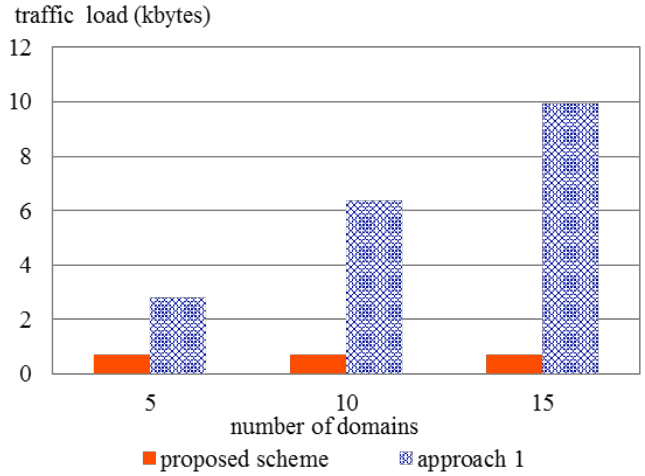


Figure 11 Experimental result 1: Traffic load of communication between SDN controllers.

In Experiment 3, we compared traffic load between SDN controllers and the RTT between the MN and the CN of the proposed scheme, approach 1, and approach 2. We set the parameters as below:

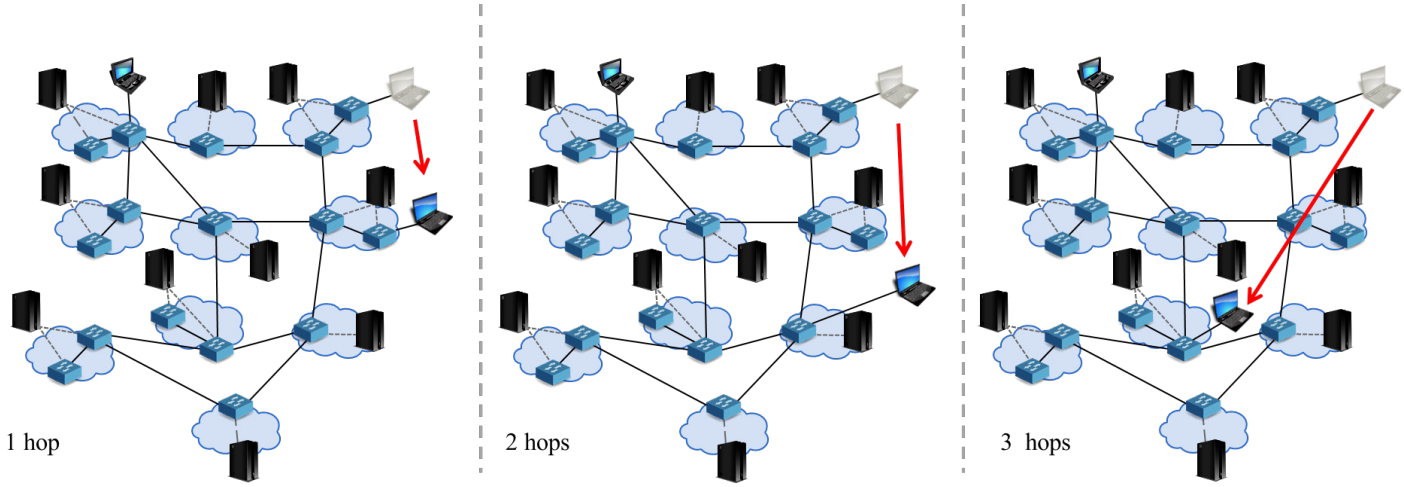


Figure 12 Experimental network in Experiment 2.

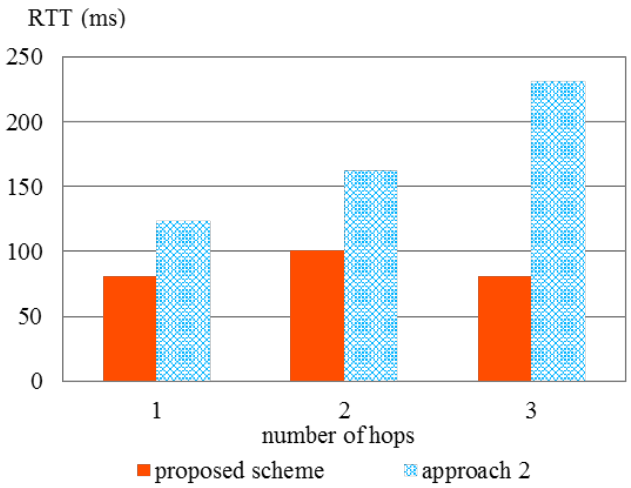


Figure 13 Experimental result 2: RTT between the MN and the CN.

- The number of inter-domain hops between the MN and the CN: 1, 2, 3
- The number of domains consisting an experimental network: 8, 9, 10

Figure 14 shows the topology of network when the number of domains are 8, 9, and 10, respectively. The MN moves as indicated by the arrows in the figure.

We show the traffic load between the SDN controllers for each number of hops between the MN and the CN in Figure 15. Traffic loads increase as D_s gets farther from D_d with the proposed scheme and the approach 2, whereas with the approach 1, traffic load was not affected by the number of hops but influenced by the number of domains.

Figure 16 shows the RTT of the communication between the MN and the CN for each number of domains. With the proposed scheme and the approach 1, RTT differs by the number of hops but is always smaller than that of the approach 2 in any case. With the approach 2, RTT increased as the number of hops between D_s and D_d got bigger. However, the results of the proposed scheme and the approach 1 showed no relevance to the number of hops between D_s and D_d.

4.5. Summary of evaluations

We discuss the results of Experiment 1. In case of the proposed scheme, the closer the source domain was to destination domain, the smaller the traffic load was because our scheme queries from neighbors.

However, the traffic load of the approach 1 increased as the number of domains increased. This is because the approach 1 announces to all the domains in the network regardless of the position of the source domain.

Next, we discuss the results of Experiment 2. In case of the proposed scheme, the number of inter-domain hops does not become larger than the number that can be taken by the approach 2. Therefore, the communication delay in proposed scheme gets smaller than approach 2. This is due to the fact that the longest route the proposed scheme chooses is the same route as the one chosen by approach 2.

Finally, we analyze the results of Experiment 3 according to the following viewpoints: traffic load between SDN controllers and RTT of communication between the MN and the CN after MN's movement.

First we talk about traffic load between SDN controllers. The proposed scheme and approach 1 exceeded approach 2 in traffic load between the SDN controllers. However, as the number of inter-domain hops between source domain and destination domain got bigger, traffic load of the proposed scheme and approach 2 tended to get bigger. For examples, we understand that when D_s and D_d were 3 hops away, traffic load of proposed scheme was pretty similar to that of approach 1 as we can see from Figure 15 (c). As D_s was 3 hops away from D_d, C_d ended up searching all the domains. When D_s and D_d was 2 hops away, traffic load of approach 2 was the biggest of all approaches. Approach 2 used MISF, the same function as the one proposed scheme uses. Thus, the traffic load of this function was the same as that of the proposed scheme. The difference came from the traffic load of IDRF. The difference in the communication route

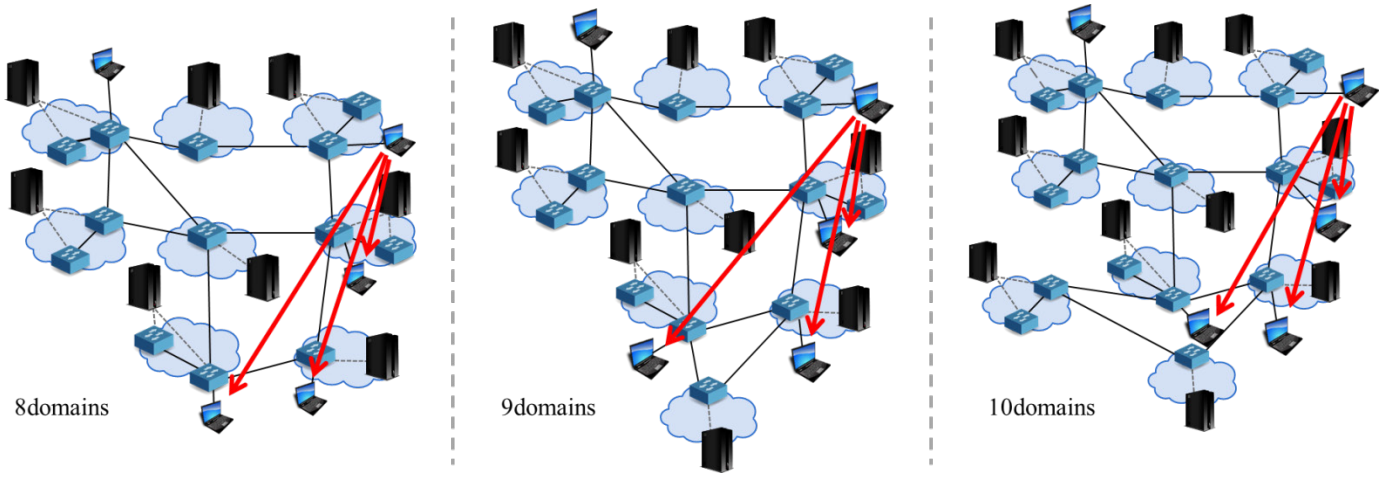


Figure 14 Experimental network topology in Experiment 3.

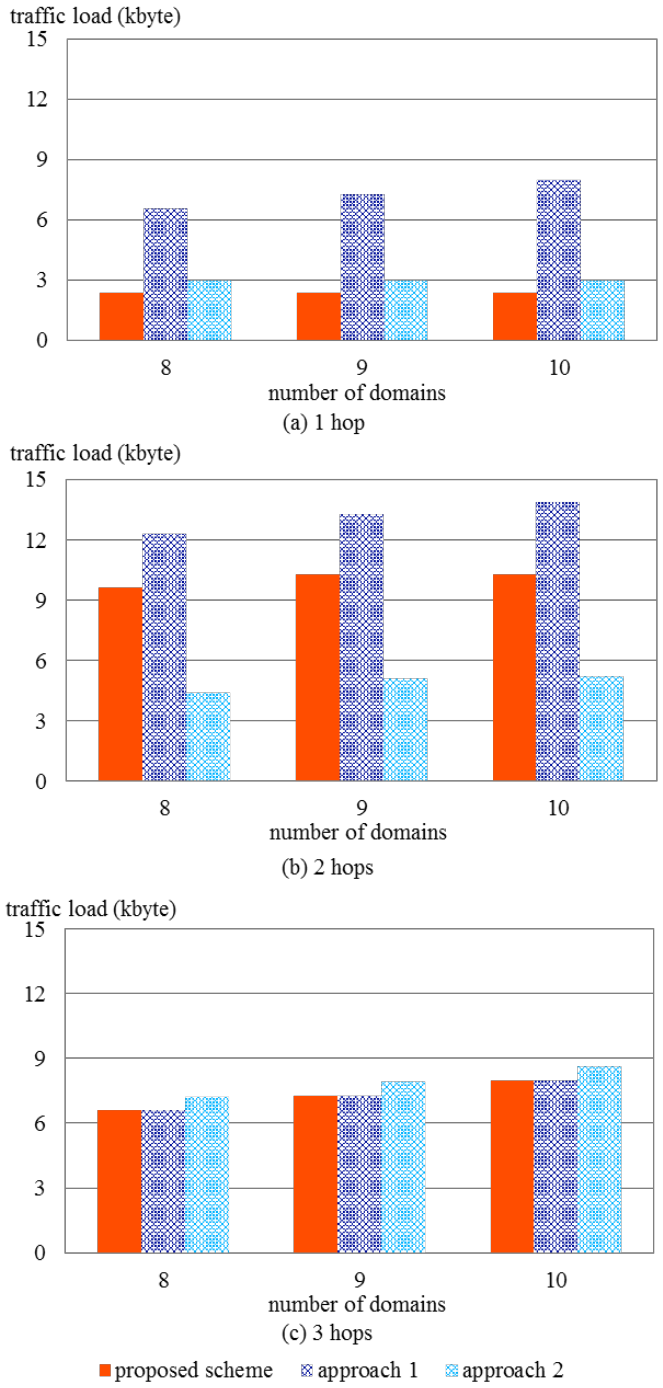


Figure 15 Experimental result 3-1: RTT between the MN and the CN.

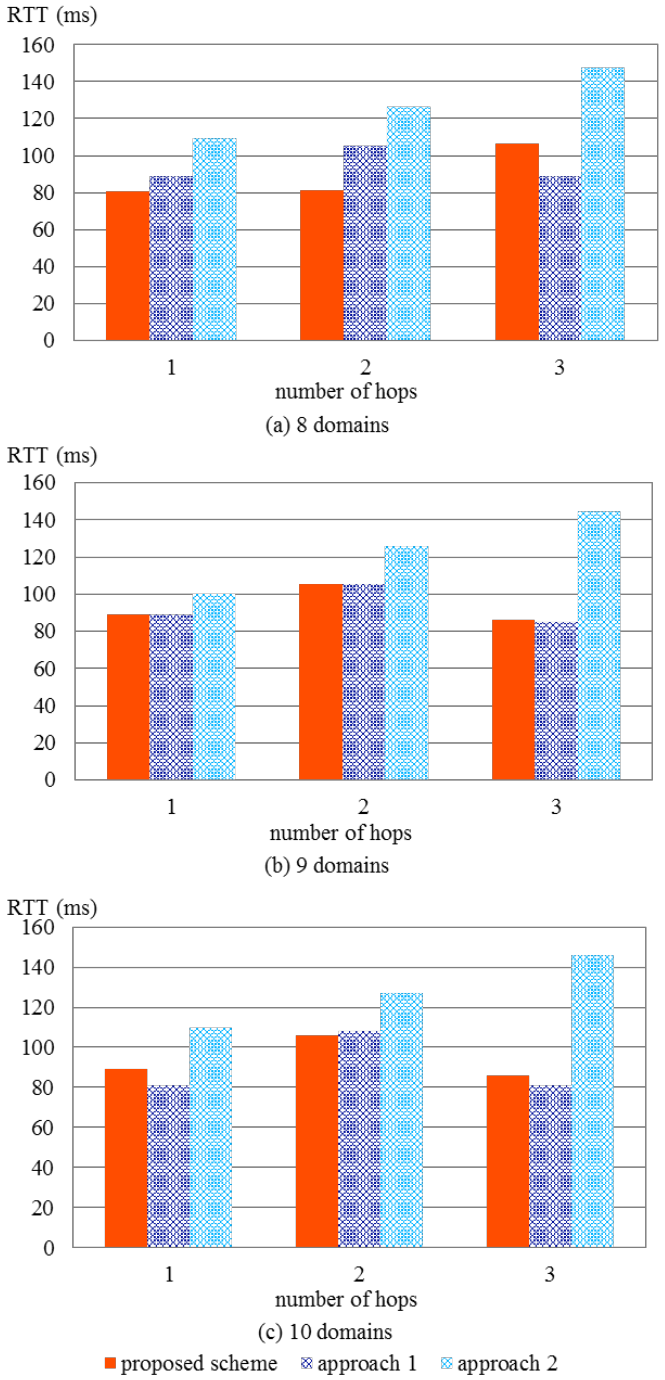


Figure 16 Experimental result 3-1: The traffic load between SDN controllers.

set after MN's movement caused the difference in amount of communication made between SDN controllers and this affected the overall traffic load. Figure 15 (b) shows that when D_s and D_d were 2 hops away, the traffic load of the proposed scheme and approach 1 exceeded the traffic load of approach 2 greatly. In this case, there were several routes with minimum inter-domain hops. Therefore, IDRF queried for the number of flow entries that was needed to be installed for each route. This querying caused the increase of the traffic load. This fact shows that by asking the number of needed flow entries, the traffic load between the SDN controllers increases.

Next, RTT of communication between the MN and the CN after MN's movement. With approach 2, RTT increased as the number of hops between D_s and D_d increased. In approach 2, packets sent to the MN were first sent to D_s and then got transferred to the MN. Thus, the communication route gets longer as the MN moves farther. In contrast, the proposed scheme and approach 1 selected the shortest route regardless of the position of D_s. Therefore, RTT did not get too big. Besides, the transferring route was the longest route that could be chosen in the proposed scheme and approach 2. Thus, these approaches could keep RTT the same as that of approach 2 or smaller.

From the results mentioned above, the approach 1 and the approach 2 can only hold back the RTT of communication between the MN and the CN after MN's move, or traffic load between SDN controllers. However, the proposed scheme is able to deal with both.

4.6. Discussion

These quantitative experimental results mentioned above reveal the contribution of our proposed scheme in practical use. This achievement enables the suppression of the consumption of limited network resources while enabling users of MNs to move around keeping seamless and fast communication even if they moved across domains. For example, consider the situation where many users work with laptop PC or tablet PC to download data from cloud servers or to upload data. The users can smoothly download/upload large data from/to cloud servers while they are moving, like during transit time. Concerning interactive-type applications such as remote login terminal or VDI environment, the users can use the services continuously when the connecting network domain changes.

Generally speaking, to avoid the termination of services or processes when MNs move across network domains, application-level handover function is usually implemented in the applications. When IP address of a MN changes according to the MN's movement, the handover function terminates the transport connection on the previous IP address, and then establishes a new transport connection on the new IP address. By this function, the data communication sessions seem to continue transparently and application logic can handle the sessions seamlessly. However, not all the applications have this feature implemented. This is because the mechanism is relatively complicated and it is necessary to deal with the server side as well. Our approach enables this functionality in network level. It means that all the Internet applications can utilize their network-dependent functions seamlessly in case that MNs move across the network domain.

Furthermore, when a user is at a place that the service areas of different access points overlap, the MN shifts its connecting access point back and forth depending on the link condition even when the user is not moving physically. Proposed scheme can manage this kind of handover and keep communication quality. Meanwhile, a standardization activity of a fast initial setup of Wi-Fi link is progressing as IEEE802.11ai [15]. This function would enable a Wi-Fi client to establish a secure link setup within 100ms. If this standard will widely be used in the above environment, switching of access point may occur very rapidly frequently in near future. There are no technologies currently available to deal with such situations, and we can expect that our proposed scheme will work effectively in such situations.

5. Conclusion

This paper aimed at realizing seamless connection and route optimization in SDN based IP mobility management on inter-domain handovers. It is difficult to realize efficient information exchange between SDN controllers and route optimization together. To solve this problem, we proposed the SDN based IP mobility management scheme considering inter-domain handovers. As simulation results show, we achieved to optimize route while keeping management information exchange traffic between SDN controllers low.

As the future works, we need to improve IDRF. By extending the algorithm to handle weight of paths, we can take more kinds of information into account, such as bandwidth of inter-domain links and network usage.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] M. Hata, S. Izumi, T. Abe, and T. Suganuma, "A proposal of SDN based mobility management in multiple domain networks," in 18th Asia-Pacific Network Oper. Manage. Symp., 1–4, 2016. https://doi.org/10.1109/APNOMS.2016.7737263

[2] D. Giusto, A. Iera, G. Morabito, and L. Atzori, Eds., The Internet of Things. Springer-Verlag GmbH, 2010. https://doi.org/10.1007/978-1-4419-1674-7

[3] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of things (IoT): A vision, architectural elements, and future directions," Future Gener. Comput. Syst., Vol. 29, No. 6, 1645–1660, 2013. https://doi.org/10.1016/j.future.2013.01.010

[4] C. E. Perkins, "RFC 5944 - IP mobility support for IPv4, revised," https://tools.ietf.org/html/rfc5944, 2010.

[5] C. E. Perkins, D. B. Johnson, and J. Arkko, "RFC 6275 - Mobility support for IPv6," https://tools.ietf.org/html/rfc6275, 2011.

[6] S. Gundavelli, K. Leung, V. Devarapalli, K. Chowdhury, and B. Patil, "RFC 5213 - Proxy mobile IPv6," https://tools.ietf.org/html/rfc5213, 2008.

[7] Y. Wang and J. Bi, "A solution for IP mobility support in software defined networks," in 23rd Int. Conf. Comput. Commun. Networks, 1–8, 2014. https://doi.org/10.1109/ICCCN.2014.6911783

[8] Y.Wang, J. Bi, and K. Zhang, "Design and implementation of a software-defined mobility architecture for IP networks," Mob. Network Appl., Vol. 20, No. 1, 40–52, 2015. https://doi.org/10.1007/s11036-015-0579-2

[9] "Software-Defined Networking (SDN) definition," Open Networking Foundation, https://www.opennetworking.org/sdn-resources/sdndefinition, (Accessed 2017).

[10] G. M. Keeni, K. Nagami, K. Koide, and S. Gundavelli, "RFC 4295 - Mobile IPv6 management information base," https://tools.ietf.org/html/rfc4295, 2006.

[11] "The OpenDaylight platform," Linux Foundation, https://www.opendaylight.org/, (Accessed 2017).

[12] "Open vSwitch," Linux Foundation, http://openvswitch.org/, (Accessed 2017).

[13] N. McKeown, T. Anderson, H. Balakrishnan, G. Parulkar, L. Peterson, J. Rexford, S. Shenker, and J. Turner, "OpenFlow: enabling innovation in campus networks," ACM SIGCOMM Comput. Commun. Rev., Vol. 38, No. 2, 69–74, 2008. https://doi.org/10.1145/1355734.1355746

[14] "Mininet: An instant virtual network on your laptop (or other pc) - mininet," Mininet Team, http://mininet.org/, (Accessed 2017).

[15] "IEEE P802.11 - TASK GROUP AI - MEETING UPDATE," http://www.ieee802.org/11/Reports/tgai_update.htm, (Accessed 2017).

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 932-936 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Computer Tomography from Micro-Electronics to Assembled Products

Keith Bryant[1,*], Ragnar Vaga[2]

[1]SMT Solutions. Johannesburg, South Africa

[2]Yxlon International, Hamburg, Germany

ARTICLE INFO

Article history:
Received: 20 April, 2017
Accepted: 06 June, 2017
Online: 30 June, 2017

Keywords:
Computer Tomography
Micro-electronics
X-ray technology

ABSTRACT

Traditional CT in our industry has been limited to Business card sized samples, due to the Cone Beam x-ray systems used by Electronics manufacturing companies. Inclined or Partial CT provides a slightly different solution showing layers or slices in 2D very well, but due to the partial nature of the scans does not produce very accurate 3D reconstructions. This seminar will look at more sophisticated x-ray systems, including dual tube units, which can image at sub-micron level and have the ability to build an accurate and detailed 3D image of a tablet or smart phone without any stitching or joining of images. With high quality reconstruction software, these images can easily be manipulated to allow key features or failure sites to be easily seen. These systems are being used in Failure Analysis but also in NPI and in the design and development process as CAD data can be overlaid and metrology is also possible with some systems.

1. Introduction

Traditional Electronics CT

Within electronics manufacturing and failure analysis inspection 2D inspection is the primary method of x-ray inspection and traditionally systems were only offering this. CT or 3D then became an option and quality varied dramatically dependant on many factors including: quality of image, stability of x-ray tube, reconstruction software, computing power etc. Time to make a 3D image was also long and to get detailed images could take many hours. This made it something which had very little use in the real world of manufacturing.

However due to the design of the systems and the technology used the biggest issue with CT was that the sample size was roughly the dimensions of a Business Card. This was due to the simple fact that to get good magnification the sample has to be close to the x-ray tube and to build the CT image the sample has to be rotated within the beam of the x-ray tube.

Cone beam technology is standard in the machines used for electronics inspection.

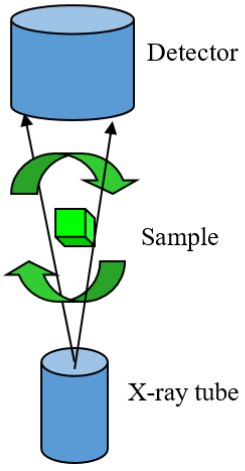


Figure 1 Conventional CT x-ray layout

The tube is at the bottom of the system, the sample held by some form of manipulator and a collector at the top of the system. Or the other way around in some systems, however this can cause issues with the sample colliding with the tube. So, in this type of systems CT is normally a failure analysis technique, unless you are working at the component level.

*Corresponding Author: Keith Bryant, SMT Solutions. Johannesburg, South Africa | Email: keith.dage@gmail.com

2. Inclined or partial CT

This is a quite recent technology, the main benefit is that it works by focussing on an area of interest on an assembly or product, the only size limit is the inspection area of the system. The area of interest is set within the x-ray cone beam and the sample or the detector is rotated.

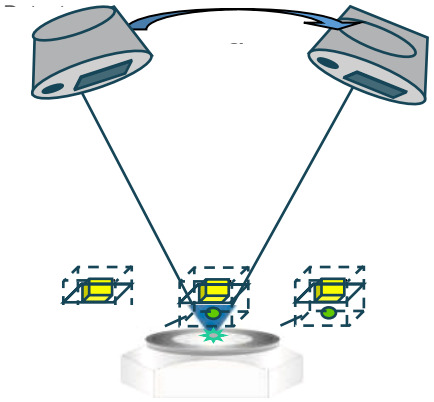


Figure 2 ICT manipulation

A larger number of images are taken, for instance one every degree, these are then processed to make a number of slices. This allows the operator to drill down through the number of slices to check interfaces and other key areas within the images. It is also used for separating components on the top and bottom sides of the board where complex components are 'mirrored' causing issues for traditional 2D x-ray systems. This is due to overlapping images of the components even at an angled view which can cause confusion or make voiding calculations impossible.

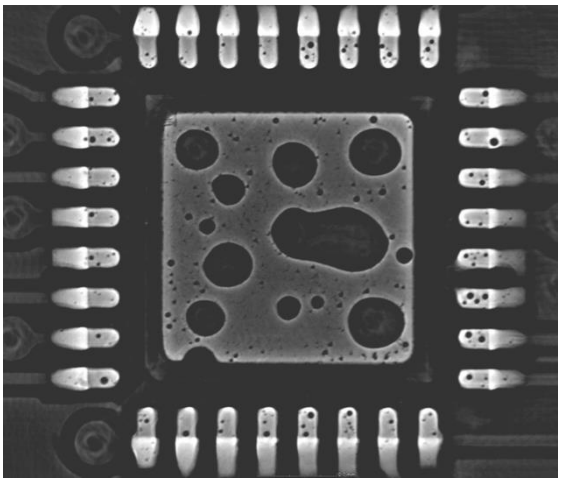


Figure 3 ICT slice showing QFN interface

However, as the captured images are taken at an angle and not by rotating the sample 360 degrees within the beam some data is missing. This means that when the 3D reconstruction is done it is not accurate in some areas as the data is incomplete. To overcome this some systems guestimate what should be there, which makes the image look more real but may not be accurate.

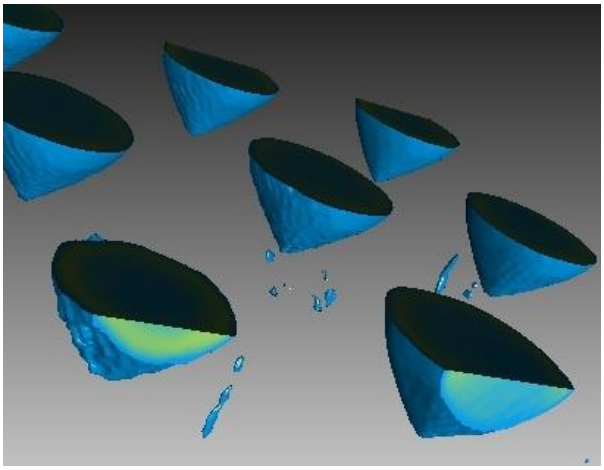


Figure 4 BGA balls showing lack of data

The figure above shows a section through some BGA balls, where the cross section is taken the 3D reconstructed image is accurate but the pointed area of the balls is inaccurate and is a result of lack of gathered data in this area.

3. Cross Over Systems

Background

The lines between NDT and Electronics x-ray inspection have become blurred as smaller features need to be seen in NDT and assembled product CT is being requested by more and more electronic manufacturing companies.

This has led to more flexible platforms being developed which can cope with sub-micron demands and large field of view applications. Certain key features are required to give a system the ability to perform well in both arenas.

Figure 5 Cross Over System

A very stable base is needed for the key components (tube, manipulator and detector), normally Granite is used for this. Very accurate motion systems, with encoder modules made by Heidenheimer or other top quality manufacturers. Rotation tables being very high precision with high end air bearings, giving accurate and repeatable results.

Figure 6 Air bearing rotation table

The higher-powered x-ray tubes also tend to be water cooled with diamond targets to give stable performance for demanding applications and over longer scanning times.

To have the ability to work at the Nano and sub-micron level and also to image larger demanding applications it is normal for these systems to have 2 tube types. This can add to the time needed, however some systems can change tubes at the touch of a button. Even keeping the same focus region when the tubes are changed, this is a great advantage

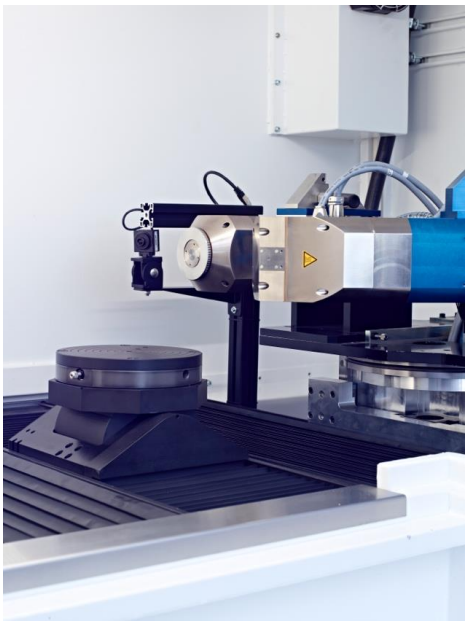


Figure 7 shows two x-ray tubes inside the system

These systems also tend to be larger than the normal off line ones seem in our industry. Partly due to the configuration which is horizontal rather than vertical, the larger and heavier sample sizes and the larger flat panel detectors.

4. New Technologies

Helical scanning

In order to increase the field of view but maintain high magnification helical scanning is used, as well as very accurate manipulation very clever software is needed to control the technology. The tube and the detector rise together allowing the field of view to move up the sample, producing very detailed large image.

The other alternative is to use stitching to join several smaller images together and produce one large image but the accuracy is poor by comparison.

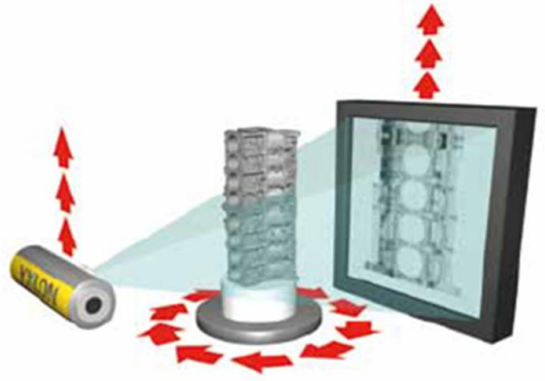


Figure 8 shows helical scanning layout

This allows very detailed images of larger objects to be made as the smart phone below illustrates

Figure 9 Smart phone imaged in a single shot

5. Virtual Rotation Access

This is a software driven technology which allows the operator to select a field of view not in the centre of the rotation table.

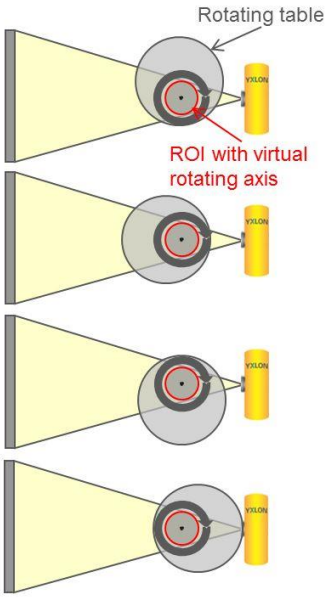


Figure 10 shows virtual rotation in action

It's ideal for hard to hold samples or to make several 3D images without having to reposition the sample multiple times to make the optimum image

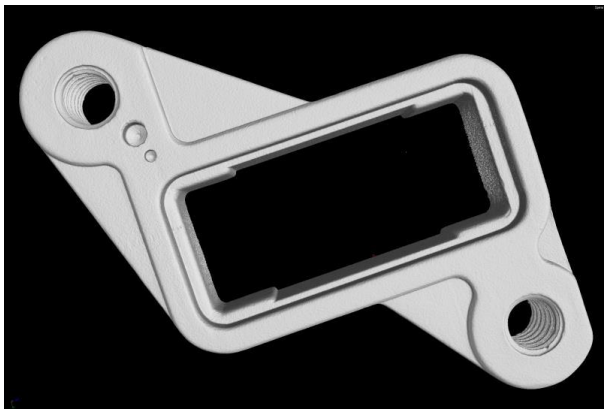


Figure 11a full 3D image of bracket

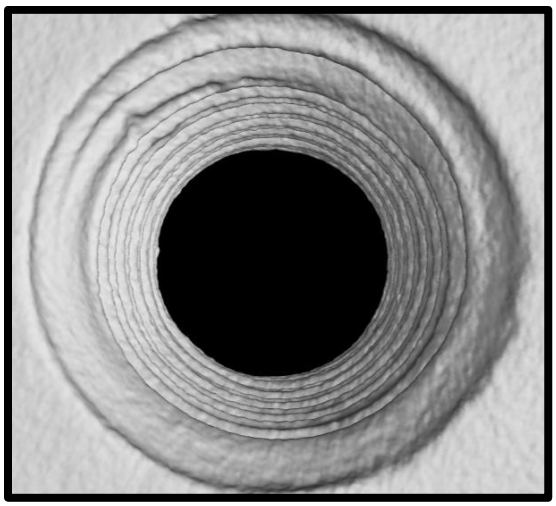


Figure 11b 3D image of mounting hole made by virtual rotation

6. Anti-collision systems

When working with large samples or with those which are offset or held in at an angle the potential for collision and damage increases. The more sophisticated systems use a camera and clever software to check the position of the sample during rotation. This allows the operator to set the sample close to the tube giving high magnification of the image without the potential for damage.

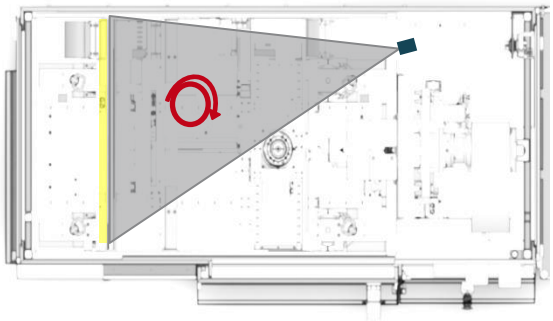


Figure 12 shows camera action for collision prevention

7. Remote Monitoring and Push Messages

This software innovation allows the operator to manage the system remotely or even manage multiple systems.

Continuous communication of inspection progress and health data gives the operator complete control. It displays the system Health Monitor, real time information and alerts. The software works on all Windows tablets and PCs; it allows remote monitoring sending push messages calling the user for direct system interaction e.g. decision expected or scan finished

Figure 13 shows system remote monitor

8. Metrology

Traditional x-ray systems have struggled with measurement accuracy and repeatability, basically they were microscope based and not measurement systems.

Cross Over systems have much more accurate components and active vibration damping, making them metrology based.

The better machines have calibration values related to the PTB certified measurement sphere gauge in a controlled environment at 20°C +/- 0.5°. Some companies offer guaranteed PTB values for a delivered system which can be ordered as optional feature.

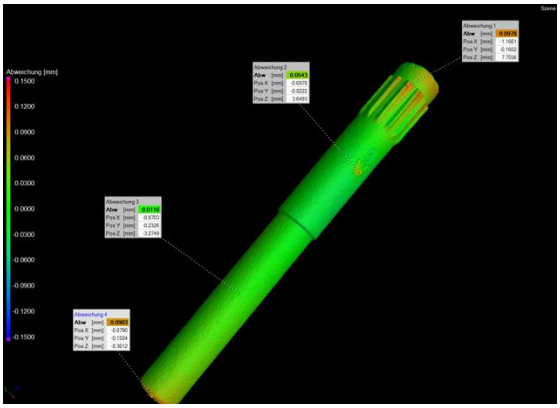


Figure 14 actual to nominal comparison measurement of component

9. CT Images from High End Systems

Below are some images demonstrating the capability of Cross Over systems, from sub-micron to full assemblies. They show features which are very challenging such as dendrites a few microns in diameter and hard to see even with high end 2D

systems. Through 17 µm diameter copper wires to batteries and relays up to smart phones

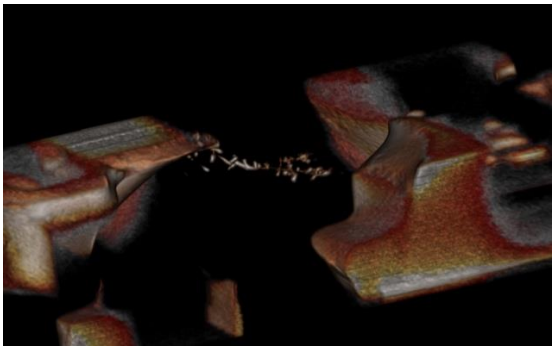


Figure 15 dendrites < 5µm

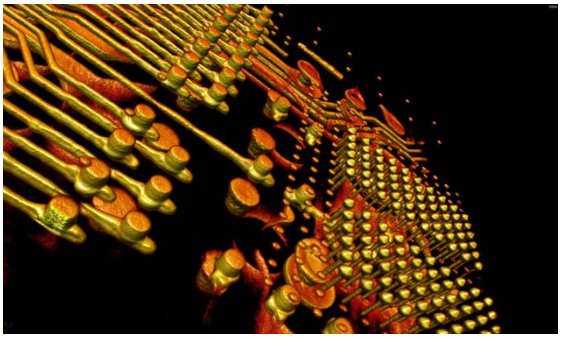


Figure 16 copper pillars

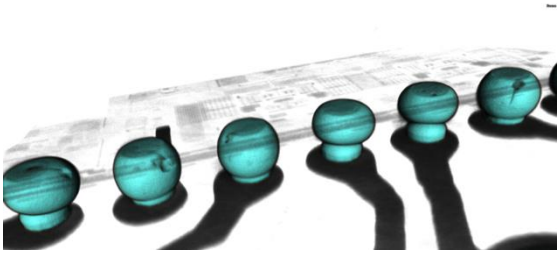


Figure 17 30µm balls

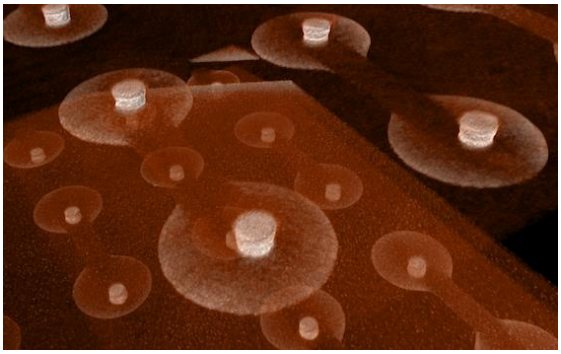


Figure 18 12µm copper pillars

Figure 19 SMART phone, complete image

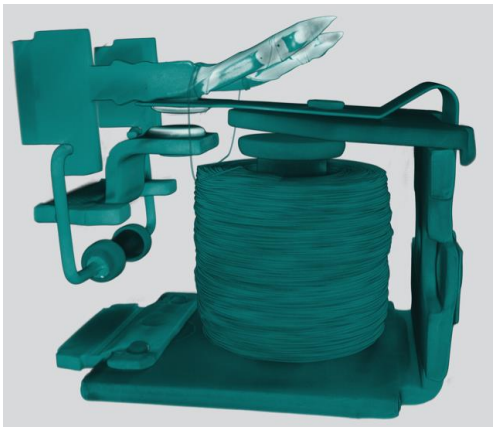


Figure 20 Relay assembly

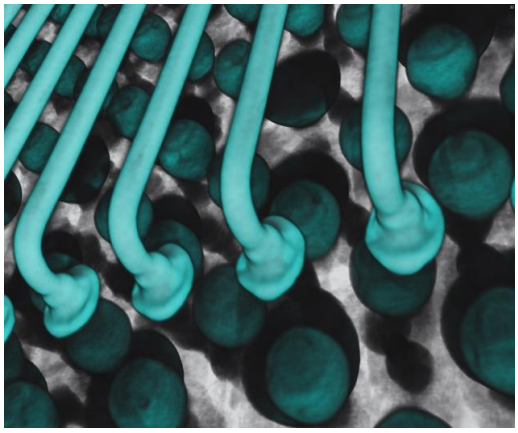


Figure 21 15µm copper wires

Acknowledgements

Rene Sommer and Thorsten Rother of Yxlon International, for many of the CT images. Dr David Bernard, for his support on this paper.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 937-945 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Mathematical Modeling and Digital Control of A Hybrid Switching Buck Converter

Muhammad Umar Abbasi*

Power Systems Division, SUPARCO LAHORE, 54000, Pakistan

ARTICLE INFO

Article history:
Received: 01May, 2017
Accepted: 11 June, 2017
Online: 30 June, 2017

Keywords:
Hybrid switching Buck Converter
Digital Control
Small Signal Model
SIMPLIS Software
8 Bit STM Controller

ABSTRACT

The aim of this paper is to describe mathematical modeling and digital control of a hybrid switching buck converter. This converter belongs to a class of so called hybrid switching converters and contains a resonant capacitor, resonant inductor and a diode in addition to original buck converter components. The dc gain of this converter is shown to be independent of resonant branch parameters. Moreover the dc conversion ratio is derived for both ideal case and including main inductor dc resistance. Small signal model of the converter is derived and is shown to be similar to conventional buck converter. Simulation results in SIMPLIS Software as well as experimental results of digital control using an 8 bit STM microcontroller are presented. The potential advantages and applications of this converter are discussed.

1. Introduction

A new class of switching power converters known as hybrid switching converters has been proposed quite recently by S. Cuk [1], [2]. These converters embody the switching characteristics of two predominant methods of power conversion; square wave PWM switching converters and resonant switching converters. In square wave PWM switching converters, the reactive elements are charged and discharged linearly. Due to this, the semiconductor devices have to be hard switched resulting in significant switching losses. In order to avoid the switching loses, quasi resonant and resonant switching converters were developed. In resonant converters, the reactive elements are charged and discharged in sinusoidal manner resulting in ZVS or ZCS of semiconductor devices depending upon mode of operation. However, the resonant converters suffer from difficult design and control issues. In order to combine the best of the two worlds, a hybrid class of switching converters is proposed by [1], [2]. These converters are essentially derivatives of PWM switching converters and contain additional components as compared to their PWM method counterparts. These additional components, referred to as resonant components, provide capabilities otherwise impossible to achieve with either of the two classes of power converters [3]. The behavior of these resonant components is a hybrid of two methods: they either charge linearly and discharge in resonant manner or vice versa. These characteristics provide improvements over conventional PWM switching converters as will be shown in this paper.

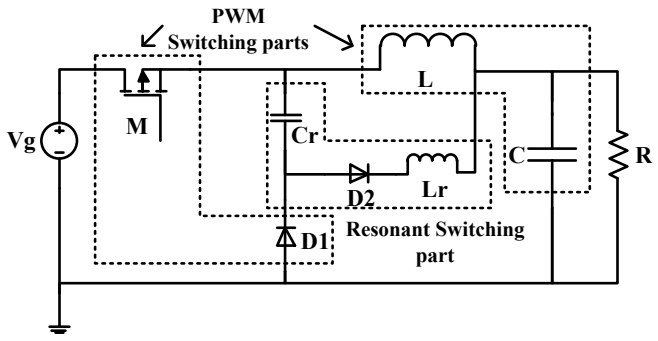


Figure 1 Hybrid Switching Buck Converter

Figure 1 shows the hybrid switching buck converter (HSBC) with PWM switching parts and resonant switching parts highlighted. It can be seen that the converter is essentially a buck converter with additional resonant reactive components i.e. resonant capacitor Cr and inductor Lr along with a diode. The resonant capacitor Cr charges in resonant manner along with the resonant inductor Lr whereas it discharges linearly. The aim of

*Corresponding Author: Muhammad Umar Abbasi, Power Systems Division, SUPARCO Lahore, +92333572393, Email: umareepower@gmail.com

this paper is to mathematically model the dc gain characteristics as well as small signal transfer function of this converter. It will be shown that this converter, though having a different dc gain ratio than a conventional buck converter, can provide similar characteristics as that of buck converter along with advantages such as magnetic size reduction of the main power inductor. Moreover, it is shown that the digital control of this converter can be achieved in a similar fashion as that of a conventional buck converter.

The paper is presented in the following manner: section 2 provides steady state operation and dc gain ratio of the hybrid switching Buck converter. Section 3 provides small signal derivation of HSBC and equivalent circuit is derived. Section 4 provides digital control implementation details around an 8 bit STM microcontroller. Section 5 gives simulation results performed in SIMPLIS software as well as practical example results. In section 6, possible applications of this converter are discussed.

2. Steady State Operation and DC Gain Characteristics

The converter is assumed to be operating in continuous conduction mode. All semiconductor devices are assumed to be ideal. Also it is assumed that the resonant components (Cr and Lr) are much smaller than the main filter components (C and L). For simplicity of analysis, we assume that the on period of the main switch is equal to half of resonant period where resonant period is defined as:

$$Tr = 2 * pi * \sqrt{Lr * Cr} \tag{1}$$

The converter operates in two modes depending on the state of MOSFET switch M. The resulting circuit waveforms along with the equations describing the reactive elements are presented next.

2.1 *MOSFET Switch ON*

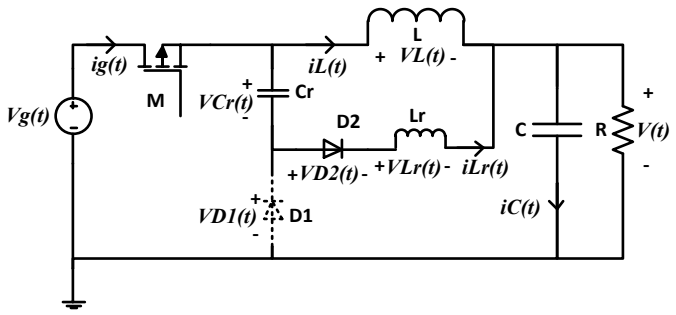


Figure 2 MOSFET ON, D2 ON, D1 OFF

When main MOSFET switch is ON, the diode D1 is reversed biased and diode D2 turns ON. The current flows from input to output charging both the main power inductor and resonant branch comprising of Lr and Cr. The capacitor Cr is charged resonantly. The inductor Lr is volt-sec balanced during this time interval only as the current through Lr flows only when D2 is conducting. This implies a different volt-sec balance approach than the main inductor L which is volt-sec balanced over complete switching period. It is assumed here that the resonant capacitor Cr is initially charged to a voltage level of Vcr1. The important waveforms and associated equations for this mode can be presented as:

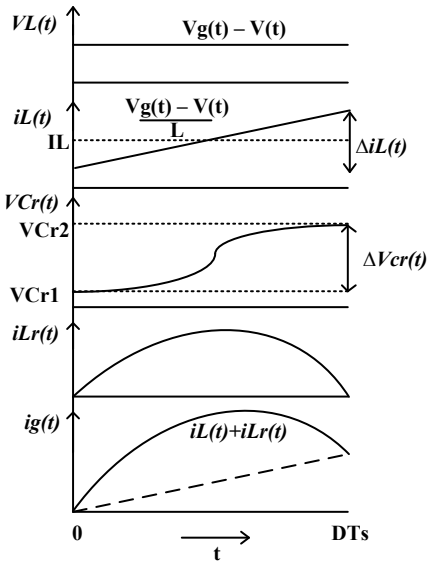


Figure 3 Waveforms in Mode I

$$L\frac{di_L(t)}{dt} = V_g(t) - V(t) \tag{2}$$

$$C\frac{dV(t)}{dt} = i_L(t) + i_{Lr}(t) - \frac{V(t)}{R} \tag{3}$$

$$Lr\frac{di_{Lr}(t)}{dt} = V_g(t) - V_{Cr}(t) - V(t) \tag{4}$$

$$Cr\frac{dV_{Cr}(t)}{dt} = i_{Lr}(t) \tag{5}$$

$$i_g(t) = i_L(t) + i_{Lr}(t) \tag{6}$$

As can be seen, the main inductor Volt-sec equation remains the same as the conventional buck converter. Although the resonant branch is charged sinusoidally, the equations can be written in similar manner as the PWM branch. The V_{Cr} and i_{Lr} can be described as:

$$V_{Cr}(t) = V_{Cr1} \times \cos(\mathrm{w_r} \times (\mathrm{t} - \mathrm{t_0})) \tag{7}$$

$$i_{Lr}(t) = \frac{V_{Cr2}}{Zn}\sin(w_r \times (t - t_0)) \tag{8}$$

Where Zn is the natural impedance and *wr* is the resonant frequency of the resonant branch given as:

$$Zn = \sqrt{\frac{Lr}{Cr}}, \quad w_r = \frac{1}{\sqrt{Lr \times Cr}} \tag{9}$$

As can be seen, the sinusoidal behavior of the resonant branch has no significant effect on the equations of the PWM branch components.

2.2 *MOSFET Switch OFF*

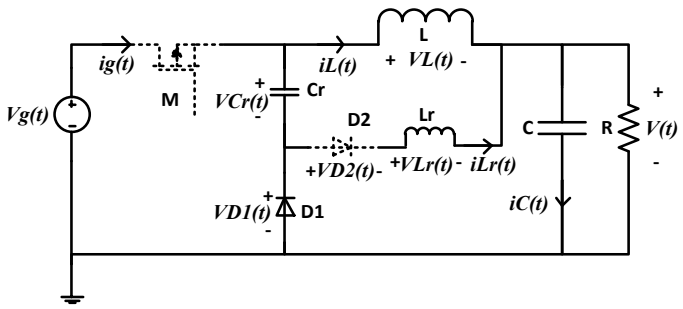


Figure 4 MOSEFT OFF, D2 OFF, D1 ON

When the MOSFET switch is OFF, the current iL(t) traverses through diode D1 which is the freewheeling diode, same as in conventional buck converter. The diode D2 turns off and the resonant capacitor Cr begins to discharge in the quasi linear fashion. The resonant inductor current is zero in this mode. The inductor current has a negative slope as shown in the figure. The important waveforms and associated equations for this mode are:

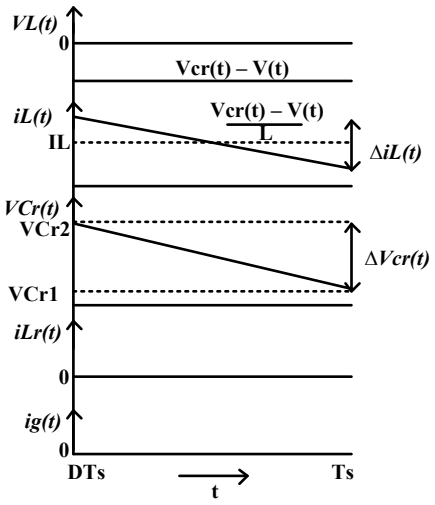


Figure 5 Waveforms in Mode II

$$L\frac{di_L(t)}{dt} = V_{Cr}(t) - V(t) \tag{10}$$

$$C\frac{dV(t)}{dt} = i_L(t) - \frac{V(t)}{R} \tag{11}$$

$$Lr\frac{di_{Lr}(t)}{dt} = 0 \tag{12}$$

$$Cr\frac{dV_{Cr}(t)}{dt} = -i_L(t) \tag{13}$$

$$i_g(t) = 0 \tag{14}$$

As can be seen, the resonant capacitor Cr is quasi linearly discharged from level Vcr2 to Vcr1. The average voltage across Cr will be shown to be equal to difference of input and output voltages. Also note that the inductor Volt-Sec balance equation for mode II is different from conventional buck converter and therein lays the advantage of present topology as will be shown later.

2.3 *DC Gain Characteristics*

The dc gain of hybrid switching buck converter can be derived by applying principle of volt-sec balance on main and resonant inductors as well as principle of charge balance on main and resonant capacitors [4].

For main filter inductor L:

$$L\frac{d\langle i_L(t)\rangle_{Ts}}{dt} = \langle V_L(t)\rangle_{Ts} = D\times(\mathrm{V_g} - \mathrm{V}) + \mathrm{D'}\times(V_{Cr} - V) = 0 \tag{15}$$

Where D is the main switch duty cycle and D'=1-D. Also we have assumed steady state so the time varying quantities are replaced by their dc values. Solving Eq. (2.14) yields:

$$D\times V_g = V - D'\times V_{Cr} \tag{16}$$

Applying Volt-sec balance on resonant inductor Lr yields:

$$Lr\frac{d\langle i_{Lr}(t)\rangle_{Ts}}{dt} = \langle V_{Lr}(t)\rangle_{Ts} = D\times(\mathrm{V_g} - V_{Cr} - \mathrm{V}) + \mathrm{D'}\times(0) = 0 \tag{17}$$

After simplification we get:

$$V_{Cr} = V_g - V \tag{18}$$

Thus the average voltage across resonant capacitor is equal to the difference of input and output voltages. After substituting Eq. (2.17) into Eq. (2.15), we get the desired dc gain as:

$$\frac{V}{V_g} = \frac{1}{2-D} \tag{19}$$

The dc gain is very different from conventional buck converter and (19) suggests that there is a lower limit on the output voltage. For duty ration D=0, we get V = Vg/2. However, in the next section it is shown that this is not the case as main inductor dc resistance reduces the dc gain ratio and thus the converter can traverse almost all the range of output voltage.

Next the principle of capacitor charge balance is applied to main filter capacitor as well as resonant capacitor. For capacitor C, we get:

$$C\frac{d\langle V(t)\rangle_{Ts}}{dt} = \langle i_C(t)\rangle_{Ts} = D\times(\mathrm{I_L}+\mathrm{I_{Lr}}-\frac{V}{R}) + D'\times(\mathrm{I_L}-\frac{V}{R}) = 0 \quad (20)$$

After simplification we get:

$$IL - \frac{V}{R} = -D'\times I_{\mathrm{Lr}} \quad (21)$$

Applying Charge balance on resonant capacitor Cr, we get:

$$Cr\frac{d\langle V_{Cr}(t)\rangle_{Ts}}{dt} = \langle i_{Cr}(t)\rangle_{Ts} = D\times(I_{Lr}) + D'\times(-I_L) = 0 \quad (22)$$

This results into:

$$I_{Lr} = \frac{D'}{D}\times I_L \quad (23)$$

Finally we get relationship between average inductor current and load resistance as follows:

$$IL = \frac{V}{R\times(2-D)} \quad (24)$$

We can also calculate the average value of input current which is the same as main MOSFET average current as follows:

$$\langle i_g(t)\rangle_{Ts} = D\times(I_L + I_{Lr}) \quad (25)$$

After substituting Eq. (2.22) we get:

$$I_g = I_L \quad (26)$$

A summary of DC characteristics of conventional buck converter and hybrid switching buck converter is presented in a tabular form.

Table 1 DC Characteristics Comparison

DC Characteristic	Conventional Buck Converter	Hybrid Switching Buck Converter
Voltage Gain	D	$\frac{1}{2-D}$
Average Inductor Current	$\frac{V}{R}$	$\frac{V}{R\times(2-D)}$
Average Input Current	$\frac{D\times V}{R}$	$\frac{V}{R\times(2-D)}$
Average Switch Voltage	V_g	V
Main Inductor ripple Δi_L	$\frac{V_g DD'Ts}{L}$	$\frac{VDD'Ts}{L}$
VD1	V_g	V

From the above table, following advantages of Hybrid Switching Buck converter can be summarized:

- The inductor current ripple is reduced. This implies that for the same switching frequency, hybrid switching buck converter requires smaller main power inductor. This advantage is achieved without increasing switching frequency.
- Both MOSFET and freewheeling diode voltage ratings are reduced. Low voltage switches with better switching features can be used.

2.4 DC Gain with Inductor DC Resistance

Finally, we derive the DC gain including main inductor dc resistance. For reference, the circuit is produced here including inductor dc resistance:

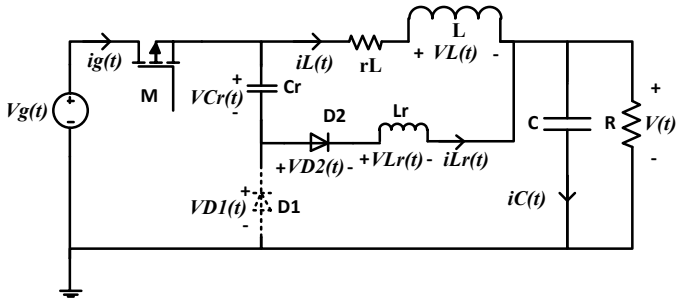


Figure 6 Hybrid Buck Converter including inductor resistance

Only the Volt-Sec balance equation of main inductor is changed and is given below:

$$\langle V_L(t)\rangle_{Ts} = D\times(\mathrm{V_g}-\mathrm{V}-\mathrm{I_L}\times\mathrm{r_L}) + \mathrm{D'}\times(V_{Cr}-V-\mathrm{I_L}\times\mathrm{r_L}) = 0 \quad (27)$$

After simplification we obtain:

$$D\times V_g - V - I_L\times r_L + D'V_{Cr} = 0 \quad (28)$$

Replacing the value of V_{Cr} from (18) and simplifying:

$$V_g - V\times(2-D) - I_L\times r_L = 0 \quad (29)$$

We replace value of IL from (24) and simplify to get final equation:

$$\frac{V}{V_g}=\frac{1}{(2-D)-\frac{r_L}{R\times(2-D)}} \tag{30}$$

This equation shows the effect of inductor dc resistance r_L on DC gain. The DC gain is reduced and thus the converter is able to traverse wider range of output voltage.

3. Small Signal Modeling of HSBC

In this section we derive small signal transfer function of hybrid switching buck converter. We can write the large signal equations

$$\begin{aligned}
&<V_g(t)>_{Ts}=V_g+\hat{v}_g(t),\ <i_g(t)>_{Ts}=I_g+\hat{i}_g(t),\\
&<V_{Cr}(t)>_{Ts}=V_{Cr}+\hat{v}_{cr}(t)\ <V(t)>_{Ts}=V+\hat{v}(t),\\
&<i_L(t)>_{Ts}=I+\hat{i}_L(t),\ <i_{Lr}(t)>_{Ts}=I_{Lr}+\hat{i}_{Lr}(t)\\
&d(t)=D+\hat{d}(t),\ d'(t)=D'-\hat{d}(t)
\end{aligned} \tag{32}$$

Where $\hat{x}(t)$ represents the small signal perturbation in steady state value of variable $x(t)$ .After replacing in (31), we get following equation:

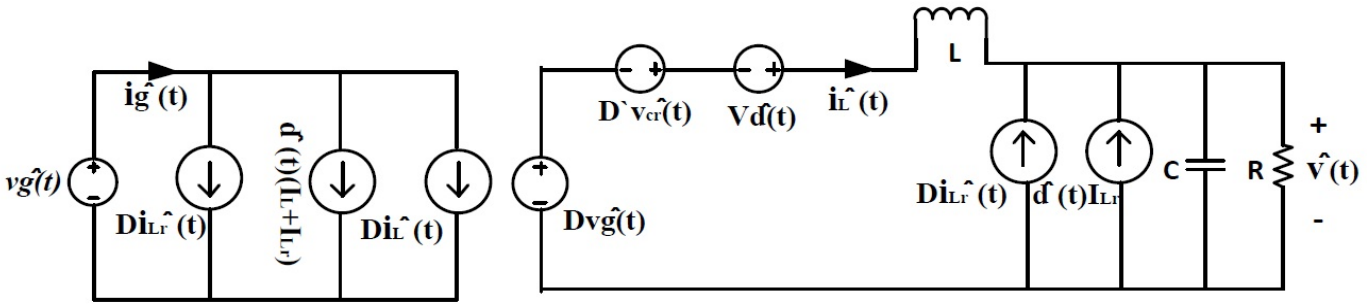


Figure 7 Small Signal Equivalent Circuit for HSBC

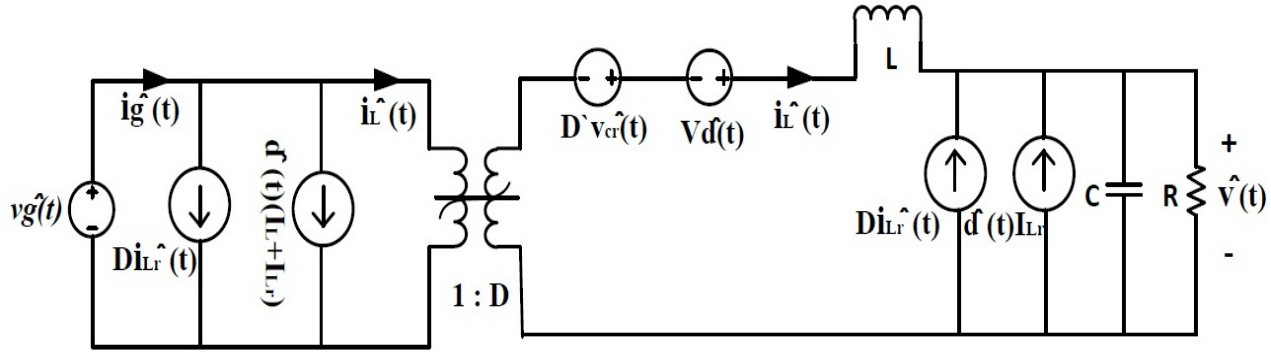


Figure 8 Equivalent Circuit with DC Transformer

for the main filter inductor and capacitor using the approach in [4]. The averaged equation for inductor L is given below:

$$\begin{aligned}
L\frac{\langle di_L(t)\rangle_{Ts}}{dt}&=d(t)[\langle V_g(t)\rangle_{Ts}-\langle V(t)\rangle_{Ts}]\\
&+d'(t)[\langle V_{Cr}(t)\rangle_{Ts}-\langle V(t)\rangle_{Ts}]
\end{aligned} \tag{31}$$

Where $<x(t)>_{Ts}$ denotes the average of x(t) over switching period Ts. We can perturb and linearize this non linear averaged equation by replacing time varying quantities with their steady state values perturbed in small quantities as shown next:

$$\begin{aligned}
L\frac{d[I+\hat{i}_L(t)]}{dt}&=[D+\hat{d}(t)]\times[V_g+\hat{v}_g(t)-V-\hat{v}(t)]\\
&+[D'-\hat{d}(t)]\times[V_{Cr}+\hat{v}_{cr}(t)-V-\hat{v}(t)]
\end{aligned} \tag{33}$$

After simplification, we ignore dc and higher order terms and arrives on following small signal equation for inductor L:

$$L\frac{d\hat{i}_L(t)}{dt}=D\hat{v}_g(t)+D'\hat{v}_{cr}(t)+\hat{d}(t)\,\mathrm{V}-\hat{v}(t) \tag{34}$$

Similarly, the non linear averaged equation can be written for the output capacitor C:

$$C\frac{d<V(t)>_{Ts}}{dt} = d(t)[<i_L(t)>_{Ts} + <i_{Lr}(t)>_{Ts} - \frac{<V(t)>_{Ts}}{R}] + d'(t)[<i_L(t)>_{Ts} - \frac{<V(t)>_{Ts}}{R}] \quad (35)$$

After perturbing and linearizing steps as shown for the inductor, we obtain following small signal equation for capacitor C:

$$C\frac{d\hat{v}(t)}{dt} = \hat{i}_L(t) + D\hat{i}_{Lr}(t) + \hat{d}(t)I_{Lr} - \frac{\hat{v}(t)}{R} \quad (36)$$

Applying the similar procedure for the input current i_g(t), we have:

$$\langle i_g(t)\rangle_{Ts} = d(t)[\langle i_L(t)\rangle_{Ts} + \langle i_{Lr}(t)\rangle_{Ts}] \quad (37)$$

After perturbation and simplification we obtain following equation:

$$\hat{i}_g(t) = \hat{d}(t)[I_L + I_{Lr}] + D[\hat{i}_L(t) + \hat{i}_{Lr}(t)] \quad (38)$$

The (33), (36) and (38) represent the small signal model of the hybrid switching buck converter. We can combine these equations to produce the equivalent small signal model of hybrid switching buck converter.

As can be observed, the small signal ac model of the hybrid switching buck converter is similar to the conventional buck converter with the inclusion of additional voltage and current sources due to the presence of resonant capacitor and inductor. However, this circuit can be solved in the manner similar to the conventional buck converter to obtain desired transfer function such as control to output transfer function [4], [5].

On the primary side of the DC transformer, two additional current sources are present. One current source is due to the effect of the change in resonant inductor current. This is because the resonant inductor current is drawn from the input port of the converter. Any change in this resonant current will be reflected in the input current ig(t). The other two sources combined as one represent the effect of change in control variable i.e. duty cycle d(t). The DC values of both the inductor current and resonant current are responsible for this effect. In case of a conventional buck converter, only DC value of inductor current is present here but in case of HSBC both the main power inductor and resonant inductor are charged by the input port and thus both values are present.

As can be observed on the secondary side, there are two voltage sources as compared to the one in case of a conventional buck converter. Moreover, the voltage source consisting of Vg in conventional buck converter is now replaced with one consisting of output voltage V. This can be attributed to the fact that the HSBC applies different volt-sec to the main power inductor as compared to the conventional buck converter. During the time when the MOSFET switch in OFF, the conventional buck converter applies minus V times the off time interval (-V*D'Ts), where HSBC applies Vcr – V time the off time interval ((Vcr-V)*D'Ts). This difference is reflected in the form of change in voltage source on the secondary side of the transformer.

An additional voltage source is due to the effect of change is resonant capacitor voltage. The resonant capacitor voltage is applied only during the off time interval and hence its small signal component is multiplied by the complement of steady state duty cycle.

One might imagine after looking at small signal equivalent circuits shown above that the HSBC is more difficult to control and that its small signal control to output transfer function will be more complex than a conventional buck converter. However, as will be shown in next sections, the small signal characteristics of HSBC closely mirror those of the conventional buck converter. The control of the HSBC can also be accomplished by the identical computational effort as that of conventional buck converter.

4. Digital Control Implementation

In order to demonstrate the viability of the converter and the potential of simple 8 bit microcontrollers to control this converter, a simple test setup was created in lab around an 8 bit STM microcontroller [5]. The simplified schematic diagram of the implementation is presented below:

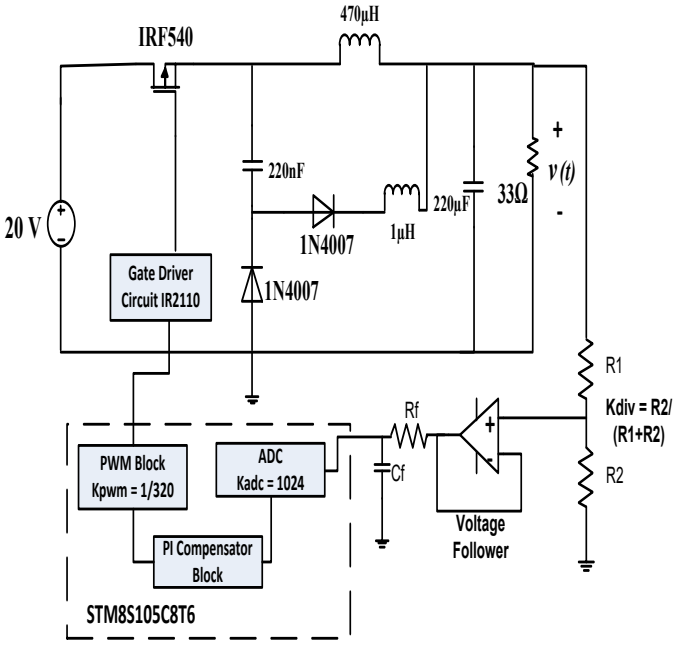


Figure 7 Digital Control Setup

The output of the converter is scaled down by the resistor divider and converted into digital domain by the ADC. The PI compensator block calculates the error and applies compensation to generate the duty cycle for the main switch. The built in PWM block of STM controller is used to generate the duty cycle controlled pulse signal for MOSFET.

The flowchart of the software used to control the converter is shown below:

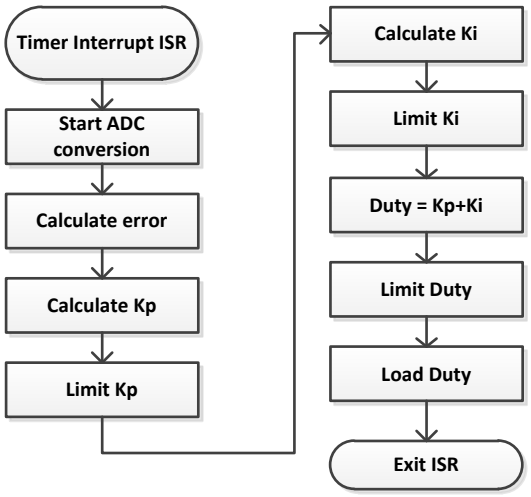


Figure 8 Software Flowchart

It should be noted here that the duty cycle calculated by the PI controller is limited to a certain minimum value to ensure that the duty cycle is always greater than half of resonant period as given by Eq. (2.1). This ensures that the resonant current cycle is completed without any interruption. The concept is illustrated in the next figure.

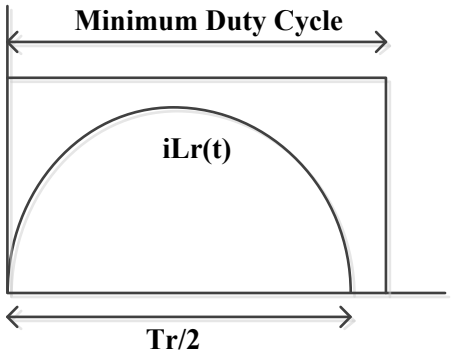


Figure 9 Minimum Duty Cycle

At first glance it would appear that this is negatively restricting duty cycle above a minimum threshold. However, in practical designs, the duty cycle generally oscillates around its designed value in small quantities and very rarely traverses the entire range of its possible values. So, if the converter is designed such that the half resonant period is some small fraction of the duty cycle, say five percent, then this restriction can be met without much problem. In practical setup used for this converter in the lab, the half of resonant period was about five percent of duty cycle and converter operated without problem.

5. Simulation and Practical Results

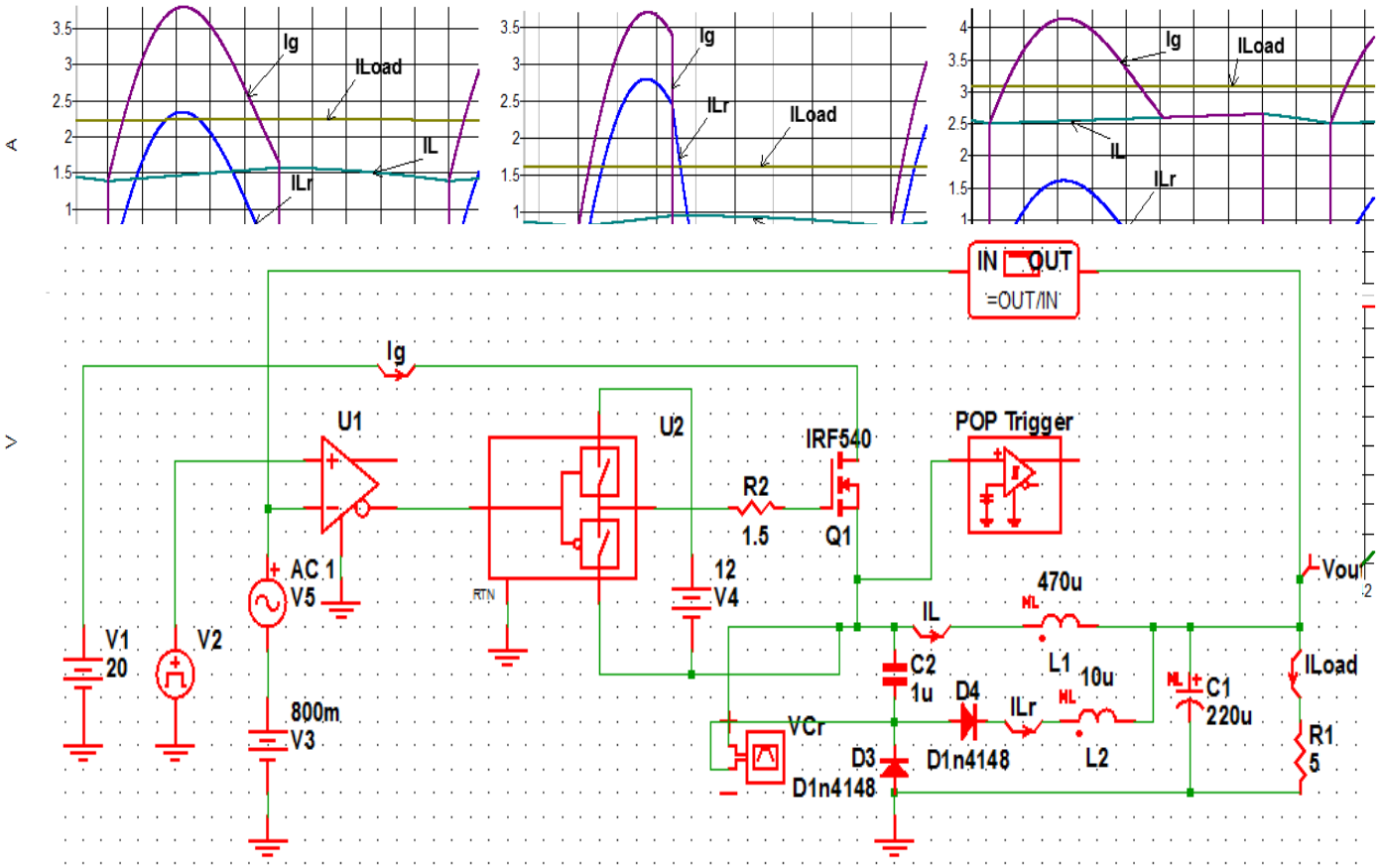


Figure 10 SIMPLIS Simulation of HSBC

The converter is simulated in SIMPLIS simulation software. The screenshot of the SIMPLIS simulation is shown in figure 12.This software provides the ability to simulate the circuit in its steady state form directly, without having to perform lengthy and time consuming transient analysis. This feature of the software is called Periodic Operating Point (POP) analysis.

Moreover, the ac analysis of this converter circuit can be performed in time domain without having to derive equivalent frequency domain models, thereby simulating a real life network analyzer in software [6].

The circuit is simulated for three different values of MOSFET duty cycle, 30%, 50% and 80%. The results of important circuit parameters for these three duty cycles are presented in figure 13.

As we can see from figure 13, the I_{Lr} becomes zero in case of duty cycle being greater than or equal to the half of resonant period. In case duty cycle is less than Tr/2, then the diode D4 will be hard switched and incurs switching losses. Therefore, the converter should be designed such that the duty cycle is limited to the values above the Tr/2.The behavior of the resonant capacitor voltage swing should also be considered here. The case in which duty cycle is less than half of resonant period, the voltage Vcr charges sinusoidally up to the point of switch transition and then discharges linearly. The behavior is consistent with the theory of hybrid switching conversion i.e. resonant charging and linear discharging. Also the case in which duty cycle is greater than half the resonant period, the voltage Vcr rises sinusoidally until the current ILr is flowing. When the current ILr tries to become negative, the diode D2 blocks the flow and the current can flow no further. This results in voltage Vcr becoming constant. This, of course, doesn't affects the converter behavior.

It can be seen that the output voltage can traverse greater range than predicted by the DC gain of (19). This is because of the fact that the simulation software includes the parasitic elements such as MOSFET on resistance and diode forward voltage drops. Such parasitic, like the dc resistance of main power inductor, work to reduce the DC gain predicted by (19), thereby allowing the converter to traverse greater output voltage range.

It was predicted in the previous section that the small signal characteristics of HSBC are very much identical to those of conventional buck converter. In this section, we present the small signal control to output transfer function characteristics simulated in SIMPLIS software. This software has the capability to simulate the circuit in time domain, thereby emulating the behavior of a network analyzer. The bode plots of the simulation are shown next:

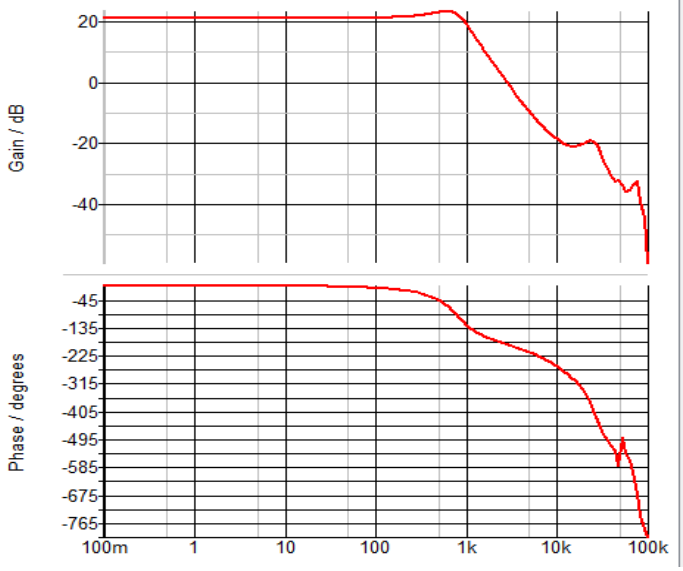


Figure 12 HSBC Bode plot

As we can clearly see, the converter exhibits almost same gain roll off as that of a conventional buck converter [4], [5]. The gain cross over frequency for the simulated circuit comes out to be 2.73 KHz. This shows that HSBC is a stable open loop converter which can be controlled easily using simple control structures.

In order to demonstrate the viability of the converter, a simple prototype was built on breadboard in lab. The setup was arranged according to the figure 9. The following table shows the parameters of this setup:

Table 2 Setup Parameters

Parameter	Value
Input Voltage Vg	20 V
Output Voltage Vout	10 V
Load Resistance	33 Ω
L	470µH
C	220µF
Lr	1 µH
Cr	220nF
MOSFET	IRF540N
Diodes	IN4007
Switching Frequency	50 KHz

The breadboard setup in the lab and the results are presented next:

The breadboard setup consists of a simple assembly of HSBC. The converter is essentially derived from a buck converter presented in [5]. The STM discovery board is used to control the converter using a PI controller. The results of the converter operation are presented next.

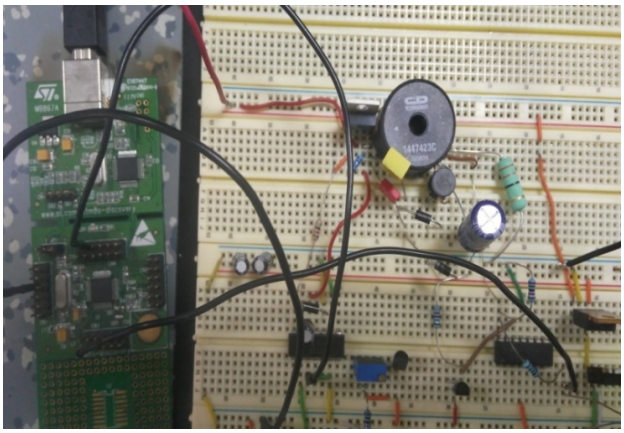


Figure 13 Breadboard Setup

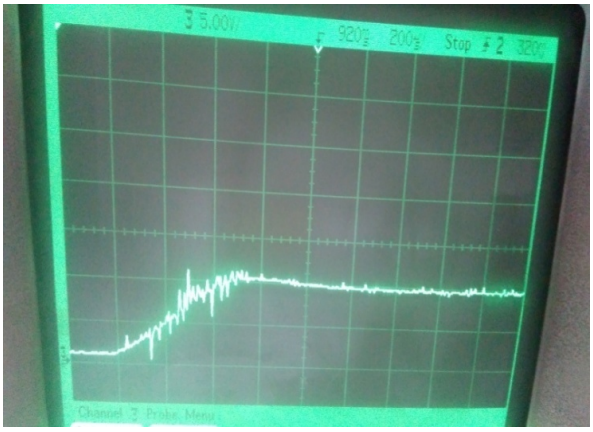


Figure 14 Voltage Startup Response

The output voltage startup response presents good picture with reasonable rise time for an 8 bit controller. The response time should be adequate for general power conversion applications not requiring stringent response constraints.

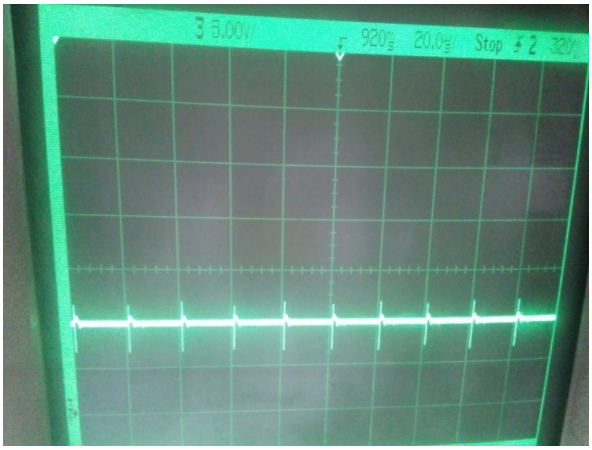


Figure 15 Steady State Output Voltage

As can be observed, over the period of various switching cycles, the converter exhibits no limit cycle oscillations. Therefore, this converter can be controlled using a simple PI

controller. The output voltage can be regulated within reasonable bounds depending on the applications.

6. Potential Applications

The hybrid switching buck converter presented in this paper belongs to a new class of converters and as such is unfamiliar to practicing engineers. Although at first glance it may appear that the part count for this converter is increased when compared to a conventional buck converter, however, the advantages presented by this converter can overcome such issues.

First, if the converter is designed carefully, the diode D2 in series with resonant inductor is soft switched thereby avoiding switching losses. Second, the most significant portion of the PCB is taken up by magnetic in a dc-dc converter. For example, in an MPPT solar battery charger dc-dc converter, the inductor core is the largest component in the power converter. The hybrid switching converter provides the reduction in the size of main power inductor and is therefore suitable for such applications.

Another area of potential application can be the converters that require reduction in size due to space constraints, such as integrated dc-dc converters for military and space applications. Such converters can benefit from this type of hybrid switching buck converter.

7. Conclusion

In this paper, a hybrid switching buck converter was presented. The concept of hybrid switching was introduced. The DC conversion ratio of this HSBC was derived and its dc characteristics were compared to conventional switching buck converter. It was shown that HSBC results in reduced magnetic size without increasing switching frequency of associated switching losses. The small signal analysis was performed on this HSBC and the result was shown to be familiar to the conventional buck converter. The simulation and practical results of digital control using 8 bit microcontroller were presented. The potential application areas of this HSBC were discussed.

CONFLICT OF INTEREST

The author declares no conflict of interest.

References

[1] S. Cuk, “Step-down converter having a resonant inductor, a resonant capacitor and a hybrid transformer,” U.S. Patent 7915874, Mar. 2011.

[2] S. Cuk, Z. Zhang, “Voltage step-up switching DC-to-DC converter field of the invention,” U.S. Patent 7778046, Aug. 2010.

[3] S. Cuk, “Bridgeless Power Factor Correction,” U.S. Patent 2010/0259240, Oct. 2010.

[4] Erickson, R.W., Maksimovic, D., “Fundamentals of Power Electronics”, 2nd Edition, Kluwer Academic Publishers, 2000.

[5] Abbasi, Muhammad Umar, “Digital Control of a Buck Converter using an 8 Bit STM Mircocontroller”, International Journal of Scientific & Technology Research Volume 6, Issue 04, April 2017, ISSBN 2277-8616

[6] SIMPLIS Technologies, www.simplistechnologies.com

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 946-950 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Effect of Ambient Temperature Variations on Particle Dimesions in Ultrasonic Nebulizers during Cold Vaporization

Ali Akpek*

Department of Bioengineering, Gebze Technical University, 44100, Turkey

ARTICLE INFO

Article history:
Received: 30 April, 2017
Accepted: 11 June, 2017
Online: 30 June, 2017

Keywords:
Ultrasonic nebulizer
Particle dimension
Cold vaporization

ABSTRACT

For human comfort and convenience, appropriate temperature values may not be sufficient enough for the environment that they live in. The moisture content of the air has an important place in this regard. 30% to 65% humidity is desireable for a relaxed and comfortable life. Humidity in the environment is fundamental for the human health, convenience and comfort. Humidity is also valid for therapeutics. The drugs that are used in medicine for diagnosis and therapy must be inhaled with cold or hot water vapor in case drug effectiveness is desired to be increased. For such achievments, cold vaporization method is being preferred to reach the alveoli of the patients, thus, the drug can be used in a more efficient manner without damaging the respiratory tract. In this project, effect of ambient temperaure in ultrasonic nebulizers is investigated. It is proved that as the temperature increases 5°C, the speed of cold vapor increases 10% which refers a decrease in particle diameter. The variations in particle dimension may severely effect the efficiency of therapeutics. Therefore, in this study, it is presented that ultrasonic nebulizers should be designed considering ambient temperature. A design rationale is proposed in this study.

1. Introduction

For the human comfort and convenience, only appropriate temperature values may not be sufficient for the environment that they live in. The moisture content of the air has an important place in this regard [2]. For a relaxed and comfortable life relative humidity is desirable in the range of %30 to %65 in the environment [3]. Humidity in the environment is fundamental for the human health, convenience and comfort [4]. The drugs that used in medicine for diagnosis and therapy must be inhaled with both cold and hot water vapor steam in the patient's body [5]. Cold vaporization method is being preferred to reach the alveoli of the patient, thus, the drug can be used in a more efficient manner without damaging the respiratory tract [6].

Ultrasonic nebulizers utilize water to carry out drugs to the alveols. Ultrasonic nebulizers dissociate water into their smallest particles by sound waves and thus compose cold vapour. Obtained cold vapour and drugs are mixed intoeach other and blown to patient through a mask or directly to the environment. Therefore, respiratory tract is moisturized and the desired drugs are carried out to the alveols of the patients efficiently [7].

Some environments may contain excessive moisture while some of them are over dried. This situation is pretty much common especially among the houses with a radiator during winters and houses with air conditioning in summers. Ideal moisture ratio should be between 30% and 65% for human beings. For humans especially for babies, dry air may cause enfections to spread faster, dehydrate pharyngonasal area, burning in the eyes and cracking in the lips. Dust formation becomes rapid and affects those with allergies. **[8]**. The spread of cold vapor provides accurate humidity in the environment and creates beneficial effects on health. Water droplets have ionizing effect for the air. Negative ions attack pollutant particles in the air, reducing it to a minimum, thereby reducing cigarette smoke, pollens and dust **[9]**. Medical studies have proved that negative ions have a positive effect on people's physical and psychological health with both their stress-relief functions and their direct effects on respiratory devices which makes them much more effective **[10]**.

*Corresponding Author: Ali Akpek, Department of Bioengineering, Gebze Technical University, 44100, Turkey | Email: aliakpek@mit.edu

Simply a nebulizer generates vapour. There are various ways to generate vapour **[11]**. The most convenient method for this is by heating the liquid. On the other hand an ultrasonic nebulizer system evaporates liquids with ultrasonic sound waves. The major problem here is not evaporating water but obtaining cold vapour.

An ultrasonic nebulizer uses a vibrating piezoelectric crystal that vibrates with 1.7 MHz or higher frequency in order to generate vapour **[12]**. The crystal transducer is made of materials like quartz-barium titanate and converts electrical signals into sound waves. The sound beam that focuses on the liquid which is located just above the transducer creates waves. When the frequency and amplitude are strong enough, the wave peak reaches the surface of the fluid as a sprinkler of water particles. Working principle of an ultrasonic nebulizer is illustrated in **Fig. 1**.

Ultrasonic nebulizers are most frequently used in intensive care units. In these units, ultrasonic nebulizers are mainly used for the humidification of the rooms, and also for the evaporation of liquid medicines to assist them to effectively inhaled by respiratory organs. In addition, it is also used to humidify the air for patients who are connected to artificial respiration and oxygen support systems **[13]**.

By the help of this study, effect of ambient temperature on ultrasonic nebulizers is investigated. In order to achieve this, ambient temperature is constantly recorded and particle dimensions of cold vapour are measured. As a result it is understood that, ambient temperature has a serious impact on particle dimension thus affecting the effectiveness of ultrasonic nebulizers therapeutic capability.

In addition, this study aims to further development of national fabrication of ultrasonic nebulizers in Turkey and creates difference in the area of biomedical device fabrication that may not be evaluated as strong **[14-22].**

2. Materials and Method

Ultrasonic nebulizers are based on the princible that disassociating water into its particles with high frequency sound waves that is generated by a vibrating piezoelectric crystal. This piezoelectric crystal that is located under the liquid, generate sound waves with a frequency of 1.7 MHz or more, and disassociating water into its particles.

It is possible to change the particle diameter of of the evaporation by changing the frequency of the piezoelectric crystal. As the frequency of the piezoelectric crystal increases, the particle diameter decreases. As the frequency decreases, the particle diameter increases. As the particle diameter increases, it will be unable for medications or humid air to reach the extreme edges of the alveoli. Same phenomenon is observed when the ambient temperature changes. Particle size decreases as the ambient temperature increases, while particle sizes increase as the ambient temperature decreases.

The quantity of the liquid is important since the piezoelectric crystal is operated inside the liquid. It is important because crystal may deteriorate due to the heat energy that occurs during the formation of the sound waves. For this reason, the quantity of the water is constantly measured and a system is designed that prevents the degradation of the crystal by cutting the energy when there is not enough water left within the device.

Block diagram of the designed ultrasonic nebülizer is illustrated in **Fig. 2**. A picture of the ultrasonic crystal is presented in **Fig. 3**. In addition, a level sensor for water measurement is shown in **Fig 5**.

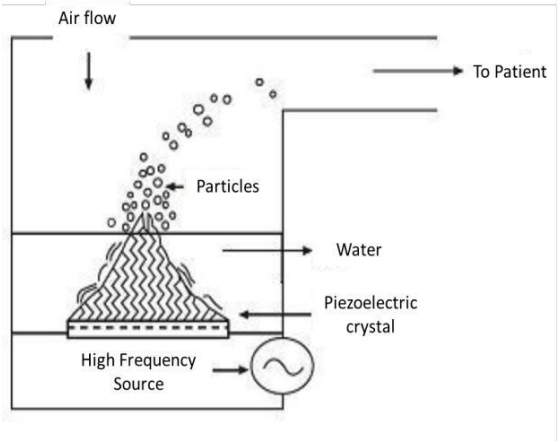


Fig. 1: Working principle of an ultrasonic nebülizer

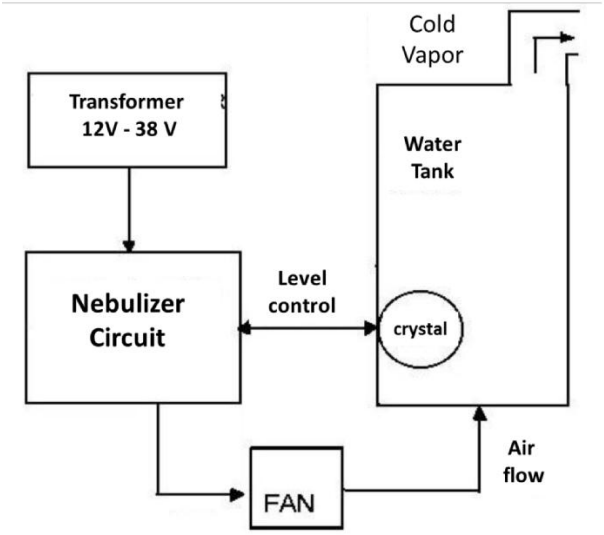


Fig. 2: Block diagram of the designed ultrasonic nebülizer

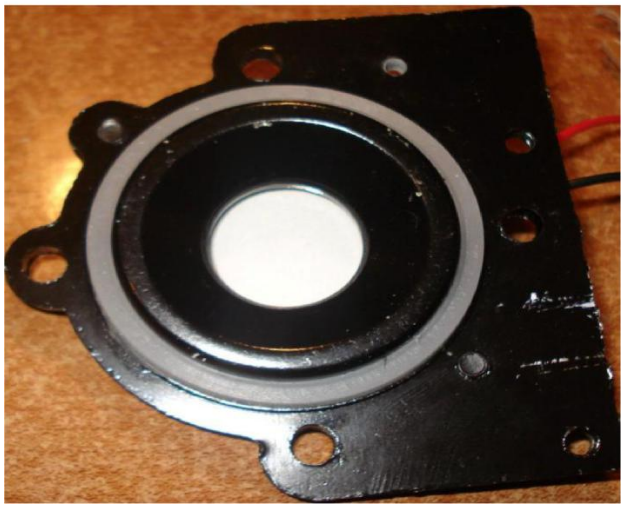


Fig. 3: Ultrasonic crystal

The design and drawing stages of the nebulizer circuit were first completed using the Proteus ISIS program. The design is simulated with the Proteus ISIS program to check the operability of the system. After being confident that the simulation process of the design is completed and the design may work flawlessly, the process of printing circuit board with Proteus ARES program is started. This process is presented in **Fig. 6**.

The design of the boxing process has initiated after the control of the nebulizer circuit and date/clock circuit are implemented and operability of the design is controled. Therefore the circuit and system could be used in maximum confort.

A crystal with a special casting design is placed in the lower part of the water tank. BU 406 power transistor that controls the ultrasonic crystal generates a lot of heat due to the high voltage and fast triggering characteristics. Since this heat may damage the transistor, the special casting that surrounds the crystal is used as a cooler. The water within the water tank cools down the casting and casting cools the crystal and the transistor. This is the reason why water level is critically important for this research. The cold vapor that is generated by the crystal is delivered to the patient by the help of a channel just located on the top of the ultrasonic nebulizer.

Finally, chassis is placed and LCD screen that displays ambient temperature, date/time and control parameters is located on it. Therefore the prototype is finalized. LCD screen and the final design of the ultrasonic nebulizer are shown in **Figs. 7, 8 and 9**.

The experiments are done by changing the ambient temperatures by several radiators that surrounds the ultrasonic nebulizer. At the beginnig the temperature is recorded as 25ºC ±0.3ºC. The process is initiated and the period until the visually detection of cold vapor out from the funnel of the nebulizer is recorded. This period provides the speed of the particles of cold vapor thus giving us the estimations for particle dimensions.

As second step the temperature is increased 5ºC to 30ºC ± 0.3ºC, the process is initiated and the period of visually detection of cold vapour is recorded. As third step the temperature is increased to 35ºC and the entire process is recorded again. As fourth and fifth step, the temperature is reduced to 20ºC ± 0.3ºC and 15ºC ± 0.3ºC, respectively. In those steps, period of visually detection of cold vapor is recorded. All steps are repeated three times to obtain statistically accurate results.

3. Results and Discussion

During the test phase, the ambient temperature is measured in different methods and double checked to ensure that the sensor and system operate properly and consistently. In order to ensure the system has taken correct measurements or not, the measurement stability was controlled with a calibrated thermometer at regular intervals.

Fig.4 presents the relationship between ambient temperature change and period for visually detection of cold vapor. As it is shown as the temperature increases the vapor speed decreases which refers an increase in particle dimensions.

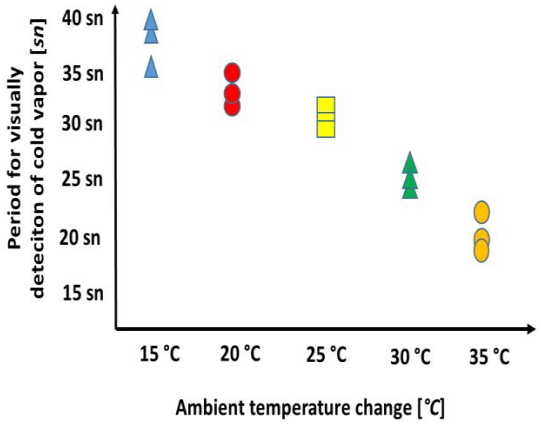


Fig. 4: The relationship between ambient temperature change and period for visually detection of cold vapor.

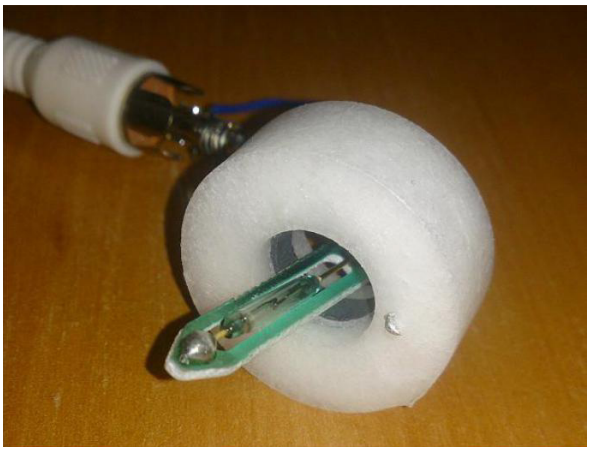


Fig. 5: Level sensor for water measurement

Fig. 6: Designing the circuit board of an ultrasonic nebulizer

Fig 7: Temperature measurement results a) 25ºC b)-11.3ºC c) +62.2 ºC

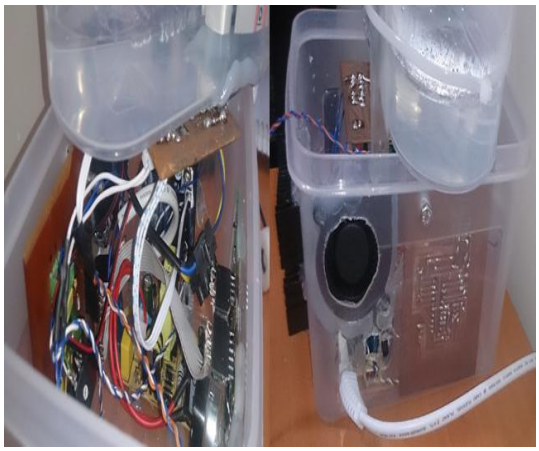


Fig. 8: Inside and back views of the designed ultrasonic nebulizer

Fig. 9: Final prototype of ultrasonic nebulizer b) During work

4. Conclusion

The purpose of this project is to prove that it is quite simple to fabricate home made low cost specially designed ultrasonic nebulizers that is mainly imported to Turkey with high prices.

In addition, it is proved that ambient temperature has serious impact on particle dimensions. It is presented that as the ambient temperature increases five celcius, the speed of cold vapour increases around 10% which refers a decrease in particle dimensions. As the temperature decreases five celcius, the speed of cold vapor also decreases 10% which refers an increase in particle dimensions.

As future researches, Particle Image Velocimetry (PIV) might be used to understand the particle dimensions in detail. A PIV may assist us to truly analysis the particle speed and dimensions of the cold vapor as the temperature increases or decreases. Also, in future the sensitivity of the research might be increased, therefore change in particle dimensions and particle speed per one celcius temperature variation might be detected. This may help us to optimize the period for therapeutics.

The most important innovation in this study is to merge a temperature sensor just outside the cover of the ultrasonic nebulizer. This assists users to estimate the particle dimensions and predict the optimum time for therapeutic.

In future, an ultrasonic nebulizer that adjusts the vibration frequency of the ultrasonic crystal based on the change in ambient temperature might be designed. This design may optimize the dimensions of particles and fix the period for therapeutic. This version of an ultrasonic nebulizer will prevent the fluctiations caused by temperature variations of the environment.

This model of an ultrasonic nebulizer is yet haven't fabricated. This should be considered as a future direction.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] This paper is an extension of work originally presented in 2016 national Conference on Electrical, Electronics and Biomedical Enginering (ELECO) under the name of
"Ortam Sicakliginin Ultrasonik Nebülizatörler Üzerindeki Etkinliğini Ölçen Sistem Tasarimi"

[2] Ulus B. Et al, Sağlık çalışanları nebülizatörle tedaviyi biliyor mu?, Çocuk Sağlığı ve Hastalıkları Dergisi 2007;
50: 174-179

[3] Güngör S. Et al, Doktorlar, Hemşireler ve Hastalar Nebülizatör Cihazlarını Ne Kadar Doğru Kullanıyor?, Solunum Dergisi t Solunum 2012; 14(3):136–140

[4] Muers MF. Overview of nebulizer treatment. Thorax 1997; 52: 25-30

[5] Uçgun İ., Ventilatördeki Hastada Aerosol Tedavisi, Yoğun Bakım Dergisi 2008;8(3):103-110

[6] Hess DR, Nebulizers: Principles and Performance, Respiratory Care June 2000 Vol 45 No 6

[7] Turner MO, Gafni A, Swan D, FitzGerald JM. A review and economic evaluation of bronchodilator delivery methods in hospitalized patients. Arch Intern Med 1996;156(18):2113–2118.

[8] Cates CJ. Holding chambers versus nebulizers for beta-agonist treatment of acute asthma. The Cochrane Database or Systematic Reviews, 1999.

[9] Smye SW, Jollie MI, Littlewood JM. A mathematical model of some aspects of jet nebuliser performance. Clin Phys Physiol Meas 1991;12(3):289–300.

[10] Nerbrink O, Dahlback M, Hansson HC. Why do medical nebulizers difffer in their output and particle size characteristics? J Aerosol Med 1994;7:259–276.

[11] Newman SP. Aerosol generators and delivery systems. Respir Care 1991;36(9):939–951

[12] Larsen KR, Svendsen UG, Molgaard F, Petersen BN. Comparability of albuterol delivered by a piezoelectric device versus metereddose inhaler in patients with chronic obstructive airways disease. J Aerosol Med 1998;11(2):81–88

[13] AARC Clinical Practice Guideline. Selection of aerosol delivery device. American Association for Respiratory Care. Respir Care 1992;37(8):891–897.

[14] Bulut S et al, A new algorithm for segmentation and fracture detection in X-ray images, Medical Technologies National Conference, 2015, doi: 10.1109/TIPTEKNO.2015.7374572

[15] Çiftçioğlu Ç., Koçak O.Akpek, A., Remote control of centrifuge and injection systems via MATLAB and ARDUINO, Medical Technologies National Conference, 2015, DOI: 10.1109/TIPTEKNO.2015.7374570

[16] Ahmet Öztarhan, Ali Akpek, Efim Oks, Alexey Nikolaev, Modifying medical textiles with antibacterial and friction resistance abilities by an alternative nanotextile technology called ion implantation technique, 15th National Biomedical Engineering Meeting, 2010, 10.1109/BIYOMUT.2010.5479859

[17] Semih Ahmet Cebeci, Çağlar Çiftçioğlu, Onur Koçak,Ali Akpek, Medical Technologies National Conference, 2015, 10.1109/TIPTEKNO.2015.7374571

[18] Ali Akpek, Chongho Youn, Toshiharu Kagawa, Temperature measurement control problem of vibrational viscometers considering heat generation and heat transfer effect of oscillators, 9th Asian Control Conference, 2013, 10.1109/ASCC.2013.6606100

[19] Ucar T., Kocak O., Akpek A., New concept design of an insulin pen for visually impaired or blind diabetius mellitus patients, 2017, 10.1109/TIPTEKNO.2016.7863074

[20] Nikolaev, A. G., et al. "Modification of anti-bacterial surface properties of textile polymers by vacuum arc ion source implantation." *Applied Surface Science* 310 (2014): 51-55, doi.org/10.1016/j.apsusc.2014.04.002

[21] Akpek, Ali, et al. "Effect of thermal convection on viscosity measurement in vibrational viscometer." *Journal of Flow Control, Measurement & Visualization*2014 (2014), 10.4236/jfcmv.2014.21003

[22] Akpek, Ali. "Effect of non-uniform temperature field in viscosity measurement." *Journal of Visualization* 19.2 (2016): 291-299. doi:10.1007/s12650-015-0311-0

[23] Akpek Ali, Youn Chongho and Kagawa Toshiharu "A study on vibrational viscometers considering temperature distribution effect." *日本フルードパワーシステム学会論文集* 45.3 (2014): 29-36. doi.org/10.5739/jfps.45.29

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 951-957 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Graphics on demand: the automatic data visualization on the WEB

Ramzi Guetari*, Maha Mallek

LIMTIC laboratory – Higher Institute of Computer Science, University of Tunis El Manar, Tunisia

ARTICLE INFO

Article history:
Received: 04 May, 2017
Accepted: 11 June, 2017
Online: 30 June, 2017

Keywords:
Structured / Unstructured data
Data visualization
Variable data extraction
Natural Language Processing

ABSTRACT

Data visualization is an effective tool for communicating the results of opinion surveys, epidemiological studies, statistics on consumer habits, etc. The graphical representation of data usually assists human information processing by reducing demands on attention, working memory, and long-term memory. It allows, among other things, a faster reading of the information (by acting on the forms, directions, colors...), the independence of the language (or culture), a better capture the attention of the audience, etc.

Data that could be graphically represented may be structured or unstructured. The unstructured data, whose volume grows exponentially, often hide important and even vital information for society and companies. It, therefore, takes a lot of work to extract valuable information from unstructured data. If it is easier to understand a message through structured data, such as a table, than through a long narrative text, it is even easier to convey a message through a graphic than a table. In our opinion, it is often very useful to synthesize the unstructured data in the form of graphical representations.

In this paper, we present an approach for processing unstructured data containing statistics in order to represent them graphically. This approach allows transforming the unstructured data into structured one which globally conveys the same countable information. The graphical representation of such a structured data is then obvious. This approach deals with both quantitative and qualitative data. It is based on Natural Language Processing Techniques and Text Mining. An application that implements this process is also presented in this paper.

1. Introduction

The English idiom "A picture is worth a thousand words" expresses an idea implicitly admitted by the human mind and indicates that it is easier to understand a complex idea through an image; But also that an image of a subject conveys its meaning or essence more effectively than a long speech or a textual description does. Data and information visualization is one of the most primitive forms of communication known to man. It meets a very basic need: telling a story. This communication way originated in the drawings found in the prehistoric caves dating back to 30,000 B.C., even before the writing that dates only 3000 B.C.

In our modern world, the visualization of information has evolved and is used to communicate information of all kinds: economic, social, military, health, etc. Graphical representations are the best way to make decisions by facilitating to grasp difficult concepts or to identify new patterns. They help to better process information by reducing the requirements on attention, working memory and long-term memory. In a business environment, visualizations can have two broad goals, which sometimes overlap: Explanatory and Exploratory. Data explanation allows solving specific problems and the exploration of large data sets allows understanding it better. Generally speaking, graphical representations are used to gain better insight of the problem we are studying.

Since the early 1990s, the emergence of the World Wide Web (WWW) [1] has facilitated the access to the information to individuals and companies. It has become the most important vehicle of a tremendous amount of data, allowing the people to share the information from the largest database repositories globally. Almost 80% of the data published on the WWW are unstructured and, despite the computer assistance for data

*Corresponding Author: Ramzi Guetari, LIMTIC laboratory – Higher Institute of Computer Science, University of Tunis El Manar, Tunisia
Email: ramzi.guetari@isi.utm.tn

processing, it remains very difficult to extract valuable knowledge from this large volume of data, when it's about countable data, especially without graphical illustrations.

When unstructured data contain statistics, they may be accompanied by graphical illustrations in the form of static images (GIF, PNG, etc.). These charts and graphics are no longer up-to-date as the data content changes. The graphs should then be reproduced in accordance with the new data.

In our opinion, it would be more appropriate to produce graphical illustrations that correspond exactly to the data processed. This can be made possible by updating the graphical representations, each time the data evolves. However, some data change instantly and reproducing the graphics that are faithful to it makes the process of analyzing it and representing it graphically too complex. The analysis and the graphical representation should, therefore, be on-demand.

The objective of this work is to propose an approach for analyzing data, especially from the WWW in order to extract the relevant countable information, organize it in a way it can be graphically represented and annotate it using the means of the Semantic Web techniques [2]. This allows users to request a graphical representation of this information whenever they want. Indeed, the "on demand graphics" allows taking into account the effective state of the information when such representation is requested.

2. The Graphical Representation of Data

William Playfair (22 September 1759 – 11 February 1823) was the first to devise and publish different types of statistical chart (most notably the time-series line graph, the bar chart, the pie chart, and the circle chart). Modern statistical graphs are almost identical to those published by Playfair and count four basic types of graphs used as well as the graphical representation of stochastic variables: Line graphs/Polygraphs, Bar graphs, Pie graphs and Flow Charts.

Not all data could be graphically represented, and not all graphic representations are of paramount importance. Only data that may vary for effects-related causes can be graphically represented. The objective of these graphical representations is to analyze the causes and predict the effects in order to make the appropriate decisions.

It is usual in the computer science domain to classify data into two well-known categories: structured and unstructured data [3].

Structured data are usually associated with a model or a schema and resides in a fixed field within a record or file such as Relational databases and Spreadsheets. The schema defines the structure of data with data types (string, integers, date, etc.) as well as the way data are stored, accessed and processed. The structured data are easily maintainable and manageable since they are based on Schema-on-write method and managed with most known Structured Query Language (SQL).

Unstructured data do not reside in a field or record, do not have a standard model and do not follow a schema. There's no specific definition of storing, accessing, and processing this kind of data. The unstructured data might follow an internal structure; however, there's no schema, tags, or markers describing the fields of data. Although the unstructured data are difficult to process using traditional computer software modules, their processing is nonetheless very important because of its important volume compared to the structured data. They have remained unidentified or ignored for a long time knowing that important business prospects are hidden in these data as well as social and political trends.

3. Processing the Data

Processing data means processing both structured and unstructured data. In the first case, when data are structured, the variable data are explicitly specified and their different values are well known. The graphical representation of such a data is an easy and very well mastered process. In the second case, when it comes to unstructured data, the amount of text is simply too large to read and analyze efficiently even with assistance of computers. The work is, therefore, more tedious. On the one hand, we must analyze the text to determine the context. On the other hand, it is necessary to extract the statistical data, when it exists, to process it and to carry out calculations in order to transform unstructured data into structured one such as a table.

3.1. Processing Structured Data

The processing of structured data does not pose any significant difficulties. The data are represented in tables, generally easy to understand and interpret. Tables are easy to read, however a visual can communicate more information than a table in a much smaller space and is, therefore more effective than tables for presenting data. For example, notice the Table 1 below, related to the French presidential elections of 2017, and try to spot the candidate with the highest estimated voting rate.

Date	Sample size	Mélenchon	Macron	Fillon	Le Pen
28 Mar 2017	2,231	14%	25.5%	17.5%	25%
24 Mar 2017	2,225	13%	26%	18%	25%
20 Mar 2017	2,245	12.5%	26%	18%	25%
18 Mar 2017	935	11.5%	25%	18%	26%

Table 1. Polls finished since the publication of the official list of candidates

When the data contained in this table are represented graphically (Fig. 1), the information is obtained very quickly. The graphical representation of a structured data set represented in the form of a table presents no particular difficulties. All data processing tools, modern spreadsheets and statistical applications do it very well. The only small problem to be solved in this case is to determine which variable will be represented in the *y-axis* and which will occupy the *x-axis*. This problem is solved from the context of the document that is determined in the same way that the data are structured or not. We will discuss the context of the document and the extraction of variables when processing unstructured data.

The unstructured data that we take into account are texts expressed in natural language. To represent graphically the unstructured data it is necessary that this data contain countable information. These countable data have to be extracted from the text and processed to be transformed into structured data.

Transforming unstructured data into structured data containing the same information requires the specification of a model or a schema for the target structured data. The structure of the countable data in relation to this work is simple and can be specified as a simple association (*variable*, *value*). This association represents what we call the "Main Variable"; e.g. (*candidate*, *voting_rate*). The Main Variable is the one whose values are

represented on the *y-axis*. The value associated with a variable may itself depend on other variable criteria. For example, the *voting_rate* may depend on the *date*, the *region*, the *age* of the voter, etc. A second association necessary is in this case: (*value*, *criterion*).

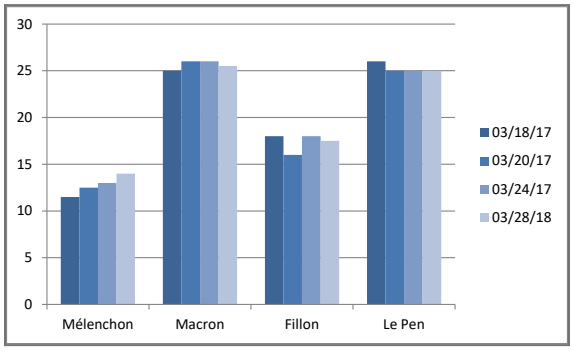


Fig. 1. Data visualization obtained from the Table 1.

3.2. Processing Unstructured Data

The schema of the targeted structured data is modeled as an Ontology that we call "*Ontology of the Context*". Indeed, the different associations specified in this ontology depend on the context. Fig. 2 shows the "*Ontology of the Context*" related to the French Presidential elections of 2017.

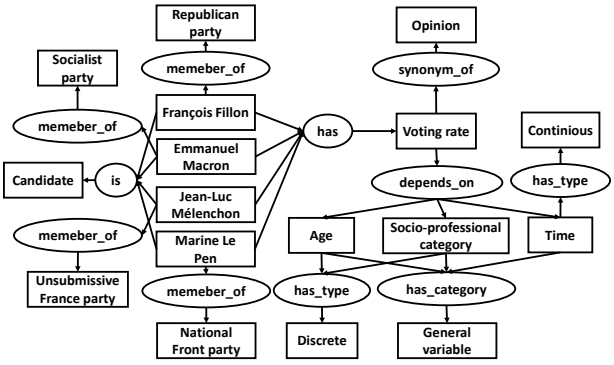


Fig. 2. Ontology of the context

At the present time, the different Ontologies of the Context are modeled manually. The ideal would be that these ontologies be created and updated automatically; this is one of the perspectives of our research works.

The processing of the data is based on two main steps: (i) the identification of variable data and the building of the corresponding structured entity; and (ii) the semantic annotation that avoids reprocessing a set of unstructured data that have already been processed.

4. Transforming unstructured into structured data

Unstructured data that can be subject to statistical calculations may appear in two ways: qualitative data which is when descriptive statements can be made about a subject based on observations, interviews or evaluations, e.g. a political forum (Fig. 3) or numerical values as in the example (*§1*) hereafter:

"*The French Presidential Election of 2017: if you are a French citizen, at least 18 years old, you can participate in the French presidential elections of 2017 and express your opinion. As for most elections in France, presidential elections are conducted in two stages. The first round, which takes place on 23rd April 2017, is open to any candidate who meets the requirements. Assuming, as is virtually certain to happen, that no single candidate acquires no absolute majority of votes in the first round, there is then a second round, a fortnight later, which is a runoff between the two candidates with the highest votes in the first round . The next French President will be whoever gets a simple majority of the votes in the second round.*"

Fig. 3. "French Politics" forum

The processing of unstructured data for the extraction of statistics is a tedious work. On the one hand, we must analyze the text to determine the context. On the other hand, it is necessary to extract the statistical data, when it exists, to process it and to carry out calculations in order to transform unstructured data into structured one such as a table. The overall context in which the data are used allows to build or to update the "Ontology of the Context" [4].

4.1. Identification of the context

The identification of the context requires five main steps as described in [5]:

- The tagging [6] which consists in dividing the text into a sequence of tokens, which roughly correspond to a words and associating to each one of them a label providing certain information (syntactic category, number, verb tense…). This step also allows identifying the different noun phrases and complex nominal groups found in a document. A noun phrase is a word or group of words in a sentence that acts like a noun. A complex nominal group is a simple noun phrase modified by one or two prepositional phrases (preposition followed by a nominal group).
- Named Entity Recognition [7] step which purpose is to label sequences of words in a text (tagged in the first step) which are the names of persons, organizations, locations,

monetary values, etc. The annotation of the different named entities allows identifying variables.

- The coreference resolution [8] occurs when multiple noun phrases refer to the same entity or object in a document. It is an important step in our process, allowing the information extraction, such as the overall context of the document and the significant variable data.
- The Identification of the main topics [9] of a document helps us to identify the most important textual units of a document. These units are formed by key-words that we refer to in this document as "key-terms candidates and are the different noun phrases in a document. The identification of the main topics is basically made thanks to the Text Rank algorithm [10] and the Word Sense Disambiguation process [5].
- The identification of the context is the last step. Once the different main topics are identified, we need to determine only one of them that will be the one representing and defining the context of the document. To do this, first, a similarity rate [11] is computed between each couple of words from the main topics set. The word that has the highest similarity rate with all the other words, is defined as the most important topic of the document. Additional information may be necessary to refine the context. Thus, in a second step, the identification of the different complex nominal groups found in the document is required in order to have a more accurate context. An algorithm allowing the identification of the context has been implemented and experimented [9].

When applying the Context Identification Process on the following paragraph (*§1*), the text raking allows selecting the different main topics found in the data: "*French presidential election*", "*round*", "*president*", "*votes*", "*candidates*", etc. The similarity between all pairs of main topics is calculated and given in Table 2.

	Round	Votes	Candidates	France	President
French Presidential Elections	0,0635	18,7456	0,0616	0,0592	0,0715
Round	×	0,0575	0,0625	0,0585	0,046
Votes	×	×	0,0618	0,0592	0,0715
Candidates	×	×	×	0,0629	0,0487
France	×	×	×	×	0,0471

Table 2. Measurement of similarity rates between all pairs of main topics

The topic that has the highest similarity rate with all the others is: "*French presidential election*". However important information is missing: the date. The algorithm presented in [9] determines the different complex nominal groups and identifies the final context: "*The French presidential election of 2017*".

The similarity rate between the topics "*French presidential election*" and "*votes*" is very high. This represents a very good indication of the variable data to be extracted from the text. Referring to the "Ontology of the Context", we see that the vote is associated with a candidate. The pair (*Variable*, *Value*) of the "Main Variable" is thus identified. Other information is easily identified in this paragraph which is the relationship between the vote and the city which represents the pair (*value*, *criterion*).

4.2. Identification of the countable information

The identification and retrieval of countable data depend on the nature of the data sources. The processing of such a data differs according to the variable data being represented by numerical values or qualitative data.

Case 1: *Data containing numeric values*

The extraction of the variables consists first of all in identifying the key words characterizing each of the identified countable data. This work is still in progress and is mainly based on the Natural Language Processing (NLP) [12] and the Text Mining [13, 14]. At the present time, it uses a controlled vocabulary and is not completely automated. The purpose of this process is to determine the identifiers of each countable data. An identifier of a variable datum is a token obtained from unstructured text and standardized according to the well-known methods used in NLP and indexing, namely stemming and lemmatization [15]. This process occurs simultaneously with the identification of the context. It uses the different named entities extracted in the data processing stage in order to group the ones that belong to the same category (the names of persons, organizations, etc.) in a "class of category". The identifier of the Main Variable is obtained from the keyword which has the highest similarity rate with the identified context, after resolving the coreferences.

The process of identification of significant variable data takes place in three main steps [9]: (*i.*) the grouping of the named entities in classes of categories, (*ii.*) the Classification of named entities of the same class of category, and (*iii.*) the variables' identification. We take the following example (*§2*) to illustrate this three-step process:

(*§2*) "*The French presidential election will be held on April 2017. The election will occur in different French cities to allow the winning candidate to access the presidency. Until now, the candidate with the estimated highest scores of votes is Marine Le Pen. She is credited with 35% of voting intentions in Paris. Polls give her 32% of voting intentions in Toulouse. However, she is ranked third in other cities such as Marseille with 18% of voting intentions and Lyon with 16.5% of voting intentions according to the most recent surveys. We asked the question "who would you vote for?" for the two childhood friends Fabien Martin and Marcel Laurent. Although they lived in the same neighborhood and in the same conditions of life, the two friends vote for two different candidates. Fabien says he will vote for François Fillon while Marcel is in favor of Marine Le Pen*"

i. Processing the named entities, this consists in grouping the named entities belonging to the same "Class of Category" such as the names of persons, organizations, etc. In the (*§2*), a Class of Category "PERSON" gathers four names (Fig.4)

ii. Processing the named entities of the same class of category. Named entities within a same category may belong to different concepts. A concept is a general idea about some distinct entities or classes of entities. Let's consider a class of category that contains different names of candidates and voters (Fig. 4), the process must be able to differentiate between both concepts "*candidate*" and "*voter*". The named entities that belong to the same class of category are classified according to their concept. The Class of Categories PERSON includes candidates for the presidential election (*François Fillon* and *Marine Le Pen*) and other names that can be considered as voters (*Fabien Martin* and *Marcel Laurent*) (Fig. 4). Two classes of concepts have been identified: "*CANDIDATE*" and "*UNKNOWN*" (Fig. 5)

Fig. 4. Building the Classes of Categories

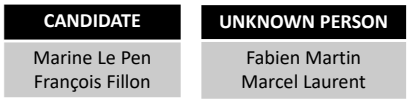


Fig. 5. Building the classes of concepts

iii. Identification of countable data. The last step consists in identifying the variable data: the "Main Variable" that depends on the context and the others according to which the Main Variable varies. We have developed an algorithm [9] allowing to extract the different variables and to "format" them in a structure such as a table. The algorithm builds a matrix *M* with *n* rows (*n* is the number of sentences in the document) and three columns containing respectively *nouns*, *verbs* and *complements*, which are obtained from each sentence. The algorithm splits up the matrix *M* into several sub-matrices (M_k). The cutting of the matrix is based on the verbs. Each row of each sub-matrix contains the triplets (*name*, *verb*, *complement*) extracted from each sentence. The rows of each sub-matrix are formed of all triplets containing the same verb or one of its synonyms.

The processing of (*§2*) using the algorithm [9] results in the following sub-matrices:

$$M_1 = \begin{pmatrix} \text{Noun} & \text{Verb} & \text{Complement} \\ \text{Marine Le Pen} & \text{is credited} & \text{with 35\% of voting intention in Paris} \\ \text{Polls} & \text{give} & \text{Marine Le Pen 32\% of voting intentions in Toulouse} \\ \text{Marine Le Pen} & \text{is ranked} & \text{third in other cities such as Marseille with 18\% of voting intentions and lyon with 16.5\% of voting intentions according to the most recent surveys} \end{pmatrix}$$

$$M_2 = \begin{pmatrix} \text{Noun} & \text{Verb} & \text{Complement} \\ \text{The French presidential election} & \text{will be held} & \text{on April 2017} \end{pmatrix}$$

$$M_3 = \begin{pmatrix} \text{Noun} & \text{Verb} & \text{Complement} \\ \text{The candidate with the highest scores of vote} & \text{is} & \text{Marine Le Pen} \end{pmatrix}$$

Among all the sub-matrices obtained from the splitting up of the matrix *M*, we keep only those with the noun and complement columns corresponding to the parameters of the main variable. In the considered example, only the sub-matrix M_1 contains the name of a candidate in the noun column or complement column. While processing the nouns or the complements one can identifies voting rates (numerical values) as well as the names of some French cities. The pair (*voting_rate*, *city*) is therefore identified. Table 3 is the result of this process:

Candidate	Votes	City
Marine Le Pen	35%	Paris
Marine Le Pen	32%	Toulouse
Marine Le Pen	18%	Marseille
Marine Le Pen	16,5%	Lyon

Table 3. Identification of countable data

Case 2: Qualitative data

The political forum contains only expressions of voting intentions and no numerical data (Fig. 3). Each row of it is an expression of opinion in which the "*speaker*" supports the *candidate* he has chosen.

In an unstructured text where people express their opinions, it is important to determine the author of the text of an opinion as well as the content of that opinion. The Opinion generally expresses a positive or negative feeling towards a particular context or situation. The recognition of the "*speaker*" prevents counting the same opinion several times.

Several works have been carried out for the recognition of the "*speaker*" in a text. K. Glass and S. Bangay [16] propose "a naïve, salience-based method for speaker identification in fiction books". They have shown that it is possible to identify different speakers from a text without using a complex machine learning or logic-based techniques. D. K. Elson et al. [17] proposed a method for extracting social networks from literary fiction. This method allows identifying all the characters in a story and determining which characters have relationships with each other. The method also allows determining whether the sentiment in each relationship is positive or negative. Other research works [18, 19] have addressed the same subject and demonstrated the feasibility of recognizing the speaker in the texts. Some of the "speaker recognition" methods have been implemented.

The sentiment analysis [20, 21], also known as opinion mining, seeks to recognize and characterize aspects of an author's opinion or appreciation on a topic from information in natural language written texts. An opinion can be expressed by a person or an organization about a product, a service, a topic, an event, etc. We use the word "*entity*" to qualify the topic that is evaluated by an expression of feeling. For example, the entity in a political forum is a candidate.

The sentiment analysis focuses today on attributing of a polarity to subjective expressions (words and phrases that express opinions, emotions, feelings, etc.) in order to decide the orientation of an opinion [22] or the positive / negative / neutral value of it in a document [23][24][25]. Most of the techniques used for the classification of sentiment in texts in natural languages are based on supervised learning, e.g. naive Bayesian classification, support vector machines (SVM), or any other supervised learning method.

In our experiments, we have used the implementation of the speaker recognition as well as the sentiment analysis in the Stanford CoreNLP framework [26].

The object that is evaluated by a sentiment in a text depends on the context. At the present time, it is determined as the Main Variable by the Ontology of the Context. The value associated with the Main Variable is calculated by summing the positive opinions expressed. Negative or neutral opinions are not taken into account. For example, in political forum (Fig. 3), voting intentions are calculated by adding together the speakers who express an opinion favorable to a given candidate. When a speaker changes his mind, only his last expression is taken into account.

5. Application

The results of our research works have been implemented as an application allowing transforming unstructured data into structured data that could be represented graphically. The architecture of the application is presented in Fig. 6.

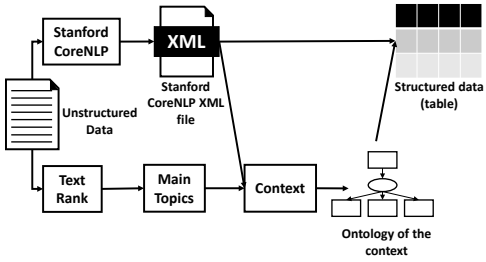


Fig. 6. Architecture of the application

The application has been developed with the help of the following open source frameworks:

- *Stanford CoreNLP* [26], a Java open source framework created by Stanford University, that implements most of the common core natural language processing (NLP) steps needed in our approach, from tokenization through to coreference resolution.
- *TextRank* [27]: a Java implementation of the TextRank algorithm allowing the identification of the most important keywords in a document.
- *WordNet* Similarity *for Java* (WS4J) [28]: a module that implements a variety of semantic similarity and relatedness measures based on information found in the lexical database WordNet.

The application has been tested on several types of data. In this paper, we present the case of the French presidential elections of 2017. The recognition of the context is precise however, we had some problems for the recognition of the "speaker". We have modified the source code, but we have not yet got the expected precision and a research work is underway to address this issue.

The forum contains only expressions of voting intentions and no numerical data (Fig. 3). Each row of it is an expression of opinion in which the "*speaker*" supports the *candidate* he has chosen. Fig. 7, Fig. 8, Fig. 9 and Fig. 10 represent the output of the application:

Voter	Age	Profession	Candidate	Date
Anna Bourgean	44	Lawyer	Marine Le Pen	4/03/2017
Isabelle Vincent	23	Student	Emmanuel Macron	4/03/2017
Stephane Roux	40	Doctor	François Fillon	4/03/2017
Joel Francois	25	Student	François Fillon	4/03/2017
Jean Martin	24	Student	Marine Le Pen	4/03/2017
Lucie Simon	40	Doctor	Marine Le Pen	4/03/2017
Michelline Vincombre	36	Lawyer	Marine Le Pen	4/03/2017
Christophe Morceau	26	Student	François Fillon	5/03/2017
Antoine Bernard	50	Lawyer	Marine Le Pen	5/03/2017
Emma Robert	23	Student	Marine Le Pen	5/03/2017
Gabriel Legrand	32	Doctor	Marine Le Pen	5/03/2017
Louise Clement	43	Doctor	Emmanuel Macron	5/03/2017
Camille Garcia	33	Lawyer	Emmanuel Macron	5/03/2017
Alice Fournier	36	Doctor	Emmanuel Macron	6/03/2017
Léo Richard	23	Student	Emmanuel Macron	6/03/2017
Nathalie Durand	25	Student	François Fillon	6/03/2017
Caroline Martinez	37	Doctor	François Fillon	6/03/2017
Tomas Andre	29	Lawyer	François Fillon	6/03/2017
Jules Dubois	47	Lawyer	Marine Le Pen	6/03/2017
Fabien Bertrand	44	Doctor	François Fillon	7/03/2017
Valérie Dupon	36	Lawyer	Emmanuel Macron	7/03/2017
Manon Girard	55	Doctor	Emmanuel Macron	7/03/2017
Michelle Vincent	36	Lawyer	François Fillon	7/03/2017

Fig. 7. Countable data recognition

Candidate	Voting Rate
Marine Le Pen	35 %
François Fillon	35 %
Emmanuel Macron	30 %

Fig. 8. Voting rate per candidate

	4/03/2017	5/03/2017	6/03/2017	7/03/2017
Marine Le Pen	57 %	54 %	42 %	35 %
François Fillon	29 %	23 %	32 %	35 %
Emmanuel Macron	14 %	23 %	26 %	30 %

Fig. 9. Voting rate vs. Date

Candidates	Doctors	Lawyers	Students
Marine Le Pen	25 %	50 %	29 %
François Fillon	37,5 %	25 %	42 %
Emmanuel Macron	37,5 %	25 %	29 %

Fig. 10. Voting rates vs. Socio-professional category

The following graphs can be obtained on request from the various calculation results performed by the application: the voting rate per candidate (Fig. 11), the voting rates vs. date (Fig. 12) and the voting rates vs. socio-professional category (Fig. 13)

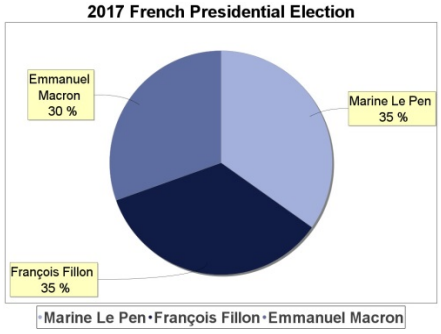


Fig. 11. Pie chart of the voting rate per candidate

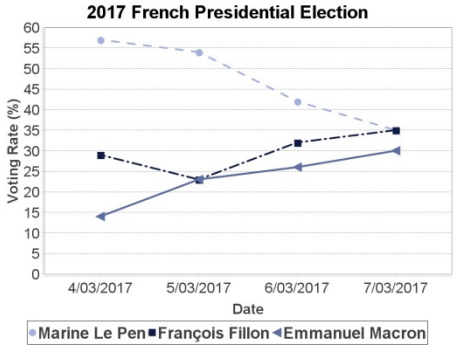


Fig. 12. Polygraph of the voting rates vs. date

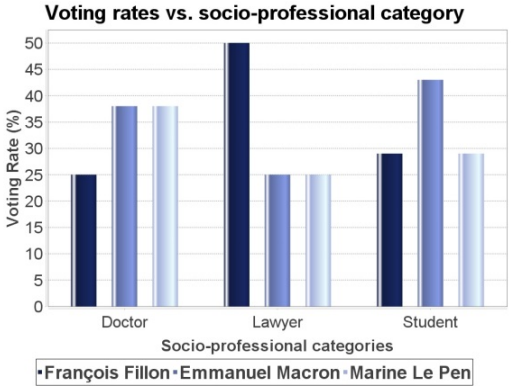


Fig. 13. Polygraph of the voting rates vs. date

6. Conclusion

Visualization provides a simple, easy-to-interpret idea of a data volume that may be too large. The graphically representable data generally contain relationships between causes and effects that may change according to different criteria (date, region, climate, etc.). Data linked by cause-and-effect relationships can be either quantitative or qualitative and may also be in the form of structured or unstructured data. When it comes to unstructured data, it is easy for a human mind to extract relationships and make the graphical representation if the volume of data provided is reasonable. Since this is a very large volume of data, computer assistance is essential. For structured data, this is hardly a problem, but when it comes to unstructured data, the work is more tedious.

In this paper, we present a method allowing the graphical representation of statistical data presented in the form of unstructured data. The first consists of transforming the unstructured data into structured data and then graphically represent this latter, which is an obvious task. We also present a framework developed from the results obtained from our research. The method is based on NLP and text mining for the analysis and identification of statistical data. An algorithm has been developed for the transformation of unstructured statistical data into structured data easily represented graphically.

The work presented in this paper is only the first step. Currently, it leaves an important number of research perspectives. Indeed, this work concerns only a single document. In our future work, we will exploit the results of a Web search query to make a graphical representation of common significant variable data. Similarly, we will intend to improve our process using the Text mining. This technique concerns the classification of multiple documents in order to enrich the "Ontology of the Context". Besides, it will allow the analyses of the documents in order to extract the variable data.

References

[1] T.J. Berners-Lee et al., "World Wide Web: the information universe", Electronic Networking: Research, Applications and Policy, 2(1), 52–58, 1992.

[2] Antoniou G. and Van Harmelen F. "A Semantic Web Primer", The MIT Press, Cambridge MA. 2008.

[3] Cardoso, J., "The Syntactic and the Semantic Web. Semantic Web services: theory, tools, and application", J. Cardoso (Ed), Information Science Reference, 2007.

[4] N. Ben Chaabene, and R. Guetari, "Semantic Annotation for the "on demand graphical representation" of variable data in Web," IEEE International conference en system, Man, and Cybernetics, Budapest, Hungary, 2016.

[5] Van Hooland, S., De Wilde, M., Verborgh, R., Steiner, T., Van de Walle, R., "Exploring entity recognition and disambiguation for cultural heritage collections". The Journal of Digital Scholarship in the Humanities, 30(2), 262-279, 2015

[6] Souvay, G., Pierrel, J.M., « LGeRM Lemmatisation des mots en moyen français », Traitement Automatique des Langues. 50(2), 149-172, 2009, "(article in French with an abstract in English)

[7] Tamilin, A., Magnini, B., Serafini, L., Girardi, C., Joseph, M., Zanoli, R., "Context-Driven Semantic Enrichment of Italian News Archive", 6th European Semantic Web Conference, Heraklion, Greece, 2009.

[8] Lee, H., Peirsman, L., Chang, A., Chambers, N., Surdeanu, M., Jurafsky, D. "Stanford's Multi-Pass Sieve Coreference Resolution System", Conference on Computational Natural Language Learning : Shared Task, Portland, Oregon, 2011

[9] M. Mallek. R. Guetari, "Automatic detection of variable data in Web document: Graphical representation on demand" in IEEE International Computer Science and Engineering Conference Web conference (ICSEC), Chiang Mai Thailand, 2016.

[10] R. Mihalcea, and P. Tarau, "TextRank: Bringing Order into Texts", The 2004 Conference on Empirical Methods in Natural Language Processing, Barcelona, Spain, 2004

[11] Jiang, J., Conrath, D., "Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy", International Conference Research on Computational Linguistics (ROCLING X), pp. 19–33 (1997)

[12] G. Chowdhury, "Natural language processing," Annual Review of Information Science and Technology, 37, pp. 51-89, ISSN 0066-4200, 2003.

[13] D. Nasa, "Text Mining Techniques- A Survey", International Journal of Advanced Research in Computer Science and Software Engineering, 2(4), 2012.

[14] A. Kaushik, S. Naithani, "A Comprehensive Study of Text Mining Approach", International Journal of Computer Science and Network Security, 16(2), 2016

[15] V. Gupta, and G.S. Lehal, "A Survey of Text Mining Techniques and Applications," Journal of emerging technologies in Web intelligence, Vol. 1 – n° 1, pp. 60 – 76, 2009.

[16] K. Glass and S. Bangay, "A naive salience-based method for speaker identification in fiction books". In Proceedings of the 18th Annual Symposium of the Pattern Recognition, 2007.

[17] N. Lie, and D.D. Wu, "Using text mining and sentiment analysis for online forums hotspot detection and forecast," International Journal of Decision Support Systems, pp. 354 – 368, 2010.

[18] H. He, D. Barbosa and G. Kondrak, "Identification of Speakers in Novels", in Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, 2013.

[19] S. Bethard, H. Yu, A. Thornton, V. Hativassiloglou and D. Jurafsky, "Automatic Extraction of Opinion Propositions and their Holders". In AAAI Spring Symposium on Exploring Attitude and Affect in Text, 2004.

[20] S.M. Shamimul Hasan, D. A. Adjeroh, "Detecting Human Sentiment from Text using a Proximity-Based Approach", In Journal of Digital Information Management 9(5), pp 206-212, 2011

[21] W. Jin, "A Novel Lexicalized HMM-based Learning Framework for Web Opinion Mining", In Proceedings of the 26th International Conference on Machine Learning, Montreal, Canada, 2009.

[22] Wilson T., Wiebe J. et Hwa R. (2004). Just how mad are you? Finding strong and weak opinion clauses, In proceedings of AAAI'04.

[23] V. Hatzivassiloglou and K. R. McKeown, "Predicting the semantic orientation of adjectives", In proceedings of ACL'97, p. 174–181, 1997.

[24] H. Yu and V. Hatzivassiloglou, "Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences", In proceedings of EMNLP'03, p. 129–136, 2003

[25] S.-M. Kim and E. Hovy, "Determining the sentiment of opinions", In Actes of COLING'04, p. 1267–1373, 2004.

[26] C.D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S.J. Bethard, and D. McClosky, "The Stanford CoreNLP Natural Language Processing Toolkit," In Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55-60, 2014.

[27] D.D Pawar, M.S Bewoor, and S.H Patil, "Text Rank: A Novel Concept for Extraction Based Text Summarization," International Journal of Computer Science and Information Technologies, 5(3), pp. 3301 – 3304, 2014.

[28] H. Shima, "WS4J : WordNet Similarity for Java," [Online]. Available at: https://code.google.com/p/ws4j/, 2013.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 958-966 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Spectral Reflectance Estimation and Color Reproduction Based on Sparse Neugebauer Model

Kohei Inoue*, Kenji Hara, Kiichi Urahama

Department of Communication Design Science, Faculty of Design, Kyushu University, 815-8540, Japan

ARTICLE INFO

Article history:
Received: 29 April, 2017
Accepted: 13 June, 2017
Online: 30 June, 2017

Keywords:
Spectral Reflectance
Color Reproduction
Sparse Neugebauer Model

ABSTRACT

This paper proposes a method for estimating the spectral reflectance of Neugebauer primaries or eight basic colors when the spectral reflectance data of several colors are given. The proposed method is based on Neugebauer model, which expresses an arbitrary color as a weighted average of Neugebauer primaries, and its sparse version called sparse Neugebauer model. We also reproduce the colors from the estimated reflectance spectra, and show experimental results of spectral reflectance estimation and color reproduction from the estimated spectral data. It is experimentally demonstrated that the sparse Neugebauer model outperforms the conventional Neugebauer model in both spectral reflectance estimation and color reproduction.

1. Introduction

The reflectance properties of object surfaces are determined by the spectral reflectance factors that are normally measured at regular intervals in the visible spectrum of radiation by reflectance spectrophotometers [1]. The integrals of the products of the spectral reflectance, the relative spectral power distribution of an illuminant and color matching functions produce the tristimulus values. On the other hand, the inverse computation from tristimulus values or colorimetric data to the spectral reflectance is a difficult problem because the computation of tristimulus values from spectral reflectance involves the dimensionality reduction of data, i.e., the spectral reflectance estimation is an ill-posed problem.

Recently, Sharma and Wang proposed a neural network-based spectrum recovery method, and evaluated the accuracy of the spectrum recovery by the spectral mean-squared error [2]. Harifi et al. presented a principal component analysis (PCA)-based method which employed six principal components [3]. Amiri and Amirshahi proposed a step by step progressive method for recovery of spectral data from the corresponding colorimetric information, where the spatial dimensions are decreased from 6D to 3D in a step-by-step manner to decrease the number of samples used for color gamut interpolation [4]. Wu et al. presented an approach that uses a new sample selection criterion based on color feature match to select a series of suitable samples for creating the adapted transformation matrix to reconstruct spectral reflectance [5]. Inoue et al. proposed two methods for recovering the reflectance spectra of colorimetric data by using the non-negative constraints in reflectance spectra, where the problem of reflectance spectra recovery is formulated as a nonnegative least squares problem [6] into which the Neugebauer model [7] is incorporated [8]. These methods can be used to estimate the spectral reflectance from general colorimetric data including CIE tristimulus values and CMYK-printed samples.

The Neugebauer model can be used to infer the gamut boundaries of printed colorants on paper [7] by printers such as conventional CMYK printers. If three inks, such as cyan, magenta and yellow, are utilized, then the paper may be covered by one ink, two inks, three inks, or may be left blank. As a result, eight combinations of overlap of inks can occur, and they correspond to eight basic colors, such as red, green, blue, black, cyan, magenta, yellow and white, which are referred to as the Neugebauer primaries [7]. Although the Neugebauer model was originally developed for modeling the printed colorants on paper, it can be used for general colorimetric data as we will prove in Subsection 2.1 that the equation of the Neugebauer model holds for any color in RGB color space.

In this paper, we propose sparse Neugebauer model for estimating spectral reflectance of Neugebauer primaries or basic eight colors when reflectance spectra of several colors are given. Experimental results show that the proposed method achieves

*Corresponding Author: Kohei Inoue, Department of Communication Design Science, Faculty of Design, Kyushu University, Japan.
Email: k-inoue@design.kyushu-u.ac.jp

lower estimation error than the previous method with the conventional Neugebauer model. We also reproduce the Neugebauer primaries and the colors of given reflectance spectra from the estimated reflectance spectra. The effectiveness of the sparse Neugebauer model in color reproduction is also demonstrated in comparison with the conventional Neugebauer model.

The rest of this paper is organized as follows. Section 2 summarizes the Neugebauer model, and then proposes the sparse Neugebauer model. Section 3 proposes a method for estimating the spectral reflectance based on the Neugebauer models, and a convergence property of the proposed estimation algorithm is proved. Section 4 describes the procedure for reproducing the colors from the estimated spectral data. Section 5 shows experimental results. Finally, Section 6 concludes this paper.

2. Neugebauer Models

In this section, we first summarize the Neugebauer model, and then propose sparse Neugebauer model to improve the performance of spectral reflectance estimation and color reproduction.

2.1. Neugebauer Model

Let $\boldsymbol{v} = [r, g, b]$ be a color vector in RGB color space (bold italic lowercase letters denote vectors), where r, g and b denote the red, green and blue components of the color, and satisfy $0 \le r \le 1, 0 \le g \le 1$ and $0 \le b \le 1$, respectively. Then $\boldsymbol{v}$ have another expression as follows:

$$\boldsymbol{v} = a_1\boldsymbol{r} + a_2\boldsymbol{g} + a_3\boldsymbol{b} + a_4\boldsymbol{k} + a_5\boldsymbol{c} + a_6\boldsymbol{m} + a_7\boldsymbol{y} + a_8\boldsymbol{w}, \quad (1)$$

where $\boldsymbol{r} = [1,0,0], \boldsymbol{g} = [0,1,0], \boldsymbol{b} = [0,0,1], \boldsymbol{k} = [0,0,0], \boldsymbol{c} = \boldsymbol{w} - \boldsymbol{r}, \boldsymbol{m} = \boldsymbol{w} - \boldsymbol{g}, \boldsymbol{y} = \boldsymbol{w} - \boldsymbol{b}$ and $\boldsymbol{w} = [1,1,1]$ denote the Neugebauer primaries [7]: red, green, blue, black, cyan, magenta, yellow and white, respectively. The coefficients in (1) are given by

$$\begin{aligned} a_1 &= r(1-g)(1-b), \quad a_5 = (1-r)gb, \\ a_2 &= (1-r)g(1-b), \quad a_6 = r(1-g)b, \\ a_3 &= (1-r)(1-g)b, \quad a_7 = rg(1-b), \\ a_4 &= (1-r)(1-g)(1-b), \quad a_8 = rgb, \end{aligned} \quad (2)$$

which are called the Demichel equations, and (1) is called the Neugebauer equation [7]. The coefficients in (2) are nonnegative and satisfy $\sum_{k=1}^{K} a_k = 1$ where $K = 8$. The equality in (1) with the coefficients in (2) can be confirmed for each element as follows: for example, the first element in (1) satisfies

$$\begin{aligned} &r(1-g)(1-b) + r(1-g)b + rg(1-b) + rgb \\ &= r[(1-g)(1-b) + (1-g)b + g(1-b) + gb] \\ &= r[(1-g)\{(1-b) + b\} + g\{(1-b) + b\}] \\ &= r[(1-g) + g][(1-b) + b] = r. \end{aligned}$$

The similar equations hold for the second and third elements in (1).

We simply call (1) with (2) the Neugebauer model, which is valid for both tristimulus values and spectral reflectance, and use it for spectral reflectance estimation.

2.2. Sparse Neugebauer Model

Let $S = \{\boldsymbol{r}, \boldsymbol{g}, \boldsymbol{b}, \boldsymbol{k}, \boldsymbol{c}, \boldsymbol{m}, \boldsymbol{y}, \boldsymbol{w}\}$ be a set of the Neugebauer primaries. Then the Neugebauer equation in (1) states that an arbitrary color vector $\boldsymbol{v}$ can be expressed as a weighted average of the eight color vectors in S. We can also express a color vector in RGB color space as a linear combination of four colors which are in general position in RGB color space because four points being in general position in a three-dimensional color span a three-dimensional linear space, and form a tetrahedron. If a color is just on a plane spanned by three vertices of RGB color cube or a line connecting two vertices of RGB color cube, then the color is expressed as the linear combinations of three or two vertices, respectively. We describe the most general four-color case as follows (in the other degenerated cases, the description can be simplified): Let $\tilde{S} = \{\boldsymbol{s}_1, \boldsymbol{s}_2, \boldsymbol{s}_3, \boldsymbol{s}_4\}$ be a subset of S the elements of which are in general position. Then, for a color vector $\boldsymbol{v}$, we have

$$\begin{aligned} \boldsymbol{v} - \boldsymbol{s}_1 &= \alpha_2(\boldsymbol{s}_2 - \boldsymbol{s}_1) + \alpha_3(\boldsymbol{s}_3 - \boldsymbol{s}_1) + \alpha_4(\boldsymbol{s}_4 - \boldsymbol{s}_1) \\ &= [\alpha_2, \alpha_3, \alpha_4]\begin{bmatrix} \boldsymbol{s}_2 - \boldsymbol{s}_1 \\ \boldsymbol{s}_3 - \boldsymbol{s}_1 \\ \boldsymbol{s}_4 - \boldsymbol{s}_1 \end{bmatrix}, \end{aligned}$$

from which the coefficients $\alpha_2, \alpha_3, \alpha_4$ are given by

$$[\alpha_2, \alpha_3, \alpha_4] = (\boldsymbol{v} - \boldsymbol{s}_1)\begin{bmatrix} \boldsymbol{s}_2 - \boldsymbol{s}_1 \\ \boldsymbol{s}_3 - \boldsymbol{s}_1 \\ \boldsymbol{s}_4 - \boldsymbol{s}_1 \end{bmatrix}^{-1},$$

where $\boldsymbol{s}_1$ is selected as a basic point. As a result, we have

$$\boldsymbol{v} = \alpha_1\boldsymbol{s}_1 + \alpha_2\boldsymbol{s}_2 + \alpha_3\boldsymbol{s}_3 + \alpha_4\boldsymbol{s}_4, \quad (3)$$

where $\alpha_1 = 1 - \alpha_2 - \alpha_3 - \alpha_4$, and the four coefficients are nonnegative if $\boldsymbol{v}$ is in the tetrahedron with the four vertices in $\tilde{S}$. We call (3) the sparse Neugebauer model since four coefficients among eight ones in (1) are zero in (3).

We adapt the minimal brightness variation criterion (MBVQ) proposed by Shaked et al. [9] for selecting four Neugebauer primaries $\boldsymbol{s}_1, \boldsymbol{s}_2, \boldsymbol{s}_3, \boldsymbol{s}_4$, which are called the minimal brightness variation quadruple (MBVQ) [9], and given by $\{\boldsymbol{r}, \boldsymbol{g}, \boldsymbol{b}, \boldsymbol{k}\}, \{\boldsymbol{w}, \boldsymbol{c}, \boldsymbol{m}, \boldsymbol{y}\}, \{\boldsymbol{m}, \boldsymbol{y}, \boldsymbol{g}, \boldsymbol{c}\}, \{\boldsymbol{r}, \boldsymbol{g}, \boldsymbol{m}, \boldsymbol{y}\}, \{\boldsymbol{r}, \boldsymbol{g}, \boldsymbol{b}, \boldsymbol{m}\}$ or $\{\boldsymbol{c}, \boldsymbol{m}, \boldsymbol{g}, \boldsymbol{b}\}$.

3. Spectral Reflectance Estimation

Assume that the reflectance spectra of m colors $\boldsymbol{v}_i = [r_i, g_i, b_i]$ for $i = 1, \dots, m$ are given and stored in $Q = [q_{ij}]$, where q_{ij} denotes the reflectance of the jth wavelength λ_j for $j = 1, \dots, n$ in the ith color $\boldsymbol{v}_i$. Then, by the above nonsparse or sparse Neugebauer model, we have

$$Q \approx AP,$$

where $A = [a_{ik}]$ denotes an $m \times K$ coefficient matrix whose elements are given by

$$a_{i,1} = r_i(1-g_i)(1-b_i),\ \ a_{i,5} = (1-r_i)g_ib_i,$$

$$a_{i,2} = (1-r_i)g_i(1-b_i),\ \ a_{i,6} = r_i(1-g_i)b_i,$$

$$a_{i,3} = (1-r_i)(1-g_i)b_i,\ \ a_{i,7} = r_ig_i(1-b_i),$$

$$a_{i,4} = (1-r_i)(1-g_i)(1-b_i),\ \ a_{i,8} = r_ig_ib_i, \qquad (4)$$

when the Neugebauer model is used, and these are the Demichel equations [7] for the ith color $\boldsymbol{v}_i$, and the number of columns in A is the number of the Neugebauer primaries, i.e., $K = 8$. For the sparse Neugebauer model, four coefficients out of the eight ones in (4) become zero. $P = [p_{kj}]$ denotes a $K \times n$ matrix whose (k, j) element p_{kj} is the reflectance of the k th Neugebauer primary at the jth wavelength λ_j. Then we formulate the problem for estimating the spectral reflectance of the Neugebauer primaries as follows:

$$\min_P \|Q - AP\|_F^2, \qquad (5)$$

$$\text{subj. to}\ \ p_{kj} \geq 0, \qquad (6)$$

where $\|\cdot\|_F$ denotes the Frobenius norm, and the constraint (6) is for $k = 1, \dots, K$ and $j = 1, \dots, n$. Let $E(P)$ be the objective function in (5), then we have

$$E(P) = \sum_{i=1}^m \sum_{j=1}^n \left(q_{ij} - \sum_{k=1}^K a_{ik}p_{kj}\right)^2$$

$$= \sum_{i=1}^m \sum_{j=1}^n \left[q_{ij}^2 - 2q_{ij}\sum_{k=1}^K a_{ik}p_{kj} + \left(\sum_{k=1}^K a_{ik}p_{kj}\right)^2\right].$$

Since $\sum_{i=1}^m \sum_{j=1}^n q_{ij}^2$ is a constant independent of p_{kj}, the minimization of $E(P)$ is equivalent to the maximization of

$$F(P) = \sum_{i=1}^m \sum_{j=1}^n \left[q_{ij}\sum_{k=1}^K a_{ik}p_{kj} - \frac{1}{2}\left(\sum_{k=1}^K a_{ik}p_{kj}\right)^2\right]. \qquad (7)$$

Applying Jensen's inequality [10] to the second term of the right side in (7), we have

$$\left(\sum_{k=1}^K a_{ik}p_{kj}\right)^2 = \left(\sum_{k=1}^K u_{ijk}\frac{a_{ik}p_{kj}}{u_{ijk}}\right)^2$$

$$\leq \sum_{k=1}^K u_{ijk}\left(\frac{a_{ik}p_{kj}}{u_{ijk}}\right)^2 = \sum_{k=1}^K \frac{(a_{ik}p_{kj})^2}{u_{ijk}}, \qquad (8)$$

where u_{ijk} are positive constants satisfying $\sum_{k=1}^K u_{ijk} = 1$. Substituting the rightmost term in (8) into (7), we have a lower bound of $F(P)$ as

$$F_{LB}(P, U) = \sum_{i=1}^m \sum_{j=1}^n \left[q_{ij}\sum_{k=1}^K a_{ik}p_{kj} - \frac{1}{2}\sum_{k=1}^K \frac{(a_{ik}p_{kj})^2}{u_{ijk}}\right], \qquad (9)$$

where $U = [u_{ijk}]$ denotes a three-dimensional array having u_{ijk} as its (i, j, k) element.

Replacing F with F_{LB}, we formulate the problem of spectral reflectance estimation as follows:

$$\max_{P,U} F_{LB}(P, U) \qquad (10)$$

$$\text{subj. to}\ \ \sum_{k=1}^K u_{ijk} = 1,\ \ u_{ijk} > 0,\ \ p_{kj} \geq 0. \qquad (11)$$

Let $L = F_{LB} - \sum_{i=1}^m \sum_{j=1}^n \mu_{ij}(\sum_{k=1}^K u_{ijk} - 1)$ be the Lagrange function for the constrained maximization problem (10) with (11), where μ_{ij} for $i = 1, \dots, m$ and $j = 1, \dots, n$ are the Lagrange multipliers. Then we have the following condition for optimality:

$$\frac{\partial L}{\partial u_{ijk}} = \frac{1}{2}\frac{(a_{ik}p_{kj})^2}{u_{ijk}^2} - \mu_{ij} = 0,$$

from which we have

$$u_{ijk} = \frac{a_{ik}p_{kj}}{\sqrt{2\mu_{ij}}}, \qquad (12)$$

which is substituted into

$$\frac{\partial L}{\partial \mu_{ij}} = \sum_{k=1}^K u_{ijk} - 1 = 0$$

to obtain

$$\sqrt{2\mu_{ij}} = \sum_{k=1}^K a_{ik}p_{kj}. \qquad (13)$$

Substitution of (13) into (12) gives

$$u_{ijk} = \frac{a_{ik}p_{kj}}{\sum_{l=1}^K a_{il}p_{lj}}. \qquad (14)$$

Next, we solve the following optimality condition for p_{kj}:

$$\frac{\partial L}{\partial p_{kj}} = \frac{\partial F_{LB}}{\partial p_{kj}} = \sum_{i=1}^m \left(a_{ik}q_{ij} - \frac{a_{ik}^2 p_{kj}}{u_{ijk}}\right) = 0,$$

and have

$$p_{kj} = \frac{\sum_{i=1}^m a_{ik}q_{ij}}{\sum_{i=1}^m \frac{a_{ik}^2}{u_{ijk}}},$$

into which we substitute (14) to have

$$p_{kj} = \frac{\sum_{i=1}^m a_{ik}q_{ij}}{\sum_{i=1}^m a_{ik}^2 \frac{\sum_{l=1}^K a_{il}p_{lj}}{a_{ik}p_{kj}}} = p_{kj}\frac{\sum_{i=1}^m a_{ik}q_{ij}}{\sum_{i=1}^m \sum_{l=1}^K a_{ik}a_{il}p_{lj}},$$

from which we have an iterative formula for p_{kj} as follows:

$$p_{kj}^{(t+1)} = p_{kj}^{(t)}\frac{\sum_{i=1}^m a_{ik}q_{ij}}{\sum_{i=1}^m \sum_{l=1}^K a_{ik}a_{il}p_{lk}^{(t)}}, \qquad (15)$$

where t denotes the number of iterations for $t = 0, 1, \dots, T$ with the maximum iteration number T. The above formula (15) can be written in a matrix form as

$$P^{(t+1)} = P^{(t)} \odot (A^T Q) \oslash \left(A^T A P^{(t)}\right), \qquad (16)$$

where $\odot$ and $\oslash$ denote the element-wise multiplication and division, respectively, because $p_{kj}^{(t)}$, $\sum_{i=1}^{m} a_{ik}q_{ij}$ and $\sum_{i=1}^{m}\sum_{l=1}^{K} a_{ik}a_{il}p_{lk}^{(t)}$ in (15) are the (k,j) elements of matrices $P^{(t)}$, A^TQ and $A^TAP^{(t)}$ in (16), respectively.

Proposition 1 *The sequence* $\{E(P^{(t)})\}_{t=0,1,\ldots}$ *converges.*

Proof. Let

$$u_{ijk}^{(t+1)} = \frac{a_{ik}p_{kj}^{(t)}}{\sum_{l=1}^{K} a_{il}p_{lj}^{(t)}} \tag{17}$$

be an instance of u_{ijk} substituted $p_{kj}^{(t)}$ for p_{kj} in (14). Then (15) can be written as

$$p_{kj}^{(t+1)} = \frac{\sum_{i=1}^{m} a_{ik}q_{ij}}{\sum_{i=1}^{m} \frac{a_{ik}^2}{u_{ijk}^{(t+1)}}}. \tag{18}$$

Since (17) is the solution to $\partial L/\partial u_{ijk} = 0$ and $\partial L/\partial \mu_{ij} = 0$ with $P = P^{(t)}$, we find that $F_{LB}(P^{(t)}, U^{(t+1)}) \geq F_{LB}(P^{(t)}, U^{(t)})$. Similarly, since (18) is the solution to $\partial L/\partial p_{kj} = 0$ with $U = U^{(t+1)}$, we find that $F_{LB}(P^{(t+1)}, U^{(t+1)}) \geq F_{LB}(P^{(t)}, U^{(t+1)})$, i.e.,

$$F_{LB}\left(P^{(t+1)}, U^{(t+1)}\right) \geq F_{LB}\left(P^{(t)}, U^{(t)}\right). \tag{19}$$

Next, substituting (14) into (9), we have

$$F_{LB}(P,U) = \sum_{i=1}^{m}\sum_{j=1}^{n}\left[q_{ij}\sum_{k=1}^{K} a_{ik}p_{kj} - \frac{1}{2}\sum_{k=1}^{K}\frac{\left(a_{ik}p_{kj}\right)^2}{a_{ik}p_{kj}}\sum_{l=1}^{K} a_{il}p_{lj}\right]$$

$$= \sum_{i=1}^{m}\sum_{j=1}^{n}\left[q_{ij}\sum_{k=1}^{K} a_{ik}p_{kj} - \frac{1}{2}\left(\sum_{k=1}^{K} a_{ik}p_{kj}\right)^2\right] = F(P),$$

from which and (19) we have

$$F\left(P^{(t+1)}\right) \geq F\left(P^{(t)}\right). \tag{20}$$

Considering the relationship between $E(P)$ and $F(P)$ given by $E(P) = \sum_{i=1}^{m}\sum_{j=1}^{n} q_{ij}^2 - 2F(P)$, we have the following inequality from (20):

$$E\left(P^{(t+1)}\right) \leq E\left(P^{(t)}\right),$$

that is, the sequence $\{E(P^{(t)})\}_{t=0,1,\ldots}$ is monotonically nonincreasing, and is also bounded below by 0, i.e., $E(P) \geq 0$ by definition. Consequently, the sequence $\{E(P^{(t)})\}_{t=0,1,\ldots}$ converges to a number greater than or equal to 0. ■

4. Color Reproduction

Let $\bar{x}(\lambda)$, $\bar{y}(\lambda)$ and $\bar{z}(\lambda)$ be the color matching functions for the 1931 2° CIE standard colorimetric observer [], where λ denotes a wavelength, and let $\bar{s}(\lambda)$ be the relative spectral power distribution of an illuminant. Then the CIE XYZ tristimulus values of a spectral reflectance $\bar{p}(\lambda)$ are given by

$$\bar{X} = \bar{k}\int \bar{s}(\lambda)\bar{p}(\lambda)\bar{x}(\lambda)d\lambda, \tag{21}$$

$$\bar{Y} = \bar{k}\int \bar{s}(\lambda)\bar{p}(\lambda)\bar{y}(\lambda)d\lambda, \tag{22}$$

$$\bar{Z} = \bar{k}\int \bar{s}(\lambda)\bar{p}(\lambda)\bar{z}(\lambda)d\lambda, \tag{23}$$

where $\bar{k}$ is a normalizing factor given by $\bar{k} = \int \bar{s}(\lambda)\bar{y}(\lambda)d\lambda$. We use the trapezoidal rule for computing the above integrals. The transformation from the CIE XYZ tristimulus values into the standard RGB (sRGB) is as follows. First, we compute the linear transformation

$$\begin{bmatrix} R \\ G \\ B \end{bmatrix} = \begin{bmatrix} 3.2406 & -1.5372 & -0.4986 \\ -0.9689 & 1.8758 & 0.0415 \\ 0.0557 & -0.2040 & 1.0570 \end{bmatrix}\begin{bmatrix} \bar{X} \\ \bar{Y} \\ \bar{Z} \end{bmatrix},$$

where note that similar but slightly different transformation matrices are also proposed and used in practice.

Next, a nonlinear luminance encoding, which is also known as gamma encoding, is calculated to obtain sRGB values as follows:

$$R_{\mathrm{sRGB}} = \begin{cases} 1.055R^{\frac{1}{2.4}} - 0.055, & \text{if } R > 0.0031308, \\ 12.92R, & \text{if } R \leq 0.0031308; \end{cases}$$

$$G_{\mathrm{sRGB}} = \begin{cases} 1.055G^{\frac{1}{2.4}} - 0.055, & \text{if } G > 0.0031308, \\ 12.92G, & \text{if } G \leq 0.0031308; \end{cases}$$

$$B_{\mathrm{sRGB}} = \begin{cases} 1.055B^{\frac{1}{2.4}} - 0.055, & \text{if } B > 0.0031308, \\ 12.92B, & \text{if } B \leq 0.0031308. \end{cases}$$

Consequently, we obtain the sRGB values $(R_{\mathrm{sRGB}}, G_{\mathrm{sRGB}}, B_{\mathrm{sRGB}})$ transformed from the CIE XYZ tristimulus values $(\bar{X}, \bar{Y}, \bar{Z})$.

5. Experimental Results

In this section, we show experimental results of spectral reflectance estimation and color reproduction. Figure 1 shows the Macbeth ColorChecker (MCC), of which the manufacturer's sRGB D65 color values are also given on the Internet (ColorChecker, Wikipedia. URL: https://en.wikipedia.org/wiki/ColorChecker). Although the 24 squares of painted samples in MCC are not materials but colorants, the proposed method can handle both of them equally because the Neugebauer model is valid for general colorimetric data as shown in Subsection 2.1.

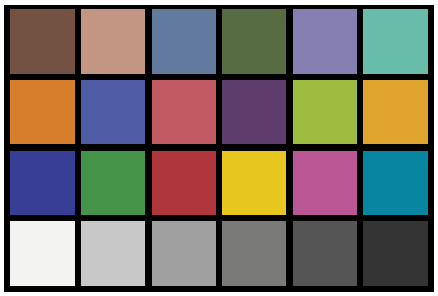


Figure 1 Macbeth ColorChecker.

The spectral reflectance data of the 24 colors in MCC are shown in Figure 2, where the line colors correspond to the MCC colors. These data are provided by the Munsell Color Science Laboratory (URL: http://www.cis.rit.edu/research/mcsl2/online/cie.php).

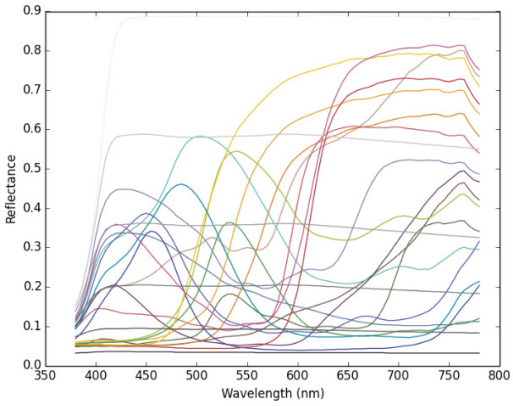


Figure 2 Reflectance spectra of Macbeth ColorChecker colors.

In the following experiments, we store the above data in a matrix $Q \in \mathbb{R}^{m \times n}$ where m denotes the number of colors in MCC, i.e., $m = 24$, and n denotes the number of spectral data in which the wavelengths between 380 nm (nanometer) and 780 nm are sampled at 5 nm intervals, i.e., $n = 81$.

Figure 3 shows the change in the value of the objective function $E(P)$ of the proposed method with the Neugebauer model described in Section 2.1, where the vertical and horizontal axes denote the objective function value and the number of iterations t with the maximum iteration number $T = 50$, respectively. The objective function value monotonically decreases with the increase in the number of iterations, that supports the claim in Proposition 1 experimentally.

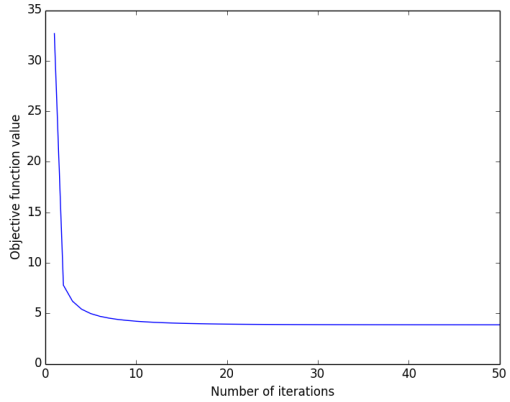


Figure 3 Change in objective function value.

Figure 4 shows the estimated spectral reflectance of the Neugebauer primaries by the sequential coordinate-wise algorithm (SCA) [6,8], where the vertical and horizontal axes denote the reflectance and wavelength, respectively, and the colors of solid lines correspond to the primary colors, and the broken line denotes white color. On the other hand, Figure 5 shows the estimated spectral reflectance by the proposed method with the Neugebauer model, and is similar to Figure 4, although the algorithms are different from each other. We measured CPU time for estimating the spectral reflectance on an Intel Core-i3-equipped PC using Portable Python 2.7.6.1. Then SCA and the proposed method required 1.12 and 0.028 seconds, respectively, from which we conclude that the proposed method is computationally more efficient than SCA.

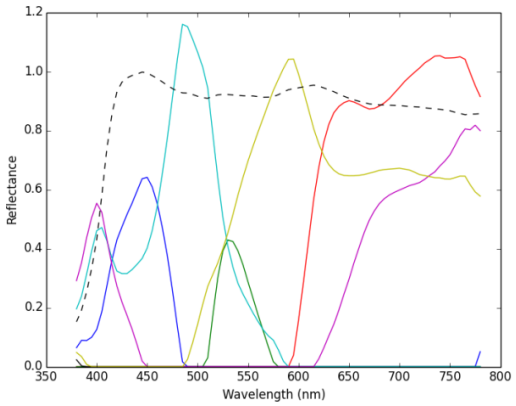


Figure 4 Estimated spectral reflectance by SCA [6,8].

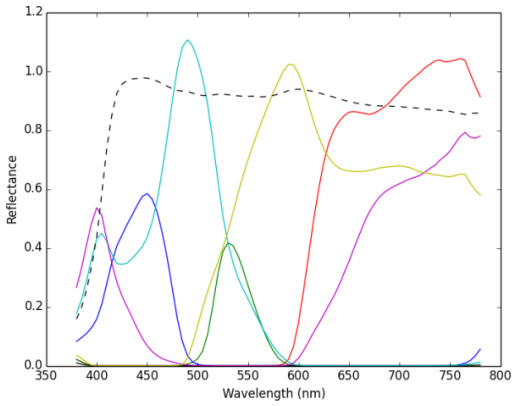


Figure 5 Estimated spectral reflectance by the proposed method with Neugebauer model.

Next, we compare sparse Neugebauer model with nonsparse or conventional Neugebauer model in the proposed method. Figure 6 shows the coefficients in two Neugebauer models for the upper left color in MCC in Figure 1, where the vertical and horizontal axes denote the value of each coefficient and the Neugebauer primaries, respectively. the blue and red bars denote conventional (nonsparse) and sparse Neugebauer models, respectively. In the conventional model (blue bars), all Neugebauer primaries have positive values. On the other hand, in sparse one (red bars), only four colors (R, G, B and M) have positive values, and the other coefficients are strictly zero.

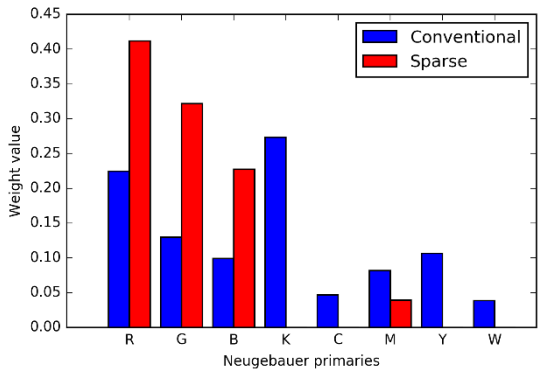


Figure 6 Coefficients in Neugebauer models.

Figure 7 shows the estimated spectral reflectance of the Neugebauer primaries by the proposed method with the sparse Neugebauer model. Each curve in Figure 7 has greater variance than that in Figure 5 or 4.

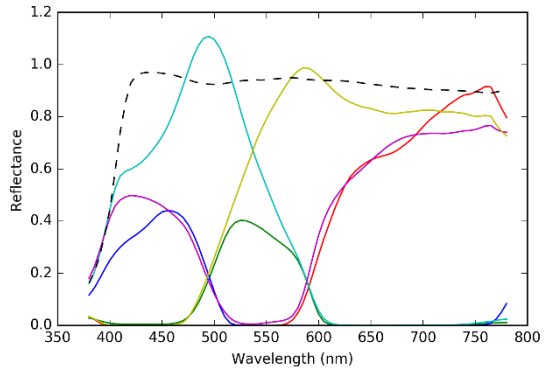


Figure 7 Estimated spectral reflectance by the proposed method with sparse Neugebauer model.

From Figure 8 to Figure 11, we show the reflectance spectra of the colors in MCC from the first to fourth rows, respectively (here we divide the 24 MCC colors into 4 figures each of which includes the spectral reflectance curves of 6 colors for visibility), where solid lines show the original reflectance spectra, and broken lines show their estimated ones by the proposed method with the Neugebauer model.

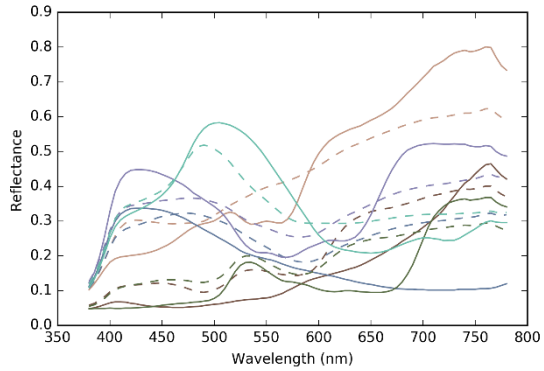


Figure 8 The original reflectance spectra of the colors in the first row of MCC (solid lines) and their estimated ones by the proposed method with Neugebauer model (broken lines).

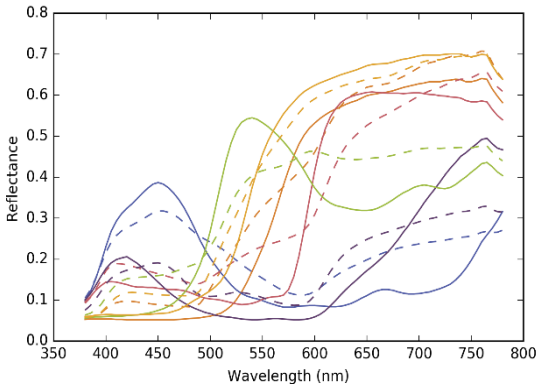


Figure 9 The original reflectance spectra of the colors in the second row of MCC (solid lines) and their estimated ones by the proposed method with Neugebauer model (broken lines).

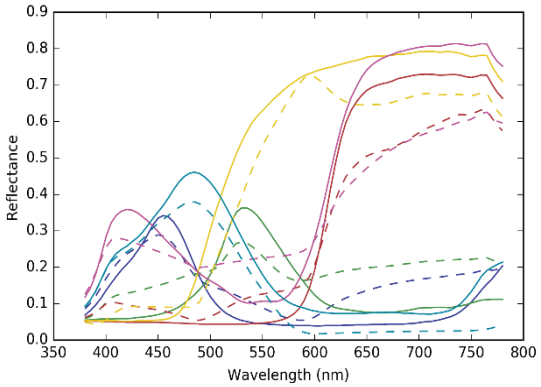


Figure 10 The original reflectance spectra of the colors in the third row of MCC (solid lines) and their estimated ones by the proposed method with Neugebauer model (broken lines).

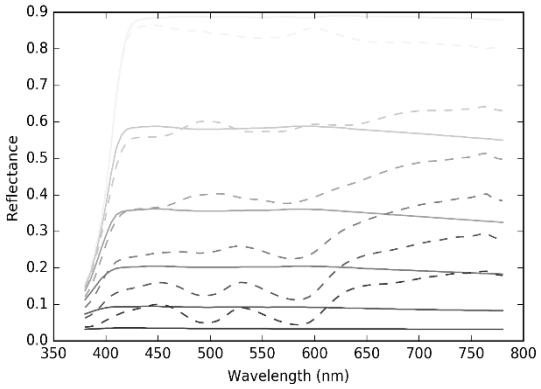


Figure 11 The original reflectance spectra of the colors in the fourth row of MCC (solid lines) and their estimated ones by the proposed method with Neugebauer model (broken lines).

On the other hand, from Figure 12 to Figure 15, we show the estimated reflectance spectra by the proposed method with sparse Neugebauer model, as well as Figures 8-11 with the Neugebauer model.

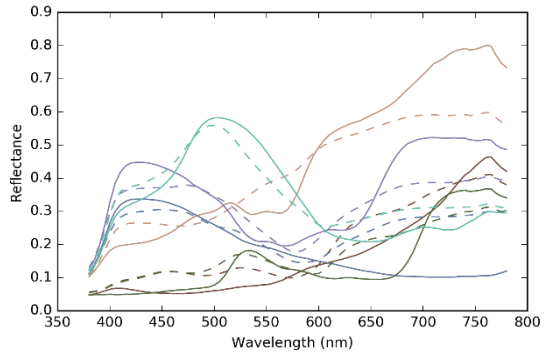


Figure 12 The original reflectance spectra of the colors in the first row of MCC (solid lines) and their estimated ones by the proposed method with sparse Neugebauer model (broken lines).

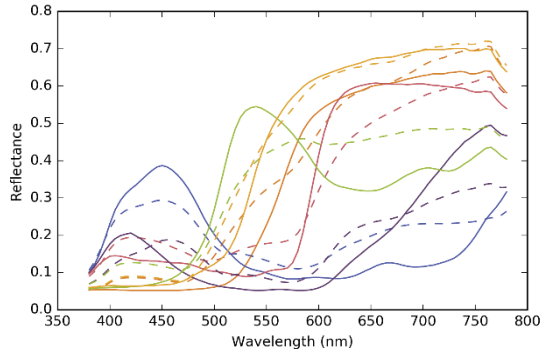


Figure 13 The original reflectance spectra of the colors in the second row of MCC (solid lines) and their estimated ones by the proposed method with sparse Neugebauer model (broken lines).

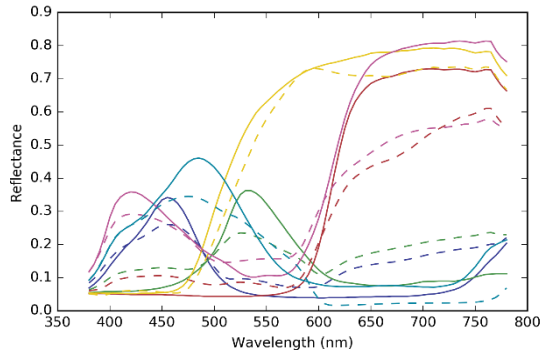


Figure 14 The original reflectance spectra of the colors in the third row of MCC (solid lines) and their estimated ones by the proposed method with sparse Neugebauer model (broken lines).

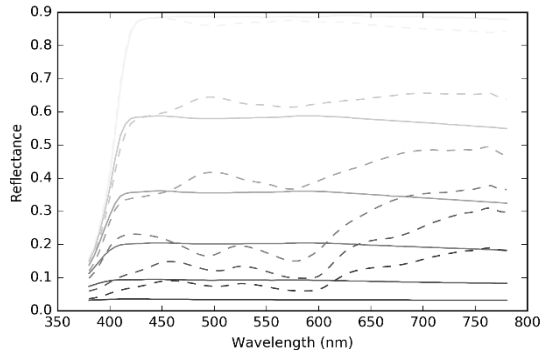


Figure 15 The original reflectance spectra of the colors in the fourth row of MCC (solid lines) and their estimated ones by the proposed method with sparse Neugebauer model (broken lines).

We evaluate the mean squared error between the original and estimated reflectance spectra given by

$$\mathrm{MSE}_i = \frac{1}{n}\sum_{j=1}^{n}\left(q_{ij} - \sum_{k=1}^{K} a_{ik}p_{kj}\right)^2$$

for the ith color in MCC. Figure 16 shows the result where the vertical and horizontal axes denote the mean squared error and the identification number of MCC colors, respectively, and the blue and red bars denote the conventional nonsparse and the sparse Neugebauer models, respectively. In 18 colors among all 24 colors, the sparse Neugebauer model achieved smaller MSE_i than the conventional one. The average of MSE_i for $i = 1, \ldots, m$ given by

$$\mathrm{MSE} = \frac{1}{m}\sum_{i=1}^{m}\mathrm{MSE}_i$$

are 7.66×10^{-3} for the conventional Neugebauer model and 6.83×10^{-3} for the sparse one.

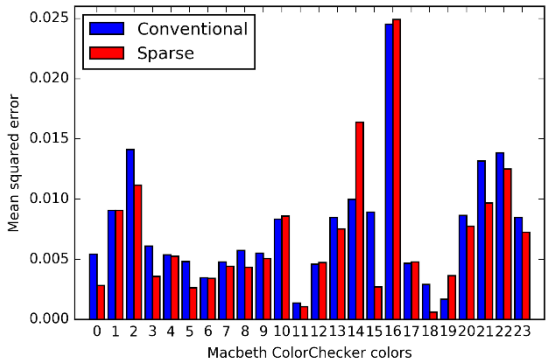


Figure 16 Mean squared error of estimated reflectance spectra.

Next, we also evaluate the same measure using the leave-one-out cross-validation [11], where the spectral reflectance of the ith color in MCC is removed from Q to evaluate MSE_i. Figure 17 shows the result, where the sparse Neugebauer model (red bars) achieved smaller MSE_i in 15 colors among all 24 colors than the conventional Neugebauer model (blue bars). The values of MSE are 1.37×10^{-2} for the conventional Neugebauer model and 1.15×10^{-2} for the sparse one.

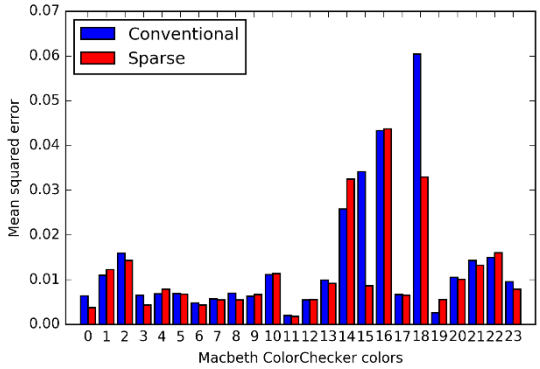


Figure 17 Mean squared error of estimated reflectance spectra evaluated by the leave-one-out cross-validation.

Finally, we show the results of color reproduction from the estimated reflectance spectra. Figure 18 shows the original Neugebauer primaries: red, green, blue, black, cyan, magenta, yellow and white.

Figure 18 Neugebauer primaries.

Figure 19 shows the reproduced Neugebauer primaries with the conventional Neugebauer model, in which blue is wrongly transformed into magenta.

Figure 19 Reproduced Neugebauer primaries from the estimated reflectance spectra with Neugebauer model.

On the other hand, the reproduced Neugebauer primaries with the sparse Neugebauer model is shown in Figure 20, where blue is correctly reproduced as well as the other colors.

Figure 20 Reproduced Neugebauer primaries from the estimated reflectance spectra with sparse Neugebauer model.

Figures 21 and 22 show the reproduced MCC colors with the conventional and sparse Neugebauer models, respectively. The reproduced colors are similar to the original ones in Figure 1. Additionally, note that the proposed method can be applied to not only the CMYK-printed colors like MCC but also more general tristimulus values including material colors.

Figure 21 Reproduced MCC colors with Neugebauer model.

Figure 22 Reproduced MCC colors with sparse Neugebauer model.

We evaluate the mean absolute error between the original and reproduced MCC colors given by

$$\mathrm{MAE}_i = \frac{|r_i - \tilde{r}_i| + |g_i - \tilde{g}_i| + |b_i - \tilde{b}_i|}{3}$$

for the ith color in MCC, where $\tilde{r}_i$, $\tilde{g}_i$ and $\tilde{b}_i$ denote the reproduced sRGB values from the estimated reflectance spectra. Figure 23 shows the result, where the sparse Neugebauer model (red bars) achieved smaller MAE_i in 20 colors among all 24 colors than the conventional Neugebauer model. The average of MAE_i for $i = 1, \ldots, m$ given by

$$\mathrm{MAE} = \frac{1}{m}\sum_{i=1}^{m} \mathrm{MAE}_i$$

are 17.65 for the conventional Neugebauer model and 13.81 for the sparse one.

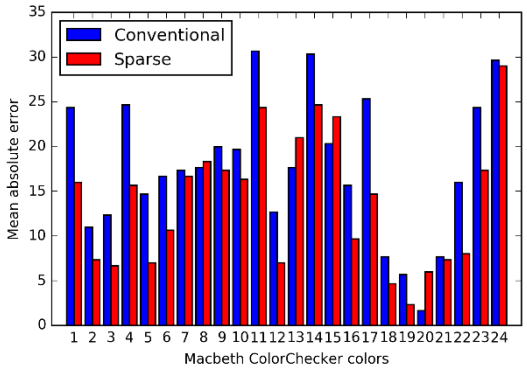


Figure 23 Mean absolute error of reproduced MCC colors.

Consequently, the sparse Neugebauer model improved the performance of both spectral reflectance estimation and color reproduction from the estimated spectral reflectance compared with the conventional Neugebauer model.

6. Conclusion

In this paper, we proposed a method for estimating spectral reflectance based on the Neugebauer model and its sparse version. The sparse Neugebauer model improved the estimation performance compared with the conventional Neugebauer model. We also proved the convergence property of the proposed algorithm for spectral reflectance estimation. Additionally, we reproduced the colors from the estimated spectral reflectance, in which the sparse Neugebauer model also improved the performance compared with the conventional one.

For future work we would like to improve the performance of both spectral reflectance estimation and color reproduction by robustifying the proposed method, and compare the proposed method with other related methods including learning-based methods. We are also planning an application of the proposed method to the design of spectral reflectance for controlled color reproduction under various illuminations.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number JP16H03019.

References

[1] S. Westland, Computational Colour Science Using MATLAB, Wiley, 2004.

[2] G. Sharma, S. Wang, "Spectrum recovery from colorimetric data for color reproductions," in Proc. SPIE 4663, Color Imaging: Device-Independent Color, Color Hardcopy, and Applications VII, San Jose, CA, 2002.

[3] T. Harifi, S. H. Amirshahi , F. Agahian, "Recovery of reflectance spectra from colorimetric data using principal component analysis embedded regression technique," Opt. Rev., **15**(6), 302–308, 2008.

[4] M. M. Amiri, S. H. Amirshahi, "A step by step recovery of spectral data from colorimetric information," J. Opt., 44(4), 373–383, 2015.

[5] G. Wu, X. Shen, Z. Liu, S. Yang, M. Zhu, "Reflectance spectra recovery from tristimulus values by extraction of color feature match," Opt. Quant. Electron., **48**(1), 48:64, 2016.

[6] V. Franc, V. Hlaváč, M. Navara, "Sequential coordinate-wise algorithm for the non-negative least squares problem," in Gagalowicz A., Philips W. (eds) Computer Analysis of Images and Patterns. CAIP 2005. Lecture Notes in Computer Science, vol 3691. Springer, Berlin, Heidelberg, 2005.

[7] E. Reinhard, E. A. Khan, A. O. Akyuz, G. Johnson, Color Imaging: Fundamentals and Applications, A K Peters/CRC Press, 2008.

[8] K. Inoue, K. Hara, K. Urahama, "Reflectance spectra recovery with non-negativity constraints," in 2016 International Symposium on Intelligent Signal Processing and Communication Systems, Phuket, Thailand, 2016.

[9] D. Shaked, N. Arad, , A. Fitzhugh, I. Sobel, "Color diffusion: Error-diffusion for color halftones," HP Labs Technical Reports, HPL-96-128R1, 1999.

[10] S. Boyd, L. Vandenberghe, Convex Optimization, Cambridge University Press, 2004.

[11] S. Arlot, A. Celisse, "A survey of cross-validation procedures for model selection," Statistics Surveys, **4**, 40–79, 2010.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 967-973 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

3D Senor-based Library Navigation System

Abdel-Mehsen Ahmad*, Zouhair Bazzal, Roba Al Majzoub, Ola Charanek

Department of Computer and Communication Engineering, Lebanese International University, Lebanon

ARTICLE INFO

Article history:
Received: 15 May, 2017
Accepted: 15 June, 2017
Online: 30 June, 2017

Keywords:
Gesture-Based Control
Kinect
Processing Software

ABSTRACT

The discussed system in this paper uses the Kinect's 3D input (acquired through IR projector and Camera) to detect gestures and then based on these gestures perform certain tasks, and display certain results on an output screen. The programming language used here is the Processing open source software and a database was implemented to manage the data required within the system. The deigned system was created for a library but the system's design is broad enough to allow its utilization in many more domains. Implementation, system screenshots, simulation and results are discussed along with testing the system in different condition of light and distance. Some simple additions have been added to this version of the program to facilitate the interaction between the user and the system further.

1. Introduction

The advancement of control systems that has taken large leaps in the past decade has raised human-machine interfaces to a whole new level where they became more and more remote from the user's body like voice control, eye control and body and hand gesture control which have changed the concept of human-machine interface from the typical and complex concept to become more natural and human-like, also known as Natural User Interfaces or NUI . Most of the new systems are trying to enhance the user interface further more so that a larger number of people can be able to use the new technology with less effort and time consumption. These types of interfaces allow users to bypass the limits of age, education and sometimes physical disabilities to have a better control over technological systems which have become an undisputable need for us on daily basis.

The system to be discussed implements gesture recognition to control the system where users interface with it remotely using simple gestures to navigate the system. With all the possibilities available for gesture detection, the most reliable method is 3D imaging, where the gesture is no longer dependent on the 2D properties of the hands, like color, shape etc. to determine and recognize the gestures of the user, which is very expensive in terms of processing and may lead to delays within the system when the user is expecting real time feedback.

Although the system was designed for a library however its possible implementations can go way beyond the walls of libraries into museums, hospitals, malls and many more applications, it can also aid elderly or people with special needs control many things from electrical appliances to electrical doors, applications etc... . It will follow the hand gestures of the user to allow him to explore different books with different genres and authors available within the library, where the books will be classified into topics, and within each topic related books available will be displayed. Further information about each book (including author's name, ISBN, a short summary about the contents of the book and its location within the library) can be displayed to the user. This paper demonstrates the design and implementation of a 3D gesture-based control system that was previously discussed in 2016 2nd International Conference on Open Source Software Computing (OSSCOM) [1].

2. Similar Systems

2.1. Maintaining the Integrity of the Specifications (Heading 2)

Similar control systems were created using different techniques. A "Wireless Gesture Controlled Robotic Arm with Vision" was presented in [2] using accelerometers that allowed the user to control the robotic arm wirelessly, by using 3-axis DOF's (Degree of Freedom) accelerometers controlled via RF signals. The robotic arm is mounted on a movable platform to signify the hand and leg of the user, where the robotic arm correspond to the user's arm and the platform corresponds to the user's leg. Accelerometers are fixed on the user's hand and leg to capture their gestures and movements allowing the robotic arm and platform to move accordingly.

*Corresponding Author: Abdel-Mehsen Ahmad, Department of Computer and Communication Engineering, Lebanese International University, Lebanon
Email: abdelmehsen.ahmad@liu.edu.lb

Another gesture controlled user interface called "Open Gesture" was proposed in [3]. A web cam is used with an open source program to implement the system in smart TVs for almost all users. Although the system can be used by most people, still it faces some drawbacks; the main tool for capturing gestures is a web cam, meaning that to identify gestures, image processing is required, and that is computationally expensive and will not allow the system to receive the same quality of information obtained by 3D scanning. Another drawback is the fact that the web cam is light-dependent and will not work in all lighting conditions, where dim lit and unlit places create a challenge for the system to identify the gestures performed. The system also assumes that all users are of a certain educational background, and that is not true, so the system is limited to these people only. Even though the system we propose in this paper is for a library, however the gesture control can be performed over anything and almost everywhere we want with some small changes.

Another approach was presented in [4], which uses the WiiMote from Nintendo to control smart homes (appliances, lights, windows…), where the 3D accelerometer will be used to detect the gestures and a classification framework composed of machine-learning algorithms like Neural Networks and Support Vector Machine (SVM) will be used for gesture recognition classification and training. Although this system has a high level of accuracy for gesture recognition due to the use of the classification framework, a setback lies in the use of the WiiMote, where it restricts the gesture recognition for the user to only the time when the controller is held by hand, which after long periods may result in inconveniences for the users especially the elderly.

Yet another gesture controlled system was presented by [5], where Kinect's depth imaging was combined with traditional Camshift tracking algorithm along with HMM (Hidden Markov Models) for dynamic gesture classification to control the motion of a bi-handed Cyton 14D-2G. A Client/Server structured robot tele-operation application system is developed, which provides a favorable function of remotely controlling a dual-arm robot by gestural commands.

3. Proposed System

3.1. Design Specifications

From the many choices of NUI we chose to utilize hand gestures because they are familiar and universal to all users where even small children learn to use gestures before they even learn to talk, and so this supports the fact that this system can be compatible with all backgrounds and ages of users.

Gesture recognition was accomplished with the help of the Kinect, which is a line of motion sensing input devices that includes a 2D camera along with a 3D scanner system called Light Coding which employs a variant of image-based 3D reconstruction that requires the work of an IR projector and IR camera.

Some of the great advantages of using the Kinect sensor are the fact that its price is affordable and that it can catch great depth images in different lighting conditions like dim lit or unlit places as well as places with good lighting. This is a property not available in any of the other 2D image capturing sensors. Another important factors is the fact that Kinect uses 3D imaging instead of 2D imaging, which is independent of the light (as previously mentioned) and thus does not require image processing and resolves the issue of computationally expensive algorithm we would have been forced to use, making the system more robust and increasing gesture detection therefore increasing the speed and dependability of the system's response to the gestures.

The Kinect will be connected to a computer with a Processing program designed to detect gestures and respond accordingly to them. Since we only needed the 3D imaging of the Kinect, 2D images were not used and thus the real time streaming of the surrounding environment was done in grey scale images of the Kinect sensor.

The system design targets the majority of people, meaning that users' age, physical abilities and educational backgrounds will vary, so we had to take into consideration the controlling gestures since they had to be performable and known to all users with ease and without resorting to difficult physical actions that on the long term may strain the users. The system was only programmed to respond to two types of gestures: hand wave and hand click, where the wave gesture was used to initiate the program and then when inside the program it was used to go back to the previous screen/menu while the click gesture was used as a selection gesture.

The programming language used was Processing, which is a free open source programming language and runs on Mac, Windows, and GNU/Linux platforms. Processing has many libraries related to Kinect, and because it was easy to learn and because it is an object oriented programming language, it allowed us to have more flexibility when dealing with Kinect input and manipulating information and output.

Data were stored in a database that can be updated according to the needs of the system where the program fetches required information from the database depending on the type of gesture detected. If the gesture is a click for example, then the system will fetch all the information related to the object clicked on by the user. And so for that purpose MySQL database was used to store the information about the books and the other information to be displayed within the program.

The following figure (Fig.1) represents the interaction between hardware components of the system designed:

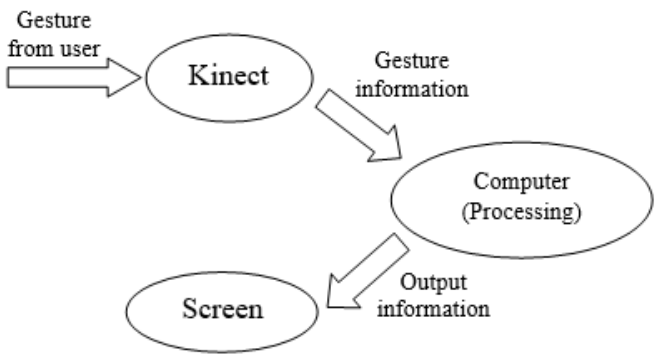


Figure 1 Hardware components of the system

4. Implementation

The system developed is an event-response application, where the user's gestures are the events that change the state of the system (the displayed information and the navigation through the system) to trigger the response to these events allowing him/her to control the system.

To allow the system to understand gestures we needed a gesture recognition algorithm to start with. So we decided to use a pre-made gesture recognition program that was available to save time and effort and to focus on the main idea of the project.

Although two approaches were available for us to control the system, one was 2 dimensional (2D) and the other was 3 dimensional (3D) depending on the algorithm used for the detection of gestures, however the 2D approach proved to have poor reliability due to the delays it showed within the system and the program's inability to follow the users hand and fingertips proficiently, thus it was halted and the 3D approach was adopted due to its good response to the users hand movements and gestures adding to the reliability of the system.

4.1. 2D Approach

The Finger Tracker library was our first alternate for gesture detection where it detects the tips of the fingers of the user. It does that through slicing the space detected by the Kinect into layers of 2D surfaces and then determining the fingertips through scanning each 2D surface and finding "inflections in the curvature of the contour" then drawing red ellipses in their positions, as seen in Figure 2.

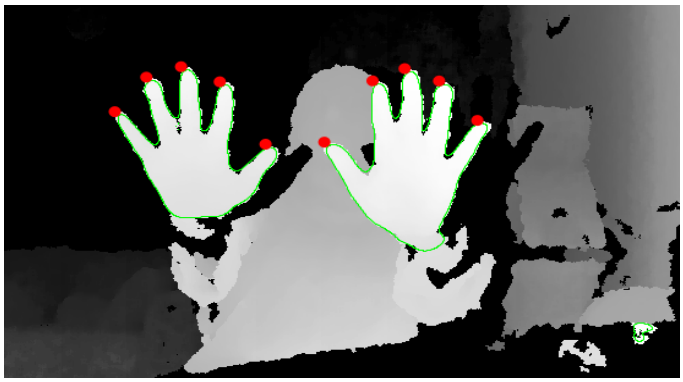


Figure 2 Fingers tracked using FingeTracker library

Finger's gestures and movements were used as the events that control the system. So we first eliminated the depth image (the grey-scaled 3D-image captured by the Kinect seen in Figure 2) and only drew the contour of the hand along with the fingertips (fingertips displayed in red and contour in green in Figure 2) with a rate approximate to 30 fps (frames per second), which is the frame rate of the Kinect, on a black background, and like that we were able to observe the real-time motion of our fingertips on screen. Next, fingertips previously drawn as 2d images were replaced by special 2D graphics known as "Sprites" as displayed in Figure 3. Sprites are 2D images that have special properties with boundaries and libraries that allow us to detect the collision between one sprite and the other, set their properties, motion, speeds and so much more. The library used here is called Sprites library, and allows us to control all the previously mentioned properties of sprites.

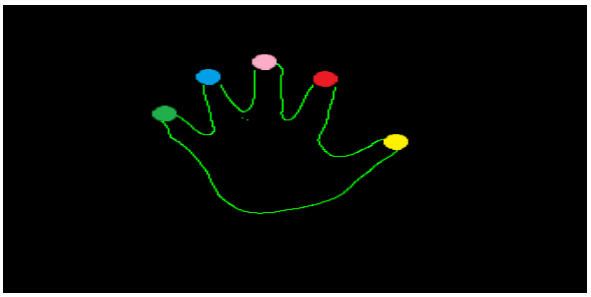


Figure 3 Sprites inserted on the fingertips

Images of books were inserted as sprites as well (see Fig.4), so to be able to detect collision between them and the fingertips sprites and the books sprites, an algorithm was created so that certain actions can be executed upon detection.

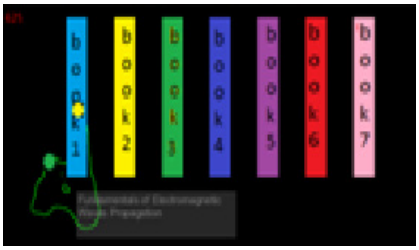


Figure 4 Book information displayed upon the detection of collision between a fingertip sprite and a book sprite

Two events were defined to control the system and change its states, a pointing gesture and a pinching gesture.A loop was used to iterate over the seven pictures given, then a condition checks for any collision between any book and a predefined fingertip (detected when the user is pointing at the book) which is shown using the yellow sprite in Figure 2. If the condition is satisfied, the name of the book that the finger sprite collided with will be printed on screen (see Figure3). Then the loop iterates over the books again to check for collisions. To select the required book two fingertips were used, such that if a pinching gesture was performed and the distance between the boundaries of the finger sprites was smaller than a certain threshold and the collision between these sprites and the book sprite was detected information about the selected book will be displayed, thus the system would transit from the initial state where the books are displayed to either state depending on the detected event; a pointing gesture results in displaying the name and a pinching gesture results in displaying the information. The following state chart displays how the system transits between states based on the events created by the user. (Figure 5)

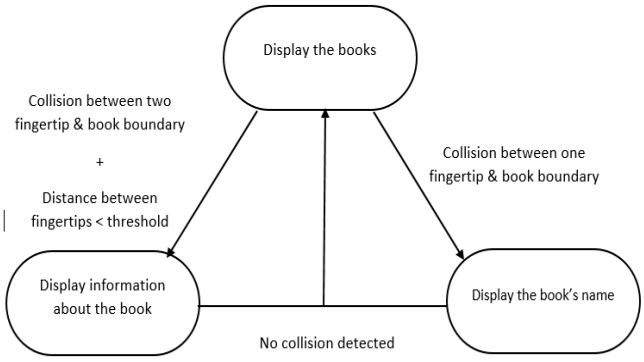


Figure 5State chart that shows the transition of the system between states based on events created by user

Figure 6 displays how the software components in this approach work with each other:

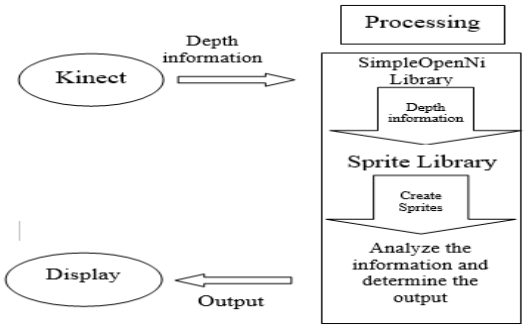


Figure 3 Software components of 2D approach

4.2. 3D Approach

The other available option for gesture implementation and detection was within the SimpleOpenNi library which is the "Hands" sample code that allows us to detect three hand gestures: Wave, Click, and Hand Raise which will be used as attributes for events to change the state of the system (and therefore change the displayed output). Upon detecting any of the three gestures, the detected hand is given an ID number between 1 and 7 and a unique colour is assigned to it where a circle of that colour is drawn in the upper middle part of the user's palm and displayed on the screen, as can be seen in Figure 7.

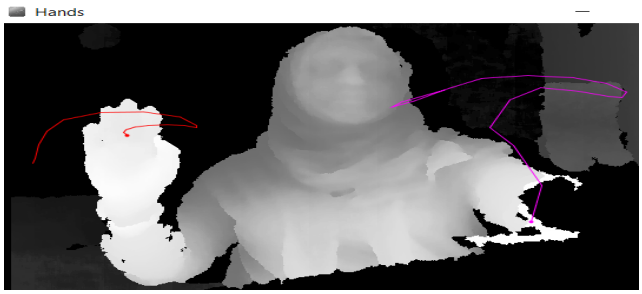


Figure 4 How hand code displays detected hands after gestures

Each gesture is assigned an integer number from 1 to 3, where Wave = 0, Click = 1, and Hand_Riase = 2. Because our system only needed the presence of two gestures (one for selection and another to go back to the previous menu) the click gesture was used for selection and the wave gesture was used for going back. Notice that the system focuses on detection of hands only, thus the user's profile, age, physical abilities and educational background are of no importance, as long as the user can use his/her hands effectively in front of the Kinect.

Then the system was programmed to respond to a certain gesture in a certain position where the position is the other attribute of the events that allow us to change the state of the system and therefore navigate forth and back through them, so that if the user wanted to select a certain topic or book all he had to do was to click on the image of that topic or book and after getting the required information, a simple wave of the hand allows the user to go back to the topics menu or to the collection of displayed books, allowing him to navigate back and forth within the system.

Meaning that the events here have two attributes; one is the gesture type and the other is the position where the gesture is performed, so upon the completion of the event when satisfying the conditions, the system transits to another state where the information matching the completed event are displayed.

A small description of the system along with some screenshots will be displayed for better understanding of how the system works.

The user faces a Kinect sensor and a Screen that will display to him the program he will interact with. The screen displays the user's depth image, this shows that the program still isn't functional till the user waves his/her hand, then a "START" button will be displayed in the middle of the screen (as seen in Figure 8) announcing that the program is ready and waiting for the click of the user.

Figure 5 After a wave of the hand the start button is displayed

Next the user will "click" within the boundaries of the start button opening the main page of the system where empty shelves and the topics of the available books are displayed on the screen. Figure 9 displays the state chart of the system, where each event allows the user to go back or forth within the system, changing the state and therefore the displayed output of the system.

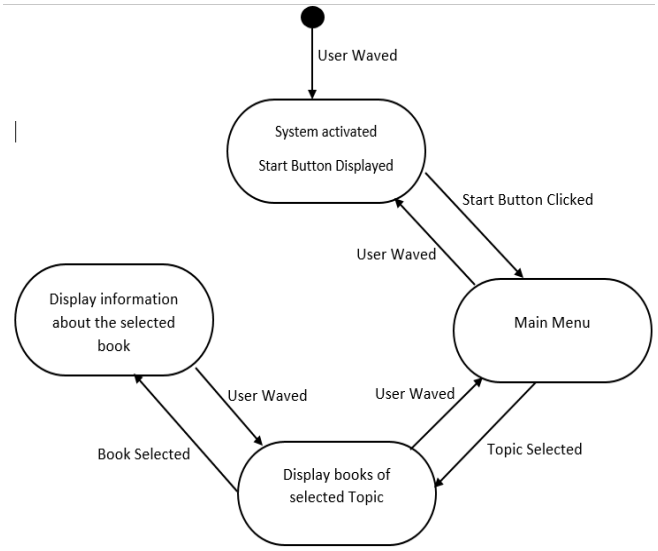


Figure 6 State chart displaying the overall states of the system and how they changed based on the user's events (click and wave)

As the user clicks the start button, a new screen will appear with the topics of the available books displayed on the left side of the screen, and empty shelves in the middle of the screen, as seen in Figure 10. Here the condition was to perform the click gesture within the boundary range of the start button.

Figure 7 Main screen displayed after clicking the start button

Next the user can select any of the displayed topics he/she is interested in through a click gesture on the required topic.

As a response to this gesture pictures of the available books concerning this topic will be displayed on the empty shelves as seen in Fig. 11.

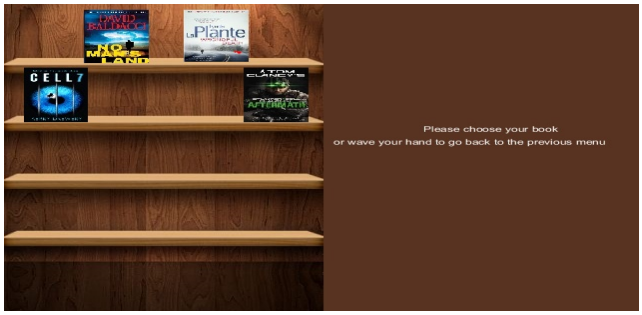


Figure 8 Books displayed on the shelves

With a simple click on the book a larger image of it is displayed along with information about it (author, ISBN, summary, position within the library…) on the right hand side of the screen as seen in Figure 12.

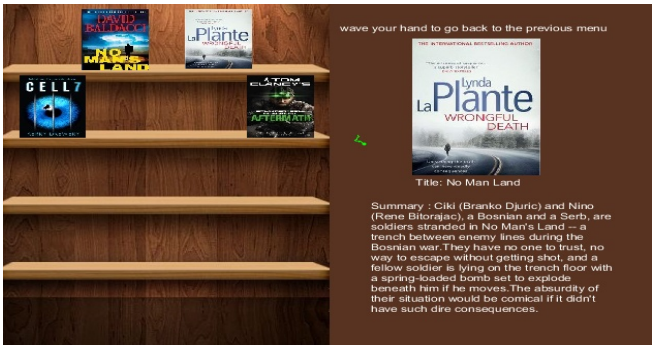


Figure 9Information displayed about the selected book

Going back to the previous menu can be done with a simple wave of the hand. The same gesture can take us back one step wherever we are within the program, as seen in Figure 11.

Figure 13 displays the software components of the 3D approach of the system.

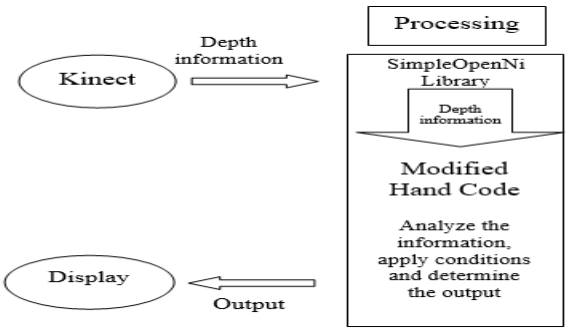


Figure 10 Software components of the 3D approach

4.3. Database Implementation

MySQL database was created to store the attributes and information and picture paths of the books to be displayed on screen, and information were retrieved from the database, using queries triggered by events created by the user.

For the system to retrieve the data a function was created which returned the desired information as a string array. The function then

- Performs a query: Selects all rows from the specified table
- Iterates over all of the selected rows
- Gets the "attribute" value of each entry
- Increments the loop to hop to the next position in the array thus retrieving the paths of pictures and books' information (summary, author…)

This implementation makes this control system quite broad where if we wanted to change the application field of the system from a library management system to another different domain, then the users must only supply the database with the new information to update the tables, so the system can then perform different actions (display images, control devices etc…) thus extending the range of the system's application field.

5. Simulations and Results

5.1. Test Cases and Acceptance Criteria

To consider the system as acceptable for interfacing it should satisfy the following criteria:

- Be able to detect gestures without too much effort from the user, and without any delays.
- Be well organized so that the user wouldn't be confused about the information or about how to select items.
- Use gestures that are easy and familiar to people so that all users would be able to perform them without effort.
- Limit the range of gesture performance such that it wouldn't be too wide, so that the hand movements wouldn't strain the user, and they must be swift and not involve too much movement.
- Have a high percentage of accuracy and fidelity.

The tests were performed on both approaches mentioned before, the 2D and 3D approaches.

5.2. Simulation Results

Although the 2D approach involved finger gestures that have become quite common among most people due to their use in touch interface (smartphones and tablets etc..), the continuous flickering of the fingertip sprites and the fact that the detected IDs of the finger sprites were inconsistent made collision detection much harder reducing the accuracy of the system significantly, leading to considering the approach as invalid.

The 3D approach underwent many tests to validate its efficiency, fidelity, real time response to users' gestures and the response of users towards the used gestures. The results proved the system to be well compatible with the criteria mentioned above. The range of gesture performance has also been detected to be good and reachable by the majority of the users who have performed this test which included users of different ages and different educational levels. The gestures used, namely the waving and clicking were very well detected and responded to, on condition that the user faces the Kinect head on without any tilting and without any obstacles and that the Kinect would remain stationary after calibration. The accuracy has increased significantly from the first prototype, and fidelity was very good, where users were able to go forth and back while navigating within the system.

The figures previously displayed were screenshots captured during the simulation of both approaches, where each of them displays exactly how the system responds to the actions of the user.

Statistics were made based on new review from test users of the system, where two surveys was conducted concerning the concept of the system as a gesture based control system, whether users had previously used similar system and if they would buy such a system themselves (see Figure 14). The survey was done on social media and on field with test users.

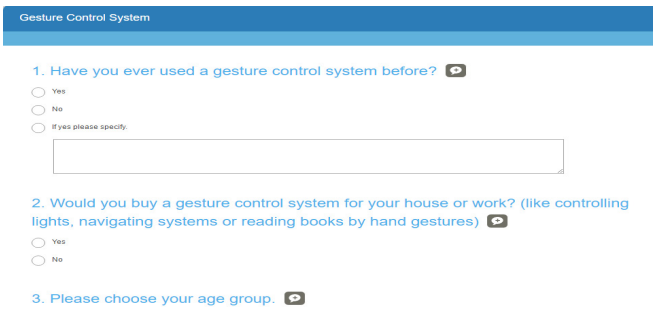


Figure 11 Online survey

The results of the survey were as follows:
64% of the users stated that this was their first time using a gesture controlled system while 36% stated that they have a similar previous experience (Xbox, telephone), 72% of them stated that they would buy such a system to use in their daily life while 28% declared that they would not. (Figure 15)

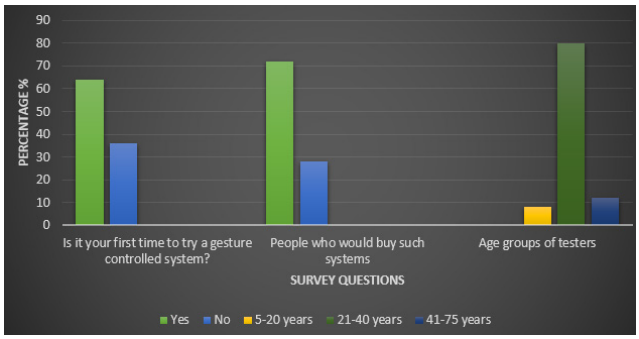


Figure 12 Statistics of Surveys

Although the feedback of users varied about buying such a system or not, but all agreed that the system was easy to use and learn, and that the gestures were fairly simple to perform. While the wave gesture was the easiest to perform with a 100% on its performance and real time response, the click gesture was more challenging where the response to the gesture was at 50% before making the Kinect stationary, but rose to 90% after fixing it to a specific location, making the gesture detection easier and the system response to users better, where the optimal distance for the user to be at is between 2 and 2,5 meters away from the Kinect with the Kinect stationed at 1.5 to 1.65 meters high.

The system was also tested for other gestures that are not defined within the program, but there was no reaction from the system whatsoever, so the system only responds to predefined gestures (the ones mentioned before).

The size of the population was increased from the previous paper, where about 20 people previously tested the system, and now the test subjects were doubled so the responses were more various and the feedback was more accurate.

The ground truth for the gestures used for us as programmers was obtained through tests, and then when the system is presented to the users, a tutorial is given to make sure the gestures and how the system works are both clear to the users, and after the tutorial, users adapted immediately to the system and the feedback was very satisfactory. These are some of the users' feedbacks on the system: "Gestures are easy to learn and perform", "The wave gesture is like saying hello and goodbye", some noted that the tutorial is very important for controlling the system properly, and so within each display window instructions were included to help first time users understand what to do while navigating the system. (See Figure 16)

Figure 13 Directions displayed to help users

A better step would be to include a video tutorial about the gestures to make sure that even people who can't read can still understand and use the system.

To test the performance of the system under different conditions we decided to perform the test with a single user but on different distances, and in different lighting conditions.

The evaluation of the response was classified as Excellent (100%), Very Good (85%), Good (70%) and Bad (0), and the distance started from 1.5 m from the Kinect to make sure that the user's hand wouldn't extend to its blind spot, where the range of detection starts from 1 m and extends to almost 6m away from the Kinect.

As for lighting conditions, the tester performed the gestures outdoors during daytime in an open space, where there was a lot of sunlight, then indoors but with sunlight as well, then the amount of sunlight was decreased and instead electric lighting was used mostly, then the lights were 85% dimmed and finally the lights were completely turned off and the sunlight was blocked to get the darkest possible environment.

Figure 17 displays the response of the system under the previously mentioned conditions. As we can see in the figure, the response of the system is optimal between 1.5 and 3 m and precisely at 2.5m which is the optimal distance for the user in all lighting conditions, however when the distance increases outdoors the response becomes very bad due to the large amounts of IR light present in sunlight which can affect the performance of the sensor and decrease its range of response. The performance gets better as the amount of sunlight decreases, however in the remaining lighting conditions it proved to have a good response till about 4m from the Kinect, then the response decreases and stops after 5m. But when all lights were blocked, the response was much better and the range increased to 6m which is the range of "vision" of the Kinect.

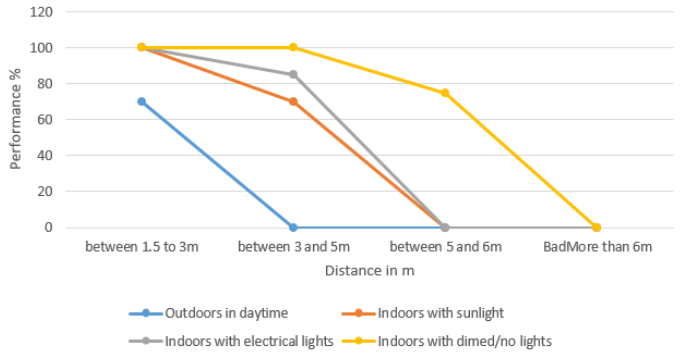


Figure 14 System response under different lighting conditions and distances

Another problem we faced in this implementation which was the calibration of the sensor distance from the user, where any change in the distance between the user and the sensor had greatly affected the system. This can be resolved by keeping the sensor stationary. Since the control system created was implemented in a library, then the educational background of the user is limited to a certain level, and age has its limits as well, however this type of control system is gesture dependent, meaning that if implemented in other domains to control things, like hardware, then it will no longer limit the educational background or age of users and it can be used by the majority of people.

The accuracy of the system was very good, however for better results new algorithms can be created specifically for the purpose of this system, which would enhance the accuracy and hence the overall system response.

6. Conclusion

The simulations of the system and the user feedback about the system have reflected a good response, as long as the Kinect remains stationary after calibration to ensure the best results. The simplicity of the system can allow it to go into various domains for control over software or hardware and so it can be implemented anywhere to control systems easier, however it can be improved through further studies and more specialized hand detection algorithms.

References

[1] Ahmad A, Al Majzoub R, Charanek O, 2016, '3D Gesture-Based Control System using Processing Open Source Software', in A Ahmad, R Majzoub and O Charanek, , *Beirut 2016 2nd International Conference on Open Source Software Computing (OSSCOM), Lebanese University.*

[2] Aggarwal L, Gaur V and Verma P"Design and Implementation of a Wireless Gesture Controlled Robotic Arm with Vision", International Journal of Computer Applications, 2013, Volume 79 - Number 13

[3] Neßelrath R., Lu C., Schulz C. H., Frey J., and Alexandersson J., "A Gesture Based System for Context-Sensitive Interaction with Smart Homes", Ambient Assisted Living, Springer, 2011, pp 209-219.

[4] Bhuiyan M. and Picking R,"A Gesture Controlled User Interface for Inclusive Design and Evaluative Study of Its Usability", Journal of Software Engineering and Applications, 2011, 4, 513-521.

[5] Qian K , Niu J and Yang H,"Developing a Gesture Based Remote Human-Robot Interaction System Using Kinect", International Journal of Smart Home, 2013,Vol. 7, No. 4

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 2, No. 3, 974-980 (2017)
www.astesj.com
Special issue on Recent Advances in Engineering Systems

ASTES Journal
ISSN: 2415-6698

Deterministic Approach to Detect Heart Sound Irregularities

Richard Mengko*, Amy Hamidah, Ricky Saputra

Electronics Lab, School of Electrical Engineering and Informatics, Institut Teknologi Bandung, 40132, Indonesia

ARTICLEINFO

Article history:
Received: 05 April, 2017
Accepted: 06 June, 2017
Online: 10 July, 2017

Keywords:
Heart sounds
Detection
S-transform

ABSTRACT

A new method to detect heart sound that does not require machine learning is proposed. The heart sound is a time series event which is generated by the heart mechanical system. From the analysis of heart sound S-transform and the understanding of how heart works, it can be deducted that each heart sound component has unique properties in terms of timing, frequency, and amplitude. Based on these facts, a deterministic method can be designed to identify each heart sound components. The recorded heart sound then can be printed with each component correctly labeled. This greatly help the physician to diagnose the heart problem. The result shows that most known heart sounds were successfully detected. There are some murmur cases where the detection failed. This can be improved by adding more heuristics including setting some initial parameters such as noise threshold accurately, taking into account the recording equipment and also the environmental condition. It is expected that this method can be integrated into an electronic stethoscope biomedical system

1 Introduction

Auscultation is an act of medical diagnosis by listening to sounds from the heart, lungs or other organs with a stethoscope. Stethoscope has been a popular tool to early diagnose a patient since its development on 1800's [1], due to its simplicity and wide availability. The outcome from a diagnose using this stethoscope could vary according to physician's hearing sensitivity and experiences [2]. The introduction of digital stethoscopes reduce the hearing sensitivity problem because the heart sound can be amplified to suit the examiner's need and furthermore noise can also be reduced. However, the digital stethoscopes still have not solved the skill and experience required to be able to diagnose heart problems accurately.

Our research group is working to develop an easy to use digital stethoscope that has some capabilities to detect various type of heart sound.To do this, we need to design a device that can detect components of heart sound such as primary heart sound; S1, S2, S3, S4, as well as murmurs and clicks. The intention is to develop a tool that can provide some suggestion to help medical practitioners in performing a quick diagnose of patient's problem. This tools should be small and simple enough so that it will be easily accepted because it is not very different than the ordinary stethoscope.

The first challenge towards this goal is to design a simple and yet effective heart sound detection technique. Most of heart sound classification techniques are based on machine learning methods such as clustering and artificial neural networks [3, 4], SVM and MCSearch [5]. They need a large amount of data to learn from and also they require a lot of computation. To overcome those problems, this study investigated a simple detection method without machine learning or statistical approach using only the time series features of heart sound.

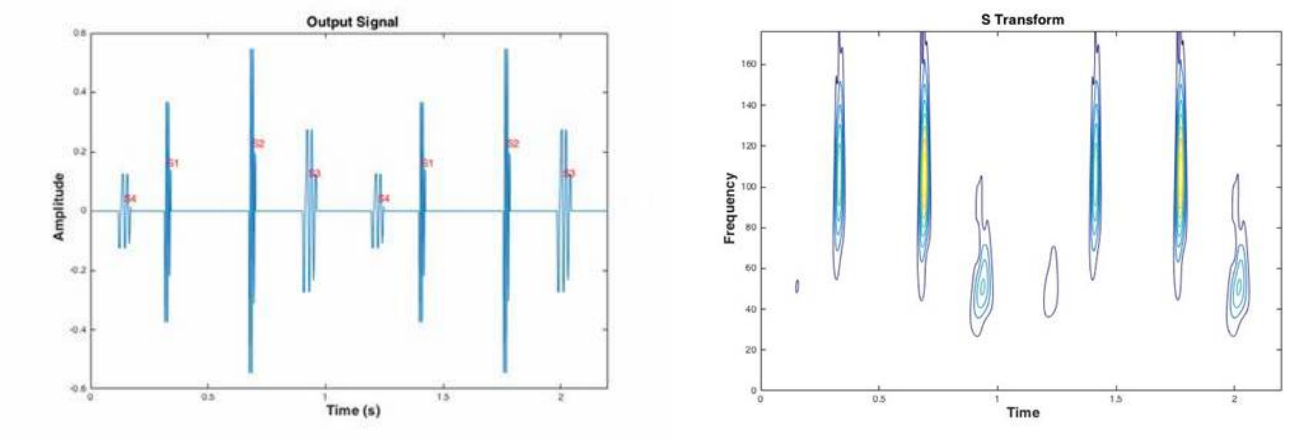
The next section takes a deeper look at heart sound signals to discover the deterministic features of each component. These features will be used in the core of the proposed detection method described in section III. Section IV presents the result and discussion of our experiment. The conclusion of this paper is written in section V.

2 Heart Sound as Time Series

Heart sound is acoustic signal produced by the mechanical activities of the heart as a result of the vibration of the valves, heart muscle tissues, and great

*Richard Mengko, Jl. Ganesha No. 10 Bandung, +628122126286 & richard@mengko.com

https://dx.doi.org/10.25046/aj0203123



(a) (b)

Figure 1: (a) Illustration of heart sound signal with S1, S2, S3 and S4 component, (b) S-transform diagram.

vessels . The closure of mitral tricuspid valves produce S1, and the second heart sound (S2) caused by the sudden closure of aortic and pulmonary valves. S1 has lower frequency with longer duration and higher intensity, while S2 has higher frequency with shorter duration. Under pathological conditions and abnormalities, additional components of the third heart sound (S3) and/or the fourth heart sound (S4) may be present. S3 occurs in the rapid filling period of early diastole after S2 due to the vibration of blood inside the ventricles, which may indicate the systolic dysfunction. S4 appears before S1 and it is caused by altered physical properties of ventricle or increase in the rate and volume of blood flow in the late diastolic interval, which may indicate the presence of acute myocardial infarction or ischemia diseases [6] . S3 and S4 are usually low amplitude and low frequency that make them difficult to be detected. These four heart sound components lie in the frequency range of 20-200 Hz [7].

A normal cardiac cycle comprises of four main segments which are S1, systolic period, S2 and diastolic period, as shown in 1. Single cardiac cycle can also be defined by the interval between the beginning of S1 and the beginning of the next S1. Systolic period is the interval between the end of S1 and the beginning of S2, while diastolic period is defined as the interval between the end of S2 and the beginning of S1 in the next cycle. The systolic period is usually shorter than the diastolic period. Figure 1 illustrates two cycles of quadruple gallop heart sound consisting of S1, S2, S3, and S4 components (each sound has been labeled by red text). By analyzing the pattern of heart sound features, the heart malfunction symptoms leading to CVD can be identified. The first and most important step towards automatic heart sound diagnosis and analysis is segmentation and detection.

According to the process how a heart sound is generated, it can be considered as a time-series events which generate a series of amplitudes values that follow a certain pattern. Hence methods commonly used in time series analysis such as Fourier Transform (FFT, DFT, STFT), wavelet, and S-transform can also be used to analyze heart sound. To further understand the heart sound behavior, S-transform analysis is applied on denoised heart sound signal to examine its time and frequency characteristics.

S-transform is introduced by Stockwell et.al and defined as:

$$S(t,f) = \int_b^a s(t) \frac{|f|}{\sqrt{2\pi}} e^{\frac{-t-\tau^2 f^2}{2}} e^{-i2\pi f\tau} d\tau \qquad (1)$$

Figure 1(a) depicts a heart sound signal in time domain, while Figure 1(b) shows its S-transform representation. From these two figures it can be deducted that each component of heart sound occupies separate area in time domain. Since there is no overlapping region we can use time as the main feature to deterministically distinguish each component (S1, S2, S3, and S4).

3 Detection Method

The steps of the proposed method are depicted in Figure 2. Description of each step is as follows:

3.1 Preprocessing

These are processes that should be done before heart sound detection can be performed:

3.1.1 Down Sampling

Normally a phonocardiogram or a heart sound record duration is at least 15 seconds. The recording can be performed using digital stethoscope. With sample rate at 44.1 kHz the amount of data generated is too large for our purpose, so this signal should be downsampled to 2000 Hz. The reason to choose this downsampling frequency is to make sure that all the heart sound which occurred in this frequency is properly represented [8].

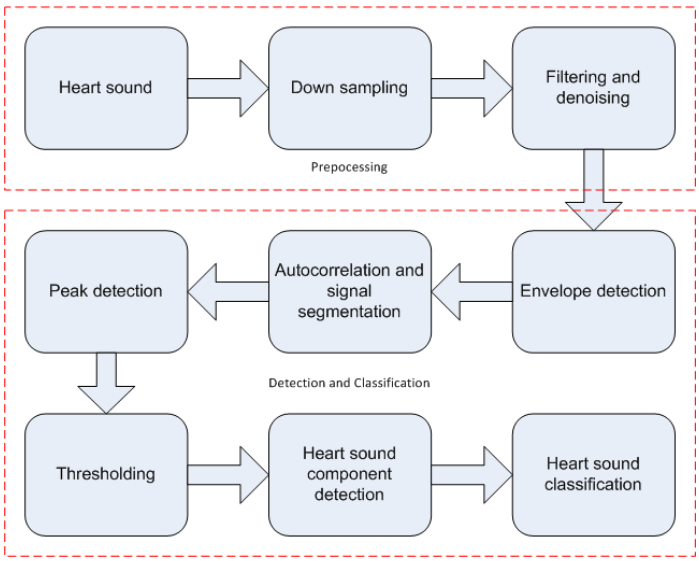


Figure 2: Heart sounds detection's steps.

3.1.2 Filtering and Denoising

To reduce high-frequency noise, the down-sampled signal $x(n)$ is filtered using low passed filter with 400 Hz cut-off frequency. To extract the remaining artifacts, signal $x(n)$ was pass through denoising process using Empirical Mode Decomposition (EMD) [9] method which produced signal $x_d(n)$. This method is chosen because it gave better results compared to other denoising technique [10].

3.2 Detection and classification

3.2.1 Envelope Extraction and Smoothing

Shannon Energy theorem was used in the envelope extraction process of denoised signal [11, 12]. Based on these envelopes heart sound components will be determined.

$$x_{env}(n) = -x_d^2(n) log x_d^2(n) \tag{2}$$

Our experiment shows that the produced envelope was not smooth. It has some ripples that might falsely be detected as peaks during peak detection process. To overcome this problem, the Finite Impulse Response (FIR) filter was used to remove the ripples. The cut-off frequency of this filter is 25 HZ.

3.2.2 Autocorrelation and Signal Segmentation

To compute cardiac cycle and cardiac number, autocorrelation method was applied to the envelopes. Cardiac cycle period was determined by calculating the average distance between highest peaks and its neighboring peaks. To calculate the number of cardiac cycle, sum of the number of highest peaks and one was divided by two. Based on our experiment, to calculate more accurate a cardiac cycle time, it was needed at least three cycles. Therefore it needs at least three seconds recording time. In order to shorten the calculation process, the calculation process, envelope signal that has been smoothed was divided into four cardiac cycle segment.

3.2.3 Peak Detection and Threshold Calculation

A peak is defined as the highest point (local maxima) between two local valley points. The formula for local maxima is:

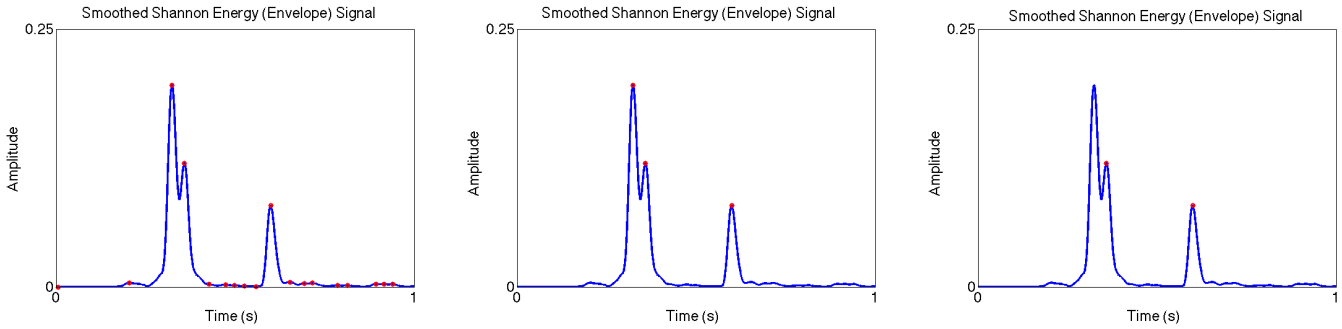
$$\frac{d(x_{env}(n))}{dn} > 0 \text{ and } \frac{d(x_{env}(n+1))}{dn} < 0 \tag{3}$$

while valley point (minima) defines as

$$\frac{d(x_{env}(n))}{dn} < 0 \text{ and } \frac{d(x_{env}(n+1))}{dn} > 0 \tag{4}$$

Some extra peaks might be detected by the program due to left over noise and artifacts. For example, a normal heart sound must have two peaks per cardiac cycle. Due to the noise, the program might detect some small peaks as shown by red dot in Figure 3a. To overcome this problem the Heron's formula was applied to calculate the threshold of the peak. With this formula the area of triangle formed by a local maximum and its two neighbor local minima was calculated. The area that is less than average of all triangles will be ignored. The results of implemented this formula is shown in Figure 3b. This procedure will not ignore the peak caused by a split.

It shows that after applying this formula, the program detects three peaks instead of two. After calculating peak's threshold, the adjustment of peak is needed. In this adjustment the difference between the peak's two minima amplitudes were checked. If the difference is greater than 0.01 mv the adjustment will ignore (not remove) the peak and its second minima. Then the new peak is produced from a combination of first peak's minima, second peak, and second peak's minima. The results of the adjustment is shown in



(a) Before using Heron's formula (b) After using Heron's formula (c) After peak adjustment

Figure 3: Graphical outcome for peak detection method.

Figure 3c.

3.2.4 Heart Sound Component Detection

to distinguish primary heart sound, extra heart sound, murmur, and other abnormalities, five features were extracted from the signal; peak frequency, peak interval, peak duration, total power, and amplitude. On the peak adjustment process, peak frequency is used to calculate each peak frequency that was detected.

The distance between two heart sound components (peaks) is determined by peak interval. The peak interval is measured from the first peak's second minima to the second peak's first minima, while the peak duration is measured from the first minima to the second minima of the peak. To calculate the power on systolic and diastolic region we use to the total power formula. Finally, the amplitude is used to differentiate heart sound abnormalities in systolic and diastolic region.

The following are the detection steps starting with wide split detection, followed by heart sound component detection, split detection, and murmur, click, and ejection click detection. Figure 4 shows the whole detection process, while the descriptions of each step is as follows:

Firstly, Wide Split Detection. Wide split is occured if S1 or S2 has more than one peaks. Wide split detection is used to distinguish whether the second peak of the closest is extra heart sound (S3 or S4) or the split sound. In this process, peak frequency feature is used. If the frequency peaks are similar, thus the peaks must be considered as a wide split sound and the number of peaks in the cycle will be reduced by one. If the frequency of peaks are quite different, it means the second peak is either S3 or S4.

Secondly, Heart Sounds Component Detection. The number of peak per cycle and the heart sounds' characteristics (TABLE 1 [13, 14, 15]) are used in this detection process as the bases to detect the heart sounds components. There are four possible number of peaks per cycle as explained below.

No	Sounds	Duration (ms)	Frequency (Hz)
1	S1	100 – 160	10 – 140
2	S2	80 – 140	10 – 400
3	S3	40 – 80	15 – 60
4	S4	30 – 60	15 – 45

Table 1: Heart sounds characteristics

- Case 1: Only one peak in a cardiac cycle is detected. It means that the peak is S1, and S2 is covered by murmur, either its continuous murmur or to-and-fro murmur.
- Case 2: There are two peaks in a cycle. It means that those peaks are S1 and S2. To determine which peak corresponds to S1 or S2, peak interval features is used. If the distance of second minima of first peak to first minima of second peak is less than the distance of second minima of second peak to first minima of first peak, it means that the first peak is S1 and the second peak is S2, vice versa.
- Case 3: There are three peaks in a cycle. It means there is an extra sound component which either S3 or S4. Before determining S3 or S4, S1 and S2 need to be decided. To distinguish extra sound from primary sound, the distinctive feature of heart sound, peak duration, is used. S3 or S4 has shorter duration than S1 or S2. Therefore, S1 and S2 can be decided by using the same method that explain before. If the distance between the third peak and S1 is shorter than the distance between this peak and S2, thus this peak is S4, otherwise this peak is S3.
- Case 4: There are four peaks in a cycle. It means that S1, S2, S4, and S4 are present in the cycle. The same procedure using peak duration and peak interval will be applied to determine each peak to corresponding heart sound.

Thirdly, Split Detection. Split only occurs on either S1 or S2. Split can be heard when the separation exceeds 20 ms [16]. It means that if there is more than one peak in either S1 or S2 and they are more than 20 ms apart, then it is considered a split Example of S1

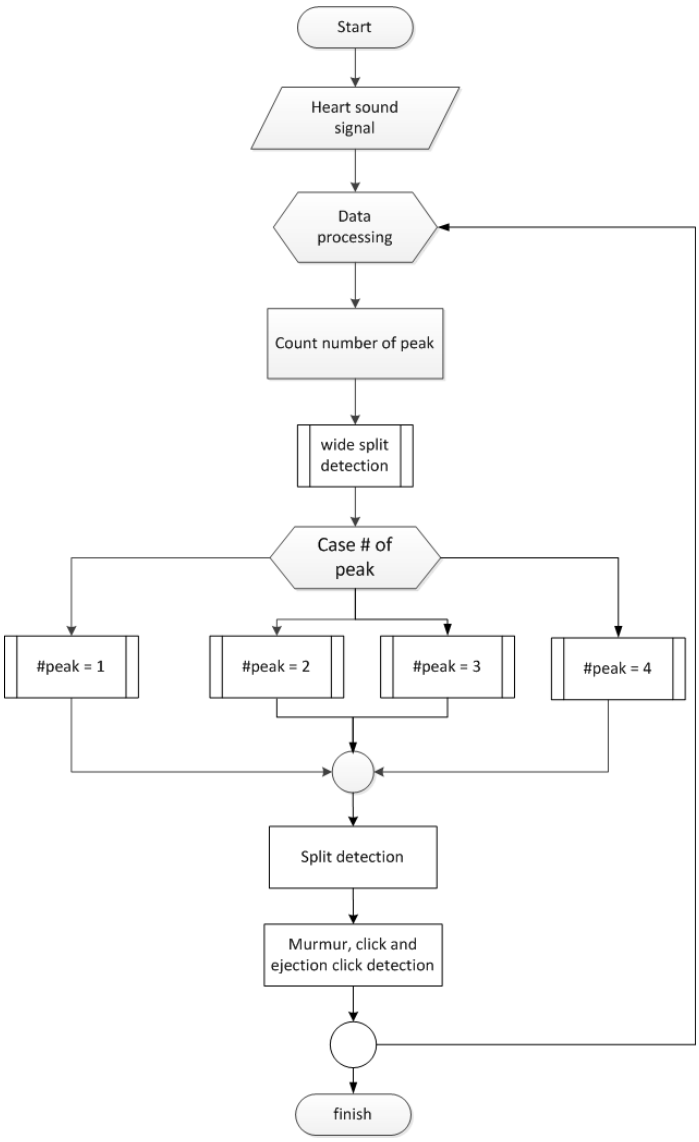
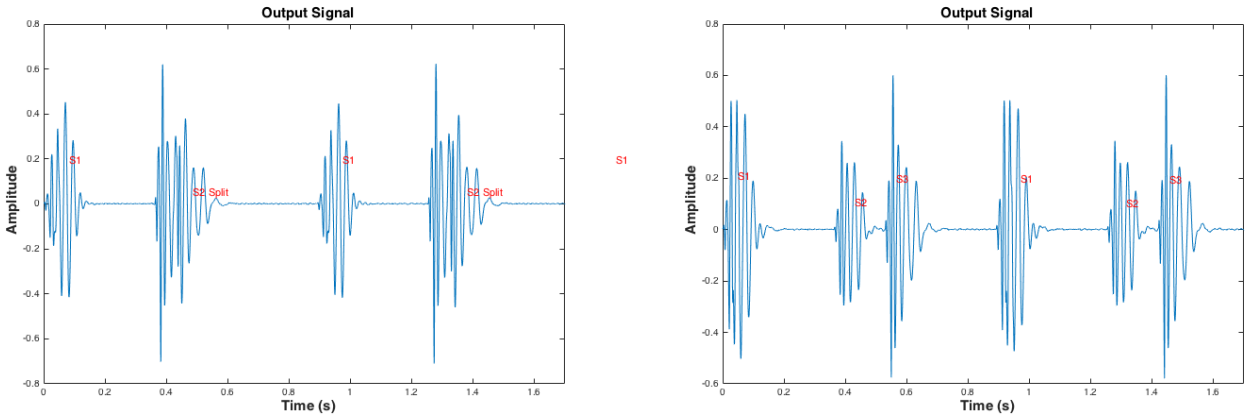


Figure 4: Flowchart of detection process.



(a) S2 Split

(b) S3

Figure 5: Graphical outcome for peak detection method.

split can be seen in Figure 3, which shows envelope signal of split S1.

Finally, Murmur, Click, and Ejection Click Detection. The avarge power feature is used to indicate whether there is murmur or click in either systolic or diastolic area. If the average power in systolic or

diastloic area greater than the average power in that area for a normal heart sound it indicates there is abnormal sound such as murmur, click or ejection click in the area [17]. In this case, the envelope signals in systolic or diastolic area are examined further to determine which abnormal sound these signals belong to. If there are several peaks that are higher than the threshold then these can be an indication of murmur. Based on our experience, the threshold for murmur should be set to half of the average amplitude of the known signal. The type of murmur is classified based on the location where the murmur occurs and its amplitude. If the peak is greater than four times of murmur's threshold then it is a Click. These threshold are obtained purely heuristically form our observations. If there is a click that followed by murmur then it is an Ejection Click.

It is expected that the above process will be able to produce information related to following heart problems [18], such as:

- The presence of systolic murmur can indicate AV valve regurgitation, while diastolic is murmur related with the problem on semilunar valves
- To-and-fro murmur indicates the problem with all valves
- Wide split of S1 and/or S2 is a symptom of right brand bundle block (RBBB)
- Fixed split on S2 indicate an atrial septal defect
- Paradoxical splitting on S2 means LBBB problem
- Ejection click is a symptom of aortic or pulmonary stenosis
- Mid systolic click indicates mitral or tricuspid prolapse
- Opening snap shows the presence of artificial valve or rheumatic mitral stenosis
- The presence of S3 or S4 indicates ventricle stiffness, the summation gallop indicates the heart failure

However the decision regarding the real heart problems of a patient is entirely left to the physician.

4 Results and Discussion

The heart sounds samples were taken form the ones that available in the internet [19, 20, 21], which signals were noiseless. To make them imitate the real sound, an artificial noise was added to the signal. In our experiment, the input signals were designed to have a 13.98 dB signal-to-noise ratio. This noise addition is intended to demonstrate the effect of filtering and denoising process.

The simulation is run on OS X El Capitan with processor 2.5 GHz intel core i5 and 8 GB RAM using MATLAB 2014a software.

There are more than 20 heart sounds recording that are used to test the detection program. The list of distinctive sounds that are successfully detected and classified are shown in Table 2. The goal of the program is to identify and produced a label on the printed heart sound signal to mark which type of sounds occur. Figure 5 shows two of the simulation results, which are Figure 5a S2 Split and Figure 5b S3

There are cases where the program does not give correct results. This is happen if and only if murmur or the abnormal sounds amplitude and frequency are similar with the heart sounds'. If this happen, the whole heart sounds will be covered by the abnormal sounds or murmur. The example of this case is continuous murmur, where S2 is covered by murmur.

Tabel 2 shows the different signals that can be identified

Name	Source
Single S1 S2	E-general Medical, Michigan, Thinklabs
Split S1	Littmann, Michigan
Mid-systolic Click	Michigan
Early Systolic Murmur	Michigan
Mid-Systolic Murmur	Michigan
Late Systolic Mumur	Michigan
Holo Systolic Murmur	Michigan
S4 Gallop	Littmann, Michigan, Thinklabs
S3 Gallop	Littmann, Michigan
Quadruple Gallop	E-General medical
Ejection Click	E-general medical , Michigan
S4 and Mid-Systolic Murmur	Michigan
S3 and Holosystolic Murmur	Michigan
Normal Split S2	Littmann, Thinklabs
S2 Wide Split	E-general medical

Table 2: Heart sounds recording

5 Conclusion

This study proposes a deterministic approach to detect heart sound and its regularity. The preprocessing apply EMD denoising and low pass filter to cancel noise and remove unwanted artifact. To extract signal's envelope Shannon Energy method is used, and follow by using FIR filter to smoothen the jagged envelope. Our main source to classify heart sound is the signal's envelope. To detect the number of peak per cardiac cycle, Heron's formula and peak adjustment were then applied. These peaks and their extracted features were the keys to detect and classify the heart sounds.

Our experiment shows that quite a few of distinctive heart sounds can be successfully detected by this proposed method. Still, some more heuristics can be added to improve accuracy. This method will be further improved so that it can be embedded as part of a digital stethoscope.

Conflict of Interest We have read and understood the policy on declaration of interests and declare that we have no competing interests.

Acknowledgment We would like to thank Electronics Department of Instiut Teknologi Bandung for the funding of our research, Mr Setiadi for helping us proof read the journal, and all others who directly or indirectly helped us to finish the project

References

[1] "History of the stethoscope." [Online]. Available: http://www.adctoday.com/learning-center/about-stethoscopes/history-stethoscope

[2] H. Liang, S. Lukkarinen, and I. Hartimo, "Heart sound segmentation algorithm based on heart sound envelogram," *Computers in Cardiology 1997*.

[3] I. N. Amit G., Gavriely N., "Cluster analysis and classification of heart sounds," *Biomedical Signal Processing and Control 4 (2009) 2636*.

[4] T. Dokur Z., lmez, "Heart sound classification using wavelet transform and incremental self-organizing map," *Digital Signal Processing 18: 951959*.

[5] P. A. Redlarski G., Gradolewski D., "A system for heart sounds classification," *PloS one, vol. 9, no. 11, p. e112673, 2014*.

[6] J. F. Tseng Y, Ko P, "Detection of the third and fourth heart sounds using hilbert-huang transform," *Biomedical engineering online, vol. 11, no. 1, pp. 113, 2012.*

[7] S. Choi and Z. Jiang, "Comparison of envelope extraction algorithms for cardiac sound signal segmentation," *Expert Systems with Applications, vol. 34, no. 2, pp. 10561069, 2008*.

[8] Z. CHOI, SJIANG, "Comparison of envelope extraction algorithms for cardiac sound signal segmentation," *Expert Systems with Applications*, vol. 34, no. 2, pp. 1056–1069, 2008.

[9] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu, "The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 454, no. 1971, pp. 903–995, 1998.

[10] A. H. Salman, N. Ahmadi, R. Mengko, A. Z. R. Langi, and T. L. R. Mengko, "Performance comparison of denoising methods for heart sound signal," *2015 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, 2015.

[11] M. Saini, *International Journal of Electronics and Communication Technology*, pp. 15–19, 2016.

[12] A. H. Salman, N. Ahmadi, R. Mengko, A. Z. R. Langi, and T. L. R. Mengko, "Automatic segmentation and detection of heart sound components s1, s2, s3 and s4," *2015 4th International Conference on Instrumentation, Communications, Information Technology, and Biomedical Engineering (ICICI-BME)*, 2015.

[13] L. James, "Heart sounds," 2011. [Online]. Available: http://www.slideshare.net/LawrenceJames/heart-sounds

[14] A. Hamidah, "Integrated biosignal processing for early detection of heart abnormality: Auscultation and visualization," 2015.

[15] G. Amit, "Heart sound analysis: Theory, technique and applications," 2004.

[16] N. Ranganathan, V. Sivaciyan, and F. B. Saksena, *The art and science of cardiac physical examination*. Humana Press, 2006.

[17] M. Singh and A. Cheema, "Heart sounds classification using feature extraction of phonocardiography signal," *International Journal of Computer Applications*, vol. 77, no. 4, pp. 13–17, 2013.

[18] L. S. Jung H., "The cardiac cycle: Mechanisms of heart sounds and murmurs," *The physiology of Heart Disease, A Collaboration Project of Medical Students and Faculty, 5th ed*.

[19] R. Judge and R. Mangrulkar, "Heart sound and murmur library," 2015. [Online]. Available: http://open.umich.edu/education/med/resources/heart-sound-murmur-library/2015/materials.html

[20] "Basic heart sounds course." [Online]. Available: http://www.littmann.ca/wps/portal/3M/en_CA/3M-Littmann-CA/stethoscope/littmann-learning-institute/heart-lung-sounds/heart-sounds/

[21] T. Bauch, "Thinklabs one - digital stethoscope." [Online]. Available: http://www.thinklabs.com/#!heart-sounds-old/c1dqs

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 2, No. 3, 981-986 (2017)
www.astesj.com
Special issue on Recent Advances in Engineering Systems

ASTES Journal
ISSN: 2415-6698

Spatial Modulation Technique For Filtered-OFDM Based Wireless Transmission

Shaddrack Yaw Nusenu[*,1,2]

[1] *University of Electronic Science and Technology of China (UESTC), School of Communication and Information Engineering, China.*

[2] *Koforidua Technical University (KTU), Electrical / Electronic Department, Faculty of Engineering, Ghana.*

ARTICLEINFO

Article history:
Received: 04 April, 2017
Accepted: 28 June, 2017
Online: 10 July, 2017

Keywords:
Spatial Modulation
F-OFDM
Maximum-likelihood (ML) detector
Spectral efficiency and Spectrum utilization

ABSTRACT

In this paper, a spatial modulation (SM) which provides no synchronization between the transmitting antennas and avoids inter-channel interference (ICI) at the receiver input while maintaining high spectral efficiency is applied to filtered-orthogonal frequency division multiplexing (F-OFDM) transmission. The SM technique maps a block of information bits into two information carrying units, namely, a symbol chosen from a signal constellation diagram and a unique antenna number selected from the set of transmit antennas. Therefore, for F-OFDM wireless transmission, each subcarrier within the subband is mapped to one of the transmitting antennas. During wireless data transmission, there is only one active antenna transmitting power on that subcarrier at an instant of time and the rest of the antennas remains silent (zero power). At the receiver, a maximum-likelihood (ML) detector recovers the transmitted block of information bits by estimating the transmitted signal and the respective transmit antenna number. The effectiveness of the proposed method is verified using numerical results.

1 Introduction

The unprecedented increase of mobile communications, wireless Internet access and multi-media applications has been driven by telecommunications companies and researchers to conceive new transmission technologies, protocols, and network infrastructure solutions in order to maximize both throughput and spectral efficiency [1]. For the next generation of wireless communication, it has become necessary to provide higher data rate and improve spectral efficiency. Employing multi–antennas at transmitter and receiver has shown to be an effective technique to achieve spectral efficiency over the past few years. The use of spatial multiplexing using multiple input multiple output (MIMO) was first proposed in [2]. MIMO technology is one of the techniques employed to attain high spectral efficiency by transmitting multiple data streams from multiple antennas [3]. However, the MIMO transmission is dependent on transmit and receive antenna spacing [4, 5], synchronization of the transmit antenna [6], as well as the algorithm needed for inter-channel interference (ICI) reduction at the receiver input. The Bell Labs Layered Space-Time Architecture (BLAST) in [7] proposed MIMO detection algorithms, named vertical BLAST (V-BLAST) [8]. In V-BLAST, multiple data streams are transmitted separately and detected successively by employing both an array processing (nulling) and interference cancelation techniques.

OFDM is another promising technique proposed for transmitting high speed data over wireless channels in future mobile communication systems [9]. OFDM has been useful in several wireless standards, for example, digital audio and video broadcasting (DAB) and (DVB-T), IEEE 802.11a [10], the IEEE 802.16a metropolitan area network (MAN), and the local area standard (LAN) [11]. Although, OFDM features seems attractive, there is still technical challenges that needs to be addressed for future wireless transmission [12, 13].

Recently, a new extension to OFDM has been proposed, namely, F-OFDM in [12, 13]. The most significant difference of F-OFDM, as opposed to the conventional OFDM, is that the former total bandwidth has been divided into multiple subband. In addition, each subband have a distinct subcarrier spacing with its own cyclic prefix (CP) and filter. On the other, in conventional OFDM, total bandwidth consists of a single block and has the same subcarrier spacing between the individual subcarriers. Note that, there is de-

[*] Shaddrack Yaw Nusenu, UESTC and KTU, Ghana, Email: nusenu2012gh@yahoo.com

https://dx.doi.org/10.25046/aj0203124

gree of flexibility in F-OFDM than its counterpart OFDM. The details can be found in [12, 13].

Therefore, in this paper, SM is applied to F-OFDM, namely, spatial modulation F-OFDM (SM F-OFDM), which provides higher spectral efficiency as a result of both SM and F-OFDM advantages. The main contributions of this work can be summarized as follows: (i) Analogue to [9], SM F-OFDM is designed for wireless transmission. (ii) Binary phase shift keying (BPSK) and quadrature amplitude modulation (QAM) modulation schemes were adopted for SM mapping rule.

The remaining sections are organized as follows: In Section II, the system model of SM-F-OFDM is presented. Simulation results are provided in Section III and concluding summaries are drawn in Section IV.

Throughout the paper, the following notations are used. Bold lowercase and bold uppercase letters denote vectors and matrices, respectively. We use $[.]^T$ and $\|.\|_F$ to denote transpose and Frobenius norm of a matrix or a vector, respectively. $\hat{i}(k)$ is the estimated transmit antenna index (number) and $\hat{x}_{\hat{i}}(k)$ is the estimated transmitted symbol. F-OFDM modulator-1 contains data belong to first antenna and so forth and $\otimes$ denotes time convolution.

2 System Model of SM F-OFDM

In this section, SM is discussed briefly. Suppose the total number of transmit and receive antennas is denoted by N_t and N_r respectively, M being the cardinality of the signal constellation. Hence, the rate of SM (bits per channel use - (bpcu)) for arbitrary N_t and M respectively, can be given by [14],[15]

$$R_{SM} = \log_2(N_t) + \log_2(M) \tag{1}$$

where $\log_2(N_t)$ defines the single active transmit antenna and $\log_2(M)$ is the transmitted QAM/BPSK modulation symbols. Note that the N_t, N_r and M, respectively, can be chosen independently [16]. The basic principle of SM is that, it maps multiple information bits into one information symbol with respect to their corresponding transmit antenna number. The number of information bits that can be transmitted in SM technique depends on the signal constellation diagram employed as well as the given number of transmit antennas. It is important to note that the SM utilizes the distinctiveness and random nature of the wireless channel for communication. The following summarized SM characteristics [14]:

1) At any signaling time instance, only one transmit antenna is active for data transmission while the other antennas remains silent.

2) The spatial position of each transmit antenna of the set is employed as an information source.

The proposed architecture of SM F-OFDM system is depicted in Figure 1, with the F-OFDM transmitter showing in Figure 2, while F-OFDM receiver is illustrated in Figure 3. Suppose the input $\mathbf{Q}(k)$ is $m \times n$ matrix which represents the transmitted symbols, where m is the total number of bits per symbol per subcarrier and n denotes the total number of subcarriers of the multiple subband of F-OFDM system. By using the SM mapping rule as listed in Table I, this matrix is mapped into $\mathbf{X}(k)$ of size $N_t \times n$.

From Table I, the SM mapping rule, maps each column in $\mathbf{Q}(k)$ into 4QAM and BPSK signal constellations respectively. In addition, one transmit antenna from a set of two (QPSK) and four (BPSK) antennas is assumed. Herein, 4QAM and BPSK modulation schemes were adopted in this work. As can be seen in Table I, the encoding mechanism are $N_t = 2$ and $M = 4$ when QAM is adopted. Note that, three information bits can be mapped onto 4-QAM with two transmit antennas selected. Similarly, for BPSK, the same spectral efficiency can be achieved with $N_t = 4$ and $M = 4$. Thus the F-OFDM system can convey three bits in each time slots. The number of information bits that can be transmitted on each F-OFDM subcarriers is given in (1). According to the SM mapping rule, the matrix $\mathbf{Q}(k)$ has one nonzero element in each column at the position of the mapped transmit antenna number. All the other elements in that column are defined as zero. It is worth mentioning that, when $N_t = 1$ simplifies SM F-OFDM to conventional single antenna communications.

Supposing in Figure 1, an input bit sequence of $[0 \;\; 0 \;\; 1]^T$ (column vector in $\mathbf{Q}(k)$) from Table I, is mapped to the QAM symbol - 1 + j and the first transmit antenna by using SM mapping. Thus only the first antenna transmits this symbol on the first F-OFDM subcarrier, whereas other antenna remains silent (zero power). As a result, the first column vector in $\mathbf{X}(k)$ is $[-1+j \quad 0]^T$. The second bit sequence is $[1 \;\; 1 \;\; 0]^T$ and is mapped to $[0 \quad -1-\mathrm{j}]^T$, and so on. Similarly, the same approach is adopted for BPSK symbols. The resultant symbols in each row vector $\mathbf{x}_v(k)$ are the data that will be transmitted on all subcarriers from transmit antenna v. And F-OFDM modulator will be used to modulate each row vector $\mathbf{x}_v(k)$.

At the F-OFDM modulator output, $\mathbf{s}_v(t)$ vector is generated. Each $\mathbf{s}_v(t)$ vector generated has a unique and disjoint set of F-OFDM subcarriers. The resulting output vectors at the F-OFDM modulator will be transmitted simultaneously from the N_t over the channel $\mathbf{H}(t)$. At the receiver, the received matrix $\mathbf{Y}(t)$ can be expressed as

$$\mathbf{Y}(t) = \mathbf{H}(t) \otimes \mathbf{S}(t) + \mathbf{G}(t) \tag{2}$$

where $\mathbf{S}(t)$ is a matrix that has all F-OFDM symbols that are transmitted from all transmit antennas. $\mathbf{G}(t)$ is the additive noise matrix, in which each element is assumed to be an independent and identically distributed (iid) zero mean complex Gaussian random variable with variance σ_N^2. $\mathbf{H}(t)$ is the $N_r \times N_t$ channel matrix between transmit antennas v and receive antennas k, respectively and can be expressed as (3)

$$\mathbf{H}(t) = \begin{pmatrix} h_{1,1}(t) & h_{1,2}(t) & \cdots & h_{1,N_r}(t) \\ h_{2,1}(t) & h_{2,2}(t) & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ h_{N_r,1}(t) & h_{N_r,2}(t) & \cdots & h_{N_r,N_t}(t) \end{pmatrix} \tag{3}$$

where $\mathbf{h}_{k,v}$ is a complex fading coefficient between the vth transmit antenna and k^{th} the receive antenna. $\mathbf{h}_{k,v}$ is as-

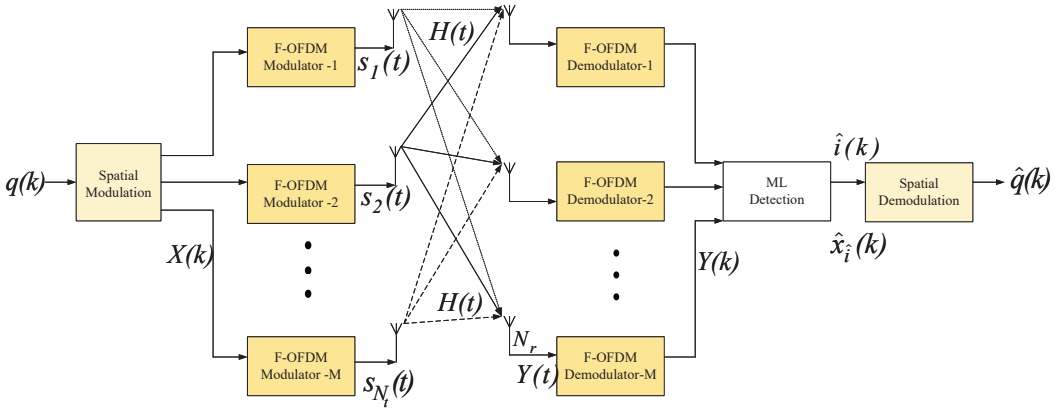


Figure 1: Proposed architecture of Spatial Modulation F-OFDM System for wireless transmission.

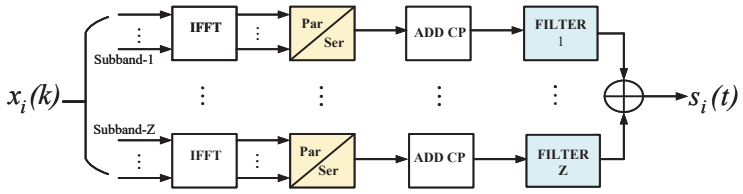


Figure 2: Block diagram of F-OFDM Transmitter.

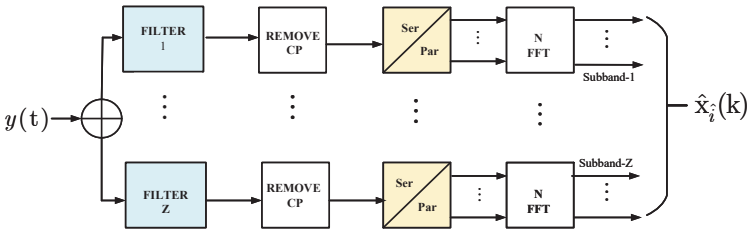


Figure 3: Block diagram of F-OFDM Receiver.

Table 1: Adopted SM F-OFDM Mapping Rule (3 bpcu).

	QAM		BPSK	
Input bits	Antenna Number	Transmit Symbol	Antenna Number	Transmit Symbol
000	1	+1+j	1	-1
001	1	-1+j	1	+1
010	1	-1-j	2	-1
011	1	+1-j	2	+1
100	2	+1+j	3	-1
101	2	-1+j	3	+1
110	2	-1-j	4	-1
111	2	+1-j	4	+1

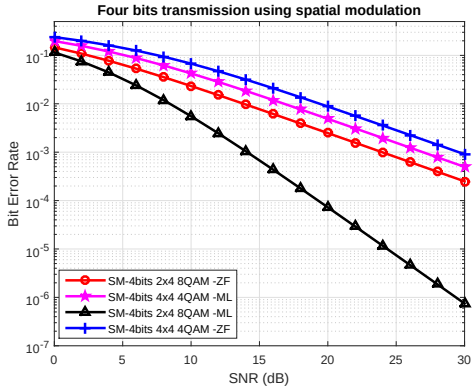


Figure 4: Illustration of BER performance for 4 bits per subcarrier employing ML and ZF detectors.

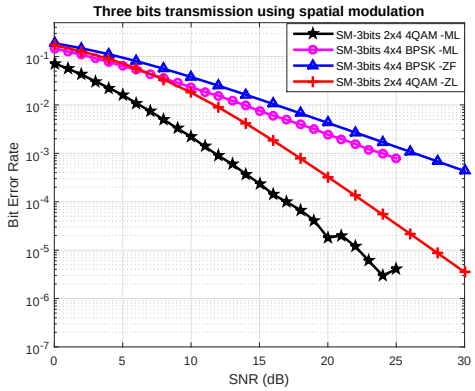


Figure 5: Illustration of BER performance for 3 bits per subcarrier employing ML and ZF detectors.

sumed to be iid complex zero mean Gaussian with variance one.

The receiver uses maximum likelihood detector to estimate the index, $\hat{i}(k)$ and the transmitted symbol, $\hat{x}_{\hat{i}}(k)$ can be expressed as

$$\begin{aligned}\left[\hat{i}^{\partial},\hat{\mathrm{x}}^{\partial}\right] &= \arg\max_{i,\mathrm{x}} \Pr\left(\mathrm{Y}_{\partial}\left|\mathrm{H}(t),\hat{\mathrm{x}}_{\hat{i}}\right.\right) \\ &= \arg\min_{i,x}\left(\left\|\mathrm{Y}_{\partial}-H(t),\hat{x}_{\hat{i}}\right\|_F^2\right)\end{aligned} \tag{4}$$

where $\mathbf{Y}_{\partial}$ is the output matrix from the ML for the subcarrier ∂, i.e. $(\partial : 1,2,...,n)$ and

$$\Pr\left(\mathbf{Y}_{\partial}\left|\mathbf{H}(t),\hat{x}_{\hat{i}}\right.\right) = \frac{1}{\pi^{N_r}\sigma_N^{2N_r}}\exp\left[\frac{\left\|\mathbf{Y}_{\partial}-\mathbf{H}(t),\hat{x}_{\hat{i}}\right\|_F^2}{\sigma_N^2}\right] \tag{5}$$

is the conditional probability density function (PDF) of $\mathbf{Y}_{\partial}$ given $\mathbf{H}(t)$ and $\mathbf{x}_{\hat{i}}$. Equation (4) estimates both the transmit antenna number and the transmitted symbols jointly. The SM demodulator, then used these two estimates to extract the transmitted information bits on this subcarrier by taking an inverse mapping process adopted by the same mapping table used at the transmitter.

Similarly, the zero forcing (ZF) detector can be written as

$$\left[\widehat{i}^{\partial},\widehat{\mathrm{x}}^{\partial}\right] = \left(\left(\mathbf{H}^T(t)\mathbf{H}(t)\right)^{-1}\mathbf{H}(t)\right) * \mathbf{x} \tag{6}$$

where $\mathbf{x}$ denotes the transmitted symbol vector, $\mathbf{H}^T(t)$ is the transpose of the channel matrix in (3) and $*$ denotes the convolution.

3 Simulation Results and Discussions

In this section, simulation results for the proposed SM F-OFDM transmission is shown. Monte Carlo simulations is used to evaluate bit error rate (BER) performance utilizing maximum likelihood (ML) detector at the receiver. In addition, zero forcing (ZF) detector was employed for comparison. Herein, it was considered that the receiver have full channel knowledge and the receive antennas are separated wide enough to avoid correlation. Flat Rayleigh fading channel was also assumed.

3.1 Three Bits Transmission

By using different configurations, a three bits per subcarrier transmission can be achieved. For SM, a 4×4 BPSK and 2×4 4QAM configuration convey three bits per subcarrier. In Figure 4, SM 4×4 BPSK shows performance degradation as compared to SM 2×4 4QAM when ML detector is used. A similar trend can be observed for ZF detector as the same degradation performance is seen with the same configurations. The results in Figure 4 shows that the ML detector estimation outperforms ZL detector estimation.

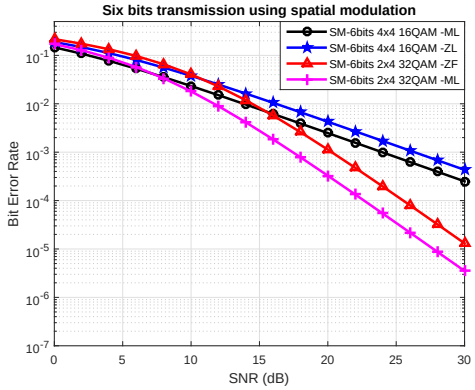


Figure 6: Illustration of BER performance for 6 bits per subcarrier employing ML and ZF detectors.

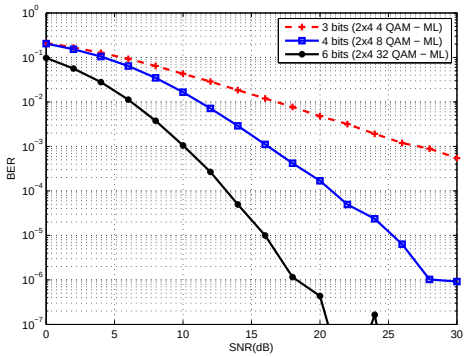


Figure 7: BER comparison performance for SM bits transmission, namely, 3, 4 and 6 bits, respectively when ML is adopted.

3.2 Four Bits Transmission

Four bits transmission can be achieved by the employment of SM 4×4 4QAM or 2×4 8QAM configuration. The comparison BER results in Figure 5 shows that the ZF demonstrates the worst performance as compared to ML detector.

3.3 Six Bits Transmission

Figure 6 results extended the simulation to a spectral efficiency of 6 bpcu for SM 4×4 16QAM or 2×4 32QAM transmissions. Again, ML detection outperforms ZF detection. It can be observed that the likelihood of errors for antenna detection increases with an increase in the number of transmit antennas. Therefore, the 2×4 SM system outperforms 4×4 system at high SNR. An interesting feature of SM can be noticed in Figure 6, where the 2×4 32QAM for ZF detector BER curve crosses the 4×4 16QAM ML detector at around 16 dB. This effect is a result of two estimation processes which is performed in SM, thus antenna number and transmitted symbol. Note that, the error level is dominated by the estimation of the transmitted symbol for low SNR, while the error level is dominated by the estimation of the transmit antenna number for a relatively high SNR. Hence, the scenario shown in Figure 6, for $SNR < 16dB$, lower constellation size and higher number of transmit antenna array is better for improved performance. On the other hand, for $SNR > 16dB$, higher constellation size and lower number of transmit antenna array is better. So there is a flexible tradeoff in this regard.

Figure 7 illustrates the BER comparison for distinct SM bits transmission, namely, 3 bits, 4 bits and 6 bits, respectively, for QAM modulation scheme. It can be observed that the 6 bits SM transmission performs better than the 3 bits and 4 bits. Thus $6bits > 4bits > 3bits$. This is because higher data rate can be achieved in the SM 6 bits transmission leading to a higher spectral efficiency.

4 Conclusion

In this paper, spectral efficient multiple antenna transmission technique, namely, spatial modulation (SM) is applied to F-OFDM for wireless transmission. The antenna pattern is considered as a spatial constellation in an innovative fashion by using SM technique to increase the spectral efficiency. By computer simulation, BER performance for the proposed scheme was evaluated for uncorrelated Rayleigh fading channel and compared with ZF detector. The proposed SM F-OFDM achieves a constant spectral efficiency in bits per subcarrier. Another significant merit of employing SM is the possibility of interchanging the constellation size and number of transmit antenna array to achieve better performance with respect to the operating SNR. In a follows up work, this feature will be further exploited. Furthermore, the performance improvement of the system increases as the transmission rate increases, which makes it a potential candidate for high data rate transmission systems. In addi-

tion, due to the estimation processes in SM approach, it is worth investigating further new algorithms for detection to improve the performance.

Conflict of Interest The author declare no conflict of interest.

Acknowledgment The author would like to thank Respected Professor Wen-Qin Wang (Senior Member - IEEE) of UESTC for his great contributions, support and encouragement.

References

[1] Marco D. R, Harald H., Ali G., Shinya S., and Lajos H., "Spatial Modulation for Generalized MIMO: Challenges, Opportunities, and Implementation ". *Proceedings of IEEE*, pp. 1–47, 2013.

[2] Mehdi Maleki Pirbazari, "Space Modulation: Signal design and performance evaluation" . *PhD dissertation – University of Akron*, 2015.

[3] E. Telatar, "Capacity of Multi-Antenna Gaussian Channels". *European Transaction Telecommunication*, vol. 10, no. 6, November/December pp. 558-595, 1999.

[4] G. J. Foschini, M. J. Gans, and J. M. Kahn, "Fading Correlation and its Effect on the Capacity of Multielement Antenna Systems". *IEEE Transactions on Communications*, vol. 48, March, pp. 502-513, 2000.

[5] S. Loyka and G. Tsoulos,, "Estimating MIMO System Performance Using the Correlation Matrix Approach" . *IEEE Communications Letters*, vol. 6, no. 1, pp. 19-21, January 2002.

[6] Marco. Chiani, Moe Z. W., and A. Zanella, "On the Capacity of Spatially Correlated MIMO Rayleigh-Fading Channels". *IEEE Transactions on Information Theory*, vol. 49, no. 10, pp. 2363-2371, October 2003.

[7] G. J. Foschini, "Layered SpaceTtime Architecture for Wireless Communication in a Fading Environment when Using Multi-Element Antennas". *Bell Labs Technical Journal*, vol. 1, no. 2, pp. 41-59, September 1996.

[8] P. W. Wolniansky, G. J. Foschini, G. D. Golden, and R. A. Valenzuela, "V-BLAST: An Architecture for Realizing Very High Data Rates Over the Rich-Scattering Wireless Channel". *Proceedings of the IEEE International Symposium on Signals, Systems, and Electronics (ISSSE)*, pp. 295-300, September/October 1998.

[9] R. Mesleh, H. Haas, C.W. Ahn and S. Yun, "Spatial Modulation OFDM". *In the Proceedings of the 11th International OFDM-Workshop 2006 (InOWO'06)*, pp. 288–292, August 30-31, 2006.

[10] IEEE, Part 11, "Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications: High Speed Physical Layer in the 5 GHz Band". *IEEE Standard 802.11a*, 1999.

[11] IEEE, Part 16, "Local and Metropolitan Area Networks: Air Interface for Fixed Broadband Wireless Access Systems". *IEEE Standard 802.16a*, 2003.

[12] Xi Zhang, Ming Jia, Lei Chen, Jianglei Ma, Jing Qiu, "Filtered-OFDM Enabler for Flexible Waveform in The 5th Generation Cellular Networks". *Accepted to IEEE GLOBECOM, San Diego, CA*, pp. 16–20, Dec. 2015.

[13] Javad Abdoli, Ming Jia, and Jianglei Ma, "Filtered OFDM: A New Waveform for Future Wireless Systems ". *IEEE 16th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pp. 60–70, 2015.

[14] M. Di Renzo, H. Haas, and P. M. Grant, "Spatial modulation for multiple-antenna wireless systems: A survey". *IEEE Communications Magazine*, vol. 49, no. 12, pp. 182-191, Dec. 2011.

[15] R. Y. Mesleh, H. Haas, S. Sinanovic, C. W. Ahn, and S. Yun, "Spatial modulation". *IEEE Transactions on Vehicular Technology*, vol. 57, no. 4, pp. 2228-2241, Jul. 2008.

[16] J. Mietzner, R. Schober, L. Lampe, W. H. Gerstacker, and P. A. Hoeher, "Multiple-antenna techniques for wireless communications – A comprehensive literature survey". *IEEE Communications Survey Tutorial*, vol. 11, no. 2, 2nd Quart. pp. 87-105, 2009.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 2, No. 3, 987-995 (2017)
www.astesj.com
Special issue on Recent Advances in Engineering Systems

ASTES Journal
ISSN: 2415-6698

Template-like Tensor Domain Operations to Enhancing Diffusion Datasets Quality

Fernando Yepes-Calderon[*,1], Flor M. Medina [2]

[1]*Children's Hospital Los Angeles, Neurosurgery Division, 90027, USA*

[2]*Fundación Valle del Lilí, Imágenes Diagnósticas, Colombia*

ARTICLEINFO

Article history:
Received: 05 April, 2017
Accepted: 05 June, 2017
Online: 10 July, 2017

Keywords:
Tractography
Image enhancement
Tensor averaging
Tensor rotation
DTI

ABSTRACT

Diffusion MRI-based tractography, which is built by connecting the principal components of the estimated water diffusion pattern, is used to elucidate neuronal connectivity. However, the anatomical accuracy of the method is affected by factors such as noise and imaging misalignments. In this manuscript, we present a method to clean diffusion datasets by rotating the diffusion patterns according to a transformation matrix, that is in turn obtained in the fractional anisotropy domain. We demonstrated the accuracy and feasibility of our method by improving brain images of a mouse and also by creating a diffusion template in a neonates framework.

1 Introduction

The diffusion-based methods of magnetic resonance imaging (dMRI) are capable of elucidating microstructural insights by estimating the water patterns of displacement into the tissue under study [1, 2]. Briefly, if there are not boundaries hindering the free movement caused by Brownian motion, the water diffuses isotropically. On the contrary, when there are restrictions, the water displacement occurs along the open spaces tracing a preference proper of the anisotropic diffusion. Thus, the images produced with dMRI are particularly useful for lighting neuronal state and trending, which has been vastly explored [3, 4, 5] and to study the integrity of muscle fiber [6, 7]. In the brain, dMRI is an outstanding tool that allows the follow up in horizontal research; this property together with the non-intrusiveness nature of magnetic resonance imaging (MRI), have lead to several studies of neurodegenerative assessments including Parkinson [8, 9], Multiple sclerosis [10, 11], Schizophrenia [12, 13] among others. However, dMRI is a time-consuming technique and image quality as any other MRI procedure deteriorates rapidly if the scanning time is reduced. The timing issue has restricted the dMRI to be fully implemented in the clinics, and currently, it is only used to diagnose stroke in ultra rapid sequences where the whole head volume is scanned three times to covering the orthogonal axes and once more to produce the reference signal (B0) needed to compute diffusion [14, 15]. Despite this clinical acquisition scheme yields images with a bad Signal-to-Noise ratio, it is enough to localize the swelling processes depictive of stroke. As for other applications in the clinics regarding dMRI, it is mandatory to use a higher angular resolution like in diffusion tensor imaging (DTI) [1, 16] or high angular resolution diffusion imaging (HARDI) [17, 18]. The required time to implement these high angular resolution techniques renders it use prohibited in the clinics. Researchers in their side, have the flexibility to plan the scans and recruit the patients. They can also apply standard methods to all samples and care for repeatability. Thanks to this flexibility humanity has faced in detail, important aspects and the core of neurodegenerative diseases. However, a full understanding requires close monitoring of the dynamic apoptotic events which calls the attention over the small animal modeling field. Here, scaning time is a concern only for ethical reasons but, spatial resolution is so exigent that the quality gained by the wider scanning periods is lost in the little spin recruitment induced when reducing the voxels' size. Additionally, at grids of cents of micrometers as used in the small animal MRI field, a slight movement in the physical setup is traduced in huge displacements or rotations in the image. In this document, we present a method that enhances the quality of dMRI datasets. Here, we turn deteriorated diffusion images into meaningful ones. More importantly, the quality is improved with every new session which is of utmost relevance in horizontal studies. The method was developed and

[*]Fernando Yepes-Calderon, 1300 North Vermont Ave, Contact No 818.930.0521 & fyepes@chla.usc.edu

https://dx.doi.org/10.25046/aj0203125

thoroughly tested on programmatically created diffusion phantoms and translated to both, the small animals and the clinics.

2 Materials and Methods

The **Fig. 1** depicts the common dMRI pipeline under the DTI modeling.

Our driving hypothesis focuses on DTI and states that a diffusion study is more cost and time effective if an averaging is done in the tensor space rather than in the Diffusion Weighted Imaging (DWI) domain. Often, quality enhancement in dMRI is achieved by incrementing the number of machine averages which increments the acquisition time. We instead propose to use several images acquired in the low SNR regime and perform the average-enhancing strategy in the tensor domain. The averaging proposed here is not an obvious one. Recall that tensors are 3 by 3 matrices representing a diffusion geometry –all forms ellipsoids including the sphere– and consequently, a corrective action is not as simple as performing a linear operation over the axes of the images. To elaborate on this aspect, we summarize a theoretical framework that illustrates our rationale on a toy example using numerical phantoms.

2.1 Creating synthetic diffusion phantoms

Synthetic DWI signals can be created and manipulated to reproduce rotational structural misalignments in noise-free situations and with added Rician noise using the framework mentioned above. For demonstrative purposes, a synthetic-diffusion phantom is created from which, one diffusion voxel or a group of them (V) are selected. The V contains several diffusion patterns as shown in **Fig.** 2. Each of them is restricted to a cylinder of $\rho = 5\mu$ and length $L = 5mm$, where the particles diffuse at $1/1500\ mm^2/s$ as in [19]. This cylindrical model for diffusion mimics the preferential direction of water motion in the free space between cells, voxel by voxel. We define an initial reference signal resting along the x-axis (D_{ref}), see **Fig.** 2-**AI**, which is a compendium of diffusion patterns in each voxel that together create a preferred global diffusion as it would do a brain tract under dMRI. Then, using this model, we create other synthetic diffusion patterns that are rotated creating the desired misalignments, as in frames AII and AIII of **Fig.** 2.

Assume then that each diffusion pattern comes from a single fiber, like in fiber tracking [20]. Each fiber orientation is then governed by **Equation. 1**.

$$fib_i = \left[cos(\alpha * \frac{pi}{180}), sin(\alpha * \frac{pi}{180}), 0\right]. \quad (1)$$

Where the parameter α will be zero in the synthetic D_{ref} signal, and will receive any value in the range [0,90] whenever non-overlapping diffusion signals are simulated among acquisitions. Here, fib_i represents the i simulated pattern. **Fig.** 2-**Panel A** illustrates this procedure.

The diffusion coefficents in the V are derived by taking the log on both sides of the Sketjal-Tanner equation as in [21]:

$$\log(\hat{S}_i) = \log(S_o) - bg_i' D g_i \quad (2)$$

Where S_o is the non-sensitized signal, b is the b-value, g_i is the gradient direction, D is the diffusion factor, and $\hat{S}_i$ is the DWI signal for the g_i direction. The $\hat{S}_i$ factor can be decomposed into the ideal signal S_i and a Rician noise component, represented as two independent Gaussians N_1 and N_2 with a variance σ:

$$\hat{S}_i = \sqrt{[S_i + N_1(0,\sigma) + N_2(0,\sigma)}} \quad (3)$$

The diffusion tensors D_t are estimated in the Log-euclidean (LE) domain (where they are represented by the parameter L) using the D contributions from all gradient directions:

$$L = \log(D) \leftrightarrow \check{D} = \exp(\check{L}) \quad (4)$$

The operations in the LE domain are computationally inexpensive when compared with other manifold-based methods such as [22, 23, 24]. To see details of how the LE outperforms other common tensor estimation strategies; please refer to the document [21]. Tensors in the LE domain can be moved back to the original tensor space through **Equation 4**.

For the framework described in this section, rotations are forced to be in the x-y plane (about the z-axis). Hence, rotational corrections under ideal conditions are generated using $T_{rot} = RTR'$, where R can be represented as:

$$R = \begin{bmatrix} \cos\frac{\alpha*pi}{180} & -\sin\frac{\alpha*pi}{180} & 0 \\ \sin\frac{\alpha*pi}{180} & \cos\frac{\alpha*pi}{180} & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (5)$$

In the presence of noise, the diffusion pattern rotates randomly in any direction; thus, the correction matrices may have rotational components along the Z axis as well. Conveniently, R in **Equation 5** is dynamically created through registration in the Fractional Anistropy (FA) domain; hence, projections in Z-axis, in the case they are present, are also accounted.

The **Fig.** 2 illustrates the synthetic data that we generated to test our method.

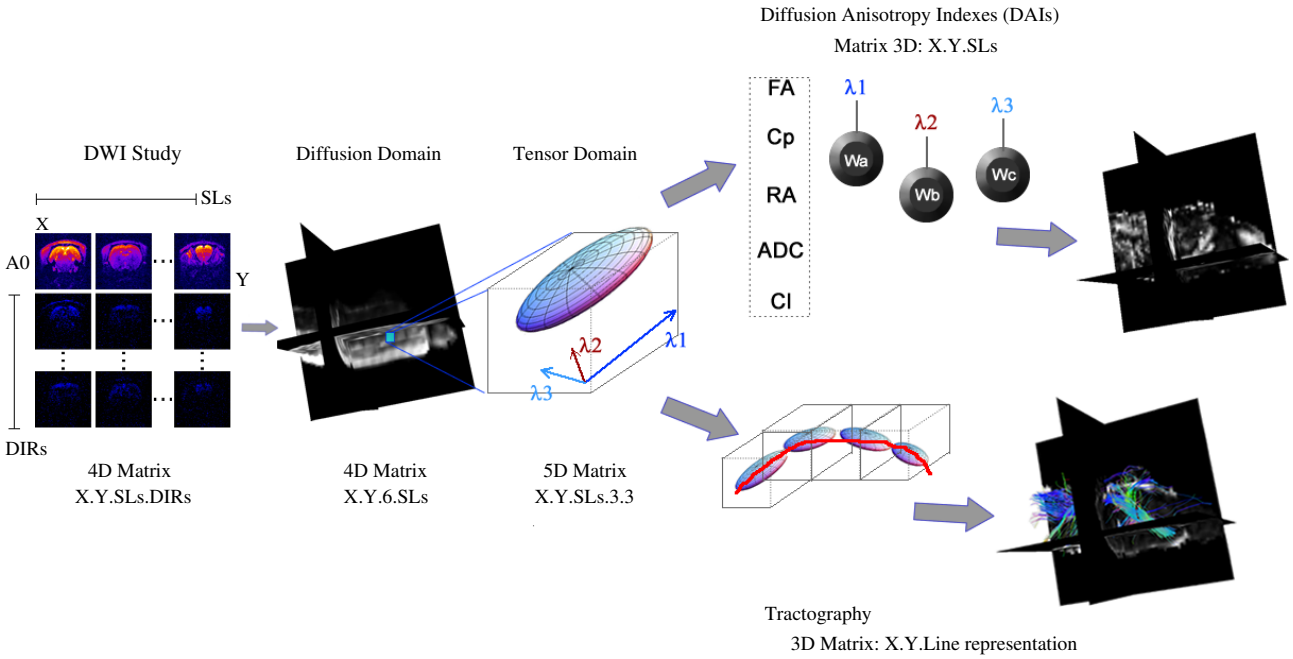


Figure 1: **Standard DTI pipeline**. In the dimensions of the data, SLs stands for slices and DIRs for gradient directions. Note how the data produces complex matrix formats along the whole path and only in the DAIs the outputs are scalars. From here the convinience to using a registration on the FA index fro correction purposes.

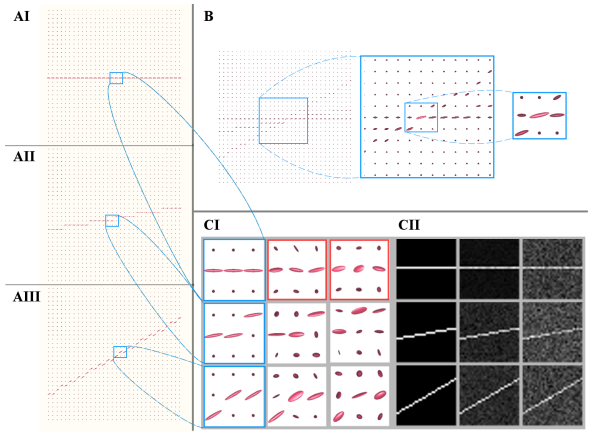


Figure 2: **Proof of concept synthetic data**

The panel A of **Fig. 2** shows ideal-synthetic-tensor fields: AI is the reference, AII and AIII are tensor fields simulating acquisition rotated 10 and 30 degrees with respect to the reference (AI). Panel B shows the tensor field that is generated when the three acquisitions (AI,AII, and AIII) are averaged in the DWI-domain. Panel C shows the same information in the tensor and FA domains. Some tensors of Panel A are zoomed-in and shown in the first column of CI. In Panel C, the columns simulate acquisitions with different levels of noise. In CI, the parameters of the tensors are as follows: rows 1, 2, and 3 represent tensors with 0, 10, and 30 degree rotations, respectively; columns 1, 2, and 3 represent tensors with 0, 0.1, and 0.2 sigma noise, respectively, with the exception of the tensors in the red boxes, in which the noise levels are 0.05 and 0.15 (left and right, respectively). Each panel of CI is the tensor field from which the FA is generated in the corresponding panel in CII.

In this proof of concept, a field of 32x32 pixels with a horizontal diffusion pattern is used as a reference. Its tensor field is extracted as shown in Panel AI. A set of rotations is induced in the tensor fields (see Panels AII and AIII) and Rician noise with different levels is added (as displayed in Panel CI). For our theoretical experiment, we generated six different datasets using a combination of angular rotations respect to the x-axis as well as various noise levels. The details of these datasets are shown in Table 1. The variance in the Rician noise of the noisy reference signals (Panel C, row 1, columns 2,3) is decreased by 0.5 to simulate the model acquisitions. The tensor fields are also used to generate FA maps for registration purposes, which generates the transformation matrices used to correct the misalignments in the tensor field. Hence, in summary, with this procedure, synthetic signals in the DWI domain, tensor fields, and FA maps are generated. The outcomes of these set of experiments will serve to demonstrate how the proposed method not only de-noises the images, but also correct structural misalignments at the tensor stage, enhancing general quality with each acquisition and, therefore, improving the results in every subsequent step.

2.2 Tensor averaging on small animal data

Small animals are often used to recreate pathogenic models. The idea of monitoring these models in time

Table 1: Angles of rotation and noise levels of 5 different datasets of synthesized signals. Sets 0 and 3 represent the ideal sets with no noise. Abbreviations: AoR (angles of rotation), σ (variance of Rician noise).

Set	Reference	1st additional acquistion	2nd additional Acquisition
Set 0	$AoR = 0°$, $\sigma = 0.00$	$AoR = 10°$, $\sigma = 0.00$	$AoR = 30°$, $\sigma = 0.00$
Set 1	$AoR = 0°$, $\sigma = 0.05$	$AoR = 10°$, $\sigma = 0.10$	$AoR = 30°$, $\sigma = 0.10$
Set 2	$AoR = 0°$, $\sigma = 0.15$	$AoR = 10°$, $\sigma = 0.20$	$AoR = 30°$, $\sigma = 0.20$
Set 3	$AoR = 0°$, $\sigma = 0.00$	$AoR = 0°$, $\sigma = 0.00$	$AoR = 0°$, $\sigma = 0.00$
Set 4	$AoR = 0°$, $\sigma = 0.05$	$AoR = 0°$, $\sigma = 0.10$	$AoR = 0°$, $\sigma = 0.10$
Set 5	$AoR = 0°$, $\sigma = 0.15$	$AoR = 0°$, $\sigma = 0.20$	$AoR = 0°$, $\sigma = 0.20$

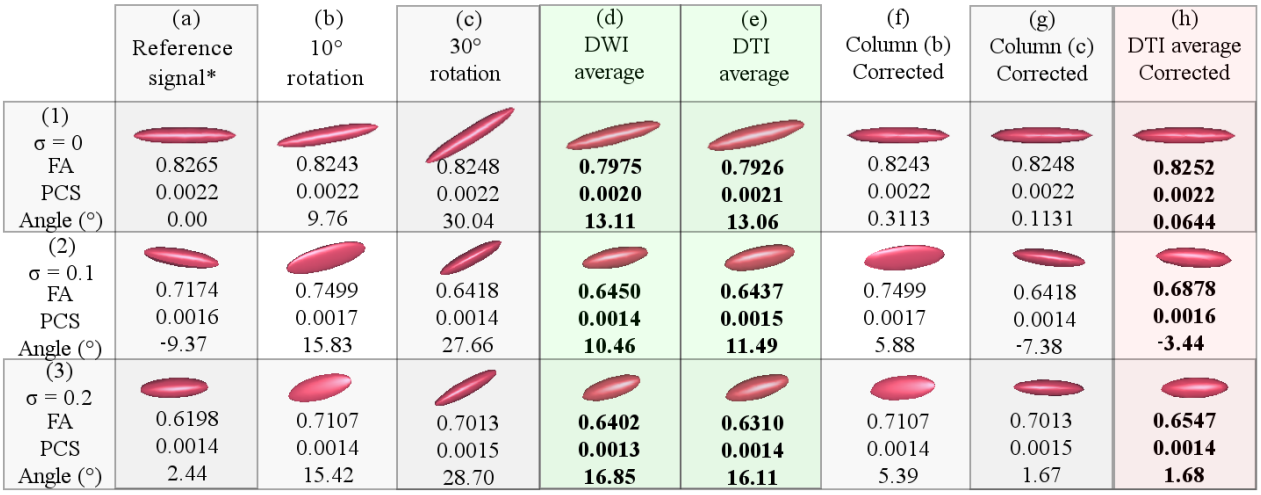


Figure 3: **Results of the theoretical experiment using the new pipeline on synthetic data with rotation induction.**

using non-intrusive means such as the MRI vanish in spite of the quality of images. Here we present a case of quality enhancement on the diffusion data of an animal model.

A mouse imaged with a 9.4 Tesla Bruker BioSpin MRI GmbH using the diffusion EPI sequence. The acquisition produced a matrix size of $106x160x88$ where each voxel has isometric side lengths of 0.08 mm, yielding a field of view of $8.5x12.8x7.0mm^3$. The diffusion parameters were set in 6 gradient directions with sensitizer b-value of $3000s/mm^2$ and one reference image without sensitizer (B0).

The mouse was scanned twice using the same protocol and once more using the DWI averaging strategy that comes built-in with scanners set in two. The added acquisition time of the two initial experiments equals the one utilized in the two DWI averaging acquisition creating a fair comparison scheme.

2.3 Templates on human neonates data

Templates are vastly used in human frameworks as references for structural assessments. Either T1 or T2 images are aligned and averaged to produce smooth models [25, 26, 27]. Recall that T1 and T2 images are 3D stacks while diffusion images are 4D stacks, from here the difficulty to apply the same aligning and averaging techniques in this kind of data. For this proof of concept, five brain DTI scans of term neonates with regular -not pathological- MRI scans images were acquired with a 1.5T GE scanner. The images are captured in a matrix of $256x256x22$ where each voxel represents a space of $0.70x0.70x4.99mm^3$; therefore, the field of view is of $180x180x110mm^3$. The diffusion parameters were set in 25 gradient directions with sensitizer b-value of $700s/mm^2$ and one reference image without sensitizer (B0). The images were acquired without DWI averaging.

2.4 Tracts extraction in Neonates

As a proof of implementation, in this section, we use the previously created template to map a limbic, some association and commissural fiber tracts in a neonate of 42 weeks conceptional age. Recall that these structures are undergoing the process of myelination at this

	(a) Reference signal*	(i) 1st additional acquisition	(j) 2nd additional acquisition	(d) DWI average	(e) DTI average	(k) Column (i) Corrected	(l) Column (j) Corrected	(h) DTI average Corrected
(1) σ = 0								
FA	0.8265	0.8265	0.8265	**0.8265**	**0.8265**	0.8265	0.8265	**0.8265**
PCS	0.0022	0.0022	0.0022	**0.0022**	**0.0022**	0.0022	0.0022	**0.0022**
Angle (°)	0.00	0.00	0.00	**0.00**	**0.00**	0.00	0.00	**0.00**
(2) σ = 0.1								
FA	0.7093	0.6648	0.6648	**0.6685**	**0.6654**	0.6648	0.6648	**0.6654**
PCS	0.0018	0.0014	0.0014	**0.0014**	**0.0016**	0.0014	0.0014	**0.0016**
Angle (°)	9.15	-5.24	-5.24	**3.35**	**0.03**	1.09	1.41	**4.25**
(3) σ = 0.2								
FA	0.8499	0.5473	0.6749	**0.6655**	**0.6861**	0.5473	0.6749	**0.6861**
PCS	0.0019	0.0012	0.0015	**0.0014**	**0.0015**	0.0012	0.0015	**0.0015**
Angle (°)	-8.88	2.58	-12.78	**-7.25**	**-8.06**	-1.79	-11.07	**-8.20**

Figure 4: Results of the theoretical experiment using the proposed pipeline on synthetic data with no rotation induction.

age according to [28, 29, 30, 31]. In this conditions, structural visualization under diffusion MRI methods is not only hardened by factors such as size and neuronal hyper-connectivity, but also by the lack of natural barriers that restrict the water displacement which is provided by the myelin sheets in adult brains. The structures shown in **Fig. 7** correspond to one of the subjects utilized to create the template.

3 Results

3.1 Synthetic data

With the synthetic material generated in **Section 2**, the analysis focuses on the behaviour of the center pixel, in which all the created signals converge (see Panel B in **Figure 2**). The eigenvalues and eigenvectors that describe the diffusion geometry are dependent on the accuracy of the overlays of concomitant diffusion patterns. The results of these experiments are shown in **Figures 3** and **4**. Recall that the transformations induced in the central voxel are also applied to the other voxels in the field of view; therefore, the outcomes of this exercise should be extrapolated to the whole image space.

All the references to columns in this paragraph belong to the **Fig. 3**. Column (a) contains the tensor of the reference DWI signal. Columns (b) and (c) hold tensors of the DWI signals that have been rotated 10°and 30°respectively, respect to the x-axis. Column (d) represents the tensor of the DWI signal that has been averaged using (a),(b) and (c) in the DWI domain without any prior rotation correction. Column (e) shows the same information as in (d), but all done in the tensor field. Column (f) and (g) contain the corrected tensors of Columns (b) and (c), respectively. Column (h) displays the result of averaging the tensors in Columns (a), (f) and (g).

All the references to columns in this paragraph belong to the **Fig. 4**. Column (a) contains the tensor of the reference DWI signal. Column (i) and (j) hold tensors of the two additional DWI signals that have been randomly generated. Column (d) represents the tensor of the DWI signal that has been averaged using (a),(i) and (j) in the DWI domain without any prior rotation correction. Column (e) is the same as (d) but all done in the tensor field. Column (k) and (l) hold the corrected tensors of Columns (i) and (j), respectively. Column (h) displays the tensor that has been averaged from the tensors in Columns (a),(k) and (l).

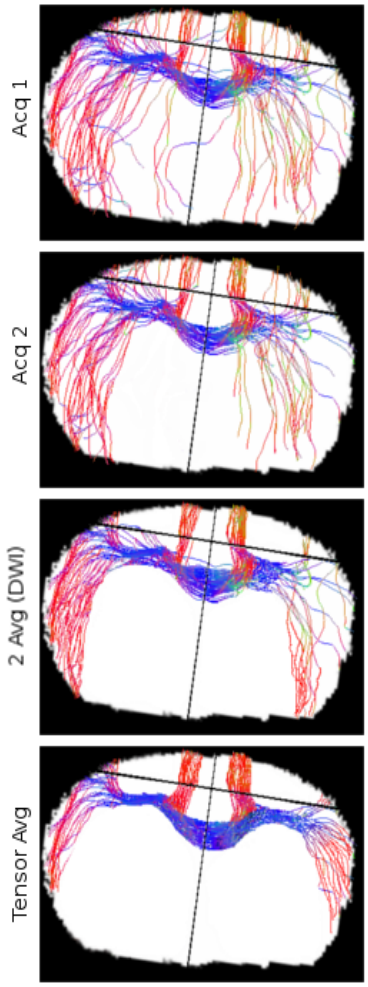


Figure 5: Diffusion image enhancement in small animal data. Tractography of the corpus callosum (CC) using a single region of interest that fully covers the structure in coronal view.

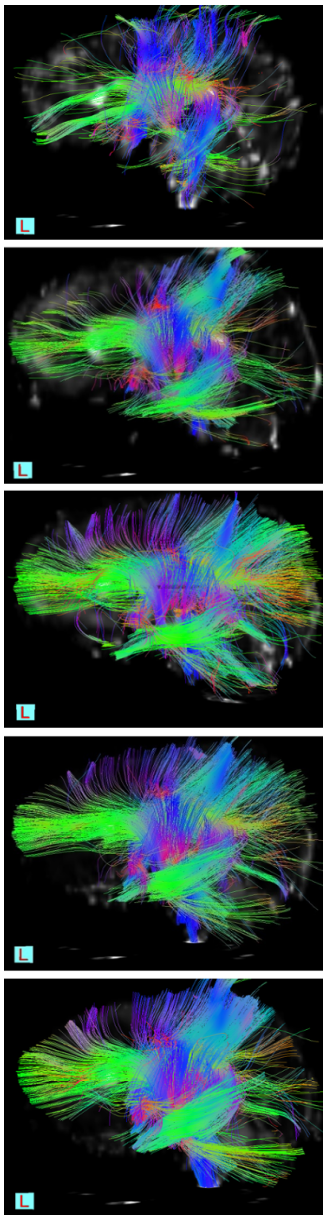


Figure 6: **Dynamic diffusion template for neonates** The tractography of the commissural and associating tracts. Using a single subject -on top- and adding one subject each time to end in a template of 5 subject at the bottom

For both Figures 3 and 4: Row (1) displays the dataset with ideal conditions (no noise), row (2) the one with $\sigma = 0.1$, and row (3) that with $\sigma = 0.2$. The provided Angle (°) is the nearest angle (in degrees) of the principle eigenvector with respect to the x-axis. Abbreviations: FA (fractional anisotropy); PCS (principal component); σ (variance of Rician noise). Also note that in sets marked with * the σ is 0.5 lower in the reference signals of the noisy datasets (row (2) and (3)). The columns with bold entries correspond to definitive results and should be use to focus the comparisons.

In the experiments with synthetic data, the proposed method always aligned the additional acquisitions to the model dataset, thus improving the overlap between structures and consequently cleaning the images while retaining the essential information. This statement remains true even in the data affected by noise. Note also how all the factors that affect the tractography are favorably adjusted.

3.2 Small animal data

The dMRI studies depicted in **Fig. 5** corresponds to only one animal scanned several times, replicating what it would happen in horizontal research.

The averaging is performed using the Acq 1 and Acq2. Note how the corpus callosum (CC) fiber are shown in blue, this is due to the position of the animal in the scanner. Recall that in humans, the coordinates are rotated 180° in the YZ plane; thus, fibers in the CC that connect the two brain hemisphere are usually shown in red.

3.3 Neonatal data

The infant data has been processed with the tensor averaging method after a rigorous re-sampling process to enable the standard space and the registration as well. The **Fig. 6** shows the progressive improvement with each newly added subject.

3.4 Tractography in neonatal datasets

With a clean template such as the one shown in **Fig. 6**, it is possible to map the tracts of a subject with highly compromised quality, as it is often the case in clinics and hospitals where the medical necessity and the comfort of the patient is prioritized over image clarity. Through non-linear registration in the factional anisotropy domain and by applying the obtained deformation fields to the tracts in the template, the tracts in the subject appear respecting the anatomy, see **Fig. 7**.

4 Dicussion

The Diffusion methods in the brain are meaningful if the extracted structures follow the neuronal trending that has been estimated through histology in postmorten samples. The tractography is an unbeatable mechanism to elucidate the structures created by neuron connectivity and thus, a visualizing method of utmost relevance for our purposes and those of any study based in dMRI. Once, the tractography is found consistent with the previous knowledge of the structure, one can start analyzing the DAIs under the tracts or out of them to enlight contrast. The results section is presented with evidence of tractography and improvement in this domain due to the relevance of this visualization mechanism in dMRI.

The proof of concept of **Section 2.1** consists of two stages. The first one is intended to demonstrate that set-up associated rotations can be dynamically corrected. In a second run, the same pipeline is used in a testing-set where no rotations are added; thus working solely with the random and unavoidable geometry modifiers introduced by noise. In both groups of experiments, ideal situations and two levels of noise are evaluated. With our method, the tractography, FA, principal components and angles are all kept within acceptable ranges of the ground truth. Tensor patterns are also closest to the original in (a), as it is evidenced if comparing columns (e),(f) and (h). Refer to **Figures 3** and **4**. Note also how the rotations do not greatly affect the diffusion geometry; the major contribution of our new pipeline consists in correcting the angle of the main principal components, therefore enabling the tractography in these datasets [32]. Also, note in **Figure 2 - Panel B**, how the averaging of tensor fields that are rotated about each other results in shrunk tensors. In contrast, our method obtains acceptable orientations and tensor sizes in spite of noise in the signal. When these specifications are moved to real data, the tractography is better defined with each new intake.

Regading the small animal data, rats and mice are preferably used for tracking pathologies associated with motor deficiencies. One big field of study is the ictus, which researchers induce in these animals. Cell swelling and apoptosis happen in the first 24 hours and diffusion normalizing processes happen along the week after the inducement. Also, consider that an adult rodent is used for these sort of experiments; thus, growing factors are not crucial. Neuron state and neuron connectivity changes, those are the ones that we need to capture. If the structures are aligned, as we proposed within our method, we do not only clean the images while averaging. Recall that as noise, any non-correlated data will be diminished. Analogously, all correlated data will be sustained. In this way neurodegenerative processes will become non-correlated information through acquisitions and thus will be manifested as disappearing structures while averaging. In a similar manner, proliferation processes will be non-correlated data along scannings, but this time, it will be displayed as appearing structures while averaging.

The above is a hypothesis and despite its veracity can be inferred from the simulations in **Section 2-A**, has not been demonstrated in the current work, overall because we do not have real data to support such as statements. What we have shown is that our method is more time-efficient that the machine averaging often used to produce better image quality. The message that we would like to transmit is stated as follows:

Using our tensor FA-based rotation strategy in n separated acquisitions that spend m minutes would yield better image quality than doing one acquisition lasting the same m minutes. This is very convenient also regarding scanning setup, recall that the animals should not be anesthetized more than 2 hours for ethical reasons. Our strategy presents a way out to this limitation which is a major drawback when working with alive subjects at such as exigent spatial resolution. This claim remains true also for the neonates field where the patient is priority and the scanning times are fastened to avoid displacement artifacts.

The procedure that we are performing does not modify the diffusion pattern (the lamdas are the same), it just re-accommodate the orientation and displacement in case present. As the diffusion anisotropy indexes (such as FA, Aparent Diffusion Coeficiente, Mean Diffusivity and so on) are rotationally independent, and we keep the lamdas unmodified, our procedure does not modify these values. Additionally, we would like to emphasize the importance of the tractography because all our efforts are focused on generating a good one. Once the structure of interest is well located and even when the rotations do not affect the DAIs – because of the facts given above –, you can always run the DAIs analysis on the voxels overlapped by the extracted tractography in the original data.

The method exposed in this manuscript has been successfully tested in a controlled environment be-

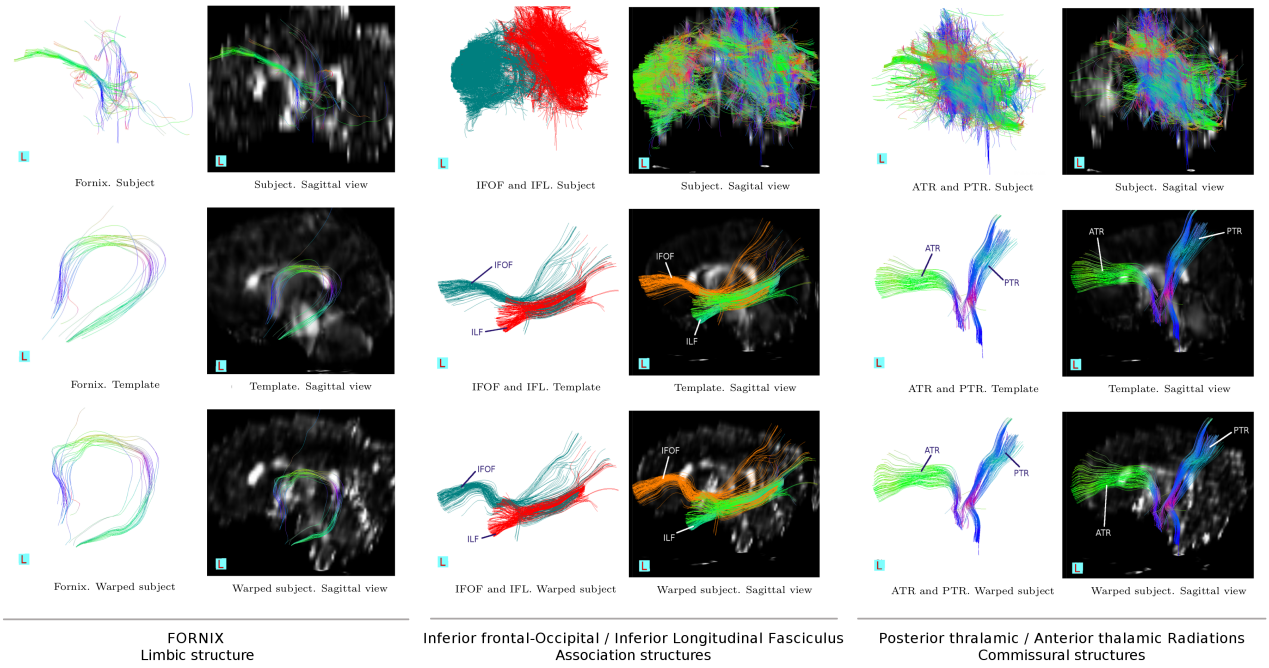


Figure 7: **Tractography in clinical datasets.** Limbic, association and commissural structures (left,center and right respectively) by standard methods (top row), in the template (center row) and in the studied clinical subject (bottom row). Each structure is presented in stand alone fashion and overlapped with its respective FA map of the brain. Top and bottom rows show, graphically, the significance of the proposed method.

fore being tested on the animal and also on clinical data. The biological variability that authors try to diminish by increasing the sample size is not an issue here because the noise and the misalignments are not produced by any metabolic or physiological reason. We are dealing with setup-related factors that are difficult to overcome due to the high resolution used in the case of the animals and also to procedimental disparities in the clinics. On the noise side, recall that making smaller voxels reduces the rate of spin recruitment while Brownian motion keeps the noise in constant levels. On the displacement side, note that we are trying to sharply locate the animal in our macro world when the scanning system is capturing the changes between 200 and 400 micrometers. As for the neonates, the diversity in the acqusition protocols and spatial resolutions is the reason for the lack of structural coherence among subjects.

In both sets of real data, the enhancement consists in better definition of structures and elimination of doubtful fibers. The better definition of paths is achieved when the diffusion information is highly correlated among acquisitions or subjects, not only reinforcing the correlated information but also weakening non-correlated one like noise. Another important aspect demonstrated in the procedure with synthetic data is the angle recovering of the tensor averaging method. The angle in which principal components of diffusion –the longest axis– coexists in neighboring voxels can create or eliminate fibers depending if this angle is smaller or bigger than 10°. The last statement applies if the fiber tracking is performed with the TensorLine algorithm [33] which is the case of the current implementation.

5 Conclusions

Diffusion MRI is a powerful tool that has not yet been implemented in full extend at the clinics due to the long lasting sessions required to achieve a good image quality. Regarding the small animal field, dMRI opens a myriad of possibilities but the spatial resolution is too exigent, and the quality of the images ends by discouraging the realization. If carefully observed, in both fields of application the quality is a concern and in MRI, quality is a matter of time. Here, we have introduced a post-scanning method that makes the inter-acquisition time to become a valued asset for image enhancing purposes. The method has been thoroughly tested in synthetic toys and successfully translated to real case scenarios. With all compelling evidence exposed in this document, we confirm that time in dMRI can be more efficiently utilized in the tensor space than in the DWI domain.

References

[1] P. J. Basser, J. Mattiello, and D. LeBihan, "Mr diffusion tensor spectroscopy and imaging," *Biophys. J.*, vol. 66, no. 1, pp. 259 – 267, 1994.

[2] D. L. Bihan, J.-F. Mangin, C. Poupon, C. A. Clark, S. Pappata, N. Molko, and H. Chabriat, "Diffusion tensor imaging: Concepts and applications," *JOURNAL OF MAGNETIC RESONANCE IMAGING*, no. 13, pp. 534–546, 2001.

[3] P. W. Schaefer, P. E. Grant, and R. G. Gonzalez, "Diffusion-weighted mr imaging of the brain.," *RSNA Radiology*, vol. 217, Nov 1999.

[4] M. Puligheddu, I. Laccu, G. Gioi, P. Congiu, M. Figorilli, and M. Peltz, "Diffusion tensor imaging tractography and mri perfusion in post traumatic brain injury hypersomnia," *Sleep Medicine*, 2017.

[5] J. Z. Dan Wu, "In vivo mapping of macroscopic neuronal projections in the mouse hippocampus using high-resolution diffusion mri," *NeuroImage*, vol. 125, pp. 84–93, 2016.

[6] A. Heemskerk, T. Sinha, K. Wilson, Z. Ding, and B. Damon, "Quantitative assessment of dti-based muscle fiber tracking and optimal tracking parameters.," *Magn Reson Med.*, vol. 2, pp. 467–72, Feb 2009.

[7] M. Froeling, J. Oudeman, G. Strijkers, M. Maas, M. Drost, K. Nicolay, and A. Nederveen, "Muscle changes detected with diffusion-tensor imaging after long-distance running.," *Radiology*, vol. 2, pp. 548–62, Feb 2015.

[8] S. Jiao, V. Gurevich, and J. A. Wolff, "Long-term correction of rat model of parkinson's disease by gene therapy," *Letters to Nature*, vol. 362, no. 1, pp. 450 – 453, 1993.

[9] J. M. Hall, K. A. E. Martens, C. C. Walton, C. O'Callaghan, P. E. Keller, S. J. Lewis, and A. A. Moustafa, "Diffusion alterations associated with parkinson's disease symptomatology: A review of the literature," *Parkinson and related disorders*, 2016.

[10] M. Ghareghani, A. Ghanbari, S. Dokoohaki, N. Farhadi, S. M. Hosseini, R. Mohammadi, and H. Sadeghi, "Methylprednisolone improves lactate metabolism through reduction of elevated serum lactate in rat model of multiple sclerosis," *Biomedicine & Pharmacotherapy*, vol. 84, pp. 1504–9, 2016.

[11] A. Weishaupt, D. Paulsen, S. Werner, N. Wolf, G. Kllner, H. Rbsamen-Schaeff, T. Hnig, T. Kerkau, and N. Beyersdorf, "The t cell-selective il-2 mutant aic284 mediates protection in a rat model of multiple sclerosis," *Journal of Neuroimmunology*, vol. 282, pp. 63–72, 2015.

[12] L. de Oliveira, D. Fraga, R. D. Luca, L. Canever, F. V. Ghedim, M. Matos, E. Streck, J. Quevedo, and A. Zugno, "Behavioral changes and mitochondrial dysfunction in a rat model of schizophrenia induced by ketamine," *Metabolic Brain Disease*, vol. 26, no. 1, pp. 69 – 77, 2010.

[13] C. Meehan, L. Harms, J. D. Frost, R. Barreto, J. Todd, U. Schall, C. S. Weickert, K. Zavitsanou, P. T. Michie, and D. M. Hodgson, "Effects of immune activation during early or late gestation on schizophrenia-related behaviour in adult rat offspring," *Brain, Behavior, and Immunity*, vol. 63, pp. 8–20, 2017.

[14] A. Kunimatsu, S. Aoki, Y. Masutani, O. Abe, H. Mori, and K. Ohtomo, "Three-dimensional white matter tractography by diffusion tensor imaging in ischaemic stroke involving the corticospinal tract," *Neuroradiology*, vol. 45, no. 8, pp. 532 – 535, 2003.

[15] Y. Yao, S. Zhang, X. Tang, S. Zhang, J. Shi, W. Zhu, and W. Zhu, "Intravoxel incoherent motion diffusion-weighted imaging in stroke patients: initial clinical experience," *Clinical Radiology*, vol. 71, p. 938.e11e938.e16, 2016.

[16] S. Sharma, J. Bekelman, A. Lin, J. N. Lukens, B. R. Roman, N. Mitra, and S. Swisher-McClure, "Clinical impact of prolonged diagnosis to treatment interval (dti) among patients with oropharyngeal squamous cell carcinoma," *Oral Oncology*, vol. 56, pp. 17–24, 2016.

[17] I. Jeffrey, R. Matthew, L. Blaskey, J. Christopher, and P. Timothy, "High angular resolution diffusion imaging (hardi) probabilistic tractography of the auditory radiation," *AJNR Am J Neuroradiol.*, vol. 34, pp. 1573–78, Aug 2013.

[18] X. Zhou, K. E. Sakaie, J. P. Debbins, J. E. Kirsch, C. Tatsuoka, R. J. Fox, and M. J. Lowe, "Quantitative quality assurance in a multicenter hardi clinical trial at 3t," *Magnetic Resonance Imaging*, vol. 35, pp. 81–90, 2017.

[19] A. Barmpoutis and B. Vemuri, "A unified framework for estimating diffusion tensors of any order with symmetric positive-definite constraints," *Proceedings of ISBI*.

[20] F.-C. Yeh, D. Timothy, Y. Wang, J. Fernndez-Miranda, and W.-Y. Tseng, "Deterministic diffusion fiber tracking improved by quantitative anisotropy," *PLOS-one*, vol. 34, Nov 2013.

[21] P. Fillard, X. Pennec, V. Arsigny, and N. Ayache, "Clinical dt-mri estimation, smoothing and fiber tracking with log-euclidean metrics," *IEEE Transactions on Medical Imaging*, vol. 26, no. 11, pp. 1472 – 1482, 2007.

[22] X. Pennec, P. Fillard, and N. Ayache, "A riemannian framework for tensor computing," *J. Comput. Vision*, vol. 66, pp. 41–66, Jan 2006.

[23] C. Lenglet, M. Rousson, and R. Deriche, "Statistics on the manifold of multivariate normal distributions: Theory and application to diffusion tensor mri processing," *J. Math. Image. Vis.*, vol. 25, pp. 423–444, Oct 2006.

[24] P. Batchelor, M. Moakher, D. Atkinson, F. Calamante, and A. Connelly, "A rigorous framework for diffusion tensor calculus," *Magn. Reson. Med.*, vol. 53, p. 221225, 2005.

[25] J. Talairach and P. Tournoux, "Co-planar stereotaxic atlas of the human brain: 3-dimensional proportional system - an approach to cerebral imaging," *Thieme Medical Publishers*, 1988.

[26] J. C. Mazziotta, A. W. Toga, A. Evans, P. Fox, and J. Lancaster, "A probablistic atlas of the human brain: Theory and rationale for its development," *NeuroImage*, pp. 89–101, 1995.

[27] L. Xie, J. B. Pluta, S. R. Das, L. E. Wisse, H. Wang, L. Mancuso, D. Kliot, B. B. Avants, S.-L. Ding, J. V. Manjn, D. A. Wolk, and P. A. Yushkevich, "Multi-template analysis of human perirhinal cortex in brain mri explicitly accounting for anatomical variability," *NeuroImage*, vol. 144, pp. 183–202, 2017.

[28] R. W. Oppenheim, "The neurotrophic theory and naturally occurring motoneuron death," *ViewPoint*, vol. 12, no. 7, 1989.

[29] B. Popko, "Myelin maintenance: Axonal support required," *Nature Neuroscience*, vol. 13, pp. 275 – 277, 2010.

[30] S. C. L. Deoni, E. Mercure, A. Blasi, D. Gasston, A. Thomson, M. Johnson, S. C. R. Williams, and D. G. M. Murphy, "Mapping Infant Brain Myelination with Magnetic Resonance Imaging," *J. Neurosci.*, vol. 31, no. 2, pp. 784 – 791, 2011.

[31] P. Yakovlev and I. Lecours, *The myelogenetic cycles of regional maturation of the brain*. Blackwell, 2011.

[32] D. M. Weinstein, G. L. Kindlmann, and E. C. Lundberg, "Tensorlines: Advection-diffusion based propagation through diffusion tensor fields," in *Proceedings of the 10th IEEE Visualization 1999 Conference (VIS '99)*, VISUALIZATION '99, (Washington, DC, USA), pp. –, IEEE Computer Society, 1999.

[33] D. weinstein, G. Kindlmann, and E. Lundber, "Tensorlines: advection-diffusion based propagation through diffusion tensor fields," *Proceedings in Conference of Visualization*, 1999.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 2, No. 3, 996-1005 (2017)
www.astesj.com
Special issue on Recent Advances in Engineering Systems

ASTES Journal
ISSN: 2415-6698

Ensemble of Neural Network Conditional Random Fields for Self-Paced Brain Computer Interfaces

Hossein Bashashati*, Rabab Kreidieh Ward

Electrical and Computer Engineering, University of British Columbia, Vancouver, Canada

ARTICLEINFO

Article history:
Received: 15 April, 2017
Accepted: 24 June, 2017
Online: 10 July, 2017

Keywords:
Brain Computer Interface
Conditional Random Fields
Neural Networks

ABSTRACT

Classification of EEG signals in self-paced Brain Computer Interfaces (BCI) is an extremely challenging task. The main difficulty stems from the fact that start time of a control task is not defined. Therefore it is imperative to exploit the characteristics of the EEG data to the extent possible. In sensory motor self-paced BCIs, while performing the mental task, the user's brain goes through several well-defined internal state changes. Applying appropriate classifiers that can capture these state changes and exploit the temporal correlation in EEG data can enhance the performance of the BCI. In this paper, we propose an ensemble learning approach for self-paced BCIs. We use Bayesian optimization to train several different classifiers on different parts of the BCI hyper-parameter space. We call each of these classifiers Neural Network Conditional Random Field (NNCRF). NNCRF is a combination of a neural network and conditional random field (CRF). As in the standard CRF, NNCRF is able to model the correlation between adjacent EEG samples. However, NNCRF can also model the nonlinear dependencies between the input and the output, which makes it more powerful than the standard CRF. We compare the performance of our algorithm to those of three popular sequence labeling algorithms (Hidden Markov Models, Hidden Markov Support Vector Machines and CRF), and to two classical classifiers (Logistic Regression and Support Vector Machines). The classifiers are compared for the two cases: when the ensemble learning approach is not used and when it is. The data used in our studies are those from the BCI competition IV and the SM2 dataset. We show that our algorithm is considerably superior to the other approaches in terms of the Area Under the Curve (AUC) of the BCI system.

1 Introduction

A brain computer interface (BCI) aims at detecting the presence of specific patterns in a person's brain activity. These patterns relate to the user's intention to control a device [1]. If such patterns are detected in the brain waves, then the BCI issues specific signals to control the device.

BCI systems can be classified into two categories: synchronous and self-paced [2]. In synchronous BCIs, the user controls the BCI output during specific short periods. Therefore, the user can only issue a control command when he/she is prompted to operate the system. Users of self-paced BCIs, on the other hand, can control the system whenever they wish. The periods during which a user is issuing a signal to control the system are called Control states, and those during which the user is not controlling the system are called No-Control (NC) states. During the NC states, the BCI system should be designed so it does not issue any control signal; otherwise a false positive output is produced.

Despite the much progress made in self-paced BCIs, they remain extremely challenging to design, compared with synchronous BCI systems. This is because (unlike the case of synchronous BCIs), a user of a self-paced BCI should be able to control the system at any time, but the BCI system does not have any

*Corresponding Author: Electrical and Computer Engineering, University of British Columbia, Vancouver, Canada, hosseinbs@ece.ubc.ca

https://dx.doi.org/10.25046/aj0203126

knowledge about the onset of any control command.

To address this property of self-paced BCIs (i.e. the onset detection of intentional control), one of the following two approaches can be used. The first approach exploits the properties of the electroencephalogram (EEG) data to design more advanced classification and/or feature extraction techniques (the focus of this paper). For instance in [3] and [4] the authors have used more advanced classifiers to design self-paced BCI. The second approach takes advantage of other cognitive tasks to detect the onset of the intentional control. For instance in [5] the authors have used high pitch and siren-like sounds to facilitate distinguishing control tasks from no-control states. In [6] the authors have used a combination of motor imagery and P300 potential to design self-paced BCIs.

The task of classifying brain signals in a self-paced BCI has been formulated as a sequential supervised learning problem [7]. In sequential learning, a sequence of observations and their corresponding labels are known, and the goal is to construct a classifier that predicts the sequence of the labels of a new sequence of observations.

The most popular way to obtain observation sequences for sequential supervised learning is to use a sliding window over the signal (these windows might overlap). This approach takes consecutive input windows of the brain signal (each of length w milliseconds), extracts features from each window and assigns a label to each window. The labels correspond to whether or not the user is issuing a command to operate the device. The sequence of the extracted feature vectors and their corresponding labels are used to train a classifier. For a new sequence, the trained classifier estimates the label of each window (i.e. the intention of the user).

The advantage of using the sliding window approach is that it converts the sequential supervised learning problem into a standard classification problem. Thus any classical supervised learning method can be used to solve the problem. The majority of the publications in the field have used this method to build different self-paced BCIs [8], [9], [10]. The disadvantage of the sliding window approach is that the sequential correlation of the labels of consecutive EEG windows is not exploited. Yet, the observations and the labels of nearby windows are usually related to each other. Thus there is a sequential correlation between adjacent EEG windows, and this knowledge could be used to improve the performance of the system.

In [11] the authors show that the brain goes through several well-defined internal state changes while the subject is carrying out a motor imagery mental task. Employing an algorithm that can model/exploit these state transitions can enhance the performance of the BCI. The class of sequence labeling algorithms are able to exploit the temporal structure of the EEG data. Among these algorithms, the most popular is the Hidden Markov Model (HMM) [12], which is a generative classifier. A generative classifier models the joint probability of observations and label sequences. Although the HMM classifier has been successful in synchronous BCIs, in [13] the authors concluded that the sliding window approach (i.e. using classical classifiers) is superior to HMM in self-paced BCIs. Due to intractability issues, HMMs assume the observations are independent given the states. This makes it difficult to incorporate knowledge about the structure of the EEG data into the model by extracting informative overlapping observations (the problem of over-counting evidence).

Another type of sequence labeling classifiers (besides HMM) are the discriminative ones. These classifiers directly maximize the conditional likelihood of the label sequence given the observations. These algorithms have yielded very promising results in the fields of natural language processing [14], and activity recognition [15] which are very similar in nature to the task of self-paced classification of BCI. The advantage of these models is that they give the user the freedom to extract many informative and overlapping features from the observation sequence [1]. These features might be extracted from previous windows in the brain signal and can be correlated.

In [16], [17] and [18], the authors applied discriminative sequence labeling algorithms to self-paced BCIs, to classify different motor imagery tasks. In this paper, we use a discriminative sequence labeling algorithms to discriminate between NC and control states. The challenge here is that during the control states, i.e. while the subject is performing the mental task, a set of well-defined state changes occur. However, during the NC states, the subject can be in any mental state, therefore finding specific patterns in the EEG signal during NC states is extremely difficult. In other words, discriminating between two predetermined mental tasks (e.g. right hand versus left hand movement) is much easier than discriminating a predetermined mental task (e.g. left hand movement) from NC states.

Our contributions in this paper are two-fold. The first is a new discriminative sequence labeling classifier for self-paced BCIs. Our approach combines the power of one of the most popular discriminative sequence labeling classifiers (Conditional Random Field) to exploit the correlation in consecutive EEG windows. It also utilizes a neural network to extract high-level features from the observations. We call our classifier Neural Network Conditional Random Field (NNCRF). We compare its performance to those of Hidden Markov Models (HMM), and two of the most popular discriminative sequence labeling classifiers i.e. Conditional Random Field (CRF) [19] and Structural Support Vector Machines (SSVM) [20]. We also compare the performance of these discriminative sequence labeling classifiers with two popu-

[1] These observations correspond to the features extracted from the raw EEG signal.

lar classical classifiers i.e. Support Vector Machines (SVM) [21] and Logistic Regression classifiers [22]. Our second contribution is discussed in the latter part of the paper, where we propose an ensemble learning approach to further improve the accuracy of the BCI system. We use Bayesian optimization [23] to train several diverse classifiers on different parts of the BCI hyper-parameter space. We evaluate the performance of different algorithms using the data from the BCI competition IV [24] and SM2 [25] dataset, and show that our algorithm is considerably superior to the existing approaches in terms of the Area Under the Curve (AUC) of the BCI system.

In the rest of this paper, in section 2, we describe the feature function in discriminative sequence labeling classifiers. In section 3, we introduce our proposed sequence labeling classifier i.e. NNCRF. In section 4 we describe our ensemble classification approach. In section 5, we explain the datasets, and in sections 6 and 7, the results of different algorithms are compared.

2 Feature Function in Discriminative Sequence Labeling

For sequence labeling algorithms, we assume that each data sample in the training set consists of a sequence of observations, [2] and its corresponding label sequence. Assuming the training set is $\{x_i, y_i\}_{i=1}^{N}$ where N is the number of training samples. x_i is the sequence of observations from K consecutive windows and y_i is the corresponding sequence of K labels. Every w milliseconds of the EEG creates a window. x_i is created by concatenating the observations from each of K consecutive windows i.e. $x_i = [x_{i1},...,x_{ik},...,x_{iK}]$ and x_{ik} corresponds to the observations from the kth window in the sequence number i [3]. Likewise, y_i is created by concatenating the labels of each of these K consecutive windows i.e. $y_i = [y_{i1},...,y_{ik},...,y_{iK}]$ and y_{ik} corresponds to the label of the kth window in the sequence number i.

The discriminative sequence labeling algorithms have the advantage of enabling a set of features to be designed based on the structure of the data. Each feature is a function of the joint observations and their labels i.e. $\Phi : X \times Y -> R$, where X is the observation space, Y is the label space and R is the real space. In classical classification the feature vector is built based on the observations (X) only. However, here the idea is to extract features from the joint observations (X) and label (Y) spaces. In this way, a feature function Φ measures the joint compatibility of x (x_i) and y (y_i) [4]. Although more complex types of features can be used, the feature functions we use here are inspired from first order HMMs.

The first feature function we used, is defined as

$$\phi_0^{\sigma} = I(y_k = \sigma),\ \sigma \in \Sigma, \tag{1}$$

where I represents the indicator function, and Σ corresponds to the set of possible label values (i.e. NC states and movement states). The above feature function is based on Y only.

The second set of features forms a vector $\phi_{k,1:L}$ with L dimensions that captures the relation between the observation vector and the kth label in the sequence. We are assuming the observation vector is of dimension L. Each element $\phi_{k,l}^{\sigma}$ of the vector $\phi_{k,1:L}^{\sigma}$ is defined as

$$\phi_{k,l}^{\sigma} = x_{k,l} I(y_k = \sigma),\ \sigma \in \Sigma, \tag{2}$$

where $x_{k,l}$ is the lth dimension of the observation vector.

The third type of features is defined as

$$\phi_{k(k-1)}^{\sigma_1,\sigma_2} = I(y_k = \sigma_1 \wedge y_{k-1} = \sigma_2),\ \sigma_1, \sigma_2 \in \Sigma, \tag{3}$$

where $\phi_{k(k-1)}^{\sigma_1,\sigma_2}$ captures the relation between the kth and $(k-1)^{th}$ labels in the sequence.

3 Neural Network Conditional Random Fields

In this section we first describe the standard Conditional Random Field (CRF), then we describe Neural Network CRF. CRF is one of the most popular discriminative sequence labeling classifiers. In general the linear chain Conditional Random Field (CRF) classifier models the posterior distribution of the form

$$\Pr(Y|X) = \frac{\exp(\sum_{j=1}^{J} \lambda_j \sum_{k=1}^{K} \Phi_j(y_{k-1}, y_k, x, k))}{Z(X)}, \tag{4}$$

where K is the sequence length, J is the number of features extracted from the joint observation labels pair, and Φ_j is the feature function. λ_js are the parameters of the model which are learned based on the training data. $Z(x)$ is the normalization factor (partition function):

$$Z(X) = \sum_{y'_1} \sum_{y'_2} ... \sum_{y'_K} \exp(\sum_{j=1}^{J} \lambda_j \sum_{k=1}^{K} \Phi_j(y'_{k-1}, y'_k, x, k)). \tag{5}$$

To calculate the partition function we use forward-backward algorithm which is a dynamic programming algorithm.

Using the above mentioned set of features (equations 1, 2 and 3) the posterior function will be of the

[2] These observations correspond to the features extracted from the raw EEG signal.

[3] In our case, x_{ik} corresponds to the band-power of the window in a specific frequency band.

[4] For the sake of clarity, we drop the index i in the remainder of the text

form

$$\Pr(y_{1:K}|x_{1:K}) = \frac{\exp\Big(\sum_{k=1}^{K}(b_{y_k} + W_{y_k,1:L}^T x_k) + \sum_{k=2}^{K} V_{y_{k-1},y_k}\Big)}{Z(X)}, \quad (6)$$

where b_{y_k}, $W_{y_k,1:L}$ and V_{y_{k-1},y_k} are the parameters of the posterior distribution and should be learned using the training dataset. The set of parameters b_{y_k}, $W_{y_k,1:L}$ and $V_{y_k,y_{k-1}}$ capture the importance of the first, second and third type of features respectively.

To learn the optimal values of the parameters, we minimize the L_2 regularized negative log likelihood

$$\min_{\lambda} \sum_{i=1}^{N} -Log\Pr(y_i|x_i) + C\lambda^T\lambda, \quad (7)$$

where the vector λ consists of the parameters of the model (i.e. b_{y_k}, $W_{y_k,1:L}$ and V_{y_{k-1},y_k}), and C is the regularization coefficient. To find the optimal value of the parameters, we use stochastic gradient descent algorithm. As the loss function is convex, the gradient descent algorithm converges to the global optimum of the loss function.

For the inference i.e. assigning a sequence of labels to x_{N+1}, we assign the most probable sequence as the predicted labels i.e.

$$y_{predicted} = \arg\max_{y} P(y|x_{N+1}). \quad (8)$$

A Viterbi [12] like algorithm is used to find the most probable sequence of labels.

The linearity of the exponent term in CRF makes this classifier to have less expressive power compared to the classifiers that exploit kernels. A good approach that makes these algorithms more powerful is to pass the data through a feed-forward neural network before applying the CRF part. Neural networks transform the observation vector into high level features which are then used as the input to the CRF. As a result, the exponent term in CRF (i.e. the exponent term in equation 6) becomes non-linear because of the several layers of non-linear activation functions that have been applied on the observation vector [26] [27]. This combination of a neural network followed by a CRF forms our proposed classifier which we denote by NNCRF (Neural Network Conditional Random Field).

The NNCRF can be viewed as a standard linear chain CRF that uses a high level representation of the observations. Therefore, the posterior function of NNCRF has the same form as in equation 6 except that the x_k terms are replaced with the $h^M(x_k)$ term which represents the output of a feed-forward neural networks with M layers.

$$\Pr(Y|X) = \frac{\exp(\sum_{k=1}^{K}(b_{y_k} + W_{y_k,1:L}^T h^{(M)}(x_k)_{y_k}) + \sum_{k=2}^{K} V_{y_{k-1},y_k})}{Z(X)}. \quad (9)$$

The output of the $(M)_{th}$ layer ($h^{(M)}$) is

$$h^{(M)}(x_k) = tanh(b^{(M-1)} + W^{(M-1)}h^{M-1}(x_k)), \quad (10)$$

where $b^{(M-1)}$ and $W^{(M-1)}$ are the weights of the $(M-1)^{th}$ layer, $h^{M-1}(x_k)$ is the output of the $(M-1)_{th}$ layer of the neural network, and $h^0(x_k) = x_k$. To avoid overfitting, the weights of the neural networks are the same for all observations x_k in a sequence i.e. the values of the weights do not depend on k.

Initialization of a neural networks is very crucial for this algorithm to converge to a good local optima. For initialization of the neural network, the value of each parameter is a sample taken from a uniform distribution $U[-\gamma, +\gamma]$ where

$$\gamma = \frac{\sqrt{6}}{\sqrt{size(M) + size(M-1)}}, \quad (11)$$

where $size(M)$ is the number of hidden units in the Mth layer of the neural networks.

The loss function is the same as the loss function of the standard CRF (Equation 7). The values of the parameters of the neural networks and of the CRF are jointly optimized using the back propagation algorithm. The learning rate of the stochastic gradient descent algorithm is adjusted using the bold driver approach. For inference the same algorithm as in standard linear chain CRF is used.

4 Ensemble of classifiers

In [28], the authors proposed an algorithm to customize the synchronous BCI based on the brain characteristics of each subject. Here, we adopt a similar approach to build an ensemble of classifiers for self-paced BCIs based on the brain characteristics of each subject.

The brain characteristics of each subject are captured in the form of some hyper-parameters of the BCI system. Hyper-parameters are the parameters of the BCI system which are selected before feature extraction and classification. We use Bayesian optimization which is an iterative algorithm to propose the values of the hyper-parameters. The proposed values of the hyper-parameters are used to train different classifiers on different parts of the input space. Then, we combine the results of the trained classifiers, using averaging, to form the output of the ensemble classifier.

Bayesian optimization [23] [29] optimizes an objective function g(x) over some bounded set X. The objective function does not have a closed form expression and its derivative is unknown. A Bayesian optimization algorithm sequentially constructs a probabilistic model for g(x) and then uses this model to select a candidate for the optimization task. In our case, the objective function is the cross-validation accuracy of the classifier. The input to the Bayesian optimization algorithm consists of the values of the hyper-parameters of the BCI system. Therefore the set X is the hyper-parameter space of the BCI system.

In this study, the hyper-parameter space of every classifier consists of the frequency ranges to filter the EEG signal, the selected channels (from which features are extracted), and the window length. For our proposed NNCRF, we have an additional hyper-parameter which is the number of neurons in the hidden layer of the neural networks. In our experiments, we use NNCRF with one hidden layer NN.

The first set of hyper-parameters are the frequency ranges to filter the EEG signal. In motor imagery BCIs, the event-related desynchronization (ERD) and the event-related synchronization (ERS) occur in the upper alpha and the lower beta rhythms of the brain [11]. Generally the range of the alpha and beta brain waves are $[8-12]Hz$ and $[16-24]Hz$ respectively. A popular approach for extracting features (i.e. building the observation vector) is to find these frequency ranges for each subject, then calculate the band-power of the brain signal in the given frequency ranges. Another approach is to apply a filter with a large bandwidth in the approximate range $\approx [4-35]Hz$ that includes both the alpha and the beta brain waves of each channel, then extract the band-power of the brain signal. But even when we apply the second approach it is better to customize the frequency range for each subject. Therefore, for frequency filtering, in each iteration of the Bayesian optimization, our algorithm chooses one of the two above-mentioned options. The first option is to apply a filter-bank with two blocks corresponding to alpha and beta frequencies of the brain on each channel. The second option is to apply a filter with a large bandwidth (in the range $[4-35]Hz$) that includes both alpha and beta brain waves of each channel.

The second hyper-parameter is the set of the channels used to extract features from. For channel selection, our algorithm selects the number of Common Spatial Pattern (CSP) [30] filters (N = 2, 4, 6). It also has the option of not applying CSP (i.e. uses all the channels without applying any spatial filtering).

The third hyper-parameter is the length w (milliseconds) of the window that we extract band-power features from. For all of the sequence labeling classifiers, we have only used the past two seconds of the EEG signal to build the observation vector. For sequence labeling classifiers, we build a chain of consecutive windows as our observation vector. The length of the chain is therefore $\frac{2}{|w|} \times 1000$. For classical classifiers we only look at the past w milliseconds of the data. As discussed above, for the NNCRF classifier we have a fourth hyper-parameter which is the number of neurons in the hidden layer of the neural networks.

The pseudo code of the ensemble of classifiers is given in Algorithm 1. At each iteration (t) of our algorithm, the optimizer suggests a new set of values (h_t) of the hyper-parameters, then a new classifier (L_t) is built based on these values. Based on the cross-validation accuracy of L_t at iteration t, another set of values of the hyper-parameters is suggested for iteration $(t+1)$ and this process continues for at most MAX iterations or when the cross-validation accuracy plateaus. After running the optimization, we have T classifiers, each trained on a different subset of the hyper-parameter space.

The outputs of the classifiers are combined into a single prediction rule which is much more accurate than the individual classifiers. We use averaging to create the output of the ensemble of classifiers. In general, we seek a model with low bias and high variance as our final classifier. When the training data is small and the classifier has high variance, and by averaging we reduce the variance of the final classifier while preserving the low bias of a single model. Especially in the case of neural networks which can get stuck in a local optimum and has a high variance, creating an ensemble of classifiers i.e. neural networks which are trained on different parts of the hyper-parameter space, may result in a better approximation of the best possible classifier.

1. t := 1;
repeat
   2. Increment t;
   3. Find candidate (h_t) from the hyper-parameter space using Bayesian optimization;
   4. Create an observation matrix X_t from the brain signals using h_t;
   5. Train a new Classifier (L_t) using observation matrix X_t;
until *t reaches maximum number of iterations (MAX) or until convergence*;
6. Combine $L_1 ... L_T$ using averaging to build an ensemble classifier;

Algorithm 1: The pseudo-code of the proposed algorithm

5 Datasets

To perform the experiments two self-paced sensory motor BCI datasets have been used. The first dataset, SM2 [25], was collected from 4 subjects attempting to activate a switch by performing a right index finger movement. At random intervals, a cue was displayed for the subjects. The subjects attempted to activate a switch by moving their right index finger after the cue appeared. The EEG was recorded from 10 channels positioned over the supplementary motor area and the primary motor cortex (i.e. FC1-4, FCz, C1-4, Cz).

The second dataset, BCICIV2a, is the dataset IIa from the BCI competition IV which is recorded from 9 subjects performing 4-class motor imagery (left hand and right hand, both feet and tongue imagery movements) tasks. The data consists of 19 channels along the scalp and recorded in a synchronous paradigm. We have treated this dataset as a self-paced BCI dataset. In other words, to evaluate the performance of the classifiers on this dataset the time of transition from previous mental task to the new one (the time cue was displayed) has not been used. We have also converted the problem into a binary classi-

fication task i.e. separating movement imagery from NC states. All 4-classes of motor-imagery are considered as movement and the periods in which the subject did not control the system are considered as the No-Control class.

6 Results and Discussion

In our experiments, we first compared the performance of our proposed NNCRF with the standard CRF classifier, Hidden Markov Support Vector Machines (HMSVM), Hidden Markov Models (HMM) and two popular classical classifiers i.e. Logistic Regression and Support Vector Machines (SVM).

A Hidden Markov Support Vector Machine [31] is a special case of Structural Support Vector Machines in which the features are designed to capture the sequential nature of the data. The set of feature vectors we used, capture the dependency between consecutive labels in a sequence (y_k and y_{k-1}), and measure the relation between the observation in the kth window (x_k) and its corresponding label y_k in a sequence.

To apply HMMs on self-paced BCIs, we trained different HMMs for different mental tasks. We trained two different HMMs, one was trained using NC data and the other one using the samples of movement (intentional control) task. Then the likelihood of a new given sequence is calculated using the forward-back algorithm, and the classification is performed by comparing the likelihoods of different HMMs. In this case, each HMM focuses on learning the structure of the mental task that it is trained on, rather than learning to discriminate between different tasks.

The emission function used for HMM is a mixture of Gaussian distributions. The parameters of the HMM include the transition probabilities, and the parameters of the mixture of Gaussian distributions. The parameters are learned by maximizing the likelihood of the training dataset. The Baum-Weltch algorithm is used to learn the parameters of the HMM model.

To select the value of the hyper-parameters for the first part of our experiments, we searched through a manually specified set of values for the hyper-parameters (default values). For frequency filtering, we applied a filter-bank with two blocks in ranges [8-12]Hz and [16-24]Hz corresponding to the typical alpha and beta frequencies of the brain. For spatial filtering, we tried CSP with two, four and six filters. The values of these hyper-parameters along with each classifier's parameters are adjusted jointly using five-fold cross-validation. The parameters with the best mean cross-validation accuracy were used to train a classifier on the training set. We used Area Under the Curve (AUC) to evaluate the performance of the classifiers on the test dataset.

In our experiments, the band power of the EEG signal is used as the extracted features (observations) for the classification phase. The window length was equal to the sampling rate of the dataset and we used the last two seconds of the data to perform the classification in the test phase.

In the second set of experiments, we used the approach explained in section 4 to create an ensemble of classifiers. For all classifiers except NNCRF, we have created the ensemble using different values of the BCI hyper-parameters as explained in section 4. For NNCRF, along with the BCI hyper-parameters, we have also used different values of the number of neurons in the hidden layer of the NNCRF algorithm to create the ensemble. The Bayesian optimization algorithm ran for at most 50 iterations or until cross-validation accuracy plateaued. We repeated our experiments five times to reduce the effect of random seed. The average number of iterations of the Bayesian optimization algorithm for all different classifiers was 27.

Table 1 shows the results of comparing different classifiers when we used an ensemble of classifiers. The columns with the label "Default" correspond to the results of not using the ensemble approach. In Table 1, the columns with the "Ensemble" label correspond to the results of using an ensemble of each of the classifiers. The results shown in Table 1 are obtained by evaluating the classifiers on the independent test dataset of each subject. Figure 1 (which is a summary of Table 1) shows the average AUC of different algorithms when we used Bayesian optimization compared to using the default value of the hyper-parameters.

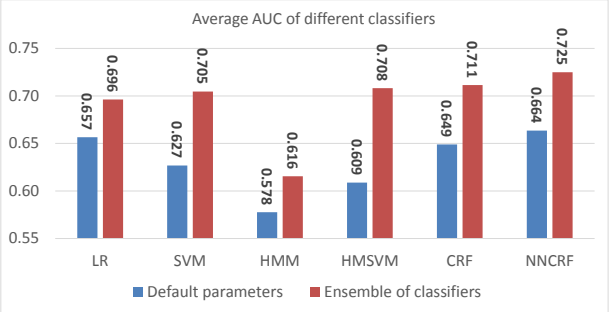


Figure 1: Average AUC of different classifiers across all subjects. The blue bars correspond to the average AUC when the default value of the hyper-parameters have been used. The red bars correspond to the average AUC after using our ensemble learning algorithm. The numbers on top of each bar correspond to the AUC of each algorithm.

Qualitative comparison of different algorithms in Tables 1 suggests the following: 1) our proposed ensemble learning approach considerably improves the performance of any classifier for almost all subjects, 2) HMM is the worst performing classifier, 3) NNCRF outperforms other algorithms in terms of average AUC across all subjects on the test dataset, and 4) interestingly, for some subjects none of the discriminative sequence labeling classifiers had a good performance - this means that these classifiers are not able to capture the temporal structure of the signal in these subjects.

Table 1: The results of comparing different algorithms on the test dataset. Default is the results of using the default values of the hyper-parameters. Ensemble corresponds to the results of the ensemble learning approach. Highlighted cells show the algorithm for which the performance is the best.

	LR		SVM		HMM		HSVM		CRF		NNCRF	
Subject	Default	Ensemble	Default	Ensemble	Default	Ensemble	Default	Ensemble	Default	Ensemble	Default	Ensemble
1	0.617162	0.708 ±0.003	0.471583	0.684 ±0.008	0.4704	0.572 ±0.003	0.566726	0.682 ±0.004	0.554054	0.651 ±0.006	0.565342	0.653 ±0.004
2	0.605175	0.663 ±0.004	0.534574	0.668 ±0.002	0.577131	0.647 ±0.014	0.602029	0.647 ±0.003	0.640085	0.66 ±0.004	0.681072	0.658 ±0.004
3	0.66646	0.709 ±0.003	0.659264	0.721 ±0.004	0.64575	0.688 ±0.003	0.642878	0.724 ±0.001	0.689771	0.735 ±0.002	0.691368	0.743 ±0.004
4	0.672782	0.769 ±0.002	0.618411	0.758 ±0.003	0.556342	0.571 ±0.004	0.592601	0.749 ±0.004	0.594204	0.725 ±0.006	0.540245	0.722 ±0.003
5	0.559119	0.629 ±0.003	0.559787	0.622 ±0.007	0.510666	0.509 ±0.003	0.509916	0.655 ±0.003	0.473098	0.618 ±0.006	0.493071	0.632 ±0.004
6	0.67886	0.692 ±0.004	0.652671	0.729 ±0.003	0.484071	0.63 ±0.015	0.629371	0.718 ±0.002	0.692138	0.731 ±0.005	0.700657	0.73 ±0.006
7	0.725141	0.762 ±0.002	0.710272	0.778 ±0.002	0.605163	0.71 ±0.01	0.68268	0.802 ±0.002	0.671861	0.751 ±0.003	0.666293	0.755 ±0.004
8	0.708607	0.729 ±0.002	0.706152	0.736 ±0.002	0.508587	0.633 ±0.034	0.541939	0.658 ±0.013	0.629435	0.697 ±0.011	0.717994	0.779 ±0.004
9	0.616434	0.626 ±0.006	0.606463	0.638 ±0.007	0.527084	0.598 ±0.013	0.564801	0.687 ±0.009	0.591498	0.669 ±0.005	0.588181	0.686 ±0.003
KT	0.766713	0.826 ±0.004	0.722982	0.818 ±0.002	0.711457	0.732 ±0.006	0.70759	0.827 ±0.004	0.789081	0.871 ±0.005	0.839713	0.881 ±0.004
CS	0.723935	0.733 ±0.004	0.727856	0.763 ±0.004	0.76225	0.637 ±0.01	0.706031	0.777 ±0.003	0.828081	0.831 ±0.004	0.797626	0.833 ±0.006
CB	0.610225	0.597 ±0.002	0.604169	0.6 ±0.003	0.547909	0.554 ±0.006	0.579983	0.617 ±0.003	0.615299	0.626 ±0.005	0.637969	0.631 ±0.003
ID	0.585436	0.605 ±0.002	0.573955	0.642 ±0.008	0.603535	0.518 ±0.004	0.588246	0.667 ±0.007	0.668705	0.682 ±0.011	0.706668	0.723±0.013
AVERAGE	0.657	0.696	0.627	0.705	0.578	0.616	0.609	0.708	0.649	0.711	0.664	0.725

To statistically compare the performance of the different classification methods, the Friedman statistical test was performed. The Friedman test [32] is a non-parametric statistical test, which ranks different classifiers for each subject separately. It then averages the ranks over all subjects. The algorithm with the lowest rank is the best performing one. Figure 2 shows the average rank of different classifiers. The best performing classifier is the ensemble of NNCRFs. In all classification algorithms, using the ensemble approach considerably improves the average rank.

In the Friedman test, the null hypothesis assumes that all algorithms have the same performance (thus, they have the same rank). After performing the Friedman test, the p-value was 4.34E-11. α was chosen to be 0.05. This p-value is low enough to reject the null hypothesis, therefore we conclude that the difference between algorithms is not random. Another set of statistical tests are performed to identify which algorithms are the source of difference. We conduct the Holm's [32] test as the post-hoc statistical test. In the post-hoc test, we perform pairwise comparison of all the other classifiers versus the best classifier (i.e. ensemble of NNCRFs) in terms of the average rank.

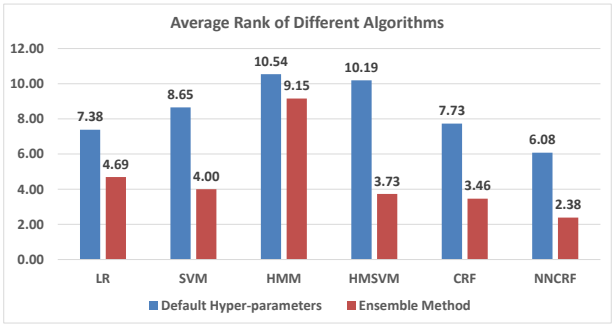


Figure 2: Average rank of different classifiers across all subjects. The blue bars correspond to the average rank when the default value of the hyper-parameters have been used. The red bars correspond to the average rank after using our ensemble learning algorithm.

	Hypothesis	P-value
11	HMM vs. NNCRF(ensemble)	8.13E-09
10	HMSVM vs. NNCRF(ensemble)	3.37E-08
9	HMM(Ensemble) vs. NNCRF(ensemble)	1.70E-06
8	SVM vs. NNCRF(ensemble)	9.29E-06
7	CRF vs. NNCRF(ensemble)	1.57E-04
6	LR vs. NNCRF(ensemble)	4.07E-04
5	NNCRF vs. NNCRF(ensemble)	9.03E-03
4	LR(Ensemble) vs. NNCRF(ensemble)	1.03E-01
3	SVM(Ensemble) vs. NNCRF(ensemble)	2.53E-01
2	HMSVM(Ensemble) vs. NNCRF(ensemble)	3.41E-01
1	CRF(Ensemble) vs. NNCRF(ensemble)	4.46E-01

Table 2: P-values corresponding to pairwise comparison of different classifiers versus the best performing classifier. α is chosen to be 0.05. All hypothesis with p-value less than 0.001 are rejected.

In this statistical test, each null hypothesis states that the best classifier and the other classifier have the same mean rank. The p-values corresponding to pairwise comparison of classifiers are shown in Table **??**. According to the Holms test results, the null hypothesis 1 through 5 are not rejected. This means that the difference between the ensemble of NNCRFs and other ensemble algorithms is not significant. However, the statistical tests show that the ensemble of NNCRFs is significantly better than non-ensemble methods.

7 Conclusion

In this study, we proposed a discriminative sequence labeling algorithm (classifier), to capture the dynamics of the EEG signal. We evaluated the performance of our algorithm (which we denote as NNCRF) on two self-paced BCI datasets and showed that it is superior, compared to classical classifiers and sequence labeling classifiers. NNCRF is a combination of a neural network and a CRF classifier. We demonstrated that CRF and NNCRF can capture the temporal properties of the EEG signal and improve the accuracy of the BCI in most of the subjects. In some subjects however, the temporal structure of the EEG data is difficult to capture by these classifiers. The neural network part of NNCRF converts the original observation vector into a new representation which is then fed to the CRF part. The non-linear transformation (by the neural network) of the observation vector helps the CRF part to easily discriminate between the different control and NC.

Overall, the reason for the poor performance of classical approaches is that they do not exploit the dynamics of the EEG signal. As for the HMM, although this algorithm models the temporal correlations in an EEG signal, its poor performance stems from the fact that it focuses on learning each mental task separately (i.e. without learning to discriminate between them). On the other hand, discriminative sequence labeling classifiers do not only model the temporal properties of each mental task, they also model the transition from one metal task to another (e.g. transition from movement to NC state).

We also showed that using an ensemble of classifiers that have been trained on different parts of the BCI hyper-parameter space can further improve the performance. We used Bayesian optimization to find the different values of the BCI hyper-parameters. Selecting different values for the hyper-parameters exposes each individual classifier to different parts of the BCI hyper-parameter space. We believe that this diversification is the key to the superiority of using the ensemble of classifiers. The performance of each individual classifier is very sensitive to the choice of the hyper-parameters of the BCI and using an ensemble of classifiers can decrease the variance of the final classifier.

The best performing algorithm was the ensemble of NNCRF classifiers both in terms of the aver-

age rank and average AUC of the final classifier. We statistically compared the performance of different classifiers to the ensemble of NNCRFs. The results showed, for SVM, LR, HMSVM and CRF, the ensemble learning approach results in significantly better performance compared to the single classifier with default value of the hyper-parameters. For NNCRF, the Holm's statistical test showed that the difference between NNCRF and the ensemble of NNCRF was not significant. However, the performance of the ensemble of NNCRFs was considerably better than the single NNCRF with default values of the hyper-parameters.

In this study, we used linear chain discriminative sequence labeling classifiers, i.e. the type of the feature functions used in this study were all inspired by Hidden Markov Models. So they are all local in nature, with each feature function only depending on the current or the previous label in the sequence. It is possible to use global features such as the ones that capture higher orders of dependency of the transition between consecutive labels, or feature functions that capture dependency between the EEG signal and labels of the EEG signal from distant past. These types of feature functions have the ability to capture more complex structures in the data and increase the power of each individual classifier.

Appendix

A. Block Diagram of the Learning

Figure 3 shows a simple block diagram of one iteration of our ensemble learning algorithm (which is explained in Section 4). In our approach the BCI is like a self-regulating system which improves itself based on the feedback that it receives from the classification block. The cross-validation accuracy of the classifier is given to the Bayesian optimization block, and this block proposes new values for the hyper-parameters for next iteration of the algorithm. The hyper-parameter values are used to extract features from the EEG signal, and a new classifier is trained using the features extracted from the training data. The extracted features are the power values in the frequency bands proposed by the Bayesian optimization block.

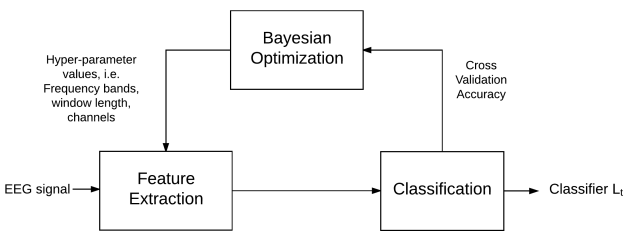


Figure 3: A simple block diagram of one iteration of our proposed ensemble learning algorithm.

Each iteration of our ensemble learning approach generates a classifier (L_t). After running the algorithm for T iterations, we will have T different classifiers which are trained on different parts of the BCI hyper-parameter space. Eventually, the classifiers $(L_1, L_2, ..., L_T)$ are combined (using averaging) to build a final classifier which is evaluated on the unseen test dataset.

Acknowledgments

This work was made possible by NPRP grant 7-684-1-127 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

References

[1] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, T. M. Vaughan, Brain–computer interfaces for communication and control, Clinical neurophysiology 113 (6) (2002) 767–791.

[2] A. Bashashati, R. K. Ward, G. E. Birch, Towards development of a 3-state self-paced brain-computer interface, Computational intelligence and neuroscience 2007.

[3] J. F. D. Saa, M. etin, A latent discriminative model-based approach for classification of imaginary motor tasks from eeg data, Journal of Neural Engineering 9 (2) (2012) 026020. URL http://stacks.iop.org/1741-2552/9/i=2/a=026020

[4] L. F. Nicolas-Alonso, R. Corralejo, J. Gomez-Pilar, D. lvarez, R. Hornero, Adaptive stacked generalization for multiclass motor imagery-based brain computer interfaces, IEEE Transactions on Neural Systems and Rehabilitation Engineering 23 (4) (2015) 702–712. doi:10.1109/TNSRE.2015.2398573.

[5] Y. Song, F. Sepulveda, A novel onset detection technique for braincomputer interfaces using sound-production related cognitive tasks in simulated-online system, Journal of Neural Engineering 14 (1) (2017) 016019. URL http://stacks.iop.org/1741-2552/14/i=1/a=016019

[6] Y. Yu, J. Jiang, Z. Zhou, E. Yin, Y. Liu, J. Wang, N. Zhang, D. Hu, A self-paced brain-computer interface speller by combining motor imagery and p300 potential, in: 2016 8th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), Vol. 02, 2016, pp. 160–163. doi:10.1109/IHMSC.2016.80.

[7] T. G. Dietterich, Machine learning for sequential data: A review, in: Structural, syntactic, and statistical pattern recognition, Springer, 2002, pp. 15–30.

[8] A. Bashashati, M. Fatourechi, R. K. Ward, G. E. Birch, A survey of signal processing algorithms in brain–computer interfaces based on electrical brain signals, Journal of Neural engineering 4 (2) (2007) R32.

[9] C. S. L. Tsui, J. Q. Gan, S. J. Roberts, A self-paced brain–computer interface for controlling a robot simulator: an online event labelling paradigm and an extended kalman filter based algorithm for online training, Medical & biological engineering & computing 47 (3) (2009) 257–265.

[10] M. Fatourechi, R. Ward, G. Birch, A self-paced brain–computer interface system with a low false positive rate, Journal of neural engineering 5 (1) (2007) 9.

[11] G. Pfurtscheller, F. H. Lopes da Silva, Event-related eeg/meg synchronization and desynchronization: basic principles, Clinical neurophysiology 110 (11) (1999) 1842–1857.

[12] L. R. Rabiner, A tutorial on hidden markov models and selected applications in speech recognition, Proceedings of the IEEE 77 (2) (1989) 257–286.

[13] S. Chiappa, N. Donckers, S. Bengio, F. Vrins, Hmm and iohmm modeling of eeg rhythms for asynchronous bci systems, in: ESANN, 2004.

[14] N. Nguyen, Y. Guo, Comparisons of sequence labeling algorithms and extensions, in: Proceedings of the 24th international conference on Machine learning, ACM, 2007, pp. 681–688.

[15] D. L. Vail, M. M. Veloso, J. D. Lafferty, Conditional random fields for activity recognition, in: Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems, ACM, 2007, p. 235.

[16] H. Bashashati, R. K. Ward, A. Bashashati, Hidden markov support vector machines for self-paced brain computer interfaces, in: 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA), 2015, pp. 382–385. doi:10.1109/ICMLA.2015.178.

[17] B. A. S. Hasan, J. Q. Gan, Conditional random fields as classifiers for three-class motor-imagery braincomputer interfaces, Journal of Neural Engineering 8 (2) (2011) 025013.
URL http://stacks.iop.org/1741-2552/8/i=2/a=025013

[18] D. Saa, F. Jaime, M. Cetin, Discriminative methods for classification of asynchronous imaginary motor tasks from eeg data, Neural Systems and Rehabilitation Engineering, IEEE Transactions on 21 (5) (2013) 716–724.

[19] J. Lafferty, A. McCallum, F. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data, in: Proceedings of the eighteenth international conference on machine learning, ICML, Vol. 1, pp. 282–289.

[20] I. Tsochantaridis, T. Hofmann, T. Joachims, Y. Altun, Support vector machine learning for interdependent and structured output spaces, in: Proceedings of the twenty-first international conference on Machine learning, ACM, 2004, p. 104.

[21] C.-W. Hsu, C.-C. Chang, C.-J. Lin, et al., A practical guide to support vector classification.

[22] C. M. Bishop, Pattern recognition and machine learning, Vol. 1, springer New York, 2006.

[23] J. Snoek, H. Larochelle, R. P. Adams, Practical bayesian optimization of machine learning algorithms, in: Advances in Neural Information Processing Systems, 2012, pp. 2951–2959.

[24] M. Tangermann, K.-R. Mller, A. Aertsen, N. Birbaumer, C. Braun, C. Brunner, R. Leeb, C. Mehring, K. Miller, G. Mueller-Putz, G. Nolte, G. Pfurtscheller, H. Preissl, G. Schalk, A. Schlgl, C. Vidaurre, S. Waldert, B. Blankertz, Review of the bci competition iv, Frontiers in Neuroscience 6 (2012) 55. doi:10.3389/fnins.2012.00055.
URL http://journal.frontiersin.org/article/10.3389/fnins.2012.00055

[25] J. F. Borisoff, S. G. Mason, A. Bashashati, G. E. Birch, Brain-computer interface design for asynchronous control applications: improvements to the lf-asd asynchronous brain switch, Biomedical Engineering, IEEE Transactions on 51 (6) (2004) 985–992.

[26] T. Do, T. Arti, et al., Neural conditional random fields, in: International Conference on Artificial Intelligence and Statistics, 2010, pp. 177–184.

[27] J. Peng, L. Bo, J. Xu, Conditional neural fields, in: Advances in neural information processing systems, 2009, pp. 1419–1427.

[28] H. Bashashati, R. K. Ward, A. Bashashati, User-customized brain computer interfaces using bayesian optimization, Journal of Neural Engineering 13 (2) (2016) 026001.
URL http://stacks.iop.org/1741-2552/13/i=2/a=026001

[29] E. Brochu, V. M. Cora, N. De Freitas, A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning, arXiv preprint arXiv:1012.2599.

[30] H. Ramoser, J. Muller-Gerking, G. Pfurtscheller, Optimal spatial filtering of single trial eeg during imagined hand movement, Rehabilitation Engineering, IEEE Transactions on 8 (4) (2000) 441–446.

[31] Y. Altun, I. Tsochantaridis, T. Hofmann, et al., Hidden markov support vector machines, in: ICML, Vol. 3, 2003, pp. 3–10.

[32] J. Demšar, Statistical comparisons of classifiers over multiple data sets, The Journal of Machine Learning Research 7 (2006) 1–30.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 2, No. 3, 1006-1012 (2017)
www.astesj.com
Special issue on Recent Advances in Engineering Systems

ASTES Journal
ISSN: 2415-6698

Optimal Plant Growth in Smart Farm Hydroponics System using the Integration of Wireless Sensor Networks into Internet of Things

Nathaphon Boonnam[*,1], Jumras Pitakphongmetha[1], Siriwan Kajornkasirat[1], Teerayut Horanont[2], Deeprom Somkiadcharoen[2], Jiranuwat Prapakornpilai[2]

[1]*Department of Applied Mathematics and Informatics, Faculty of Science and Industrial Technology, Prince of Songkla University, Surat Thani Campus, 84000, Thailand*

[2]*School of Information, Communication and Computer Technologies, Sirindhorn International Institute of Technology, Thammasat University, 12120, Thailand*

ARTICLEINFO

Article history:
Received: 30 April, 2017
Accepted: 24 June, 2017
Online: 10 July, 2017

Keywords:
Wireless Sensor Networks
Internet of Things
Hydroponics

ABSTRACT

Greenhouse cultivation is easy to keep up and control important factors such as light, temperature, and humidity. Using of sensors and actuators in the greenhouse to capture different values allows for the control of the equipment, it can also be optimized for growth at optimal temperature and humidity of various crops planted. We use wireless sensor networks' system by sending results to the cloud service, monitoring values, and devices's controlling via smart phone. The results of this study are useful for growing crops not only in technical parts, but also in physical part; it was evaluated by questionnaire using technology acceptance model.

1 Introduction

Technology can link everything to the Internet called the Internet of Things (IoTs). This means that the human beings control any object using a wireless sensor network (WSN) which makes convenience to everyday life. Technology can be applied in many fields such as home, city, wearable, industrial, health, retail, farming, etc. IoTs refer to the network of objects, vehicles, buildings, and other electronic detectors, as well as software for connecting networks to exchange data from sensors or actuators. IoTs can also make people aware of the environment and remote control over the existing network infrastructure. We can integrate the physical world into the computer system and virtual resources available on the internet to provide both value-added data and functionality for users. The trend of 2020 IoTs usage will increase by 50 billion object were expected to be connected in the IoT [1].

IoTs' devices include RFID, sensors, and computer nodes. The system must have an internet connection that the devices can send data and receive results to communicate with each other. If there is no connection to the Internet but there are also some models that can be connected to other IoTs' devices. Bluetooth is also an option in many ways to connect devices to physical phenomena on networks [2].

At present, farmers need access to the existence of agricultural information and they have relevant knowledge in making decision and respond to information needs. In the agricultural sector, through the development of knowledge management systems and farmers' information inquiries, they can be answered with the helpfulness of accessible multimedia as the application of technology. Agriculture was interested in improving access to technology. Promoting agriculture on several aspects in developing countries, technology had crossed handles using wireless technology and networking to utilize energy and power consumption by equipment, which is helpful in the agricultural development. The development of technology in various domains had driven substantial interest according to rising investments by private sectors towards the growth of technology [3].

In this research, The greenhouse is the best experimental place of our study with the IoTs system as well as a closed viewing area. It was equipped with sensors and internal devices for easy operation to analyze [4].

[*]Nathaphon Boonnam, 31 Moo 6 Surat-Nasarn Rd., Makhamthae, Muang, Suratthani 84000 Thailand, +66(0)77-355-040 & nathaphon.b@psu.ac.th

https://dx.doi.org/10.25046/aj0203127

2 Related Work

Greenhouses can control easily the environment inside their houses and disease outbreak in plants. It can reduce energy consumption for the growth of the plant fully. The benefits of greenhouses are high, but the investment is also high. The original purpose of the greenhouse is able to grow crops that cannot normally grow in the climate of that area. If a good greenhouse system was managed, the greenhouse system can grow all kinds of crops in the right habitat.

Hydroponics is a plant that does not use soil, but that uses water with nutrients dissolved or planted in nutrient solution. This is a new way to grow crops especially growing vegetables and plants used as food. Because of saving space and non-contaminated with chemicals in the soil, we get clean and fresh vegetables. Nowadays, hydroponics has two main benefits; first, it allows for a more controlled environment for planting growth instead of using the same soil eliminating many unknown variables out of many experiments. Second, many plants produce much less time and sometimes better quality in a certain economic environment and condition. Hydroponics will bring more profit to farmers, have no disease in the soil, and it can grow crops very close together. For this reason, the plants produce more volume while using limited space. There is also very little water use because of the use of containers or closed loop water system to circulate water compared with traditional agriculture.

Hydroponics technology uses it gently such as measuring ions, pH, and temperature in water. By adopting sensors with the microcontroller, we get the bills in the greenhouse and analyze the information to control the water supply from this paper. Making WSNs work helps us saving costs in many areas, such as energy, labor costs, and chemical costs. And it can make long-term value.

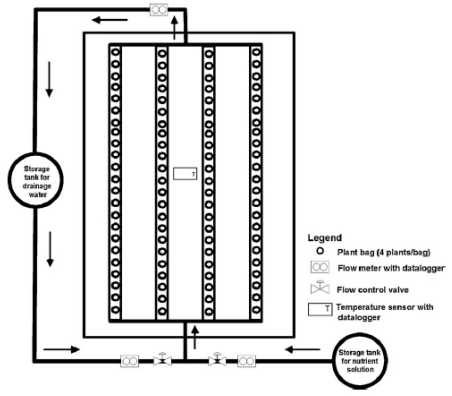


Figure 1: Schematic of the hydroponic greenhouse system used in this study [6]

From the beginning step of our research, we build a greenhouse and take the sensor to control the crops of Hydroponics freely as automation [5] which requires modeling as we want to fit the environmental management by placing the sensors in order to provide humidity control in relation to the ambient air temperature. The conditions were met by installing a color bulb to feed the plants LED representing of red (R) and blue (B) in emission either warm (4700K+2700K) or cool white in parenting [6] and managing water to save water as shown in Figure 1.

The development of today agriculture brings more and more applied technology have a reference and understanding of the environment that is important in the economic production. The yield is sometimes uncertain because of lack of environmental planting's understanding. It creates the WSNs' technology for environmental monitoring and theft detection; simple, cheap, and easy to use the system the detect the data of temperature and humidity in the air covered on the greenhouse properly.

Since we do not want to control by our own, we have to look for the way that technology can control itself automatically. The technology that we mentioned above called Message Queuing Telemetry Transport (MQTT). This is communication technology between sensors as M2M (machine-to-machine) connected with the lightweight protocol. MQTT is designed for the small electronics, small traffic data, and low-bandwidth networks. It can be not only linked to devices simultaneously but also changed the information together. This means that we can make your devices into the bigger network allowing it to control the device where it was needed to turn on-off the devices [5] as shown in Figure 2.

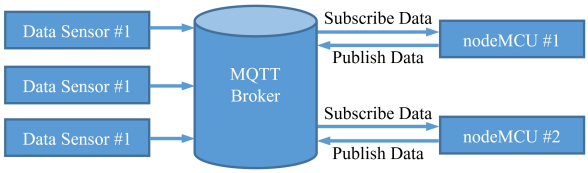


Figure 2: Flow Diagram of MQTT with nodeMCU

Nevertheless, there are many types of protocols, such as HTTP and XMPP. MQTT on IoTs will have an MQTT Broker to manage the storage that can send information to hardware. MQTT clients and servers handle complex messages for mobile applications and reduce network management costs. MQTT applications run on mobile devices, such as smartphones and tablets. MQTT was used for sensor and remote control data reception.

3 System Design and Realization

Adaptation of agriculture is difficult. We have to invite an agency to improve agriculture and make the agricultural sector more productive. Therefore, we make one more comfortable and efficient technology provided that farmers understand for controlling the greenhouse. We can use one of the mobile application called Blynk. It can be used the control system via API from the application into the operation of various devices and also tracked the data via mobile without

technology knowledge. It is easy for people who do not have the ability for using technology. Moreover, the Blynk's user interface system is very easy to use. This following is the devices in details used in this work as shown in Table 1.

Parameter	Value	Voltage
Temperature & Humidity	DHT-11	2.5mA
Ultrasonic Sensor	HC-SR04	15mA
Relay Board	4 Channel	10mA
Grow light	-	1.7mA
Solenoid	Plastic 1/4	1.3mA
Architecture	Greenhouse	-

Table 1: Deployment parameters

3.1 Monitoring Interface

This work was divided into two major topics: hardware and software requirements. The first topic shows in details such sensors, actuators, and components in the system. The second topic deals with software that was used for connecting hardware controls such as turning on-and-off grow light and water pump.

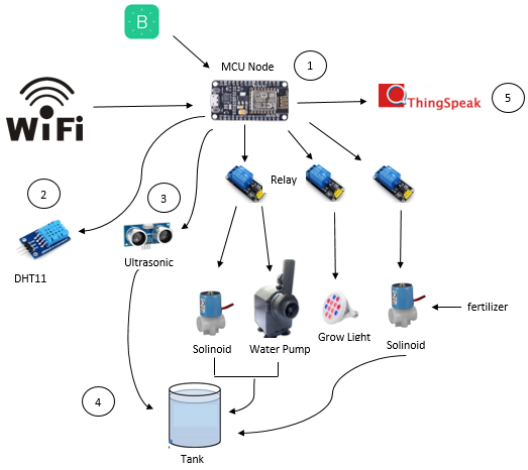


Figure 3: Phototype of system

Hardware Requirements: The main components of the prototype as shown in the Figure 3. The details of each section are as follows:

- NodeMCU is use to implement the monitoring modules. The following sensors and other peripherals were used to collect real time data from the field.
- DHT11 is a relatively cheap sensor for measuring temperature and humidity values.
- Four Channel Relay Board (5V) for switching AC/DC is used to high level trigger with AC motor (220V) to operate the electricity devices.
- Ultrasonic Sensor Module (HC-SR04) includes an ultrasonic transmitter, receiver, and circuit. There are four pins measuring 5-700 cm. used to detect the water level in tank.
- Grow light is an artificial light source used to keep plant growth at night by emitting an electromagnetic.
- Solenoid is the device controls the opening and closing of the water flow.
- Water pump is water pushers move forward from the tank to the pipe hydroponics.

Sensor data was connected to the ThingSpeak™as API to store all results to the cloud results to the database as shown in Figure 4. In order to analyze the future information, we retrieve the information on the smart phone device control which can be controlled by grow light and water pumps.

Software Requirements: Sending data from hardware to smart phone by using the Arduino IDE as a communication medium. Here, the smart phone application was called Blynk.

Blynk is an open-source electronics prototyping platform which is a modified version of Wiring/Arduino IDE with iOS and Android apps to control Arduino board using the Internet as shown in Figure 4.

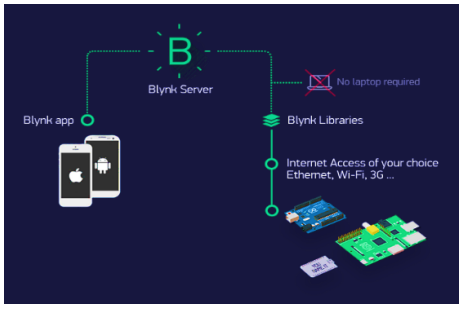


Figure 4: Blynk platform

Blynk platform with iOS and Android apps to control Arduino, Raspberry Pi, and the likes over the Internet. It is a digital dashboard where you can build a graphic user interface for your research by simply dragging and dropping widgets.

3.2 Greenhouse Specification

- *Area:* The greenhouse is mounted on the roof of the building with the size of 2.4 * 5.3 * 2.0 (length * width * height: m).
- *Plants:* The plants used in this study is Cantonese that season pale green succulent stems, petioles 25-40 cm. wide, 5-15 cm. long. Vegetable consumption grows rapidly; it can be harvested just 35-45 days. We acknowledged the seeds coming from Mae Jo University as an organic grain.

4 Results

4.1 Productivity receive

First generation's Productivity: Due to the high temperature of the plant, 48 plants dying in the first week since maximum temperature is 44°C measured from the sensor via the IoT. The DHT11 sensor was used to measure in the survival rate 0% as shown in Figure 5.

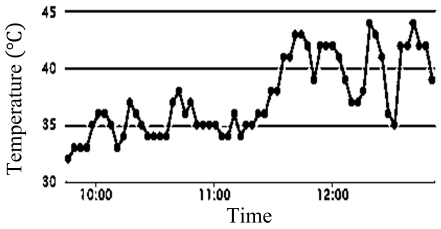


Figure 5: Temperature in the first week

Next step, we created a 50% black SLAN UV filter mesh filter under the roof. Then the heat in the greenhouse is noticeably reduced, as shown in Figure 6.

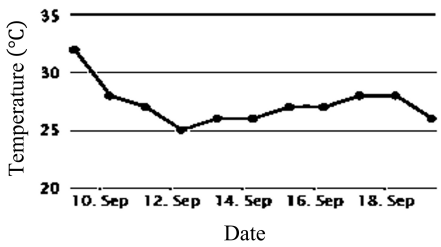


Figure 6: Temperature after installing filter UV 50%

Second generation's productivity: After reducing temperature by UV filtration inside a greenhouse, 48 plants can be survived 22 plants. This means that the survival rate is 45.83%. It can be seen that from the lower temperatures, the plants were more survived. Hence, this is better productivity for growing as well. The ultrasonic sensor monitors the amount of water in the tank if it is lower than normal as we set up for the suitable value of water level in the tank. After instant processing, the solenoid valve was working to add water until full of water. Both devices work reliably Ultrasonic sensor sends the up-to-date data which we can desire the duration time for collecting data into ThingSpeak™. The simulation of the water level can be represented by graph plotted by the relationship between water usage and time as shown in Figure 7.

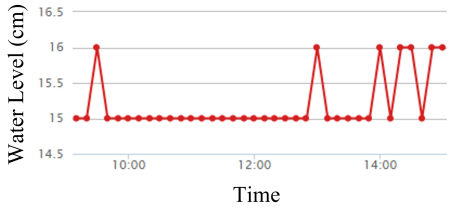


Figure 7: Water usage represented to ThingSpeak™

Electronic devices, such as the solenoid and UV light, were controlled by nodeMCU which makes the connection with the sensor and sends data to ThinkSpeak™for processing devices work of vegetables growth as well as shown in Figure 8.

Figure 8: Vegetables growth in the greenhouse

The greenhouses can work automatically and control via the applications controlled by the mobile device. The connections were able to control by NodeMCU. Freely, no matter where we are around the world, you can control via Blynk, the top and middle part displays the current temperature and the status of the device, respectively, as shown in Figure 9. Remark that it is really simple to set up everything and start customizing. Blynk is one application instead of supports hardware of your choices whether your Arduino was linked to the Internet over Wi-Fi, Ethernet, or this new ESP8266 chip.

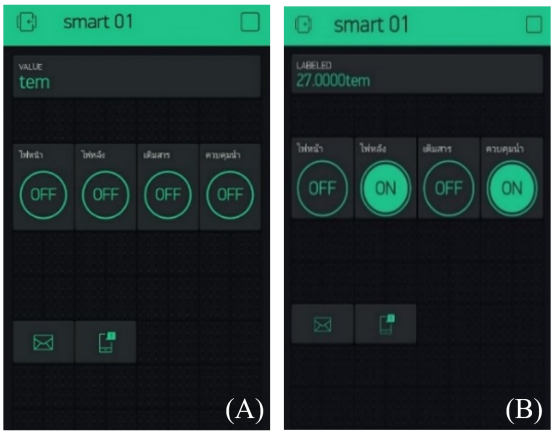


Figure 9: Blnk Applications: (A) Draft display screen (B) Display screen after processing

The sensor data including temperature and relative humidity were recorded to the ThingSpeak™every 3 minutes frequently. Not only all data from sensors were processed in a reasonable manner for the next plotted graph, but also the location of this study from Google API service of ThingSpeak was represented from all tasks can get the appropriate information of the survival plant. It can be seen that the temperature and humidity in the air inverse variation as shown in Figures 10 by analyzing with principal component analysis.

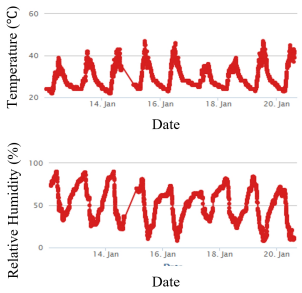


Figure 10: All data of temperature and relative humidity

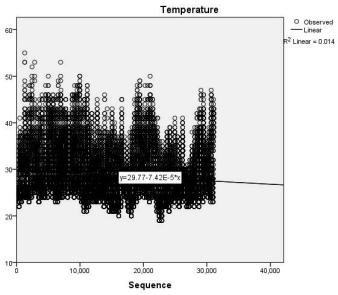


Figure 11: Relationship between temperature and time

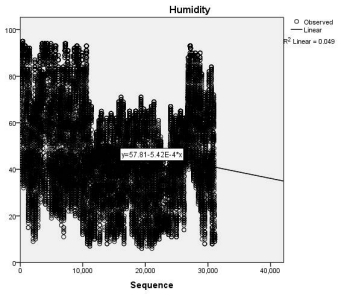


Figure 12: Relationship between relative humidity and time

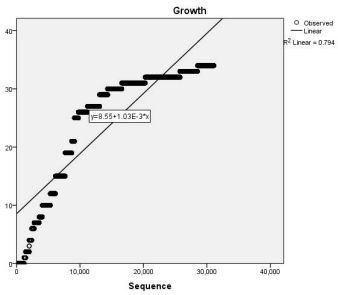


Figure 13: Relationship between plant growth and time

We brought all the data to create the chart for finding the appropriate value through the SPSS program. We plotted the graph for finding the correlation between temperature, humidity, growth with time as shown in Figures 11-13, respectively. We can arrange the correlations observation by linear regression represented by

$$\frac{\nabla y - 96.13}{0.4128E^{-}3} = x; \quad \nabla y = \sum_{i=1}^{n} y_i,$$

where $y_i, i = 1,2,3$ denotes temperature, humidity, and growth variables, respectively.

Since we analyze from principal component analysis, we can consider the communalities and total variance explained as initial eigenvalues as shown in Tables 2-3.

Component	Initial Extraction	Time
Time	1.000	0.730
Temperature	1.000	0.858
Humidity	1.000	0.857
Growth	1.000	0.971

Table 2: Factors Processed by Principal component analysis

Component	Total	Variance %	Cumulative %
Time	2.301	57.531	57.531
Temperature	1.115	27.871	85.402
Humidity	0.404	10.089	95.491
Growth	0.180	4.509	100.000

Table 3: Total variance explained with initial eigenvalues

Notice that the temperature reverse variation with humidity. From the total variance explained the acceptance of correlation's observation with 70%, we can see the humidity was adapted growth extremely as well. However, the temperature is one of factor reacted with growth as necessary condition.

4.2 Control Testing

In our evaluation, we tested 10 times for checking the efficiency of delay controlling. The Statistical tool is mean or average (AVG). Hence, the result is not significant as shown in Tables 4-5. This means that both testing control inside and outside the LAN Delay were equated.

NO.	Inside the LAN Delay(s)			
	LED		Solenoid	
	Turn on	Turn off	Turn on	Turn off
1	5.0	5.0	4.0	0*
2	5.0	5.0	6.0	0*
3	4.0	7.0	4.0	0*
4	5.0	6.0	5.0	0*
5	6.0	7.0	5.0	0*
6	6.0	8.0	3.0	0*
7	7.0	10.0	6.0	0*
8	8.0	14.0	4.0	0*
9	10.0	12.0	8.0	0*
10	14.0	13.0	3.0	0*
AVG	7.0	8.7	4.8	0

Table 4: Moblie Control Testing via Blynk in case of inside the LAN delay

NO.	Outside the LAN Delay(s)			
	LED		Solenoid	
	Turn on	Turn off	Turn on	Turn off
1	6.0	10.0	8.0	0*
2	5.0	9.0	6.0	0*
3	6.0	10.0	7.0	0*
4	4.0	10.0	7.0	0*
5	5.0	8.0	8.0	0*
6	6.0	7.0	9.0	0*
7	7.0	9.0	10.0	0*
8	8.0	6.0	14.0	0*
9	9.0	7.0	19.0	0*
10	10.0	9.0	14.0	0*
AVG	6.6	8.5	10.2	0

Table 5: Moblie Control Testing via Blynk in case of outside the LAN delay

Remark the 0* that the sensor stops the water to overflow the tank makes Solenoid stop working.

4.3 Evaluation Method

Technology Acceptance Model (TAM) was a recognized and well-known theory as a measure of success. Using TRA technology to develop a TAM model and use it to study in the context of the adoption of information systems. Without adopting the norms of the surrounding people. Behavioral display it was used as a factor in forecasting actual usage behavior.

Evaluation methods in how we will test our prototype show control over let the enthusiasts play mobile apps users will need more time 5-10 minutes per person. The rating 5 points 4 scale questionnaires based in TAM provides the theory by being one, there are external variables, how the influence Attitude and willingness to use in the research, the model was designed as follows:

- Study on the use of TAM as the principle to create a 4-D combination of forms 4 main factors of TAM 16 articles,
- Satisfaction and meaning 5 agree that Strongly agree 4 means Agree , 3 means 'Neutral,2 means 'Disagree', and 1 means to ' Strongly disagree',
- Those interested in the quiz after our presentation Android application,
- Collect results Assessment, and
- Analysis of Cronbach (performance by processing using the SPSS program standard the achievement of Cronbach) must be 0.7 or higher.

We consider four factor of TAM as follows:

- Perceived Usefulness (PU) is a perception of the technology use. The technology will make the job better. Make more revenue.
- Perceived Ease of Use (PEU) is a perception that the technology is easy to use. Users are more comfortable than ever.
- Online Cognitions (OC) is a perception data on the Internet through online for helping people recognize information through social media, and viable online communities.
- Behavior Intention (BI) is behavior in technology that influences the intended and actual use of technology.

Our questionnaire used all factors of TAM such as Perceived Usefulness, Perceived Ease of Use, Online Cognitions and Behavioral Intention to Use in Figure 14.

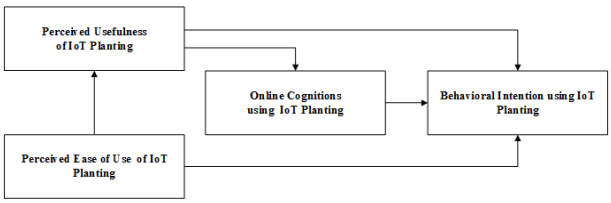


Figure 14: TAM for system Smart Farm Hydropronics Style

4.4 Sample Storage

The sample size for this study was 30, totaling 30 Thai persons in the age range of 20-40 years living in Suratthani as shown in Table 7.

Attrinute	Value	N	Percentage
Gender	Male	11	36.3%
	Female	19	63.6%
Age (years)	20-24	10	33.3%
	25-29	10	33.3%
	30-34	7	23.3%
	35-39	3	33.5%
IT usage ability	Nothing	0	0.0%
	Moderate	18	60.0%
	High	12	40.0%

Table 6: Details of Samples

4.5 User Satisfaction

Analyzing the data using SPSS22. The Cronbach's alpha Coefficient and Pearson Correlation of evaluation as shown in Tables 8-11.

Factors	Cronbach's α Coefficient
PU	0.744
PEU	0.763
OC	0.816
BI	0.721

Table 7: Reliability Coefficient of Constructs

From the result in Table 8, the value of all factors is greater than 0.7, which was considered reliable and accepted average satisfaction of each factor the result

and analyze the average of evaluation satisfaction of each factor as shown in Table 9.

Factors	Average	S.D.
PU	3.6754	0.51003
PEU	3.8653	0.57487
OC	3.8667	0.54756
BI	3.5671	0.63521

Table 8: Average Satisfaction of Each Factor

After that, we find the relationship between two weighted variables in the level measurement section as Independent or dependent factors. Correlation coefficient is a measure of dependence between two factors to interpret the TAM by using Pearson Correlation as shown in Table 10.

Factors Relationship		P.C.	R^2(%)
Independence	Dependence		
PEU	PU	0.731	0.559
PEU	BI	0.724	0.603
PU	OC	0.873	0.758
PU	BI	0.746	0.724
OC	BI	0.623	0.454

Table 9: Factors Correlation

Suggest that the correlation of PEU and OC is the most related together. For another correlation, almost factors have relation closer as shown in Figure 15.

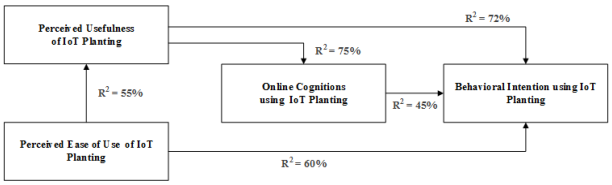


Figure 15: Relational Model of Factors Relationship

5 Conclusion and Future Work

Sensor technology was intended for the automatic control devices via mobile accurate correctly. This technology will save time to look for vegetables which people can eat healthily. Farmers need some helps in the different stages of crop growth and the guidance should be given at the right time. Farmers are suffering a lot of problem; economy, social, and politics. Various challenges in the agricultural domain were identified. The architecture of the challenges mentioned above, knowledge-based structure have various details about agriculture, flow, various input such as market availability, geospatial information and weather prediction. Monitoring system contains modules like remainder, monitoring plant growth in various stages, irrigation planner, and the water need of a plant per day with devised algorithms help

The future work of this study will be developed the sensor for using in public actually. We will study the process of using machinery or machinery to determine the growth of crops and harvest the crops in order to save time and labor. Each vegetable was calculated to get the right temperature for the vegetable.

Acknowledgment This research was financially supported by Prince of Songkla University, Surat Thani Campus, 2015. Furthermore, the present author would like to deeply express thanks to Intelligence and Intelligent Informatics (AI3) Laboratory, Sirindhorn International Institute of Technology, Thammasat University for making the experiment and giving him various comments.

References

(1) G. L. Atzori, A. Iera, G. Morabito, The Internet of Things: a survey Comput.Network, Computer Network,54,27872805, 2010.

(2) H. Sundmaeker, P. Guillemin, P. Friess, S. Woelffl (Eds.) Publications Office of the European Union, Luxembourg, 88, 299309.,2016.

(3) Sjaak Wolfert, Lan Gea, Cor Verdouw, Marc-Jeroen Bogaardt Big Data in Smart Farming , Agricultural Systems,153,6980, 2017.

(4) Harsharn S. Grewala, Basant Maheshwaria, Sophie E. Parksb,Water and nutrient use efficiency of a low-cost hydroponic greenhouse for a cucumber crop: An Australian case study, Agricultural Water Management,98, 841846,2011.

(5) J.delSagrado, J.A.Snchez, F.Rodrguez, M.Berenguel, networks for greenhouse temperature control Journal of Applied Logic, 17, 2535,2016.

(6) Olga Dugosz-Grochowska, Anna Koton, Renata Wojciechowska, Modifying folate and polyphenol concentrations in Lambs lettuce by the use of LED supplemental lighting during cultivation in greenhouses, Journal of Functional Foods,.26 ,228237, 2016.

(7) A.W. Al-Kayssi, Spatial variability of soil temperature under Green-house conditions,, Renewable Energy,27, 453462,2002.

(8) I. Mohanraj, Field, Monitoring and Automation Using IoT in Agriculture Domain, Procedia Computer Science,93, 931939, 2016.

(9) M. Azaza , C.Tanougast , E.Fabrizio , A.Mami, Smart greenhouse fuzzy logic based control system en hanced with wireless data monitoring, 61, 297307, 2016.

(10) L. Atzori, A. Iera, G. Morabito, M. Nitti, The social internet of things (siot) when social networks meet the internet of things: concept, architecture and network characterization, Comput. Netw. 56 ,35943608,2012.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 2, No. 3, 1013-1018 (2017)
www.astesj.com
Special Issue on Recent Advances in Engineering Systems

ASTES Journal
ISSN: 2415-6698

Skeletonization in Natural Image using Delaunay Triangulation

Vicky Sintunata[*,1], Terumasa Aoki[1,2]

[1] *Tohoku University, Graduate School of Information Science, 980-8579, Japan*

[2] *Tohoku University, New Industry Creation Hatchery Center (NICHe), 980-8579, Japan*

ARTICLEINFO

Article history:
Received: 18 May, 2017
Accepted: 15 June, 2017
Online: 11 July, 2017

Keywords:
Grey-scale skeletonization
Delaunay triangulation
Natural image
2D skeleton

ABSTRACT

In this paper a novel approach to extract 2D skeleton information (skeletonization) from natural image is proposed. The work presented here is the extension of our previous paper presented at the International Sympsosium on Multimedia 2016. In the past work, a threshold based method is utilized. Here the algorithm is further improved by using a better edge points detection and skeleton extraction. Furthermore the proposed method is compared with the Skeleton Strength Map (SSM) and shows better result visually and numerically (F-measure comparison).

1 Introduction

This paper is an extension of the work [1] originally presented at International Symposium on Multimedia 2016 (ISM'16). In this paper we provide a more detail and some improvement from the previous paper presented at ISM 2016.

Skeletonization is a process to extract the simplest, the most compact form of an object (skeleton) from 2D or 3D objects. Although mostly used in animation[2], skeleton information has also been used in symmetry detection [3], shape analysis [4], etc. Based on the input to the skeletonization, we can classify it into two types; 2D skeletonization and 3D skeletonization. In this paper, we are dealing with the first type, i.e. the 2D skeletonization or skeletonization from an image.

Corneat et al.[5] classify skeletonization into 4 classes, i.e. thinning and boundary propagation, distance field-based, geometric and general-field functions. While most of these approach works well, it only applies to the binary image. In order to get the binary image, of course one can use a foreground-background separation, but unfortunately foreground-background separation is still an open problem in image processing field. Another way to see the problem is that conventional approaches need a closed boundary information to extract the skeleton information. It holds when the object and the background information has a high contrast, but does not hold for object with a low contrast value with the background (such as similar color and texture).

When the input image is not a binary image, the skeletonization is often referred as grey-scale skeletonization. Grey-scale skeletonization can be regarded as a relaxation of the conventional skeletonization. Instead of a binary image, the input to grey-scale skeletonization is usually the brightness image (grey-scale image) or the color images. Skeleton is related closely to the symmetry information, therefore several authors are utilizing the symmetricity. You and Tang[6] proposed a method to detect the skeleton of characters using the wavelet. Widynski et al.[3] proposed a method based on particle filter approach. Levinshtein et al.[7] proposed a method based on superpixel segmentation and learned affinity function. In the last two methods, a supervised learning step is incorporated.

Du et al.[8] proposed a method based on a structure adaptive anisotropic filtering. The method is extracting the skeleton by updating the scale orientation and kernel of every pixel. The process is done iteratively including the update step therefore requiring a high computational cost. Another method is proposed by Li et al.[9] using a distance transform method [10]. By first extracting the edge information using Canny edge extraction algorithm and then calculating the inverted of the distance transform, this method can extract the skeleton of the object and its surrounding. To further increase the performance, a learned context model is utilized. Since this method relies on the distance transform, connected edge pixel is needed to

[*] Vicky Sintunata, vicky@riec.tohoku.ac.jp

https://dx.doi.org/10.25046/aj0203128

some extend. It is important to note again here that while the connected edge pixel can be retrieved or extracted in an image where the object has high contrast, the opposite is not true. This condition results in a sparse or unconnected edge pixels in the image.

Most of the papers mentioned above has high computational cost or need a learning step in one of their method. In this paper we tried to solve the grey-scale skeletonization without a learning step. Furthermore the proposal can be used to relax the unconnected (sparse) boundary pixel condition. The rest of the paper is as follows: in section 2 the more detail explanation of the proposed method is presented. In section 3 some examples of the result and discussion will be presented. Finally section 4 will conclude this paper.

2 Proposed Method

The general framework for the proposed method can be seen in Figure 1. There are three major steps in the method: boundary points extraction, skeleton extraction and pruning step.

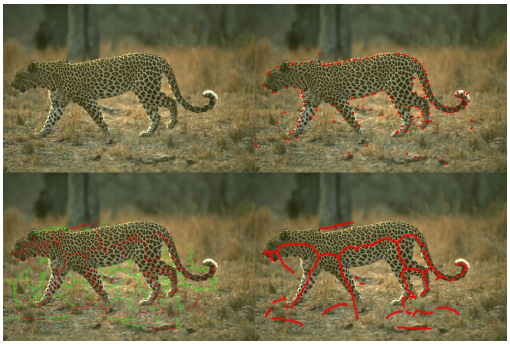


Figure 1: Proposed Method Framework. Upper-left: input; Upper-right: edge points extraction; Lower-left: skeleton extraction; Lower-right: pruning

2.1 Boundary Points Extraction

The method starts by making a grid of the input image and finding the edge points inside each grid. This edge points are selected by choosing the highest intensity among the pixels inside the grid. The question is then how are we suppose to find these edge points. It is very common to get the edge information from the gradient of the image. The higher the intensity of the gradient image is, the higher the probability it is considered to be the boundary object. Instead of using Canny edge detection which relies on the brightness image only, we calculate the edge (or gradient) information from both the brightness image and the color images.

First, the input image is converted into CIE color space and then split to each corresponding channels (i.e. brightness or luminance (L*) and two color channels (u* and v*)). Let $e_c(i,j)$ denotes the value of the gradient image at channel c ($c \in$ L*, u*, v*) at (i,j) position. The value of $e_c(i,j)$ is calculated from the standard deviation of a pixel at (i,j) with its surrounding ($n \times n$ patch) at its respective channels (1). The benefit of calculating the standard deviation is that textured object will have a higher deviation which results in an easier way to locate the boundary pixel between two different objects. Finally we can combine them by averaging the values of these edge images (2) at each pixel.

$$e_c(i,j) = \sigma_{i,j,n} = \sqrt{\frac{\sum_{a,b=-\frac{n-1}{2}}^{\frac{n-1}{2}} (x_c(i+a,j+b) - \bar{x})^2}{n^2}} \quad (1)$$

$$e(i,j) = \frac{e_{L*}(i,j) + e_{u*}(i,j) + e_{v*}(i,j)}{3} \quad (2)$$

Since we are only interested on the boundary of the object, we can further reduce the texture effect by redoing the previous step onto the combined image (generated from 2). This will reduce the texture inside the object, but will also make the boundary of the object to lies inside a "valley". This can be avoided by applying the resulting image with the Laplacian operator. Let $I_{\mathcal{L}}$ be the image after applying the Laplacian operator. $I_{\mathcal{L}}$ will then be convoluted with the standard deviation image of the combined image, to construct the final edge image. Figure 2 shows the example of this approach.

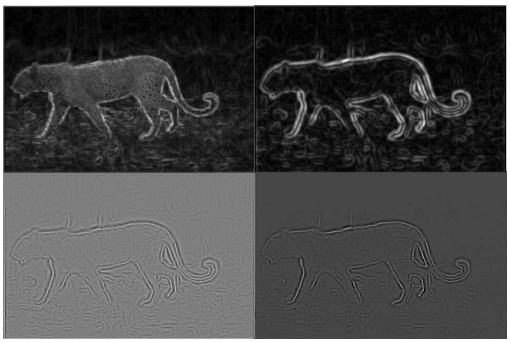


Figure 2: Edge Image Generation. Upper-left: combined image, Upper-right: Standard deviation image; Lower-left: Laplacian image; Lower-right: Final Edge image.

The edge points are then extracted from this edge image. Since we are selecting the edge points from the grid, there will be points which lie in the non-edge position. Filtering these points (candidate points) is done by using a method similar to a method proposed in [11,12]. The idea is very intuitive: at each pixel in the image, generate a patch and divide it into two halves along an orientation. At each section calculate its histogram (intensity, colors, etc.) and compare both of the sections. The higher the distance between the histograms (χ^2 distance) the more likely the examined pixel to be an edge pixels.

Notice that we need to have an orientation at each candidate point. To approximate the orientation at each candidate point, the phase calculated from the difference of horizontal and vertical neighboring pixels of the edge image is used (Sobel operator). Let $d_x(i,j)$ and $d_y(i,j)$ denote the gradient in x and y direction respectively and $\theta(i,j)$ defined in (3), denotes the

phase at position (*i,j*). Let $p_\theta(i,j)$ be the approximation of orientation in a candidate point with coordinate (*i,j*) and is defined by (4). Equation 4 means that the angle approximation is determined by the average of the phase surrounding the point. Finally the filtering is done by using the k-means (k = 2) algorithm with the histograms' difference as features along with the intensity value at the corresponding pixel in the edge image.

$$\theta(i,j) = \frac{d_y(i,j)}{d_x(i,j)} \tag{3}$$

$$p_\theta(i,j) = \frac{1}{n^2} \sum_{a,b=-\frac{n-1}{2}}^{\frac{n-1}{2}} \theta(i+a, j+b) \tag{4}$$

Some of the points will lie very near with each other and it will affect the triangulation in the next step. Therefore a merging step is also applied to the filtered points, i.e. merging the points which lie near each other. The merged new position is calculated based on the average position of the merge points.

2.2 Skeleton Extraction

Skeleton extraction is done by first applying the Delaunay Triangulation to the edge points acquired in previous step. In order to do the skeleton extraction, a triangle grouping should be done. In [13,14] the authors classify triangles into 3 types: terminal triangle, junction triangle and sleeve triangle. The terminal triangle is defined as a triangle which has two outer contour as its edges. Junction triangle is defined as a triangle which does not have outer contour as its edges. Sleeve triangle is defined as a triangle which has only one outer contour as one of its edges. Note that in their application, which can be classified as the binary image skeletonization, the information regarding the outer edge is readily available. In our approach, we do not have that kind of information.

Note that in the previous step during clustering using k-means, we do have the idea to determine a point whether it belongs to the boundary points or non-boundary points based on its histogram distance feature. Therefore, we used this information to cluster the edge as follows. Let point m_i be the middle point of the *i*-th edge. By applying the same approach as in section 2, i.e. orientation approximation and histogram calculation, we can determine point m_i to be an edge point based on its (feature-) vector distance to the center points of each cluster. Another way to see this is, we are not recalculating the cluster centers since the cluster centers (which will determine whether a point is a boundary or non-boundary point) has already been calculated in the previous step, the evaluation of these edge points (*m*) can be done just by calculating the distance to the cluster centers. If $dist(m_i, C_k)$ is defined as the L2 distance between point m_i and the *k*-th center, then point m_i belongs to the the *k*-th cluster if it has the minimum distance compared to another cluster's center.

In this case we can similarly define the triangles as in [13,14]. In sleeve triangle, two of its non-boundary *m* points will be connected together and no connection will be established in terminal triangle. Two types of junction triangles are defined: obtuse triangle and acute triangle. In case of acute triangle, an additional point (the center of triangle) will be added and every *m* points will be connected to this additional point. In obtuse triangle type, instead of generating an additional point, we connect the points on two of the edges of the largest angle to the point lying in front of the angle. See Figure 3 for visualization. Note that the yellow point (e.g. m_1 in Figure 3.a) represents *m* belonging to the boundary cluster.

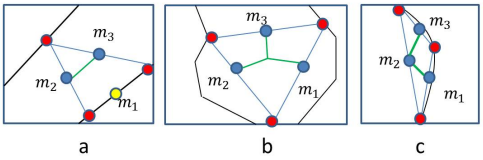


Figure 3: Three types of triangles: a. Sleeve; b. Acute Junction; c. Obtuse Junction .

Given an initial skeleton (point to point connection) from the previous step, the problem now is to determine which of the connection belongs to the same object. Let *T* denotes the set of pixels traversed by the skeleton's edge (Figure 4). By calculating the mean pixel value of T we can get a set of feature vector (f) to represent the skeleton for the clustering algorithm (5). The subscript l denotes the type of image the pixel belongs to, i.e. color images, brightness image, etc. Figure 5 shows the result of the clustering algorithm. Note that since the contrast of the color images might be low for some images and resulting in a non-discriminative cluster, a contrast limited adaptive histogram equalization (CLAHE) is applied to the color images.

$$f_l = \frac{1}{n} \sum_{i=1}^{n} T_i, \qquad T_i \in T_l \tag{5}$$

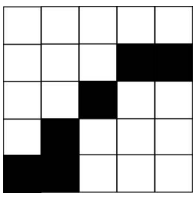


Figure 4: Pixels' value traversed by the skeleton (black tile) will be included in the mean value calculation (5)

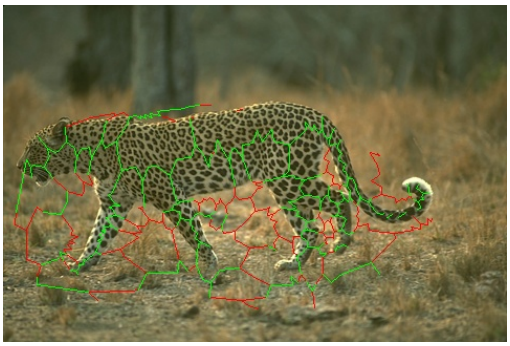


Figure 5: Example result of skeleton grouping

There is a possibility that the generated skeleton has a loop. In order to prevent such condition, loop removal step is applied. This is done by merging all the points of the loop into a point c, where c is the mean position of the points belonging to that loop (Figure 6). If a loop consists of a smaller loop, only the largest loop is considered (Figure 7).

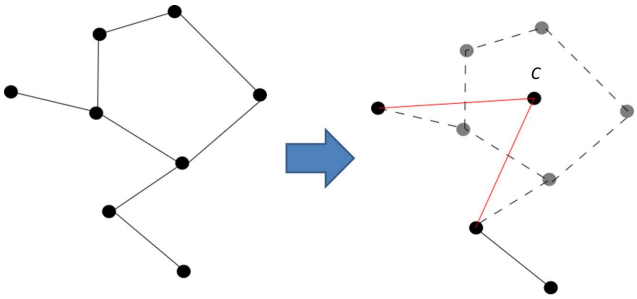


Figure 6: Loop removal step example. A new connection between the generated middle point c and remaining points are constructed.

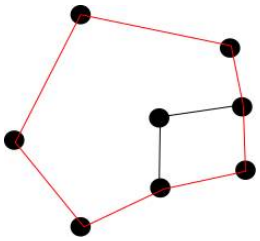


Figure 7: When there is a smaller loop inside a greater loop, only the outer loop will be considered (red lines)

2.3 Skeleton Pruning

Some might notice some redundant skeleton branches from Figure 5. Skeleton pruning is the step required to remove these branches. Borrowing the triangle types term, depending on the number of connection it has, skeleton points can be classified into three types: terminal, sleeve, and junction point. Terminal point means that the point only has one connection with another skeleton point. Sleeve point means that it has two connections with two other skeleton points. Junction point defines a point with more than two connections with other points. In this paper, skeleton branches are defined as the terminal points which connect to junction point. If there are such connection, then the connection will be deleted (Figure 8).

Finally, a bezier smoothing is applied to the skeleton. A curve skeleton is defined as the set of points starting from the terminal point until a junction point, therefore taking Figure 8 as an example, there are two curve skeletons in Figure 8. The sleeve points will be used as the control points for the bezier curve generation.

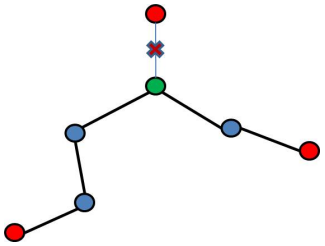


Figure 8: Colored points represent types of skeleton point. Red: terminal; Blue: sleeve: Green: junction

3 Experimental Result and Discussion

All experiments were done using a C++ language and OpenCV 2.4.9 library. There are several parameters which are fixed in this paper, i.e. the number of histogram bins is set to 16 and the size of the grid is set to 5 (refer to section 2.1). CIE Luv is used in this paper. Five images taken from the Berkeley dataset [15] were used for the experiment.

The results of the proposed method are then compared with a method based on [9]. Figure 9 shows the comparison, including the ground truth image generated by human observer. The first column is the ground truth skeleton image, the second column is the proposed method and the third column is skeleton generated based on [9]. Notice that although the proposed method did not generate the skeleton perfectly, it still represents the object in the image better than the method based on [9].

In order to further assess the result, F-measure of the ground truth and the methods were calculated (6). The precision and recall is calculated following (7) and (8). The number of true positive (tp) is calculated by counting the number of pixel where both the ground truth (GT) image (with a $\sigma = 1$ allowance) and the generated skeleton image have the same value. σ here means that if the generated skeleton differs only slightly, i.e. still inside the allowance σ then it is still considered to be true positive and hence counted. Notice again that when comparing both the image to calculate the F-measure, skeleton image generated has binary color where the pixels will have a 1 (white colored) value if the pixel is not traversed by the skeleton edge and 0 (black colored) if it is traversed by the skeleton edge. For the false negative (fn), it is calculated as the number of pixel where it has a value of 1 in the ground truth image, but 0 in the skeleton im-

age. Vice versa, for false positive (fp), we are counting the number of pixel where it has value of 0 in the ground truth but 1 in the skeleton image. Table 1 shows the result of the F-measure calculation.

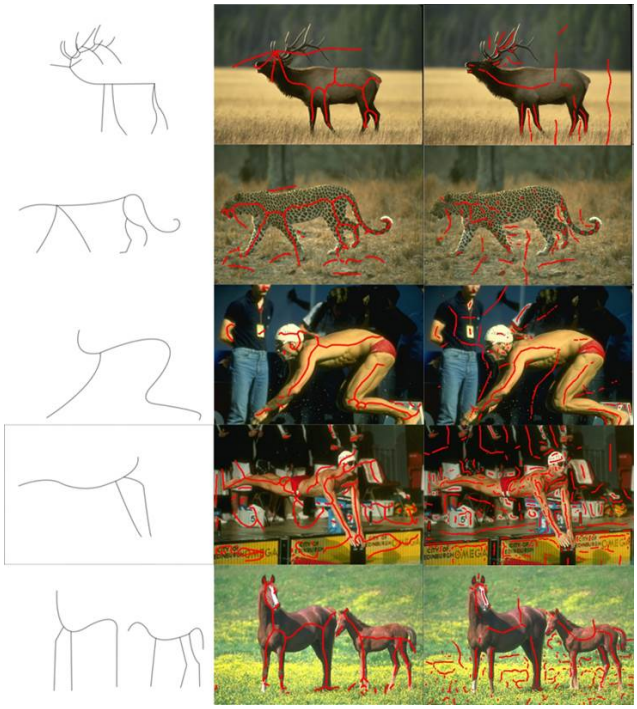


Figure 9: The skeleton is overlaid on top of the input image. For visualization purpose, the skeleton is colored red.

$$F = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{6}$$

$$Precision = \frac{tp}{tp + fp} \tag{7}$$

$$Recall = \frac{tp}{tp + fn} \tag{8}$$

Table 1: F-measure Results

Image	Proposed	SSM[9]
Deer	**0.250**	0.180
Cheetah	**0.345**	0.104
Swimmer1	**0.299**	0.154
Swimmer2	**0.175**	0.092
Horse	**0.359**	0.215

Although every image has its own best parameters, for general purpose, we try to find the parameter values which gave the best average F-measure from the test images. Four window sizes, which is used for histogram calculation, were tested (n = 7, 9, 11, and 15) along with the CLAHE limit parameter (l = 2 and 4). The result is shown in Figure 10. The x-axis is the combination parameters, i.e. (n,l). As shown in Figure 10 the combination of n = 9 and l = 4 has higher average F-measure among the other combinations.

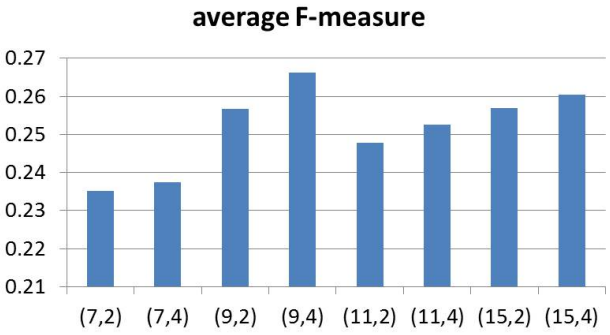


Figure 10: Average F-measure of parameters combination

4 Conclusion

In this paper a novel method to extract 2D skeleton information from a natural image is proposed. Based on the results shown in section 3, the proposed method shows a higher performance compared to the SSM method. Not only compared visually, in order to have a better comparison, the proposed method and SSM method is also compared by calculating their F-measures, which is shown to outperform the SSM method. There are several limitations for the proposed method, e.g. some images still have redundant skeleton branches and user interference (manually select) when choosing the final skeleton cluster (subsection 2.2). Therefore for our future work, we are going to solve these problems.

Conflict of Interest The authors declare no conflict of interest.

References

(1) V.Sintunata, T.Aoki. "Skeleton Extraction in Cluttered Image based on Delaunay Triangulation", in International Symposium on Multimedia (ISM'16), 2016.

(2) I. Baran, J. Popović. "Automatic Rigging and Animation of 3D Characters" ACM Trans. Graphics, 26(3), 72-1 – 72-8, 2007.

(3) N.Widynski, A.Moevus, M.Mignotte. "Local Symmetry Detection in Natural Images Using a Particle Filtering Approach" IEEE Trans. Image Processing, 23(12), 5309-5322, 2014.

(4) X.Bai, L.J.Latecki. "Path Similarity Skeleton Graph Matching" IEEE Trans. PAMI, 30(7), 1282-1292, 2008.

(5) N.D.Cornea, D.Silver, P.Min. "Curve-Skeleton Properties, Applications, and Algorithms" IEEE Trans. Visualization and Computer Graphics, 13(3), 530-548, 2007.

(6) X.You, T.Y.Tang. "Wavelet-based approach to character skeleton" IEEE Trans. Image Processing, 16(5), 1220-1231, 2007.

(7) A.Levinshtein, C.Sminchisescu, S.Dickinson. "Multiscale Symmetric Part Detection and Grouping" Int.J.Comput. Vis. 104(2), 117-134. 2013.

(8) X.Du, S.Zhu, J.Li. "Skeleton Extraction via Structure-Adaptive Anisotropic Filtering" in International Conference on Internet and Multimedia Computing and Service (ICIMCS'14), 2014.

(9) Q.Li, X.Bai, W.Y.Liu. "Skeletonization of Gray-Scale Image from Incomplete Boundaries" in International Conference on Image Processing (ICIP'08), 2008.

(10) L.J.Latecki, Q.Li, X.Bai, W.Y.Liu. "Skeletonization Using SSM of the Distance Transform" in International Conference on Image Processing (ICIP'07), 2007.

(11) D.R.Martin, C.Fowlkes, J.Malik. "Learning to Detect Natural Image Boundaries Using Local Brightness, Color and Texture Cues" IEEE Trans. PAMI, 26(5), 530-549, 2004.

(12) M.Maire, P.Arbelaex, C.Fowlkes, J.Malik. "Using Contours to Detect and Localize Junctions in Natural Images"in Conference on Computer Vision and Pattern Recognition, 2008.

(13) L.Prasad. "Rectification of the chordal axis transform and a new criterion for shape decomposition" Image and Vision Computing, 25, 1557-1571, 2007.

(14) T.Igarashi, S.Matsuoka, H.Tanaka. "Teddy: A Sketching Interface for 3D Freeform Design" ACM SIGGRAPH'99, 1999.

(15) Berkeley Dataset. https://www2.eecs.berkeley.edu/Research/Projects/CS/vision/bsds/

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 2, No. 3, 1019-1025 (2017)
www.astesj.com
Special issue on Recent Advances in Engineering Systems

ASTES Journal
ISSN: 2415-6698

A High Efficiency 0.13μm CMOS Full Wave Active Rectifier with Comparators for Implanted Medical Devices

João Ricardo de Castilho Louzada, Leonardo Breseghello Zoccal, Robson Luiz Moreno, Tales Cleber Pimenta

Federal University of Itajubá, IEST, Itajubá - MG, Brazil

ARTICLEINFO

Article history:
Received: 30 April, 2017
Accepted: 29 June, 2017
Online: 11 July, 2017

Keywords:
Active Rectifier
CMOS Comparator
Medical Devices

ABSTRACT

This paper presents a full wave active rectifier for biomedical implanted devices using a new comparator in order to reduce the rectifier transistors reverse current. The rectifier was designed in 0.13µm CMOS process and it can deliver 1.2Vdc for a minimum signal of 1.3Vac. It achieves a power conversion efficiency of 92% at the 13.56MHz Scientific and Medical (ISM) band.

1 Introduction

This paper is an extension of work originally presented in The 15th International Symposium on Integrated Circuits, Sigapure, 2016 [1]. It presents some details not described in the original paper. This paper also brings new results obtained after a new calibration of the system that improved the circuit efficiency.

Implanted medical device is a technology that offers several possibilities to improve medicine and it can change the paradigm of how doctors and patients deal with treatments and drugs. In order to be implanted, a medical devices need a source of power that will provide energy to the circuit. As used in RFID technology, one of the ways to supply power to implanted medical device is by using an inductive link [2] that can power up a circuit through the skin without to use transcutaneous wires, which could expose the patient to bacterias and viruses [3]. The other solution is by using batteries that have limited life and can cause heavy metal contamination in case of leakage.

Although this work is focused on inductive links, there are other ways to transfer energy. There are some applications that use ultrasound waves to excite a piezoelectric device to generate a current. Another way to provide power to medical implants is by using a thermoelectric generator that makes use of the temperature gradient between inside and outside of body to create energy by Seebeck Effect.

An inductive link requires an AC-DC rectifier that is responsible to transform the signal wave into a power supply. Rectifiers are usually designed to use conventional diodes that have a built drop voltage of 0.7V, which is suitable for conventional applications but it is unusable for circuits running on power supplies of 1V or less. Alternatively to the conventional diodes, Schottky diodes can be used, but the manufacturing process of low forward drop voltage can be more expensive than the traditional processes [4] because they might not be present in some regular processes and require extra fabrication steps [5]. An alternative is the use of diode connected MOS transistors to replace the conventional diodes in a bridge connection. CMOS transistors connected as diodes can work at high frequencies, such as the Scientific and Medical - ISM band (13.56MHZ) that has been widely used [4, 5, 6]. One problem of this solutions is that diode connected MOS transistors in a bridge connection offers low ON-OFF switching speed which in turn creates an undesirable reverse current. A cross-coupled connected structure (Figure 1) can make the operation more efficient and minimizes the reverse current, but in the ideal case the ON-OFF switch of the CMOS transistors would be controlled by another circuit that can speed up the operation and can create a dynamic rectifier to make the operation more efficient, thus minimizing the reverse current. Some dynamic rectifiers have been reported in the last few years [6, 7, 8] as solutions for powering medical devices under low power.

One important characteristic of power supply is the power conversion efficiency (PCE), equation 1, which is defined as the rate of the energy dissipated by the load by the power consumed by the entire circuit [10], where η_{pa}, η_{link}, $\eta_{rectifier}$, $\eta_{regulator}$ represents the power amplifier, inductive link, rectifier, and regulator efficiency respectively. The $\eta_{rectifier}$ represents the amount of power provided to the output divided by the amount of power that enters in the rectifier (equa-

https://dx.doi.org/10.25046/aj0203129

tion 2). For a PCE of 30% it is expected to have around 80% of efficiency of the AC-DC converter [6].

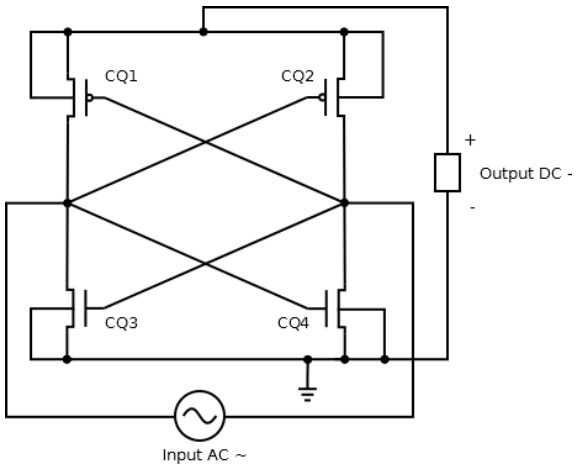


Figure 1: Cross-Coupled Rectifier [9].

$$PCE = \eta_{pa} * \eta_{link} * \eta_{rectifier} * \eta_{regulator} \quad (1)$$

$$\eta_{rectifier} = \frac{Pout\_DC}{Pout\_AC} \quad (2)$$

This paper presents one active rectifier using diode connected CMOS transistors, in 0.13μm process, controlled by two original comparators that are responsible for turning ON-OFF the circuit and minimize the reverse current. This work was intended to provide an output voltage of at least 1V.

2 Cross Couple Structure

The cross-coupled structure shown in Figure 1 connects the transistors gates in a cross mode. That configuration puts the gates of transistors CQ1 and CQ2 biased at a positive signal when their drain are biased by the negative signal. On the other hand, when the circle change, the drains are biased by the positive signal and the gates by the negative signal. The transistors are made to turn ON at different voltage levels in their gates. NMOS needs to have high voltage and PMOS needs to have low voltage. At the same time the cross coupled circuit is designed so that the PMOS transistors provide higher voltage to the load while NMOS provide the lower voltage or GND voltage.

3 Active Rectifier

As mentioned in the introduction section, the desirable rectifier should have a minimal reverse current, and one widely used way to achieve it is by using a comparator that can turn the rectifier ON-OFF faster enough. The Figure 2 shows the schematic of the rectifier [6].

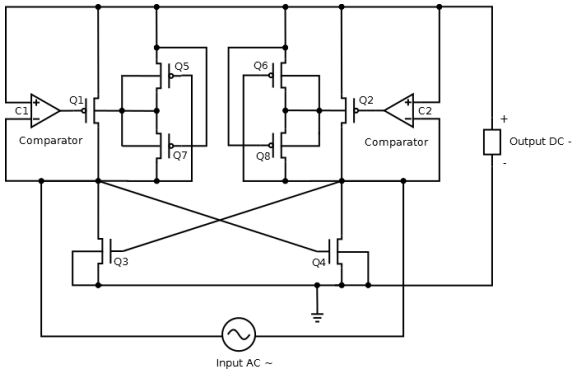


Figure 2: Active Rectifier Circuit [6].

The rectifier of Figure 2 is comprised of PMOS transistors Q1 and Q2, and NMOS transistors Q3 and Q4. Transistors Q5, Q6, Q7, and Q8 work in order to tie the PMOS base transistor always at the higher voltage [11]. Transistors Q3 and Q4 are connected as a cross-coupled gate structure while Q1 and Q2 are driven by comparators C1 and C2.

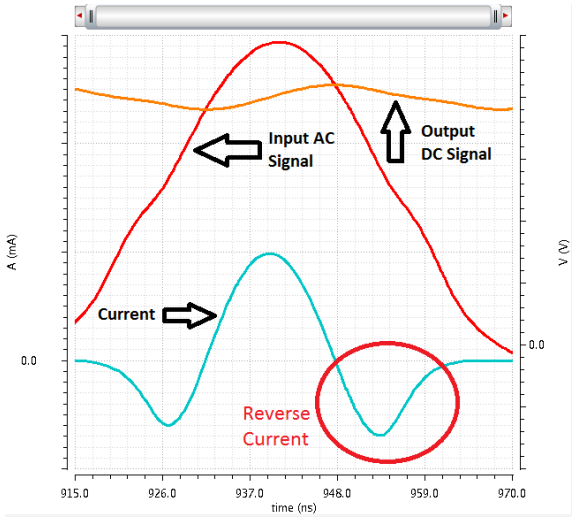


Figure 3: Cross-Coupled Rectifier wave form.

While the cross-coupled rectifier has a considerable reverse current (Figure 3) the two comparators shown in Figure 2 allow the active rectifier to switch faster and minimize it. The challenge faced by this technology is to make the comparators work fast enough. Figure 4 shows the waveform signals and highlights the comparator signal delay by circles C and B, as well as the reverse current, by circle A.

In the ideal case, a perfect comparator would start its output signal a moment before the AC signal became larger than DC signal and would finish the pulse in the exactly moment the AC signal becames smaller than the DC signal. It is possible to see in the Figure 4 the delay showed inside of circle B which creates the reverse current in circle A.

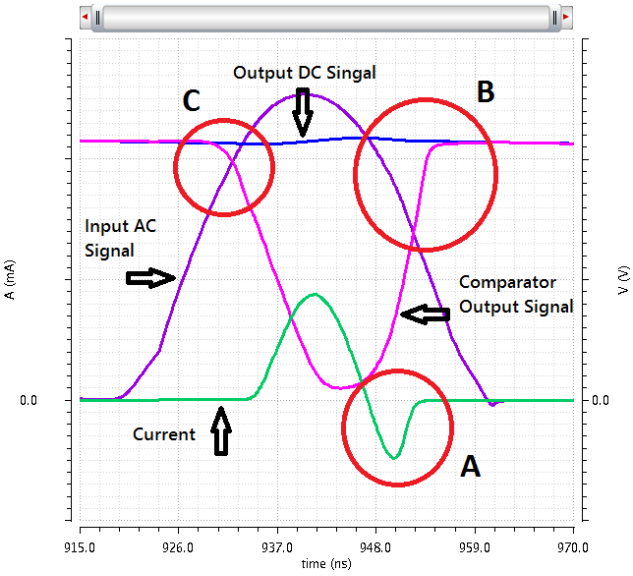


Figure 4: Active Rectifier waveform - (A) reverse current, (B) output delay, (C) entry delay.

4 Two Input Common Gate

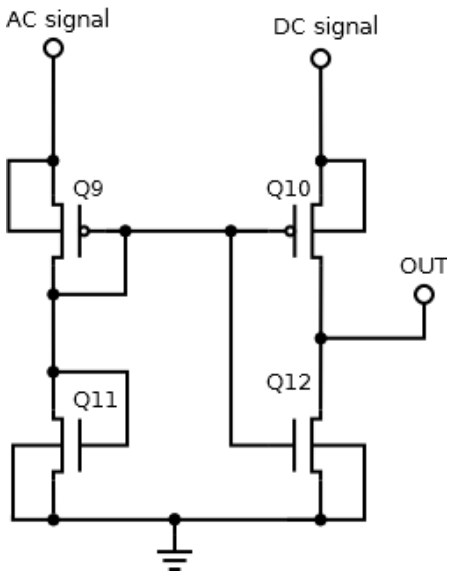


Figure 5: Two input common gate.

Figure 5 shows the Two Input Common Gate Comparator, which is the comparator used to develop the proposed comparator. This comparator is used in order to monitor AC and DC signals, and to switch ON and OFF the rectifier transistors always when the AC signal becomes larger or smaller than the DC signal. The Two Input Common Gate uses only four transistors, Q10 and Q12 that form an inverter, and Q9 and Q11 that form a voltage divider responsible by create an (AC signal)/2. That comparator has been used since it is small and simple, it does not require external power source, it consumes low power and provides faster switch for the rectifier transistors than a cross couple structure. Although that comparator is relatively fast, we propose a new comparator based on the Two Input Common Gate, with some modifications that speeds up the operation of the switching transistors.

5 Proposed Comparator

The proposed comparator shown in Figure 6 is a modified version of the two-input common gate, with an extra capacitor and an extra voltage divider that are responsible for speeding up the comparator.

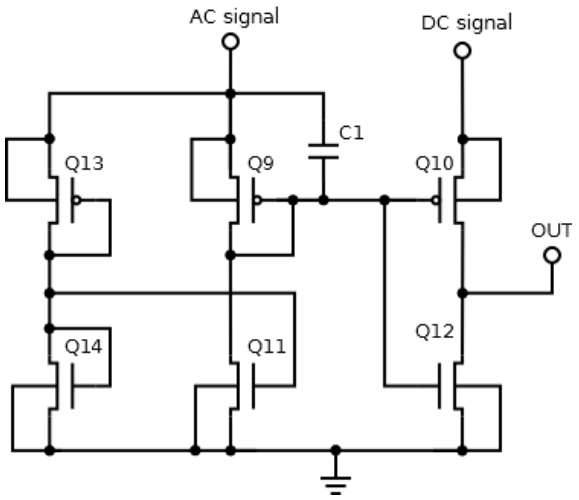


Figure 6: Proposed Comparator.

Transistors Q9, Q10, Q11, and Q12 form a modified common gate structure. The gate of transistor Q11 is connected to Q13 and Q14. Transistors Q10 and Q12 compose an inverter that can bring the OUT pin to low or high level (DC signal). The transition between high and low level happens when the voltage, at Q10 and Q12 gates, is larger than half of DC signal. Our comparator aims to detect when AC signal is larger than DC signal. In order to accomplish it, transistors Q9 and Q11 form a voltage divider designed to put the gates of Q10 and Q12 at a voltage larger than DC/2 in the moment that AC signal is larger than the DC signal.

In the proposed comparator, capacitor C1 works in order to improve the rise time of the signal provided by the voltage divider formed by Q9-Q11. While C1 is not charged, it works as a short circuit and allows the AC signal to drive the Q10-Q12 inverter. When C1 is charged, it becomes an open circuit, and the voltage between Q9 and Q11 starts to drive the inverter. Figure 7 Q10-Q12 gates voltage waveform inside of the circle. A lump appears when C1 gets full and disconnects the AC signal from driving the inverter. The lump can occur above or below in the signal according to the value of capacitor C1. If C1 is larger than a certain value it will stop decreasing the reverse current and will start increasing it. Figure 8 (a) shows a waveform with an acceptable capacitor and (b) with a capacitor larger than the optimum point. Figure 8

(b) does not have a lump since its capacitor C1 is so big that it never gets fully charged, and the AC signal drives the inverter the whole time.

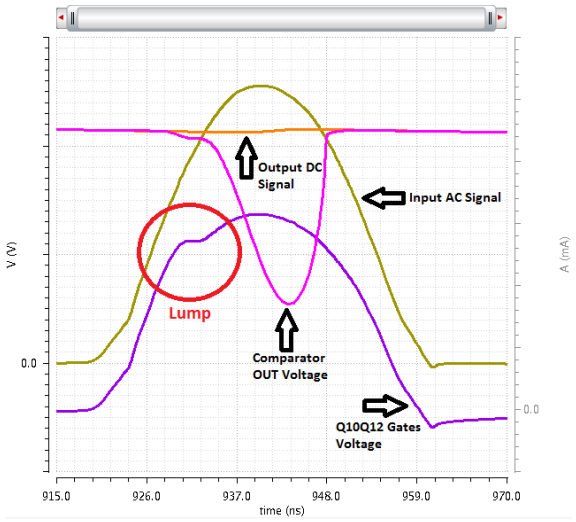


Figure 7: Proposed Comparator waveform. Lump's highlight.

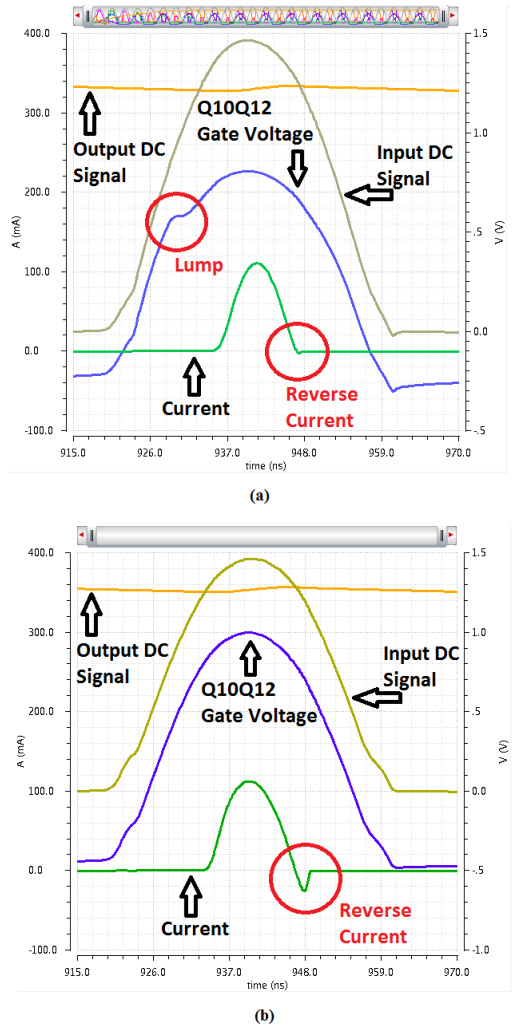


Figure 8: Comparative between waveforms using an acceptable capacitor size (a) and an over sized capacitor (b).

Capacitor C1 helps the circuit to work faster but the use of integrated capacitor can be a problem since it may take a lot of silicon area. In order to avoid it, a second voltage divider is used. The voltage between transistors Q13 and Q14 is designed to be just a little higher than the voltage between transistors Q9 and Q11. It allows transistor Q11 to turn ON a little earlier and to provide the delay shown in Figure 4 circle B to be a little shorter. Just as a comparison, if capacitor C1 is 10pF at 1.3Vac input signal voltage, the maximum reverse current, without the voltage divider, is almost 2 times larger than the use of a voltage divider. In order to have the same result without to use the voltage divider, capacitor C1 would need to be 97pF.

6 Details About the Capacitor

In order to choose an ideal size of capacitor, some considerations taken into account. The capacitor size depends on the resistance and on the frequency operation, which defines the time the capacitor needs to get fully charged. Since the circuit resistance changes during the circuit operation, the maximum capacitance will be calculated, and later a smaller capacitance value needs to be calibrated through simulations. The τ constant shown in equation 3 relates capacitor load time to the circuit resistance and capacitance

$$\tau = RC \tag{3}$$

During the time given by the RC constant, a capacitor gets 63% charged/discharged [12]. For a 99.3% load, it takes 5τ [13]. If a charge time of $T = 5\tau$ replace τ at equation 3 then the equation 4 will provide capacitance that can be used in the circuit.

$$C = \frac{T}{5R} \tag{4}$$

In order to obtain the time T on equation 4, the circuit frequency needs to be analyzed. The time for the capacitor be fully charged needs to be less than the time to signal AC changes from zero to the maximum voltage value. That time is exactly 1/4 of the cycle time. For 13.56MHz, it is 0.0184μs.

In order to determine the resistance R, the Thevenin method for equivalent resistance can be used. The R resistance applied to equation 4 results in a reference capacitance that needs to be calibrated by simulations. Its occurs because the R resistance is not constant in this circuit.

7 Type of Transistor

In this work halo implanted transistors of small channel were used. Although this kind of transistor has some problems such as threshold voltage mismatch, caused by the halo implant, [14] they offer the advantage of having less internal resistance [14]. This smaller internal resistance helps the circuit to waste less energy on its own operation.

8 Transistors Sizing

The transistors sizing takes 4 steps. The first step is used to obtain the current to be provided by the rectifier transistors. This amount of current sets the minimum size of the four rectifier transistors according to the technology used. After selecting the size of the four rectifier transistors, the others three steps are related with the comparator.

In the second step 2, transistors Q10 and Q12, which form the inverter are chosen. Those transistors need to be large enough to supply current to the rectifier. Besides that, Q10 and Q12 need to be designed to switch the inverter exactly at half of the DC signal voltage.

The next step would be the design of transistors Q9 and Q11 that form the main voltage divider and provide voltage to the inverter. Nevertheless, since the gate of Q11 is controlled by Q13 and Q14, the auxiliary voltage divider, therefore the voltage divider Q13Q14 needs to be designed first.

While the main voltage divider needs to provide a voltage of (AC signal)/2 to the inverter, the auxiliary voltage divider needs to provide a voltage larger than (AC signal)/2 to the gate of Q11 in order to turn it on a little earlier. Thus, in step 3 the voltage divider Q13Q14 needs to be designed to provide voltage ((AC signal)/2 + δ). In our work we have chosen $\delta = 100$mV, but a different voltage is acceptable.

The fourth step, after designing the inverted and the auxiliary voltage divisor, the main divisor needs to be designed to provide (AC signal)/2 voltage. Since the auxiliary voltage divider controls the gate of Q11 it can be a little difficult to obtain the Q9 and Q11 sizes. A simulation tool can help this task.

After the 4 steps, the capacitor needs to be calculated, as described in section 6.

Figure 9 summarizes all the steps described in this section.

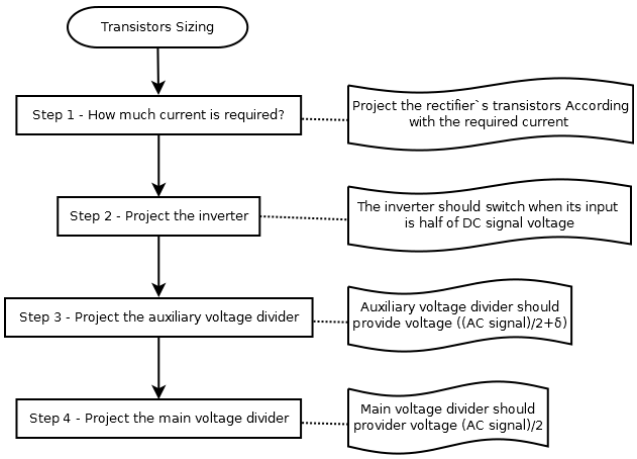


Figure 9: Transistors Sizing Flowchart.

9 Results

In order to simulate an input sine-wave at 13.56MHz, of 1.3V peak was used.

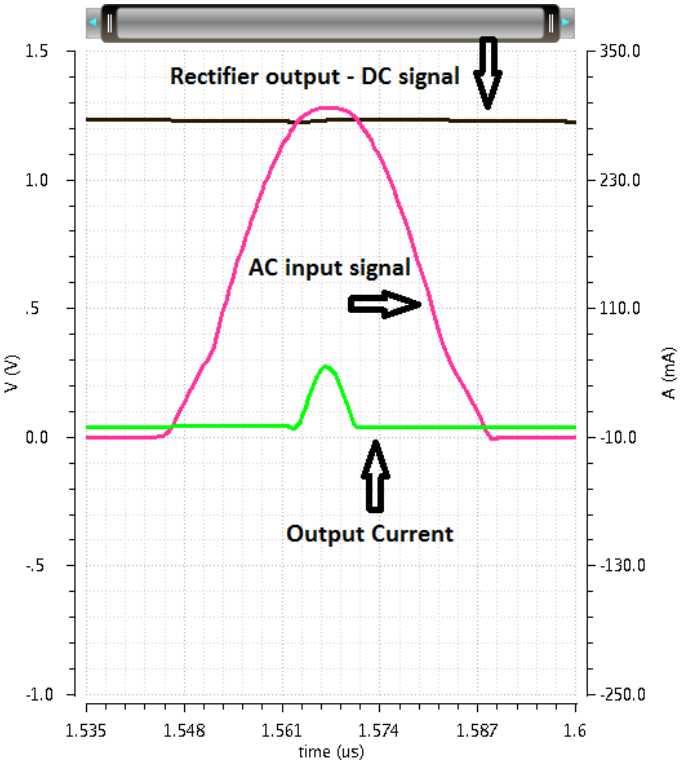


Figure 10: Output of wave form using the proposed comparator.

Figure 10 shows the output waveform. The output DC voltage is 1.2V and the output peak current is 55.55mA under a reverse current of -1.47mA. In order to trace this curve, a capacitor of 20nF and a resistor of 193 ohms were connected in parallel as load [6].

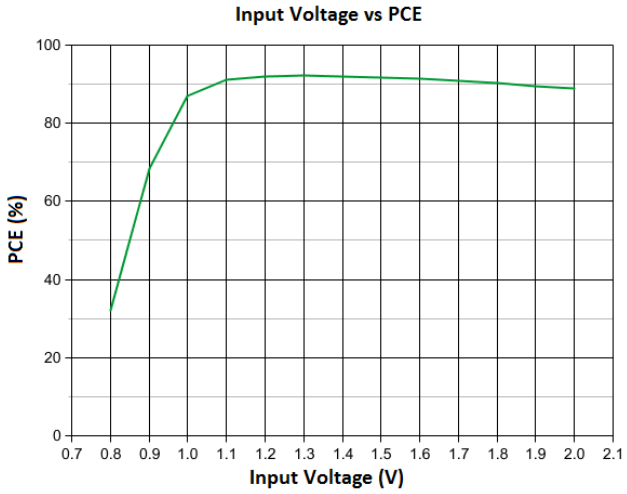


Figure 11: Relationship between input voltage and PCE.

The maximum $\eta_{rectifier}$ obtained is 0.92 and it was

Parameter	This Work	[11]	[5]	[4]
Input Frequency (MHz)	13.56	13.56	13.56	13.56
Input Amplitude (|Vac|)	1.3	1.5	3.8	3.5
Output Voltage (V)	1.2	1.33	3.1	3.2
R Load (KΩ)	0.193	1	0.1	1.8
Output Power (mW)	44.93	1.8	96	5.7
PCE (%)	92	81.9	80.2	87
Process Technology (μm)	0.13	0.18	0.5	0.35

Table 1: Comparison with previous works.

computed using the same method described in [6] that can be obtained from equation 2. The curve of input voltage vs PCE is shown in Figure 11.

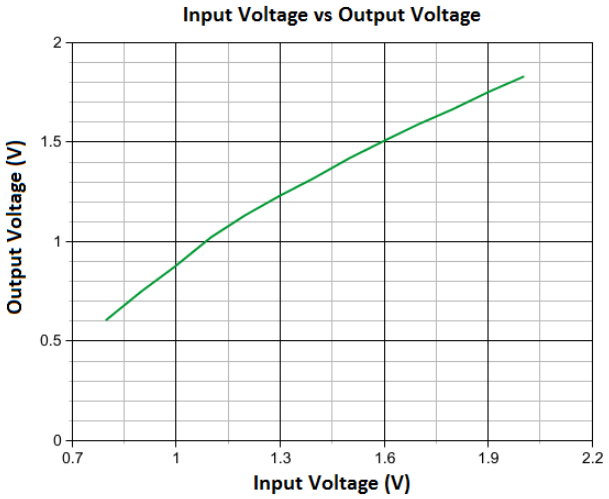


Figure 12: Relationship between input and output voltage.

Figure 12 shows the relationship between input and output voltage.

Table 1 shows a comparison to other works. It is possible to verify the PCE achieved was compatible with other reported rectifiers and that the output voltage is compatible with implantable devices.

10 Conclusion

In this extended version it is presented a low power rectifier, on a new operating condition, using a proposed comparator. The comparator is intended to reduce the switching delay of the rectifier transistors, in order to decrease the reverse current. It was accomplished by using an extra capacitor and an extra voltage divider as compared to the original two-input common gate comparator. Those comparator and rectifier were designed and simulated on 0.13μm CMOS process. The circuit was designed to work at 13.56MHz with an input signal of at least 1.3Vac. The rectifier efficiency increased by 10% when compared to the original paper and it can achieve a 92% of maximum efficiency at 1.2V.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This work was supported by FAPEMIG, CNPq and CAPES.

References

[1] J. R. C. Louzada, L. B. Zoccal, R. L. Moreno, and T. C. Pimenta, "A 0.13 um CMOS full wave active rectifier with comparators for implanted medical devices," in *2016 International Symposium on Integrated Circuits (ISIC)*, Dec 2016, pp. 1–4.

[2] M. Ghovanloo and K. Najafi, "A wideband frequency-shift keying wireless link for inductively powered biomedical implants," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 51, no. 12, pp. 2374–2383, Dec 2004.

[3] P. T. Theilmann, "Wireless power transfer for scaled electronic biomedical implants," Thesis, University of California San Diego, 2012.

[4] Y. H. Lam, W. H. Ki, and C. Y. Tsui, "Integrated low-loss CMOS active rectifier for wirelessly powered devices," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 53, no. 12, pp. 1378–1382, Dec 2006.

[5] H. M. Lee and M. Ghovanloo, "An integrated power-efficient active rectifier with offset-controlled high speed comparators for inductively powered applications," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 58, no. 8, pp. 1749–1760, Aug 2011.

[6] H.-K. Cha and M. Je, "A single-input dual-output 13.56 MHz CMOS AC–DC converter with comparator-driven rectifiers for implantable devices," *Microelectronics Journal*, vol. 45, no. 3, pp. 277 – 281, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0026269214000135

[7] Y. Lu and W.-H. Ki, "A 13.56 MHz CMOS active rectifier with switched-offset and compensated biasing for biomedical wireless power transfer systems," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 8, no. 3, pp. 334–344, June 2014.

[8] R. Ramzan and F. Zafar, "High-efficiency fully cmos vco rectifier for microwatt resonant wireless power transfer," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 62, no. 2, pp. 134–138, Feb 2015.

[9] L. Yao, J. H. Cheong, R.-F. Xue, and M. Je, "An inductive coupling based CMOS wireless powering link for implantable biomedical application," *world academy of science, engineering and technology*, vol. 69, pp. 456–459, 2012.

[10] A. Khalifa, "Study of CMOS rectifers for wireless energy scavenging," Thesis, Linköping UniversityLinköping University, Electronic Devices, The Institute of Technology, 2010.

[11] H. K. Cha, W. T. Park, and M. Je, "A CMOS rectifier with a cross-coupled latched comparator for wireless power transfer in biomedical applications," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 59, no. 7, pp. 409–413, July 2012.

[12] Federal University of Uberlândia, "Charging and Discharge of a Capacitor," http://www.facip.ufu.br/sites/facip.ufu.br/files/Anexos/Bookpage/fe3-06-carga-e-descarga-de-um-capacitor.pdf, acessado: 13 jan. 2017.

[13] M. A. Cavalcante, C. R. C. Tavolaro, and E. Molisani, "Physics with Arduino for beginners," *Revista Brasileira de Ensino de Física*, vol. 33, pp. 4503 – 4503, 12 2011.

[14] G. D. C. Colletta, "An asynchronous sigma-delta modulation architecture in ultra-low power for biomedical applications." Theses, Federal University of Itajubá, 2015.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1026-1033 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Multiclass Myoelectric Identification of Five Fingers Motion using Artificial Neural Network and Support Vector Machine

Jawad Ahmad[*1], Ammar Mohsin Butt[2], Muhammad Tanveer Riaz[1], Shoaib Bhutta[1], Muhammad Zeeshan Khan[1], Inam-Ul-Haq[3]

[1]*State Key Laboratory of Power Transmission and System Security, School of Electrical Engineering, Chongqing University, Chongqing 400044, China.*

[2]*Department of Electrical Engineering, International Islamic University, Islamabad 44000, Pakistan*

[3]*Department of Electrical Engineering, National University of Science and Technology (NUST), Islamabad 44000, Pakistan*

ARTICLE INFO

Article history:
Received: 30 May, 2017
Accepted: 28 June, 2017
Online: 15 July, 2017

Keywords:
Electromyography
Discrete Wavelet Transform
Artificial Neural Network
Support Vector Machine

ABSTRACT

The research in Neuro-Prosthetics is gaining more significance and popularity as the advancement in prosthetics control allows amputees to perform even more tasks. Indeed, the improvement of classification accuracy is a challenge in prosthetics control. In this research, a system is developed in order to improve the multiclass classification rate. Two classifiers namely Artificial Neural Network(ANN) and Support Vector Machine(SVM) are trained to recognize five different myoelectric motions of hand fingers. The Electromyography(EMG) signals are acquired using surface electrodes placed on the forearm at specific nodes. The signal conditioning is performed using two stage filtering and amplification followed by digitization process. The final version of EMG signals is correlated in joint time and frequency domain for best feature vectors done via Discrete Wavelet Transform (DWT). The feature vectors are used to train the ANN and SVM. The classification results show an exceptional performance of ANN with classification accuracy of 98.7%. over the SVM, which is 96.7%.

1. Introduction

A lot of work is done in the field of bionics and kinematics. The purpose is to identify the finger movements. Electromyography is a procedure commonly used to assess the muscular activity created by skeletal muscles when fired by neurons. These signals can be used to detect muscular diseases and disorders etc. Some of the applications are found in prosthetics control where EMG signals are used for the robotic control of hands, arms and fingers. The most frequently used approaches for acquiring EMG signals are surface and intramuscular methods. In first case, surface electrodes are placed on specific nodes on forearm in pair form. A reference or a neutral point is defined usually taken as bone and a single electrode is placed on it. The EMG signal picked up from nodes have a small amplitude usually in millivolts and contains noise of line and skin impedance. Signal conditioning of these EMGs is needed to further processing. Feature extraction is a significant step following signal acquisition. It helps to identify and extract useful information present in signals. These feature vectors are used for training, validation and testing of the classifier used for classification. Thus, the pattern recognition method has an important role.

A technique used for multi-channel classification of surface EMG using support vector machine (SVM) and signal based wavelet is reported in [1]. The algorithm is trained to classify EMG signals of six hand movement. The technique showed an overall classification accuracy of 95%.

There are some works which describe the use of SVM either in medical findings [2] or related to myoelectric prosthetics control. Crawford *et al.* [3] described the use of SVM for Myoelectric classification. The purpose is to control a robotic arm with 4 degrees of freedom. A set of seven single electrodes placed on the forearm were employed to distinguish eight hand gestures. The results showed a classification rate was over 90%,

The hand gesture EMG recognition based on ANN technique has been reported in [4,5]. A back-propagation ANN classifier with Levenberg-Marquardt training algorithm was used. The network used features in time and frequency domain. The results

[*]Corresponding Author: Jawad Ahmad, Chongqing University, Contact: +86-157-36123585, Email: Jawad.ahmad@cqu.edu.cn

showed that ANN can classify six different hand movements (left, right, up and down) with an error rate of 11.6 %.

Similarly, Bitzer and Smagt have also used SVMs to identify six different finger actions with ten double differential electrodes placed near to the wrist [6]. The accuracy was between 90-94% for relax and pronation postures of arm. Another work reported by Lucas *et al.* [7], shows a multi-channel classification problem where Myoelectric identification was performed in order to distinguish six different movements of the hand based on SVM and wavelet coefficients. Final results show that the optimal wavelet selection favors the performance of classifier, giving an average error rate of 5%.

Some work related to the five fingers motion identification has been reported in [8]. The two-layered feed forward neural network was trained to identify 16 features from five fingers motion. The ANN was able to classify the features with the maximum classification rate of 96.7%.

The main aim of this research is to improve the classification rate by designing a hardware and software based system. A hardware module consists of filters and amplifiers is designed for capturing an optimum EMG signal in meaningful frequency range. Firstly, the EMGs are acquired using electrodes placed on nodes and are conditioned by filters and amplifiers. The conditioned signals are digitized and read in MATLAB which are then subjected to transformation in joint time and frequency domain by Discrete Wavelet Transform (DWT). The features returned by DWT are availed to trained the ANN used to classify five finger movements. A total of 500 utilized feature vectors will be fed up to the input of neural network for training, validation and testing. To compare the performance results, SVM is also trained. Furthermore, the classification performance of ANN and SVM is compared when classifiers are subjected to testing data. Finally, the efficiency and robustness of SVM and ANN is also discussed.

The classification results of ANN show an improved classification rate of 98.7% achieved in five fingers classification. The performance results of ANN showed an edge over the performance of SVM which is 96.7%. The network robustness is depicted by the superb performance plots of the neural network.

2. Material and Method

This research is carried out in flow stated in Figure 1 below. The block diagram shows the phases for EMGs processing.

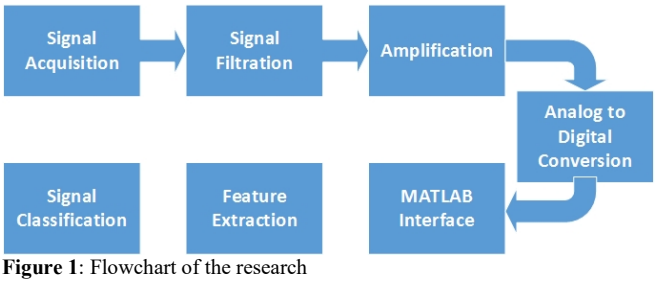


Figure 1: Flowchart of the research

2.1. Electrode Placement

Signals are acquired from five fingers using surface EMG Electrodes. Gelled Ag/AgCl electrodes are used as it helps in creating good contact between skin and electrode thus mitigating skin impedance. To avoid crosstalk and distortion suitable size and distance of electrodes is chosen. The size of electrode is chosen 10mm as suggested by SENIAM [9] and European inventory. For each finger two electrodes are used in pair configuration. These electrodes are placed between motor units and tendon insertion of muscles with the inter-electrode distance of 20mm between them. For each finger, the electrodes are placed on forearm on specific nodes as shown in Figure 2. The reference electrode is placed at back side of hand on bone giving a reference signal to differential amplifier used for filtration in next step.

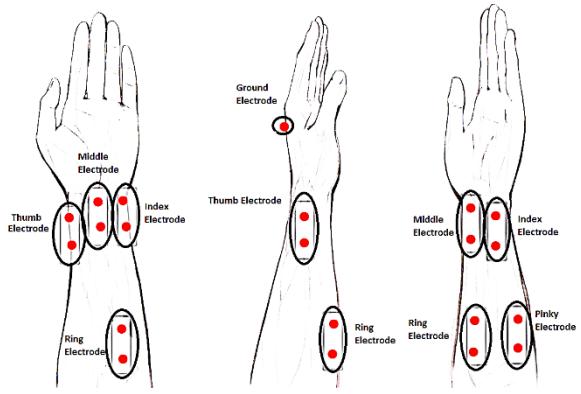


Figure 2: Electrode placement on forearm for signal acquisition

A stereo cable is used to connect electrode and the signal conditioning circuit. One end of the cable has a dual-channel stereo jack and other end consists of alligator clips.

2.2. Signal Conditioning Circuit

In order to remove noise from EMGs, a signal conditioning circuit is implemented as shown in Figure 3. It consists of five channels for five fingers respectively. Stereo jack sockets are utilized to connect the connecting wires.

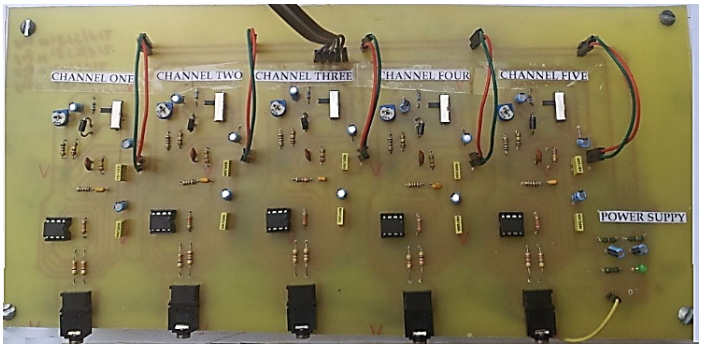


Figure 3: Hardware module used for signal conditioning

2.3. Differential Filtration

Signals acquired from electrodes contain noise due to skin Impedance and connecting leads. An instrumentation amplifier AD623 is used to remove noise. The signals are read with respect to reference electrode and their difference is amplified as shown in (1). The common mode rejection capability of the amplifier removes noise from signals. The read signals are free of noises such as power lines noise. Indeed, the CMMR is an important factor for any differential amplifier as it eliminates any correlated signals that come from power sources or electromagnetic devices.

The CMMR determines how well the input signals have subtracted. AD623 provides better CMRR, high input impedance and strong DC voltage suppression, therefore, keeping the error to a minimum. The CMRR for amplifier remains constant up to 200 Hz. Figure 4 shows the schematic of the instrumentation amplifier used for differential filtration.

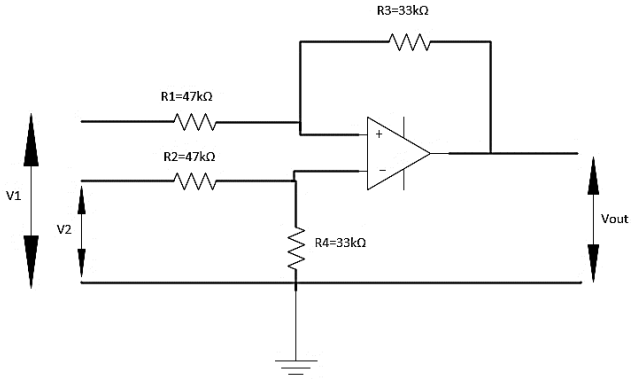


Figure 4: Circuit diagram for differential amplifier

$$V_{out} = \frac{R_3}{R_1}(V_2 - V_1) \quad (1)$$

Where R1=R2=47 kΩ and R3=R4=33 kΩ

2.4. Bandpass Filtration and Amplification

To select the useful range of frequencies and amplification, a low voltage operational amplifier LMV358 having voltage range between 2.7 V and 5.5 V is used. LMV358 is dual low voltage amplifier. First op-amp is used as frequency selective bandpass filtering and amplification. The frequency range is selected between 100-500 Hz. Figure 5 shows the schematics of the bandpass filter and amplifier.

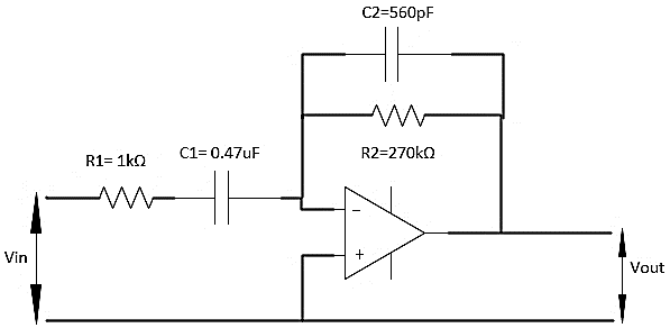


Figure 5: Circuit diagram for bandpass filter

In figure given above R1 = 1 k, C1 = 0.47uf, R2 = 270 k, C2 = 560pf. The equations for bandpass filter are given as follows:

$$Gain(Av) = -\frac{R_2}{R_1}, f_{c1} = \frac{1}{2\pi R_1 C_1}, f_{c2} = \frac{1}{2\pi R_2 C_2} \quad (2)$$

To enhance the resolution of recoded signals, amplification is also required. Amplifier of high quality having adjustable gains is used. The second operational amplifier in LMV358 is used as inverting amplifier with adjustable gain. The values of resistors are selected such that the op-amp provides negative gain of 250, therefore, during each acquisition of EMG signal the signal to noise ratio gets maximized and EMG signals are amplified to appropriate levels. Figure 6 shows the schematic of an inverting amplifier used for amplification.

$$Gain(Av) = \frac{V_{out}}{V_{in}} = -\frac{R_f}{R_{in}} \quad (3)$$

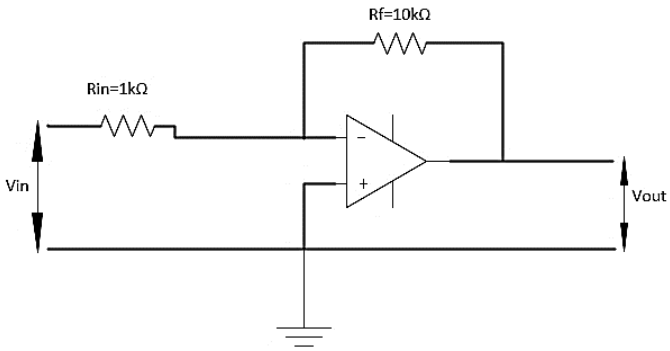


Figure 6: Circuit diagram for an inverting amplifier

After amplification, EMG signals are subjected to a peak detector circuit. It consists of diode which renders only positive deflections of the signal. The random nature in amplitude of EMG signals exists even after rectification. To extract amplitude related information, the concept of smoothing is adopted by using the RC combination as a low pass filter to suppress high frequency fluctuations in the EMG signal. The deflections in EMG signal appear smooth and the peak value of EMG signal is easily captured by RC combination.

After passing the EMG signals through various steps of signal conditioning, filtration, amplification and rectification, the final signal is obtained as shown in Figure 7. This is the positive signal with average amplitude lies in range of 0-5V as ensured through previous amplification.

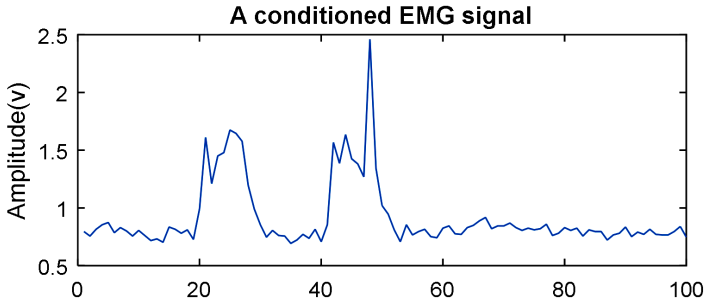


Figure 7: The final waveform of a conditioned EMG signal

2.5. Analogue to Digital Conversion

Signal conditioning steps remove noise and amplify the signals to appropriate levels. These signals then need to be used in Matlab for further processing. An Arduino Uno board based on ATmega328P is used for analog to digital conversion. It uses a built-in 10-bit ADC at a sampling rate of 8.6 kHz according to the Nyquist criteria. The analog voltages are mapped between 0 and 3 V with resolution of 1024.

Digital signals from the Arduino are read in Matlab through serial interface of Arduino. 9600 baud rate is used. For ease of coding a hardware support package of Arduino is used in Matlab.

2.6. Feature Extraction

Feature extraction is important step for any classification problem as it helps to extract important information from the signals. The features acquired in this step are used as training data to classifier. Wavelet transform transforms the stochastic EMG signals into time and frequency based information which is helpful in finding the harmonic contents and better diagnostic possibilities as compared to the conventional Fourier transform. The rapid variations within a signal can be accurately studied and analyzed using wavelet transform. As can be seen in the Figure 9, the Fourier transform forms a constant frequency resolution as the signal is divided into constant time window. However, in the case of wavelet transform the signals are divided into different time windows with different range of frequencies which provides better allocation for high frequency components in the signal. Another advantage of wavelet transform is simple computation and less complexity as compared to the Fourier transform.

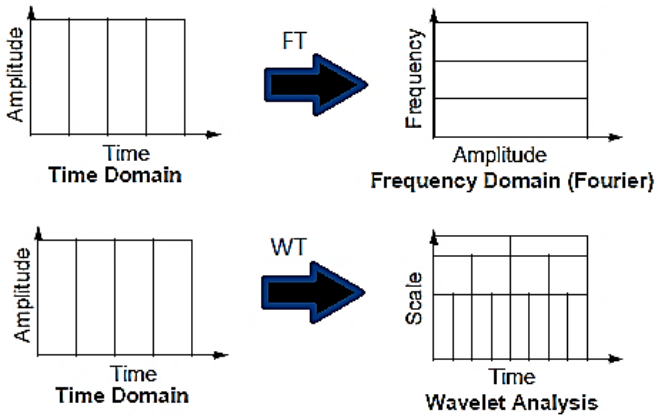


Figure 9: Comparison between Fourier and wavelet transform

Since, the DWT exhibits good frequency resolution at low frequencies and good time resolution at high frequencies [10], therefore, the DWT is chosen as a feature extraction technique for surface EMG signal. The DWT coefficients of a signal *x(n)* may be obtained as:

$$C(a,b)=\sum_{n=z} x[n]\Psi_{a,b}[n] \tag{4}$$

$$\Psi_{a,b}[n]=\frac{1}{\sqrt{a}}\Psi\left\{\frac{n-b}{n}\right\} \tag{5}$$

Where *a* is scale, *b* is translation and ψ(n) is a discrete wavelet. The Daubechies 4 or db4 wavelet window is used as it provides best correlation with the EMGs, therefore, giving best feature vectors for training the classifiers. The DWT is achieved by passing the signal *x[n]* through two complementary filters, a low pass filter having impulse response *g[n]* and a high pass filter having impulse response *h[n]* as shown below:

$$y_g[n]=\sum_{k=-\infty}^{\infty} x[k]g[2n-k] \tag{6}$$

$$y_h[n]=\sum_{k=-\infty}^{\infty} x[k]h[2n-k] \tag{7}$$

The application of two filters gives the approximation and detailed coefficients containing information about the shape and sharp variations of the signal respectively. For capturing sharp variations in the five fingers EMG signals, the detailed coefficients are used as features for training the classifiers.

Hence, the DWT helps to analyze the EMG signals at different frequency band with different resolution which is helpful to extract best feature vectors for training ANN and SVM.

2.7. Feature Classification

The features extracted from EMGs are utilized to train ANN and SVM.

2.8. Support Vector Machine (SVM)

Support Vector Machine (SVM) was first introduced by Vapnik and Chervonenkis at 1965. SVM is a supervised learning technique used in the field of machine learning. In SVM, the training is reformulated in such a way to achieve a quadratic programming problem whose solution is unique and global. The detailed information regarding the SVM can be found in [11]. SVM is a binary classifier whose output is estimated by the given empirical data:

$$(x_1,y_1),....,(x_m,y_m)\in R^N\times\{\pm 1\} \tag{8}$$

The problems related to non-linear problems are solved by mapping the original data into a feature space. The mapping data is linearly separable in feature space. Function $\varphi(x)$, which maps the training vector x_i into a higher dimensional space, requires to belong to a dot product space. The dot product of mapping function is termed as kernel:

$$k(x_i,x_j)=\varphi(x_i)^T\varphi(x_j) \tag{9}$$

The group of data that is mapped in high dimensional linear space is divided in two labeled classes by a hyperplane. The labeled classes are shown as:

$$w^T\varphi(x)+b=0 \tag{10}$$

$$w\in R^N, b\in R$$

To ensure maximum margin between separating classes, the selection of a suitable hyperplane is necessary. To construct this hyperplane, we need to solve the following quadratic problem:

$$\min\frac{1}{2}\|w\|^2+C\sum_{i=1}^{m}\xi_i \tag{11}$$

$$y_i(w^T\theta(x_i)+b)\geq 1-\xi_i,\xi_i\geq 0, i=1,2.....m$$

where ξ is called the slack variable, which takes values for each x_i. It is related to the concept of soft margin, whereas, C is the tuning parameter used to tolerate errors and balance the training data. Equation (11) can be solved by the introduction of a Lagrange multipliers α_i through next expression as:

$$\max \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j k(x_i, x_j) \tag{12}$$

$$\sum_{i=1}^{m} \alpha_i y_i = 0, 0 \le \alpha_i \le C, \forall i \tag{13}$$

Solving (12) leads to an optimal decision function as:

$$f(x) = \operatorname{sgn}\{\sum_{i=1}^{m} \alpha_i y_i k(x_i, x) + b\} \tag{14}$$

The decision function shown in (14) depicts that the classifier has an extension in terms of a subset of the training data. Those patterns whose α_i is non-zero, are termed as support vectors. The non-linear classification can be achieved by introducing different kernel functions, which computes the inner product of vectors $\varphi_{(x_i)}$ and $\varphi_{(x_j)}$. These typical kernels functions include lineal, quadratic and radial functions. The SVM inherently is a binary classifier and performs classifications on two class problems. To solve multiclass problems, like in our case for five fingers motion classification, we need to extend the capabilities of SVM to multiclass approach. Therefore, for multiclass SVM, (8) can be represented as:

$$(x_1, y_1)_1, ..., (x_m, y_m)_k \in R^N \times \{1, ...k\} \tag{15}$$

The theme is to train *5* independent binary classifiers so that each one is trained to discriminate the training samples in only one class against all the remaining classes. For classifying a new sample, *5* classifiers work separately, the one giving largest output is chosen as an estimated class. This scheme is referred to as the *one-against-all*. It is simple, fast and accurate method to implement. The kernel function used for separating patterns is a quadratic function having box constraint level equal to 1.

2.9. Artificial Neural Network

In this paper, the pattern recognition of the features is done using the ANN as shown in Figure 10.

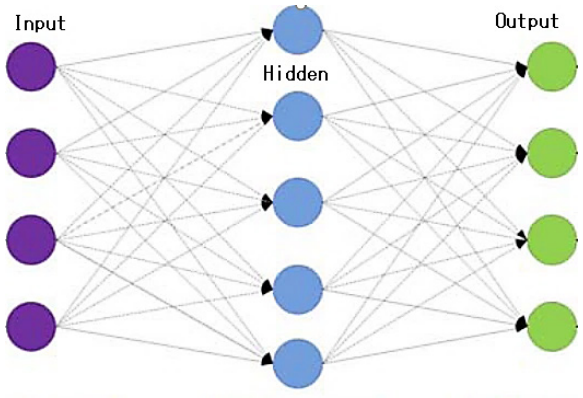


Figure 10: The structure of Neural Network

The first output neuron in the hidden layer is given by (16).

$$a^1 = f^1(IWp + b^1) \tag{16}$$

Whereas a^1 is the output vector from the input layer, *IW* is the weight matrix of input, f^1 is the hidden layer transfer function and b^1 is the hidden layer bias function. Similarly, the first output neuron of the output layer can be expressed as:

$$a^2 = f^2(LW(f^1(IWp + b^1)) + b^2) \tag{17}$$

Whereas a^2 is the output vector of output layer, LW is the weight matrix of output, f^2 is the output layer transfer function and b^2 is the bias vector of output layer. The network is a two-layered feed forward network having with a sigmoid transfer function in hidden layer and a SoftMax transfer function in the output layer as depicted in (18) and (19) respectively.

$$f(n) = \frac{1}{1 + e^{-n}} \tag{18}$$

$$h_\theta(x) = \frac{1}{1 + \exp(-\theta^T x)} \tag{19}$$

The network has 10 neurons in the hidden layer. The network is trained using scaled conjugate gradient backpropagation. The backpropagation is performed on the whole feedforward network in order to improve the results of ANN. This technique is referred to as fine tuning of a whole ANN. The theme is to retrain the whole network in a supervised fashion. In backpropagation training, the goal of training is to mitigate $\lambda(w)$ which is the square error taken over all n vectors.

$$\lambda(w) = \sum_{p=1}^{n} \sum_{k=1}^{k} (d_{kp} - o_{kp})^2 \tag{20}$$

Where d_{kp} and o_{kp} is the desired output and network output for node k and p respectively During training, the backpropagation algorithm fine-tunes the weights that are randomized initially respective to the steepest gradient along the surface of error. Weights are set as per their influence to the output which is done by the method of reusing the squared error signal through weights layers. Training is proceeded by the presentation of each normalized input vector z and its equivalent output target vector d, until each $\lambda(w)$ becomes smaller than the maximum allowable error $\lambda(w)_{max}$. The presentation of complete set of training data to the network is called an epoch.

The performance of ANN is evaluated based on cross entropy. During the learning process, the ANN goes through various stages in which the reduction rate of error is very slow which can influence the learning process. For solution, the mean square error (MSE) is replaced by cross entropy error function. The results of

error function show a better performance having a shorter stagnation period. The original MSE function for all training patterns is given by

$$E_m = \frac{1}{m}\{\sum_{k=1}^{m}\left(t_k - y_k\right)^2\} \quad (21)$$

where t_k represents the target value and t_y is the actual value of network. In the backpropagation technique, the error through the iterative updates of weights is minimized for all training patterns. In practice, this approach allows the network to have a good performance but convergence rate is slow. Therefore, in order to accelerate the backpropagation, we propose the following cross entropy error function to be minimized:

$$E_m = \frac{1}{m}\{\sum_{k=1}^{m}[t_k \ln y_k + (1-t_k)\ln(1-y_k)]\} \quad (22)$$

To minimize the error E_m, each weight w_{jk} is updated proportional to the partial derivative of E_m, with respect to weight. Using the mean square error, the partial derivative of E_m with respect to w_{jk} is

$$\frac{\partial E_m}{\partial w_{jk}} = \sigma.(y_k - t_k)y_k(1-y_k).z_j \quad (23)$$

By using the cross-entropy error function, (20) becomes:

$$\frac{\partial E_m}{\partial w_{jk}} = \sigma.(y_k - t_k).z_j \quad (24)$$

Hence, the error signal that propagates back from each output unit becomes proportional to the difference between target value and actual value. This leads to a better network performance along with shorter stagnation period. A total of 350 samples are used to trained the ANN. The network performance is tested using 150 samples of EMG data.

3. Experiments

3.1. Simulation of Signal Conditioning Circuit

Schematics and simulations of the circuit are carried out using Proteus Software. The differential filter removes the noise by amplifying difference of two signals. This removes the noise due to the skin impedance and terminals of the connecting wires. The bandpass filter selects the frequency range of 100-500 Hz and provides 250x gain, thereby, removing the power line noise and giving an optimum signal in meaningful frequency range. The millivolt signal is amplified to more than 2 volts which is appropriate level for further processing. The impedance of wires interfacing the electrodes and the circuit board is matched to, that of the board mainly, the input of differential amplifier for ensuring the maximum power transfer.

The circuit's performance shows that this configuration of hardware components is suitable for EMG signal acquisition. The fingers movement is depicted in Figure 11. The EMG signal is recorded as a result of individual finger movement. A total of 500 samples are taken from finger patterns which is then used for feature extraction followed by the training, validation and testing of the classifier (ANN).

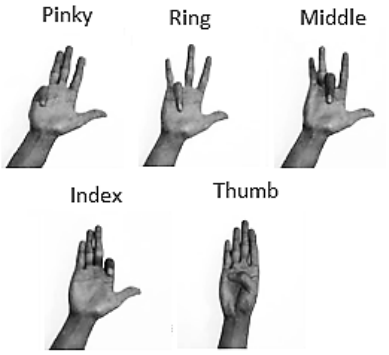


Figure 11: The five fingers movement pattern for EMG signal acquisition

The EMG data acquired from five fingers are sent via serial interface to the Matlab workspace. Shown in Figure 12 are signals for five fingers read through serial port of Arduino in Matlab. The Arduino Hardware support package is used in Matlab for having simplicity in serial interface. This package allows to import the libraries of Arduino in Matlab, therefore, giving a sophisticated way to program the controller in Matlab and to read serial data from its serial interface. The data read in Matlab is stored in an array form in the workspace of Matlab.

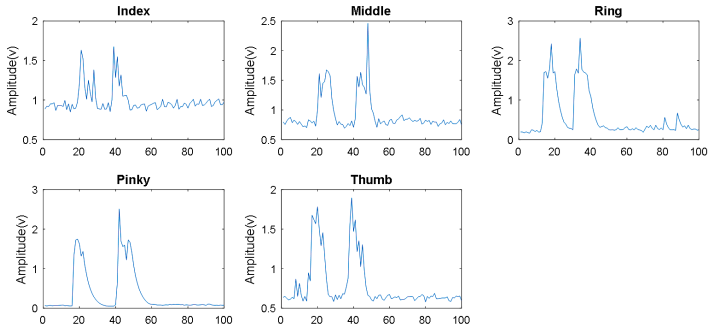


Figure 12: The five fingers EMG signals read in Matlab

3.2. Feature Extraction and Classification

The feature extraction and classification was performed in Matlab 2017a software. Discrete Wavelet Transform (DWT) was used for feature extraction from EMG signals. The Daubechies 4 or db4 wavelet window is used for best correlation with EMGs as discussed previously. It provides approximation and detailed coefficients containing information related to shape of signal and sharp variations in the signal respectively. The detailed coefficients or high frequency components as shown in Figure 13 are used as feature vectors for training of the classifier.

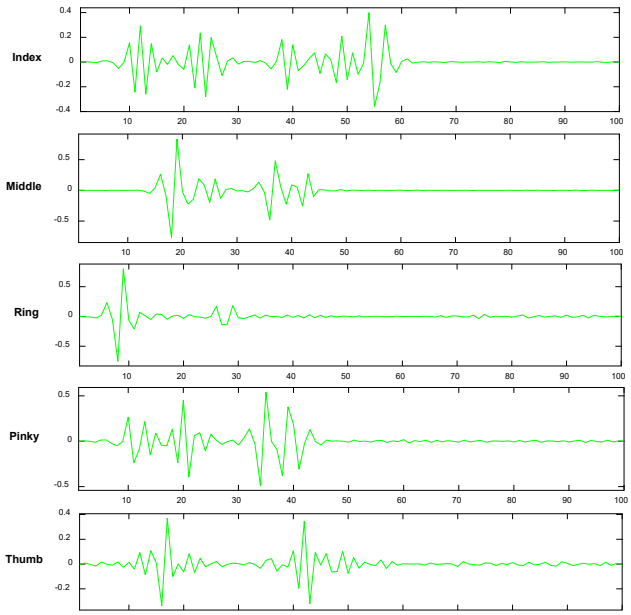


Figure 13: The DWT detailed coefficients of five fingers respectively.

3.3. Cross Entropy

During training the ANN, there is a measure of error between computed outputs and the desired target outputs of the training data. The most common measure of error is called mean squared error. The performance of ANN based on cross entropy is shown in Figure 14(A). The results show best performance of 0.024758 at epoch 27. The steep fall of training curve shows best performance of ANN achieved during training. The performance of ANN is tested by 150 samples of data which is depicted in the curve of testing. The ANN converges to output in a very small stagnation time as can be seen in Figure 14(A).

3.4. The Confusion Matrix

The confusion matrix is used to estimate the quality of the output of a classifier on the specified data set, providing an insight of the errors and specially the types of errors made during the classification. The diagonal elements characterize the number of samples correctly classified by the classifier, while non-diagonal elements are those that are misclassified by the classifier. The higher values in diagonal depicts the better accuracy of the classifier. The theme is to divide the correct predictions by total no of predictions made by the classifier to output the percentage accuracy as given below.

$$Accuracy = \frac{\sum_{p=1}^{n}(y_p)}{n} * 100 \quad (25)$$

Whereas y_p corresponds to correct classification taken over n patterns. The confusion matrix showing the results of ANN classification using 10 hidden neurons are shown in Figure 14(B). The confusion matrix of SVM is shown in Figure 15.

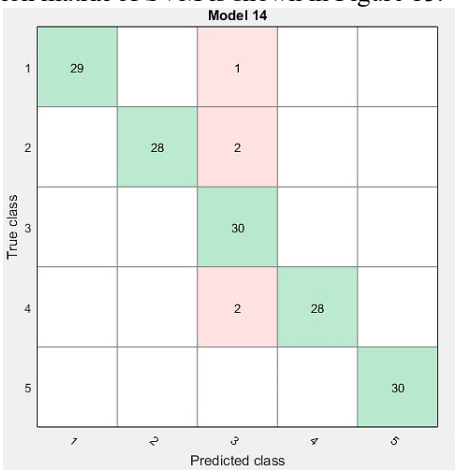


Figure 15: The confusion matrix of SVM

The overall accuracy of classification is calculated to be 96.7%. The ANN matrix shows an excellent classification accuracy of 98.7% which is an edge over the accuracy of SVM. The classification accuracies using SVM for each finger are given in Table 1 whereas the classification results of ANN are listed in Table 2.

3.5. The Receiver Operating Characteristics

Receiver Operating Characteristic (ROC) curve is also used to assess the classifier output performance. ROC curve plots the True Positive Rate (TPR) on Y axis, and False Positive Rate (FPR) on X axis. This characterizes an ideal point, a false positive rate, to lie on the top left corner of the plot, giving a FPR of zero, and a TPR of one. This type of characteristic gives a greater area under the curve (AUC) for achieving best performance. The Figure 14(C) shows the ROC of ANN output. Figure 16 shows the ROC of

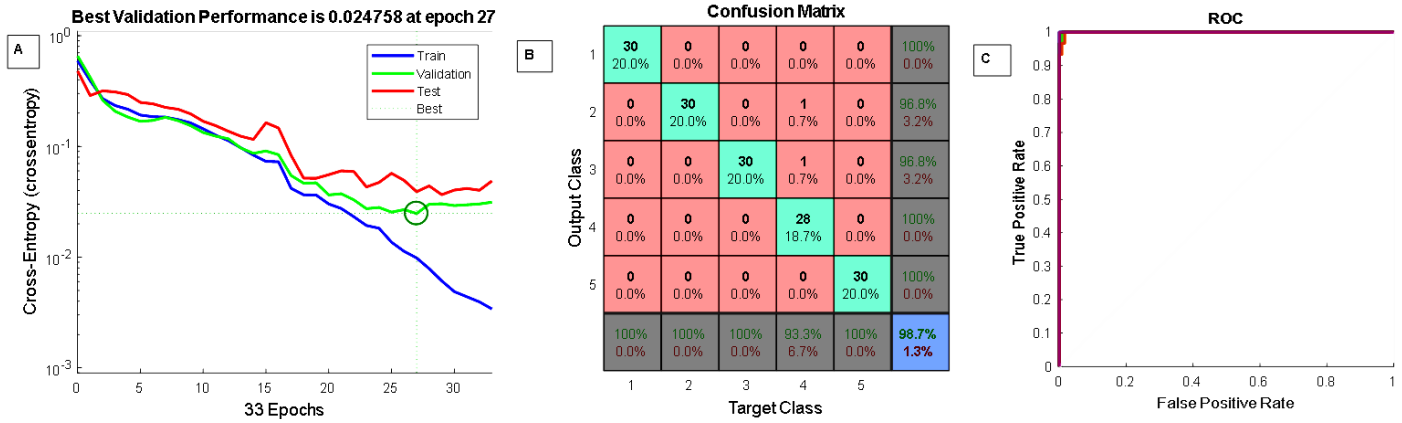


Figure 14: The plots showing the performance of ANN. (A) shows the performance of ANN training, validation and testing respectively. (B) shows the confusion matrix classification results of ANN obtained using 10 hidden neurons. (C) shows the ROC for five classes respectively.

SVM. The ROC curve of ANN is the representation of an excellent performance, giving AUC = 0.983, whereas, the ROC for SVM is 0.978. The steepness observed in curves of both ANN and SVM is also very significant, thereby, maximizing the TPR while minimizing the FPR.

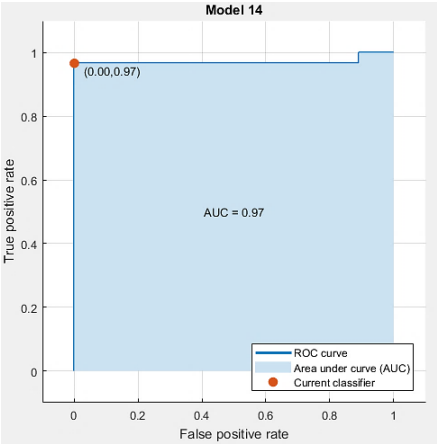


Figure 16: The ROC of SVM

Table 1 Classification accuracy of five fingers using SVM

Finger Name	Total Samples	Classified samples	Accuracy	Error
Index	30	29	96.67%	3.33%
Middle	30	28	93.3%	6.7%
Ring	30	30	100%	0%
Pinky	30	28	93.3%	6.7%
Thumb	30	30	100%	0%
Overall Accuracy & Error			96.7%	3.3%

Table 2 Classification accuracy of five fingers using ANN

Finger Name	Total Samples	Classified samples	Accuracy	Error
Index	30	30	100%	0%
Middle	30	30	100%	0%
Ring	30	30	100%	0%
Pinky	30	28	93.3%	6.7%
Thumb	30	30	100%	0%
Overall Accuracy & Error			98.7%	1.3%

4. Conclusion

A noise free and an optimum EMG signal is captured by the signal conditioning circuit. The feature vectors obtained through DWT is best for feeding the classifier as it provides maximum correlation with EMGs. The binary SVM is made capable of solving five class finger recognition problem via multiclass approach. The SVM showed an overall accuracy of 96.7% with an average error rate of 3.3%. The two-layered feed forward network showed an excellent performance over the SVM in the classification of EMG signals showing a minimum error of 1.3%. The accuracy is improved to 98.7% with 10 neurons being used in hidden layer. The network is converged in a small stagnation time. In future work, the two-layered feed forward network can be extended to multilayered deep networks, for improving the classification accuracy. Moreover, the multiclass SVM and ANN network proposed can be used to recognize several hand gestures at certain degrees of freedom proving more robustness in the prosthetics control specially for amputees.

Conflict of Interest

No any.

Acknowledgment

We thank Ammar Mohsin Butt, Inam-Ul-Haq, Muhammad Tanveer Riaz, Shoaib Butta and Muhammad Zeeshan for their assistance.

References

[1] M. -Franc¸oise Lucas, A. Gaufriau, S. Pascual, C. Doncarli, D. Farina," Multi channel surface EMG classification using support vector machines and signal-based wavelet optimization," Biomedical Signal Processing and Control, vol.3(2), pp. 169–174, November 2007.

[2] G. Kaur, A. S. Arora, and V. K. Jain, "Multi-class support vector machine classifier in EMG diagnosis," WSEAS Trans. Sig. Proc., vol. 5, pp. 379-389, 2009.

[3] B. Crawford, K. Miller, P. Shenoy, and R. Rao, "Real-time classification of electromyographic signals for robotic control," Proc. 20th Nat. Conf. Artificial Intell. (AAAI 05), 2006, pp. 523-528.

[4] M.R. Ahsan, M.I. Ibrahimy, O.O. Khalifa, "Hand Motion Detection from EMG Signals by Using ANN based Classifier for Human," Computer Interaction, Modeling, Simulation and Applied Optimization (ICMSAO), 2011 4th International Conference on , pp.1-6, April 2011.

[5] M.R. Ahsan, M.I. Ibrahimy, O.O. Khalifa, "Electromygraphy (EMG) Signal based Hand Gesture Recognition using Artificial Neural Network (ANN),"Mechatronics (ICOM), 2011 4th International Conference On , pp.1-6, May 2011.

[6] S. Bitzer and P. van der Smagt, "Learning EMG control of a robotic hand: towards active prostheses," Proc. IEEE Int. Conf. Robotics and Automation (ICRA 06), 2006, pp. 2819-2823.

[7] M.-F. Lucas, A. Gaufriau, S. Pascual, C. Doncarli, and D. Farina, "Multi-channel surface EMG classification using support vector machines and signal-based wavelet optimization," Biomed. Signal Process. Control, vol. 3, pp. 169-174, 2008.

[8] M. Ariyanto *et al*., "Finger movement pattern recognition method using artificial neural network based on electromyography (EMG) sensor," 2015 International Conference on Automation, Cognitive Science, Optics, Micro Electro-Mechanical System, and Information Technology (ICACOMIT), Bandung, 2015, pp. 12-17. doi: 10.1109/ICACOMIT.2015.7440146

[9] Freriks, B., Hermens, H., Merletti, R., Stegeman, D., Blok, J., & Rau, G. European Recommendations for Surface Electromyography Results of SENIAM Project. Roessingh Research and Development, (2000) pp. 13-54.

[10] Eldin Henry Shroffe D, P. Manimegalai, "Hand Gesture Recognition Based on EMG Signals Using ANN," International Journal of Computer Application Issue 3, Vol. 2, April 2013.

[11] V. N. Vapnik, Statistical Learning Theory, Danvers, MA: Wiley Interscience,1998.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1034-1041 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Group of UAVs Moving on Smooth Control Law with Fixed Obstacles

Dmytro Pavlovich Kucherov[1*], Andrei Nikolayevich Kozub[2], Oleg Nikolayevich Kostyna[3]

[1]*National Aviation University, computerized control system Department, Educational and Scientific Institute of Computer Information Technologies, 03680, Ukraine*

[2]*National University of Defense of Ukraine, usage space system and GIS Department, Institute, 03049, Ukraine*

[3]*Central Research Institute of Armament and Military Equipment of the Armed Forces of Ukraine, 03049, Ukraine*

ARTICLE INFO

Article history:
Received: 02 April, 2017
Accepted: 24 June, 2017
Online: 15 July, 2017

Keywords:
group of UAVs
aggregation
potential field
dynamic model

ABSTRACT

In this paper considered the movement of multi-agent system that consists of several UAVs that carry out monitoring ground surface. The multi-agent system includes a lead agent and several agents-members of the group. The motion of this system occurs along a trajectory, which is determined by the initial conditions, its mathematical model and obstacles on the route. Only the leader of the group knows the ultimate goal of the movement. The motion of this structure is considered in the potential field, which determined the forces of attraction and repulsion and created control signals by measuring the distances to the nearest neighbors. This allows the UAV group to consider an aggregate that has some size and to describe its motion the system of differential equations of second-order. As UAV selected Quadrotor. In this investigation, the stability conditions of such motion are considered, and simulation of approach is proposed.

1. Introduction

UAV is a popular air reconnaissance platform due to developed technologies that it were allowed to use a large number of constructive elements and motion control components for its creation. Now developers of unmanned vehicles pay attention to air platforms with the possibility of vertical launch and landing and existence a programmable control such as a Quadrotors that were developed recently.

Compared with other types of UAVs the Quadrotor differs accuracy of positioning, fly stability, allowable flight time, payload, and reliability. Such platforms can not only carry a payload, but also fly a flexible program. Their main tasks can be ground surveillance, execution of search and rescue operations, fire detection, cargo operations and military tasks. The advantage of the UAV is a low cost operation, no pilot that rules out the risk of loss of human life, existence software for performance of variety problems.

Intuitively, we can assume that application of the group of UAVs is more effective than the application of one aircraft. As parameters of evaluation of the effectiveness of the group except for accuracy and time performance can be the number of assignments, objectivity and reliability of its performance. However, group behavior is more difficult than the behavior of one UAV.

An important task, as in the case of one UAV, is path planning of a group. This task complicated the flight safety requirement. The distance between UAVs should be neither small, nor very large in order to keep the integrity of the group. Usually minimum distance between elements of group must not less than 2-3 maximal length of aircraft. The requirement of flight safety is achieved by increasing the distance between the UAVs. Maximum distance must to accord stable communication between elements of group. Thus for maintaining the integrity of the group distance between agents should not be too large.

Group's elements should know how to interact in order to achieve mission goal. To solve it, we consider methods that are based on behavior of herds of animals, birds and insects. Unfortunately, not all of them are suitable for application in technical means because it needs definition the man-operator role, group of UAVs should not be too large otherwise it technical and calculated complexity realization of this system will lead to expensive of managing system. Flight control of big UAVs team is need transmission the significant data flow and increasing time on decisions making.

*Corresponding Author: Dmytro Pavlovich Kucherov, National Aviation University, Educational and Scientific Institute of Computer Information Technologies, 03680, Ukraine, Email: d_kucherov@ukr.net

For solving main problems, the behavior of group used agent-based approach as usually. Tasks performance also should be provided by mechanisms and procedures of interaction for achieving common goal. The classic approach to solving this problem is the use in scientific works of methods based on the analogies with sociality in animals and especially in insect societies. However, not all of these methods are suitable for use in technical means, as a group of technical agents usually manages by human operator, the group consists of a limited number of UAVs, and it must operate in a real environment where there is interference and obstacles movement. Considering these properties increases the flow of data in the system, leads to an increase computational and technical complexity of a system.

In some cases, the performance of tasks limited interference of artificial or natural origin. Impact of power interference on the group flight leads to restructure of a group or change the route in order to achieve flight objectives. Thus, the interference signals should be known for synthesis of a control law.

The problem formulation involves the traditional description of group behavior as a dynamic system with restrictions on the intervals and the distance between the individual elements of the group. The group performs the movement on a route. There are fixed obstacles that located on the path of its movement. The mission of a group is to reach the goal and to avoid collision on the route.

In this paper considered movement of Quadrotors group that carry out monitoring of ground surface. Main problem of this task we consider the maintenance the group size for following of the route. Group consists of an agent-leader and agent-members that are fulfil their task in some closed region, where located an obstacles. One of problems that should be solved is to overcome an obstacles type of «bottleneck» on a route of UAVs.

The structure of this paper is organized as follows: in the next section, related work on the behavior of UAVs group is reviewed. Section 3 gives the system model and problem formulation. Section 4 provides movement stability study in different condition. The control laws for movement with and without obstacles are also proposed. Analysis of a Quadrotor model as a group element conducts in Section 5. Simulations conducts in Section 6. Section 7 concludes the paper.

2. Related work

This paper is an extension of work originally presented in International Conference “Methods and Systems of Navigation and Motion Control (MSNMC)” [1], which considered a description of group dynamics UAV in the presence of fixed obstacles.

Despite the fact, that currently drawn a lot of reviews and publications on multi-agent systems, the most common questions about this type of systems given in [2]. This monograph introduces the concept of autonomous agents capable of performing independent activities. Generalized architecture of agent system has three levels of management that are answer for the planning, conduct and interaction of the system.

The results of the use of multi-agent systems when performing practical tasks is given in [3-6]. In [3], for example, using the principles of ant colony for wireless sensor networks and new routing functions is considered. The proposed algorithm of optimization synthesized by on a probabilistic procedure, whereby authors cannot reduce the amount of computational operations that occur during the delivery of information packets. The positive results obtained by the authors using known approaches based on search algorithms, evaporation and aggregation.

In [4] proposed algorithms coordination and optimization of multi-agent system RobotCup to perform rescue operations. Coordination of agents carried out on protocol that provided by some agreements and optimization by the way modified A* algorithm and search Nilsson’s algorithm. The work demonstrates the advantages of decentralized management for these tasks.

The authors [5] consider UAV swarm consisting of Quadrotors whose actions modeled by algorithms of behavior of animals and insects swarm. Conducted experiments revealed that quality of multi-agent system depends on the number of agents, but increases the number of links in the group, making it difficult to manage the system.

The problem of robot control and dynamic goal on surface coordinates is represented in [6]. Robot movement occurs in additive potential field of attractive and repulsive forces. These functions depend on mutual positions, velocities, and accelerators between robot and goal. If deviation of position, velocity and acceleration of the robot from the corresponding goal coordinates is equal to zero, robot is stopped.

Tracking the robots of a moving goal are presented in [6], [7]. The movement of vehicles [6] and UAVs swarm [7] occurs in the additive potential field forces of attraction and repulsion. The resulting force depends on the relative position, speed and acceleration, developed between robot and objective. Stopping robots is performed on speed and acceleration coordinates.

In [7] considered movement of swarm of UAVs for target tracking. The UAVs swarm moving to the target in artificial potential attractive and repulsive field. Attractive field will lead any free agent to the swarm. Repulsive field helps to avoid collisions and to save constant distance between the systems. Authors propose main ideas with respect to analytical relations for description of potential field.

The authors [8] introduced the concept of aggregation for movement and control operations of units with kinematic links. Agents described by integrators and transition to another positions performer in sliding mode. The main goal this movement lies in the avoidance of "unfavorable" positions on the way. Motion carried in potential field, which is a closed area some size. The disadvantage of this presentation is to use a large amplitude control signals which cause the sliding mode of agents, leading to at least to delay the process control.

The authors [9, 10] offer a description of the actions the group of agents for computer network. Agents’ actions are based on the protocol of negotiations between neighbors, according to them the agent network is created on principles of self-organization. To support cooperation agents and avoid delays in the process of organizing the authors suggest a mechanism of mutual concessions and the management on algorithm Q-learning.

In solving practical problems the most interesting is route planning. Recent results on the topic presented in [11-13]. The combination of advanced heuristic search algorithm A^* and stochastic dynamic programming approach in planning the trajectory of an unmanned ground vehicle is proposed in [11]. To

avoid the pits, ditches and congestion on the route, the unmanned vehicle uses information from optoelectronic sensors, which it is equipped, to optimize the route in real time.

Planning of the route with fixed wing of UAVs by using differential geometry solved in [12]. For carrying out the maneuver type of slew for a minimum time with constant velocity authors propose to build a route on well known Dubins path, arcs curves submitted semi-circles and segments of straight lines, but the lateral acceleration becomes intermittent. Continuousness of acceleration achieved using Pythagorean hodograph. In [13] authors solve a similar problem by Bezier curves.

For a stable and reliable work group is responsible communication subsystem. In [14, 15] addresses the problem of providing subsystem reliability data. Common approaches to the design of the subsystem presented in [14], especially that arising in data and algorithms for their consideration could be found in [15]. Planning methods trajectory shown in graph theory [16]. The motion work is carried out within the given trajectories. In this paper, movement is viewed as completely determined task at a constant speed and without interference.

Mission of path planning can be solved by methods of graph theory as shown in [16]. Robots are moving between spots of interest. Authors assume robot movement to determinate the order with the constant velocity and without obstacles.

An approach of tracking UAVs for the moving target is presented in [17]. The movement of a UAVs swarm occurs in attractive and repulsive fields that created regarding robot-leader by state feedback controller. However, Lyapunov criterion that is applied in the work, assumes stability only in the asymptotic behavior at large time interval, and it may not be applicable in the presence of obstacles.

The problem of surveillance by swarm of UAV of moving target given in [17]. Movement of swarm occurs in a potential field created robot's controllers relatively leader by feedback. Proposed the Lyapunov criterion for evaluating the stability of motion has effect only on a large interval of a time, and it cannot used in presence of interference. Practical realization where robot follow to target with radio interference was given in [18]. Unfortunately, the proposed method is not suitable for controlling the UAV.

Compared to [1], in this work is developed synthesis UAV traffic control for a limited size group, which is not fixed and it can increase with rounding maneuvers and obstacles.

Compared to previous work, are proposed in this paper the algorithms of explicitly solution of the task moving in bounded conditions. The main algorithm creates two tasks. First task is to save dimension of the group when naturals or imitation obstacles of increase its dimension. Second task is to guarantee performance of the main task of group. To solve these tasks, much used potential field between robot-members within group and feedback control for following for robot-leader of group. Correctness of the proposed algorithm is proved. The details of the algorithms are presented in Section IV.

3. Problem formulation

Let a swarm of N UAVs (further agents) that are indicated as $A_1, \ldots, A_n$, which carry out some task in Euclidean space (x, y, z), for example monitoring or reconnaissance areas of the earth's surface. For this purpose, agents can use known tactics such as "bearing", "wedge" or other, or perform autonomous flight. On Figure 1 we can see the flight by "wedge" the group of UAVS, N=4.

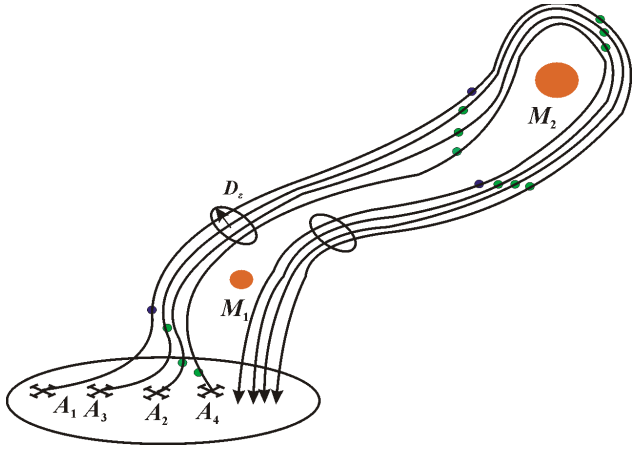


Figure 1. Group of UAVs, N =4

Figure 1 shows the trajectory of UAVs group that forms multi-agent system from the 4-agents and performs the task of monitoring a predetermined route; where there are obstacles M_1, M_2 with known coordinates (x_M, y_M, z_M). To avoid collisions and saving group configuration when performing motion is assumed approximately equal velocity agents $v_i \approx v_j, \ i \neq j$.

Such formation of the agents' behavior we can describe as aggregation with some centroid. In order to perform the task of group, each agent of aggregation is provided with means of communication and motion control. If members act in a limited size, additional members may be joined to this formation. Such formation is a swarm.

Let imagine structure swarm of UAVs as some formation of limited size. Using the terminology of [8], this is formation called aggregate with a center. We assume that each agent provided of communications and control means to carry themselves mission.

Each agent is moving in a three-dimensional Euclidean space. On its rout it can meet obstacles. The agent has vertical takeoff and landing (VTOL) and it therefore is can to bypass obstacles. Thus, agent i-th dynamic is described in space position as p_i with coordinates $p_i = (x_i, y_i, z_i)$ and it has control signal u_i. We will describe the dynamic model of agent's motion as an integrator of the second order in form [1, 5, 7]

Agents move in three-dimensional Euclidean space, in which can presented obstacles. Existing controls allow them to make the maneuver and get around obstacles. Let p_i is a position of i-th agent and u_i is a control signal in space, $p_i = (x_i, y_i, z_i)$. Then the dynamic model of movement of i-th agent can be described by differential equations of the second order as [1, 5, 7]

$$\ddot{p}_i = k_i u_i. \tag{1}$$

In (1) p_i, u_i are functions of a time t. As in [8] we enter parameter $D_\varepsilon > 0$, which characterizes the size of aggregate that for $\forall t \to \infty$ maintained control signals u_i

$$p_i \to D_\varepsilon(p_c), \tag{2}$$

where position of central agent (leader) in aggregator

$$p_c(t) = \frac{1}{n}\sum_{i=1}^{n} p_i(t), \tag{3}$$

and

$$D_\varepsilon(p_c) = \{p \in \Re^3 : d = \| p - p_c \| < \varepsilon\}. \tag{4}$$

In (4) ε is a constructive parameter, which is determined by the system developer. Equation (4) is a criterion to ensure the integrity of the group.

If on the route of UAVs is obstacles, group configuration is change, it selects a new type of configuration, which ensures their overcoming, in general, it slightly increases the parameter $D(p_c)$ so as to ensure inequality $D(p_c) \le D_\varepsilon(p_c)$. This assume reconfiguring the formation in "a small" and does not change the overall objective of the group.

This paper concerns and solves the problem of the synthesis of the control of group from n UAV, which moves along the route, where there are natural obstacles fixed type, to ensure the fairness of condition $D(p_c) \le D_\varepsilon(p_c)$.

4. Method of potential field

As in [6-8, 17], we submit multi-agent system a field of agents, which is formed by forces of attractive and repulsion. Attractive energy supports the chosen configuration, and the repulsion energy keeps agents from collisions among themselves and with obstacles.

As known [19], the energy of a potential attractive field is described by

$$E_{att}(\Delta p) = \frac{1}{2}\mu(\Delta p)^2, \tag{5}$$

where $\mu > 0$ is scaling factor, and $\Delta p_i = \| p_i - p_c \|$ is the Euclidian distance between the i-th agent and the agent centric. The view of relation to centric agent A_c in uniform attractive field is show on Figure 2. This field is weaker in the center and stronger on edges.

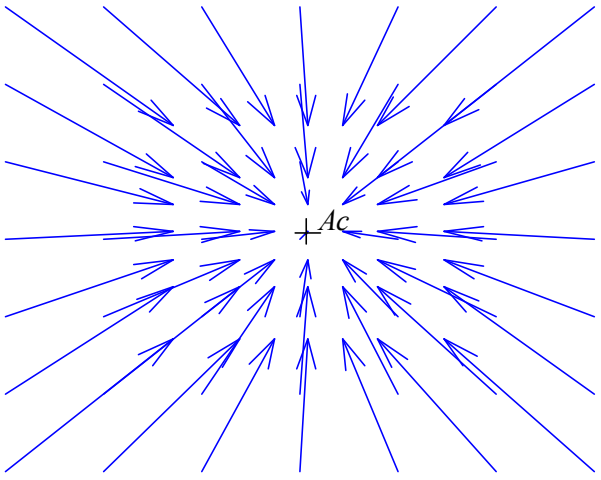


Figure 2. Attractive field acting on agent A_c

Forces of attractive field is constructed as [20]

$$u_{att}(\Delta p) = -grad\ E_{att}(\Delta p) = -\ \mu(\Delta p). \tag{6}$$

The repulsive energy is calculated as [19]

$$E_{rep}(\Delta p_i) = \begin{cases} \frac{1}{2}\nu\left(\frac{1}{\Delta p_i} - \frac{1}{d_0}\right)^2, & if\ \Delta p_i \le d_0, \\ 0, & if\ \Delta p_i > d_0, \end{cases} \tag{7}$$

where $\nu > 0$ is scaling factor, $d_0 > 0$ is constant, which determines the maximum distance of action on the closest agent. Veiw of repulsive field built in accordance to (7) shown on Figure 3. This field is stronger in the center unlike attractive field.

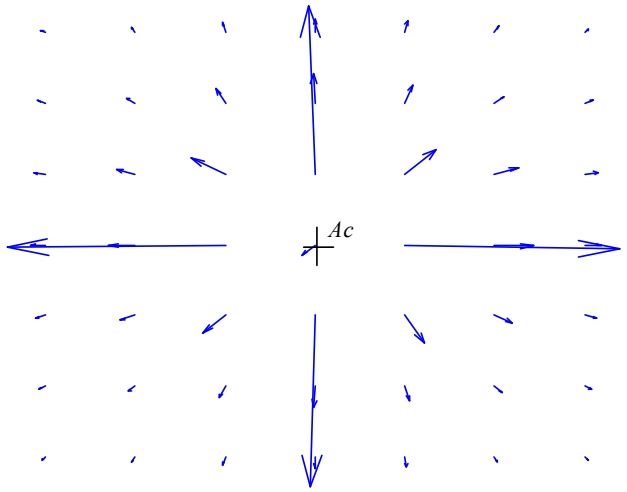


Figure 3. Repulsive field acting on agent A_c

Repulsive forces similarly (6) we can represent as

$$u_{rep}(\Delta p_i) = -grad\ E_{rep}(\Delta p_i) = \begin{cases} \nu\left(\frac{1}{\Delta p_i} - \frac{1}{d_0}\right)\frac{1}{(\Delta p_i)^2}\frac{p_i - p_0}{d_0}, & if\ \Delta p_i \le d_0, \\ 0, & if\ \Delta p_i > d_0. \end{cases} \tag{8}$$

In (8) p_0 is the coordinates to closest agent p_i or obstacle p_M, and

$$d_0 = \max_{1 \le i \le N} \| p_i - p_0 \|,$$

is constant equal of maximum from all of their Euclidean distances.

Then, the general control law, which allows agents to achieve the objectives is the sum of components

$$u_i^g = u^i + u_i^{att} + u_i^{rep}, \tag{9}$$

where u_i is law of motion of i-th agent in the aggregation, and u_i^{att}, u_i^{rep} components, which are determined by formulas (6), (8).

If the i-th agent is the leader of the group, then $u^i = u^l$. Trajectory of leader determined the dynamics of its motion and described by equations, like (1)

$$\begin{aligned} \dot{p}_1^l &= p_2^l, \\ \dot{p}_2^l &= k^l u^l, \end{aligned} \tag{10}$$

where u^l is known law, the trajectories of other agents are being built respect to him. This approach allows to enter two terminate state such as start state (0, 0, 0) and final state (p_k, 0, 0) for each agent system.

Now consider the three types motion of the aggregation to follow on the route: along line, along curve, and with obstacles.

4.1. Motion along line

In this case, the leader moves without acceleration. Then (10) is transformed into one equation

$$\dot{p}_1^l = p_2^l = const. \tag{11}$$

The mission of i-th agent is to save its position in a group, and to prevent exposure on other agents. The control law formed as a control

$$u^i = \begin{cases} -k_{p_1}(p_{1i} - p_{1j}) - k_{p_2}(p_{2i} - p_{2j}), & if\ i,\ \ j\ \ is\ \ memder, \\ 0, & if\ i\ \ is\ \ not\ \ member. \end{cases} \tag{12}$$

For agent leader is set $j = l$. Here $k > 0$ is the positive factor that define velocity of agent-member. In addition, there are a couple of small $\xi_{p1} \neq 0$, $\xi_{p2} \neq 0$ such that deviations $|p_{1i} - p_{1j}| \leq \xi_{p1}$ and $|p_{2i} - p_{2j}| \leq \xi_{p2}$, which occur in (12) and is needed to maintain stable movement.

If the leader is in the center of the group, its coordinates determined by equation (3), namely

$$p^l = \frac{1}{N^*}\sum_{i=1}^{N^*} p_i,\ \ \dot{p}^l = \frac{1}{N^*}\sum_{i=1}^{N^*} \dot{p}_i\ , \tag{13}$$

where $N^* < N$ is the quantity of neighbors or agents in a group. Such type of motion is shown on Figure 4.

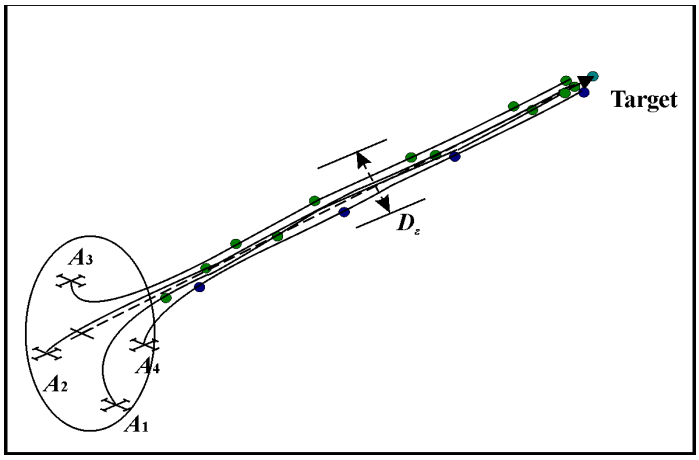


Figure 4. Motion UAVs swarm along line

LEMMA 1. The movement of each agent of group that consist of N agents is described by system of differential equations of second order in form

$$\begin{aligned} \dot{p}_1 &= p_2, \\ \dot{p}_2 &= \overline{u} = -k_{p_1}(p_{1i} - p_{1j}) - k_{p_2}(p_{2i} - p_{2j}). \end{aligned} \tag{14}$$

where $p = p_i - p_j$ is the appropriate coordinates of i-th and j-th agents, $i \neq j$, $i \leq n$, and $j \leq n$. If $j = l$, is a leader of the group member.

Proof. This statement follows from (9), (11).

Corollary 1. Closed loop dynamic system (14) converges to the equilibrium position

$$p_{1i} - p_{1j} \neq 0,\ \ p_{2i} - p_{2j} \neq 0. \tag{15}$$

Corollary 2. Agent-members velocity asymptotically converges to the agent-leader velocity.

Corollary 3. In system (14) coordinates p_i is $p_i \neq p_j$ for any $i \neq j$, $t \geq 0$.

Corollary 4. If $d = \| p_i - p_j \| < \varepsilon$ is maximal size between UAVs in group, the size of formation is $D(p) < D_\varepsilon(p)$.

4.2. Motion along curve

This type of movement is performed with an acceleration, condition (11) is not satisfied, and dynamic system (14) is transformed into

$$\begin{aligned} \dot{p}_1 &= p_2, \\ \dot{p}_2 &= \dot{p}_2^l + \overline{u}. \end{aligned} \tag{16}$$

Coordinates p_1, p_2 in (16) coincide with the same coordinates in equation system (14), but control signal is value $|\overline{u}| < U_{\max}$. This motion type is shown on Figure 5.

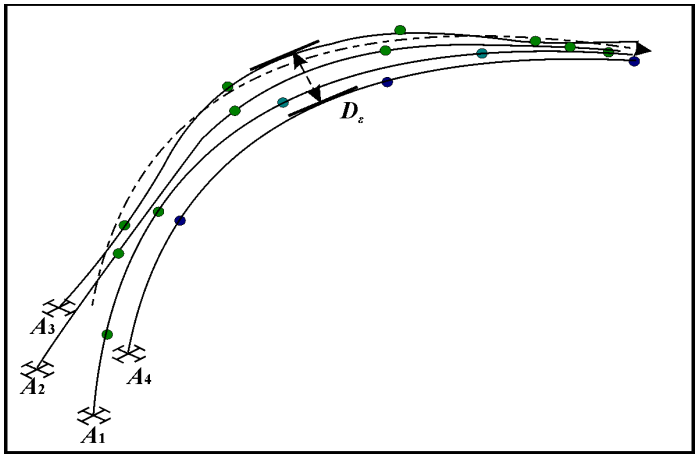


Figure 5. Motion UAVs along curve

LEMMA 2. Let given factors k_y and k_p such as $k_y > 0$, $k_p > 0$ is true inequality $0{,}5k_y / k_p < 0$, system (15) has exponential stability.

Proof. We represent the system of differential equations (16) in the form of a differential equation

$$k_p \ddot{e}_p(t) + k_y \dot{e}_y(t) + k_{\dot{y}} e_{\dot{y}}(t) = 0, \tag{17}$$

where $e_p = p_i - p_j$, $e_y = y_i - y_j$, $e_{\dot{y}} = \dot{y}_i - \dot{y}_l$. If the $k_p \neq 0$, and the (k_p, k_y, $k_{\dot{y}}$) are positive factors, this equation can be represented as

$$s^2(t) + k_y / k_p s(t) + k_{\dot{y}} / k_p = 0. \tag{18}$$

Then condition of exponential stability of (15) is a negative real part of the roots (18) that is satisfied for $k_y/(2k_p) < 0$.

Corollary 1. If $k_y = 0$, we have oscillation movement around leader with constant amplitude and it frequency is equal to $\sqrt{k_{\dot{y}}/k_p}$.

Corollary 2. If $k_y/\sqrt{4k_{\dot{y}}k_p} > 0$, we have asymptotically stability movement.

Corollary3. If $0 < k_y/\sqrt{4k_{\dot{y}}k_p} < 1$, we have oscillation with error that is decreasing.

Corollary4. If $k_y/\sqrt{4k_{\dot{y}}k_p} > 1$, we have aperiodic movement of agent-member.

Corollary 5. If maximal size between UAVs in group is $d = \| p_i - p_j \| \le \varepsilon$, size is $D(p) \le D_\varepsilon(p)$.

4.3. Motion with obstacles

Existing approaches to the consideration of the motion agents with obstacles assume a repulsive force from the obstacle that is located on the route to goal. Action of this forces increases when agent is approaching to obstacle (deceleration of agent), and diminishes with distance. In order to bypass the obstacles, agent should make additional efforts that will lead to the deflection of the original route. It can be made by introducing repulsive force of the additive components of the agent respect to velocity and acceleration of motion

$$u_i = \begin{cases} u_{rep\ p_i} + u_{rep\ p_{1i}} + u_{rep\ p_{2i}}, \\ \quad if\ (d-2r) \le d_0,\ \ p_{1i} > 0,\ \ p_{2i} > 0 \\ u_{rep\ p_i} + u_{rep\ p_{1i}}, \\ \quad if\ (d-2r) \le d_0,\ \ p_{1i} > 0,\ \ p_{2i} \le 0, \\ 0,\ \ if\ (d-2r) > d_0,\ \ or\ p_{2i} \le 0. \end{cases} \tag{19}$$

In equation system (19) uses

$$u_{rep\ p_i} = k_1\left(\frac{1}{d-2r} - \frac{1}{d_0}\right), \tag{20}$$

$$u_{rep\ p_{1i}} = k_2 p_{1iobs}, \tag{21}$$

$$u_{rep\ p_{2i}} = k_3 p_{2iobs}, \tag{22}$$

value r is the maximal geometric size of an agent,

$$p_{1iobs} = p_{1i}^T n_{aobs}, \tag{23}$$

and

$$p_{2iobs} = p_{2iobs}^{\ T} n_{aobs}, \tag{24}$$

where n_{aobs} is a unit vector pointing from i-th agent to obstacle. The value p_{1iobs} shows agent movement, if $p_{1iobs} > 0$ is agent moving to obstacle and $p_{1iobs} \le 0$ moving from obstacle. Motion of N=4 agent in the presence of obstacles is shown on Figure 6.

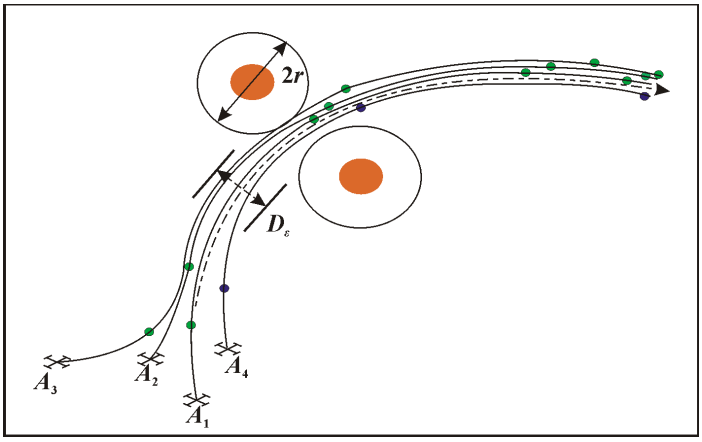


Figure 6. Motion N=4 UAVs in obstacles

The order to overcome the obstacles type of bottlenecks, as shown in Figure 6, is determined by the control signal, that is defined by the equation (19), and UAVs are sorts in order decreasing its velocity, $p_{1N} < p_{1(N-1)} < \ldots < p_{12} < p_{11}$. Sorting can be done otherwise if you choose group leader as a reference.

5. Quadrotor model

As a dynamic model of UAV selected model Quadrotor that management is simple, stable movement, the ability to hang in the air. A complete description of the model is described in [21]. Model of Quadrotor is presented by following equations system

$$\begin{cases} \ddot{\phi} = \dot{\theta}\dot{\psi}\left(\frac{I_y - I_x}{I_x}\right) - \frac{J_r}{I_x}\dot{\theta}w + \frac{l}{I_x}U_2, \\ \ddot{\theta} = \dot{\phi}\dot{\psi}\left(\frac{I_z - I_x}{I_y}\right) - \frac{J_r}{I_y}\dot{\phi}w + \frac{l}{I_y}U_3, \\ \ddot{\psi} = \dot{\phi}\dot{\theta}\left(\frac{I_x - I_y}{I_z}\right) + \frac{1}{I_z}U_4, \\ \ddot{z} = -g + (\cos\phi\cos\theta)kU_1, \\ \ddot{x} = (\cos\phi\cos\theta\cos\psi + \sin\phi\sin\psi)kU_1, \\ \ddot{y} = (\cos\phi\cos\theta\cos\psi + \sin\phi\cos\psi)kU_1 \end{cases} \tag{25}$$

Here (ϕ, θ, ψ) is the roll, pitch and yaw angles, $I_{x,y,z}$ is the inertia UAV, J_r is the inertia rotor, l is the lever in coordinate system with the Quadrotor, k=1/m, m is the mass UAV, w is the rotor speed, and U_1, U_2, U_3, U_4, are calculated as

$$\begin{cases} w = w_2 + w_4 - w_1 - w_3, \\ U_1 = b(w_1^2 + w_2^2 + w_3^2 + w_4^2), \\ U_2 = b(w_4^2 - w_2^2), \\ U_3 = b(w_3^2 - w_1^2), \\ U_4 = d(w_2^2 + w_4^2 - w_1^2 - w_3^2), \end{cases} \tag{26}$$

where b is the thrust factor, and d is the drag factor.

For simulation the model (25) is transformed to a space-state form $\dot{X} = f(X,U)$, where $X \in \Re^{12}$ is the state vector with the coordinates

$$\begin{aligned} &x_1 = \phi,\ x_2 = \dot{x}_1 = \dot{\phi},\ x_3 = \theta,\ x_4 = \dot{x}_3 = \dot{\theta},\\ &x_5 = \psi,\ x_6 = \dot{x}_5 = \dot{\psi},\ x_7 = z,\ x_8 = \dot{x}_8 = \dot{z},\\ &x_9 = x,\ x_{10} = \dot{x}_9 = \dot{x},\ x_{11} = y,\ x_{12} = \dot{x}_{11} = \dot{y}. \end{aligned} \tag{27}$$

Then, we can present equations system as

$$\dot{X} = \begin{cases} x_2, \\ x_4 x_6 a_1 + x_4 a_2 w + b_1 U_2, \\ x_4, \\ x_2 x_6 a_3 + x_2 a_4 w + b_2 U_3, \\ x_6, \\ x_4 x_2 a_5 + b_3 U_4, \\ x_8, \\ -g + (\cos x_1 \cos x_3) k U_1, \\ x_{10}, \\ u_x k U_1, \\ x_{12}, \\ u_y k U_1, \end{cases} \tag{28}$$

where

$$\begin{aligned} &a_1 = \frac{(I_y - I_z)}{I_x},\ a_2 = -\frac{J_r}{I_x},\ a_3 = \frac{(I_z - I_x)}{I_y},\ a_4 = \frac{J_r}{I_y},\\ &a_5 = \frac{(I_x - I_y)}{I_z},\ b_1 = \frac{l}{I_x},\ b_2 = \frac{l}{I_y},\ b_3 = \frac{l}{I_z}, \end{aligned} \tag{29}$$

$$\begin{aligned} u_x &= \cos x_1 \sin x_3 \cos x_5 + \sin x_1 \sin x_5, \\ u_y &= \cos x_1 \sin x_3 \cos x_5 - \sin x_1 \sin x_5. \end{aligned} \tag{30}$$

In contrast to [21], which used to create a sliding mode for control, the control actions (U_1, U_2, U_3, U_4, u_x, u_y) of each agent are smooth functions of time t.

6. Simulation

The motion of group of Quadrotors that moves in the potential field towards to goal on the route on which there are obstacles, is considered.

We assume that this group consists of three agent-members and one agent-leader. The route of agents gives start positions and goal position: $p_{A1}=(100,\ 30,\ 0)^T$, $p_{A2}=(75,\ 20,\ 0)^T$, $p_{A3}=(130,\ 50,\ 0)^T$, $p_{A4l}=(110,\ 70,\ 0)^T$, $p_f=(600,\ 270,\ 0)^T$. The coordinates of the located fixed obstacles on the route are known as $p_{o1} = (270,\ 130)^T$, $p_{o2} = (220,\ 210)^T$, and $D_\varepsilon(p_A) \le 50$. The altitude of obstacles is higher than altitude of Quadrotors that lead to their maneuver in horizontal plane. The altitude for calculations is taken z = 10. The movement of agent-members and agent-leader to target is shown on the Figure 7.

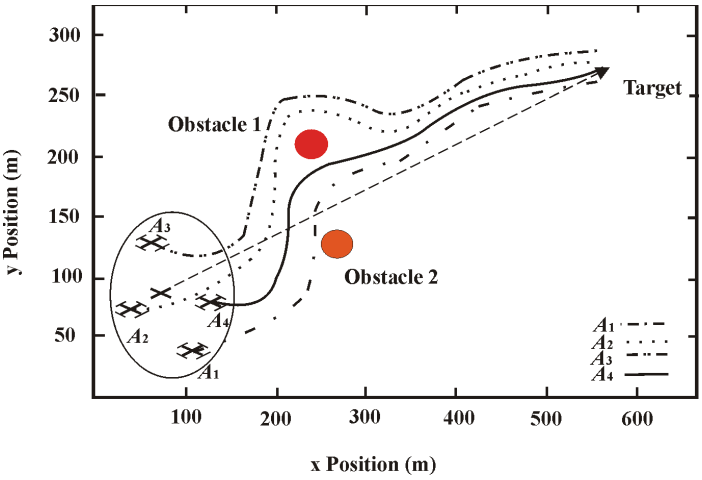


Figure 7. Bypass movement of agent-leader and agent-members to the target

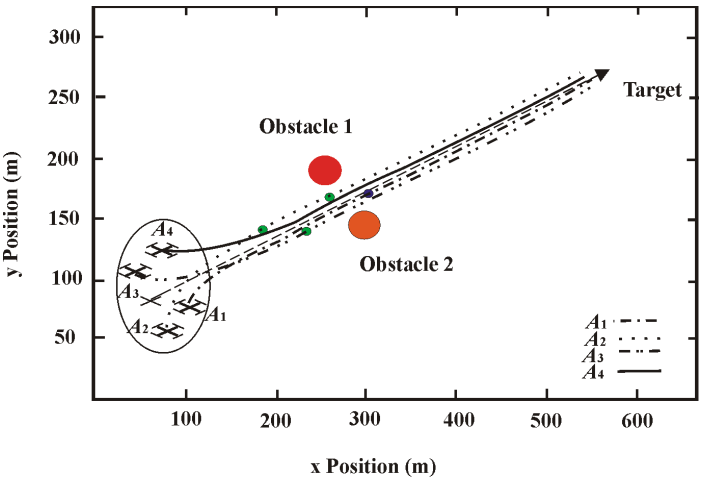


Figure 8. Group of UAVs overcomes side-by-side obstacles by line type motion

Figure 7 shows that group of agents bypasses obstacles and continues its movement to target. Sometimes, when overcoming obstacles group of agents can change its formation size and the formation of agents is remains without changes, if there are no obstacles on the route,.

Figure 8 represents reconfiguration of the group of UAVs in line type motion, if there are located obstacles on the route side-by-side. In this case, the tactic of UAVs is similar to column and UAVs follow in rigidly defined order: $A_4 \to A_3 \to A_2 \to A_1$, where A_1 is the first and A_4 is the last.

7. Conclusions

The paper presents a synthesis of control law of motion of multi-agent system to follow on the route on which possible the obstacles of movement. The feature of the proposed approach is synthesized control law with a smooth function of a time. Two ways of overcoming obstacles, one of them based at the bypassing and the other on reconfiguration are also shown. Control law is formed by two special potential functions, one of which creates an attraction force agents to the center of the structure, the other repels its from neighbors and obstacles. Configuration flight group in terms of obstacles to change, suggesting appropriate changes size and shape.

On the basis of the proposed control law created algorithm by which the simulation was conducted multi-agent system.

Examples of the system behavior modeling show stable movement of the agents. Study of system parameters allow to ensure its stable movement and revealed relevant system parameters. Its behavior includes such stages as gathering into group, tracking on the given trajectory and avoidance of the obstacles. Research shows that successful simulation depends on such parameters as μ and ν of potential function.

In view of the fact that in the conditions of obstacles the movement along the curve is typical, because of which the signal in the control channels increase and it is switch it at a high speed, which is the so-called «sliding mode». The investigation of a multi-agent system in slide mode and effectiveness of proposed potential functions for movement one is the topic of our future research.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

Authors thank both the authorities of National Aviation University, National University of Defence of Ukraine and Central Research Institute of Armaments and Military Equipment of the Armed Forces of Ukraine for their support during the preparation of this paper.

References

[1] D.P. Kucherov, A.N. Kozub, O.N. Kostyna "Group Behavior of UAVs in Obstacles Presence", proc. IEEE 4th Int. Conf. "Methods and Systems of Navigation and Motion Control (MSNMC)", Kyiv, Ukraine, 2016.

[2] M. Wooldridge, An Introduction to multi-agent systems, John Wiley & Sons Ltd, 2009.

[3] Benu, C. Goel, Sheenam, "Ant Colony Optimization for Wireless Sensor Network: A Review", IOSR-JCE, **17**(3), 89-92, 2015.

[4] W.Y.J. Chou, L. Marsh and D. Gossink "Multi-Agent Coordination and Optimisation in the RoboCup Rescue Project", proc. 18th World IMACS / MODSIM Congress, Cairns, Australia, 2009.

[5] A. A. Bandala, E. P. Dadios, R. R. P. Vicerra, L. A. G. Lim "Swarming Algorithm for Unmanned Aerial Vehicle (UAV) Quadrotors – Swarm Behavior for Aggregation, Foraging, Formation, and Tracking", J. of Advanced Comput. Intell. and Intell. Inform., **18**(5), 745 – 751, 2014.

[6] L. Yin, Y. Yin, and C.-J. Lin, "A new potential field method for mobile robot path planning in the dynamic environments", Asian J. of Control, **11**(2), 214–225, 2009.

[7] A. D. Dang, J. Horn, "Intelligent Swarm-Finding in Formation Control of Multi-Robots to Track a Moving Target", Int. J. of Computer, Electr., Autom., Control and Inf. Eng., **8**(4), 561 – 567, 2014.

[8] V. Gazi, B. Fidan, Y. S. Hanay, M.I. Köksal, "Aggregation, Foraging, and Formation Control of Swarms with Non-Holonomic Agents Using Potential Functions and Sliding Mode Techniques", Turk. J. of Electr. Eng., **15** (2), 149-168, 2007.

[9] L. Zhang, H. Gui, "Optimal Strategies for Agents in an Alternating Offers Negotiation Protocol Considering Time Constraint", J. of Autom. and Control Eng., **4** (4), 313 – 318, 2016.

[10] D.K. Jesintha, J.P. Anandh, M. Ramnath "Dynamic analysis of agent network in self organization using service level agreement technique", Int. J. of Eng. Sci. Invent., **4** (3), 44 – 49, 2015.

[11] L. Marsh, G. Calbert, J. Tu, D. Gossink, H. Kwok, "Multi-Agent UAV Path Planning", Int. Congress on Modelling and Simulation, MODSIM05. Modelling and Simulation Society of Australia and New Zealand, 2005.

[12] M. Shanmugavel, A. Tsourdos, B. A. White, R. Zbikowski "Differential Geometric Path Planning of Multiple UAVs", Trans. of ASME J. of Dyn. Syst., Meas., and Control, **129**(3), 620-632, 2007.

[13] D.P. Kucherov, "Agent-based approach to the problem movements by the route", Visnik of the Volodymyr Dahl East Urk. National University, **4** (211), 168 – 171, 2014.

[14] İ. Çuhadar, M. Dursun "Unmanned Air Vehicle System's Data Links", J. of Autom. and Control Eng., **4** (3), 189-193, 2016.

[15] D.P. Kucherov, A.N. Kozub "Control System Objects with Multiple Stream of Information", proc. IEEE 3rd Int. Conf. Actual Problems of UAV Developments (APUAVD)", Kyiv, Ukraine, 2015.

[16] Y. Wang, C. Hu, "Moving as a whole: multirobot traveling problem constrained by connectivity", Turk. J. of Electr. Eng. & Comput. Sci., 23, 769 – 788, 2015.

[17] A. D. Dang, J. Horn,"Formation Control of Leader-Following UAVs to Track a Moving Target in a Dynamic Environment", J. of Autom. and Control Eng., **3** (1), 1-8, 2015.

[18] MD. A. A. Ansary, M. M. Hasan, "An Intelligent Line Following Robot with Obstacle Detection", Int. J. of Sci. and Eng. Investig., **5** (53), 44-48, 2016.

[19] M. C. Jain, Textbook of Engineering Physics, (Part I): PHI Learning Private Ltd, 2009.

[20] G.A. Korn, T. M. Korn, Mathematical handbook for sceintist and engineers: McGraw-Hill Company, 1968.

[21] S. Bouabdallah, R. Siegwart, "Backstepping and Sliding-mode Techniques Applied to an Indoor Micro Quadrotor", proc. IEEE Int. Conf. on Robot. and Autom., Barcelona, Spain, 2005.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1042-1049 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Predictive Technology Management for the Identification of Future Development Trends and the Maximum Achievable Potential Based on a Quantitative Analysis

Michael Fries*, Markus Lienkamp

Technical University of Munich, Department of Mechanical Engineering, Institute of Automotive Technology, 85748, Germany

ARTICLE INFO

Article history:
Received: 26 April, 2017
Accepted: 11 June, 2017
Online: 15 July, 2017

Keywords:
Innovation management
Technology management
Technology assessment
Exponential growth
Logistic growth
Computational intelligence
Performance periods
Big data
Data mining

ABSTRACT

A company's ability to find the most profitable technology is based on a precise forecast of achievement potential. Technology Management (TM) uses forecasting models to analyse future potentials, e.g. the Gartner Hype Cycle, Arthur D. Little's technology lifecycle or McKinsey's S-curve model. All these methods are useful for qualitative analysis in the planning of strategic research and development (R&D) expenses. In a new approach, exponential and logistic growth functions are used to identify and quantify characteristic stages of technology development. Innovations from electrical, mechanical and computer engineering are observed and projected until the year 2025. Datasets from different industry sectors are analysed, as the number of active Facebook users worldwide, the tensile yield point of flat bar steel, the number of transistors per unit area on integrated circuits, the fuel efficiency per dimension of passenger cars, and the energy density of Lithium-Ion cells. Results show the period of performance doubling and the forecast for the end of the technological achievement potential. The methodology can help to answer key entrepreneurial questions such as the search for alternatives to applied technologies, as well as identifying the risk of substitution technology.

1. Introduction

This paper is an extension of work originally presented at the IEEE International Conference on Industrial Engineering and Engineering Management, Indonesia in December 2016 [1]. The positive feedback from the audience helped in writing a more detailed mathematical description of the growth functions. Furthermore, the energy density of Li-Ion cells was analysed and added to the case studies, because of its relevance in future automotive driveline systems. The summary section includes an overview of the technology assessment for this newly added energy storage technology. An extended conclusion shows the influence of the approach introduced on strategical Technology Management. The newly added outlook paragraph discusses the benefits of a fundamental figure-based approach in combination with data mining for strategic decision making in research and development departments. With the help of a reporting system and individually derived performance indicators the future possibilities of making the right decisions will rise. In a big data driven, rapidly growing digital environment, new fields of application in many engineering branches will appear for the integration of the method.

Due to the rapid arrival of substitution technologies, companies are under growing pressure to make fast and sustainable decisions. In the past, even companies with market leader positions such as AOL, Kodak, Nokia and Grundig [2]have disappeared due to their reliance on outdated technology. According to methods such as the Gartner Hype Cycle and the technological lifecycle of Arthur D. Little [3], different stages of maturity describe the acceleration and deceleration of the achievable potential. The beginning and end of technological growth is qualitatively described by the S-curve model [4]. Figure 1 shows a qualitative example of technology substitution due to different growth phases. The objective of this paper is to forecast performance development and maximisation of the achievable potential. Therefore, logistic and exponential growth functions are used to quantify different growth patterns. This paper proposes a mathematical approach for technology forecasting and assessment.

*Corresponding Author: Michael Fries, Boltzmannstr. 15, 85748 Garching, Germany | Email: fries@ftm.mw.tum.de

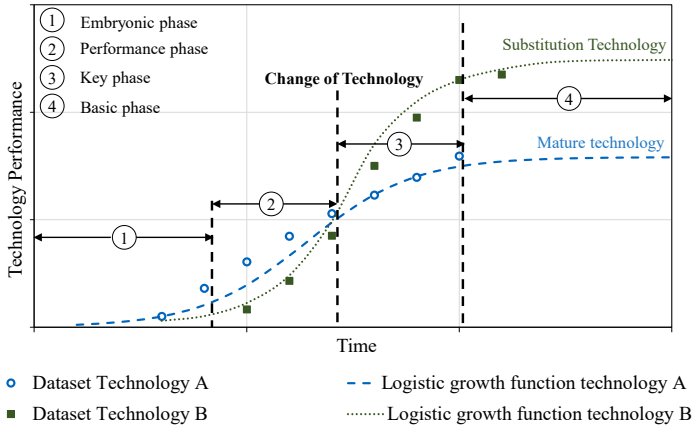


Figure 1: S-curve model, Technology A is substituted by Technology B, the growth phases refer to Technology A [5].

2. Methodology

Data collected from different branches of industry are analysed with the logistic growth function from [6] and [7] to determine the growth phases. If the technology is in its performance phase, the data are fitted with an exponential growth function from [8] to identify the future performance development. The exponential growth is similar to the performance phase of the S-curve model. The overall shape of the S-curve follows the logistic growth function. Figure 2 displays the overlay of the performance phases of a mature technology and a substitution technology that leads to extended exponential performance growth.

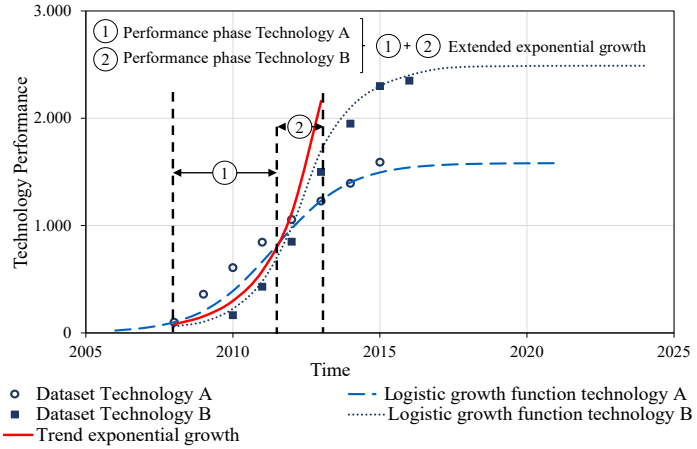


Figure. 2: Extended performance phase by the overlap of two logistic growth functions.

2.1. Mathematical description of logistic growth

Logistic growth is characterised by the growth coefficient r changing over time x, with y representing the performance of the technology and K the maximum value to be approximated, [6], [7]. Equation (1) gives the formula for the growth rate:

$$\frac{dy}{dx} = ry\left(1-\frac{y}{K}\right) \quad (1)$$

A positive slope in the growth rate leads to a continuous rise in technology performance. A negative slope of the growth rate corresponds to a slowdown in performance until the maximum value is approximated, Figure 3.

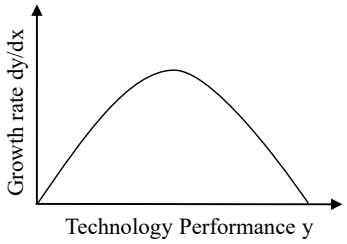


Figure 3: Slope of the technology growth rate

The logistic growth equation (2) follows from integrating (1) with the initial performance value $y = y_0$ at time $x_0 = 0$:

$$y = \frac{K}{1+\left(\frac{K}{y_0}-1\right)e^{-rx}} \quad (2)$$

The relative growth of the performance of a technology divided by the absolute performance y is linear relative to the absolute performance y. The linear equation can be derived from Figure 4.

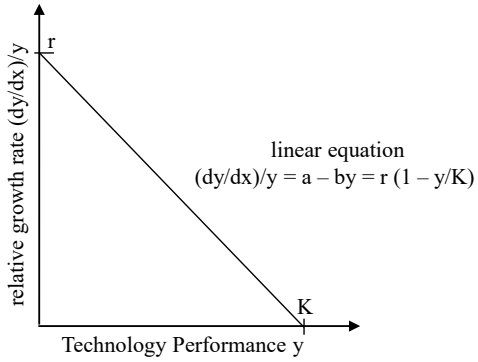


Figure 4: Determination of the growth coefficient r and the max. value K

Relative growth in relation to the absolute technology performance is described by the linear equation (3).

$$\frac{(dy/dx)}{y} = a - by \quad (3)$$

The maximum value K and the growth coefficient r, from (2), can be determined by solving the linear equation (3).

$$a - by = r\left(1-\frac{y}{K}\right) \quad (4)$$

The growth rate r and the maximum value K follow from Equation (4):

$$r = a \quad \text{and} \quad (5)$$

$$K = \frac{a}{b} \quad (6)$$

Given the maximum value K and the growth coefficient r, the logistic growth function from (2) is described completely (see Figure 5.)

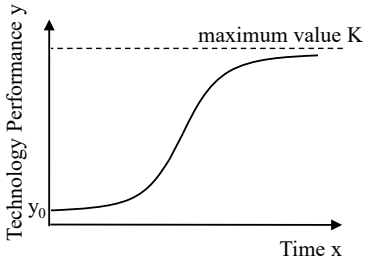


Figure 5: Logistic growth function from Equation (2)

2.2. Mathematical description of exponential growth

As shown in Figure. 2 the sum of the performance phases of multiple logistic growth functions leads to the continuous exponential growth of the technology performance. The effect is valid for many market developments.

Exponential growth is characterised by the growth coefficient r, the starting technology performance y_0 in the embryonic phase and the time x shown in (7) and [8]:

$$y = y_0 e^{rx} \quad . \tag{7}$$

In exponential regression analysis, the existing data y are compared with the values calculated from the exponential curve equation (7). Equation (7) is logarithmised in Equation (8) to apply the method of least squares to the performance phase of a technology, [9].

$$\ln(\hat{y}) = \ln(y_0\, e^{r\,\hat{x}}) = r\,\hat{x} + \ln(y_0). \tag{8}$$

ŷ

$$V = \sum_{i=1}^{k} (y_i - \hat{y}_i)^2 \tag{8}$$

and

$$V\,(r, \ln(y_0)) = \sum_{i=1}^{k} \Big(\ln(y_i) - r\,x_i - \ln(y_0 \tag{9}$$

To determine the constants r and y_0 as two unknowns with two equations, (10) is modified and the partial derivatives are set to zero to obtain the minimum.

Partially differentiating with respect to y_0:

$$\frac{\delta V\big(r, \ln(y_0)\big)}{\delta V\big(\ln(y_0)\big)} = 2 \sum_{i=1}^{k} \big(\ln(y_i) - r\,x_i - \ln(y_0)\big)(-1) = 0 \tag{10}$$

$$\rightarrow \quad \sum_{i=1}^{k} y_i = r \sum_{i=1}^{k} x_i + \ln(y_0) k \tag{11}$$

,

$$\rightarrow \quad \overline{\ln(y_i)} = r\,\overline{x}_i + \ln(y_0) \tag{12}$$

$$\rightarrow \quad \overline{\ln(y_0)} = \overline{\ln(y_i)}\, r\,\overline{x}_i. \tag{13}$$

Partially differentiating with respect to r:

$$\frac{\delta V\big(r, \ln(y_0)\big)}{\delta V\big(\ln(y_0)\big)} = 2 \sum_{i=1}^{k} \big(\ln(y_i) - r\,x_i - \ln(y_0)\big)(-1) = 0 \tag{14}$$

.

$$\rightarrow \sum_{i=1}^{k} (x_i\, y_i) = r \sum_{i=1}^{k} x_i^2 + \ln(y_0) \sum_{i=1}^{k} x_i \tag{15}$$

$$\rightarrow \quad \overline{x_i\, \ln(y_i)} = r\,\overline{x_i^2} + \ln(y_0)\overline{x_i} \tag{16}$$

$$\rightarrow r = \frac{\overline{x_i\, \ln(y_i)} - \overline{\ln(y_i)}\,\overline{x_i}}{\overline{x_i^2} - \overline{x_i}^2} \tag{17}$$

ŷ

(7)

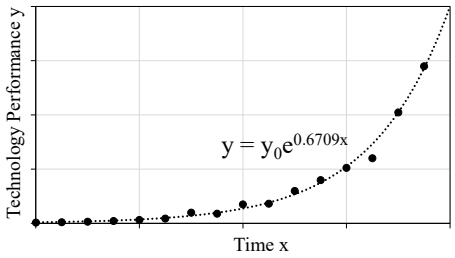


Figure 6: Determination of the growth coefficient r, in the graph r =0.6709, which corresponds to an annual technology growth of 67.09%.

Given the growth coefficient r, the exponential growth function from (7) is described completely. The period during which the performance of technology doubles is calculated with the equations (19) and (20):

$$x = \ln\left(\frac{y}{y_0}\right)\frac{1}{r} \tag{19}$$

Doubling the performance implies $y/y_0 = 2$:

$$\ln(2) = 0.69314718 \approx 0.70 \tag{20}$$

The fact that the natural logarithm of 2 equals 0.7 is the reason behind the 70-Rule, cf. (20), which is used in the automotive industry as a rule of thumb for rapid technology forecast and assessment.

$$x = 0.7\frac{1}{r} = 70\frac{1}{100\ r} \quad (21)$$

In the example of Figure 6, the period of doubling of the technology performance is 70 divided by 67.09, resulting in 1.04 years.

3. Analysed cases

In this section, the mathematical description of the technology performance is compared with actual data from technology undergoing continuous development. Five examples show the dependence of technology performance on growth functions. Each of the examples considered was individually analysed to identify logistic or exponential growth rates. The chosen examples are:

- Number of active Facebook users worldwide
- Tensile yield point of flat bar steel
- Number of transistors per unit area on integrated circuits
- Fuel efficiency per dimension of the VW Golf series
- Energy density of Li- Ion cells

3.1. Number of active Facebook users worldwide

Table I gives the number of active Facebook users from 2008 to 2015 [10]. The data considered for the analysis begins in 2008, when Facebook launched French, German and Spanish versions, leading to a surge in users worldwide [11].

Table 1: Annual Number Of Active Facebook Users Worldwide, [10]

Active Facebook users worldwide	
x	***y***
Quarter / Year	Number in millions
Q4 / 2008	100
Q4 / 2009	360
Q4 / 2010	608
Q4 / 2011	845
Q4 / 2012	1,056
Q4 / 2013	1,228
Q4 / 2014	1,393
Q4 / 2015	1,591

Figure 7 shows the determination of the growth coefficient r = 0.791 and the maximum value K = 1,582 using the linear equation (4).

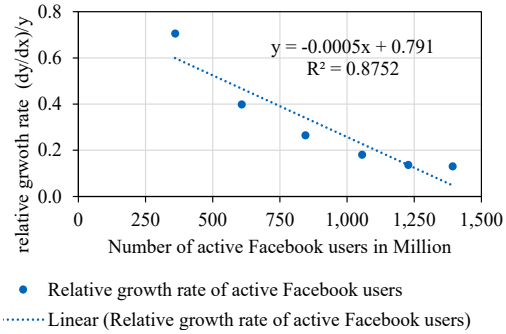


Figure 7: Relative growth rate in relation to the absolute technology performance

Figure 8 displays the S-curve including the coefficients of the logistic growth from equation (2).

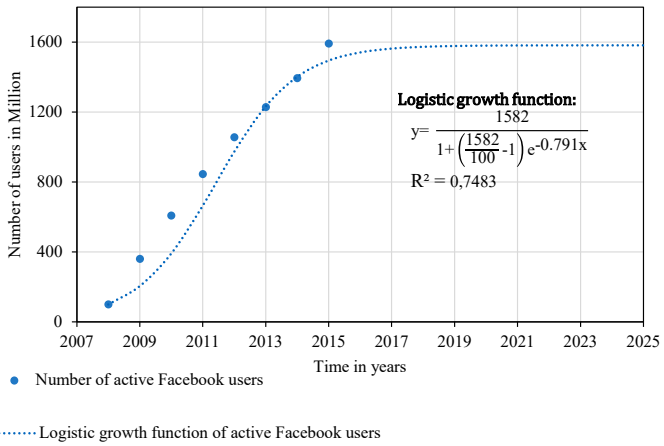


Figure 8: Active Facebook users worldwide, described by logistic growth.

Conclusion: The relative growth rate of active Facebook users worldwide slowed down after the years 2009 and 2010. People without internet access represent the sleeping potential of the technology performance. Despite the slowdown in user growth, no substitute technology has been identified thus far.

3.2. Tensile yield point of flat bar steel

The tensile yield point is a material parameter and designates the tension limit of non-permanent plastic deformation in the material. Table II gives an overview of the technological progression of the achievement potential.

Table 2: Tensile yield point of flat bar steel [12]

Tensile yield point of flat bar steel	
x	***y***
Year	*Tensile yield point in* MPa
1926	355
1955	460
1965	690
1972	890
1980	960
1996	1,100
2013	1,200

Figure 9 shows the linear equation of the relative growth rate in relation to the absolute technology performance. Solving the linear Equation with Equation (4), the values r and K are found to be equal to r = 0.0411, K = 1,500.

Figure 10 displays the logistic growth function adapted from r = 0.0411 to 0.03 due to a minimum set of data pairs relative to the period.

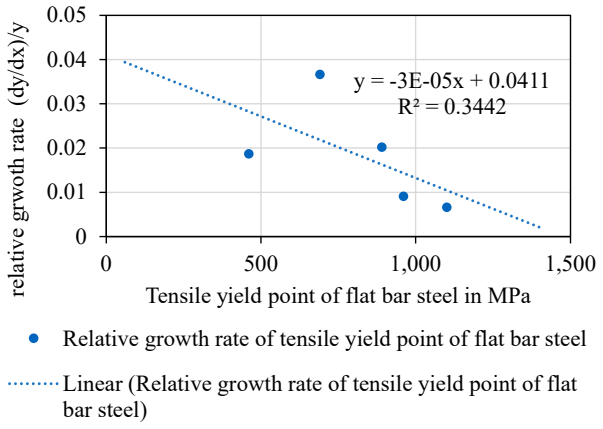


Figure 9: Relative growth rate in relation to the absolute technology performance of flat bar steel

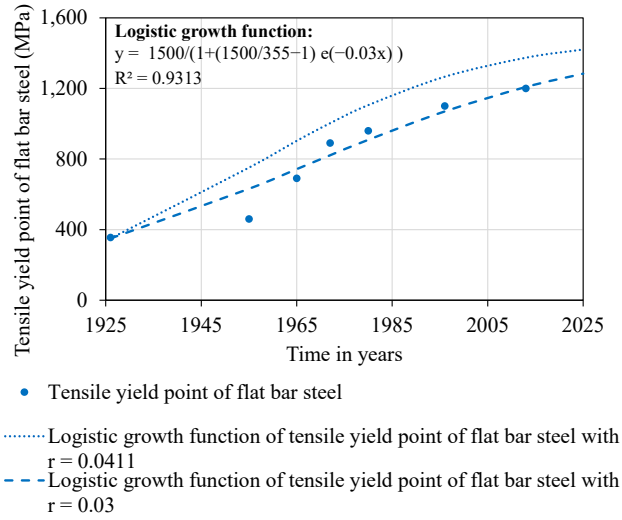


Figure 10: Logistic growth function of the tensile yield point of flat bar steel with adapted growth rate from 0.0411 to 0.03.

Conclusion: The tensile yield point of flat bar steel from [12] has left the performance phase and is now in its key phase.

3.3. Number of transistors per unit area on integrated chips

The number of transistors per unit area on integrated circuits is designated as the integration density, [13]. In 1965 Gordon Moore, Co-founder of Intel, predicted that the integration density will double every 12 to 24 months. This observation is known as Moore's law. In [14] the executive Vice President of Intel William M. Holt, showed that there are leading technology options in the near future beyond complementary metal-oxide-semiconductors (CMOS) that can advance the exponential prediction of Gordon Moore. Table III depicts the integration density development.

Table 3: Overview of Integration density development on integrated chips [13]

Number of transistors per unit area on integrated circuits (integration density)		
Processor	***x***	***y***
	Year	Integration density
Intel 4004	1971	2,300
Intel 8008	1972	3,500
Intel 8080	1974	4,500
Intel 8088	1979	29,000
Intel 80286	1982	134,000
Intel 80386	1985	275,000
Intel 80486	1989	1,180,235
Pentium	1993	3,100,000
AMD K5	1996	4,300,000
Pentium II	1997	7,500,000
Pentium 4	2000	42,000,000
AMD K8	2003	105,900,000
Itanium 2 with 9MB cache	2004	592,000,000
Cell	2006	241,000,000
Core 2 Duo	2006	291,000,000
Six core opteron 2400	2009	904,000,000
8-core Xeon nehalem-ex	2010	2,300,000,000
8-core Itanium Poulson	2012	3,100,000,000
Xbox One main SoC	2013	5,000,000,000
22-core Xeon Broadwell-E5	2016	7,200,000,000

Figure 11 shows that the mathematical determination of the growth coefficient r and the maximum value K with (4) is not possible. The reason is the continuous relative growth rate of the integration density, which has not slowed down over the years.

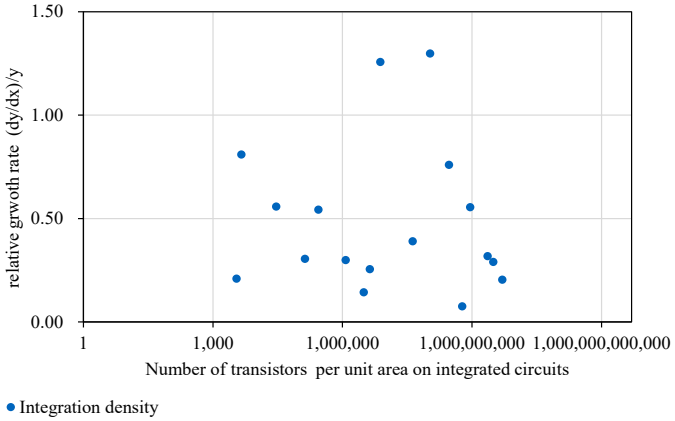


Figure 11: Relative growth rate over a number of transistors per unit area on integrated circuits.

According to [14] the exponential growth of the integration density is based on the substitution of the process and the device technology. This leads to an extension of the performance phase. Figure 12 gives an overview of the innovations leading to the performance increase. The exponential fit uses the method of least squares from [9].

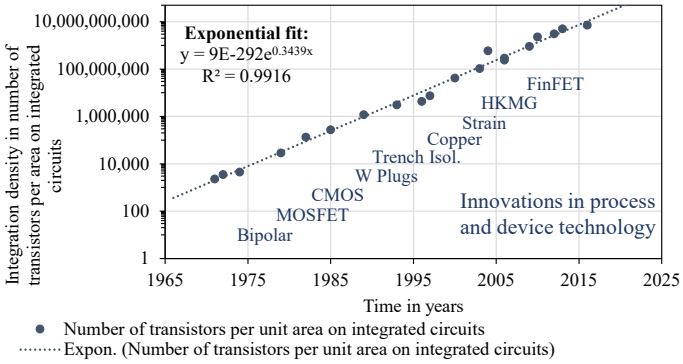


Figure 12: Exponential growth of integration density on logarithmic scale

Conclusion: The trend demonstrated in Figure 12 shows the substitution of process and device technology leading to an extended performance phase. The time for doubling the number of transistors per unit area is around 2.04 years, taking into account an annual growth rate of 34.39 %. Moore's prediction from 1965 is still valid.

3.4. Fuel Efficiency per dimension of VW Golf series

The automotive industry faces new challenges such as carbon dioxide (CO_2) legislation in Europe with an upper limit of 95 g/km for every newly licensed passenger car from the beginning of 2020 [20]. In the United States, the Environmental Protection Agency (EPA) tightened the regulatory programme to 225 gram CO_2 per mile for passenger cars [21] this year. As a contrary trend, vehicle dimensions (length, width, height) are still increasing. Table V shows the fuel efficiency in kilometre per litre and the growing dimensions of the VW Golf series since 1976. That year, the VW Golf had its start of production (SOP) with its first diesel engine.

Table 4: Fuel efficiency Per vehicle dimensionS of VW Golf series I to Vii [22], [23], [24]

Fuel efficiency per vehicle dimension of VW Golf series (diesel engine)							
	x						***y***
Series	Year of SOP	l/100 km	km/l	hp	ccm	*l w h in* m³	*(*km/l) m³
Golf I	1976	5	20.0	50	1,500	8.4	167
Golf II	1983	5	20.0	54	1,577	9.4	188
Golf III	1996	5	20.0	110	1,896	9.7	194
Golf IV	1997	4.9	20.4	110	1,896	10.4	211
Golf V	2007	4.5	22.2	105	1,896	11.0	244
Golf VI	2009	4.1	24.4	105	1,598	11.0	269
Golf VII	2012	3.8	26.3	105	1,598	11.1	292

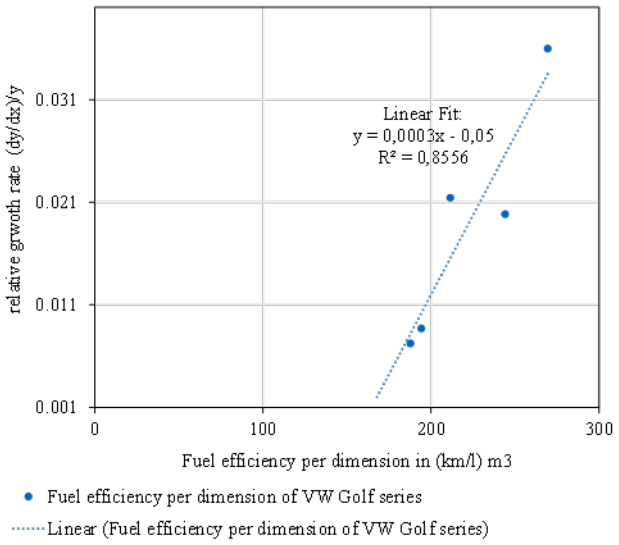


Figure 14 displays the exponential growth of the fuel efficiency per dimension. The coefficient of determination shows a correspondence of $R^2 = 0.9062$.

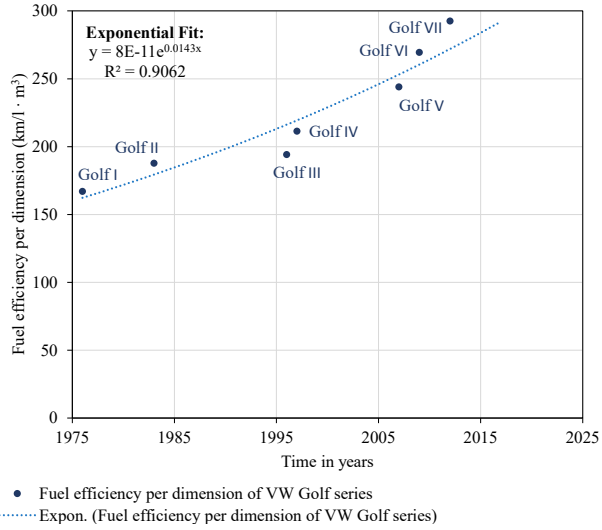


Figure 14: Exponential growth of fuel efficiency per vehicle dimension of VW Golf series.

Conclusion: The fuel efficiency per vehicle dimension of the VW Golf series is exponentially increasing. The fuel efficiency of successive generations of vehicles has improved despite the increasing vehicle dimensions (length, width and height). The substitution of technologies in each vehicle generation has caused this effect, leading to the extension of the performance phase. The exponential annual growth rate determined amounts to 1.43 %. The 70 rule predicts a doubling of performance in 48.95 years.

3.5. Energy density of Li-Ion cells

Over the next ten years, Li-Ion cells will not be only used in stationary and portable applications. According to [20], about 50 % to 70 % will be used in electro-mobility.

Table 5: Energy density of Li-ion cells with 18650 format [21], [22], [23] and [24]

Energy Density of LI-Ion cells	
x	***y***
Year	*Density in* Wh/kg
1992	100
1996	110
1998	130
2000	140
2001	150
2002	160
2004	180
2005	200
2009	210
2010	230
2012	270

Figure 15 displays the relative growth of Li-Ion cells compared to technology performance in Wh/kg. The linear equation from (4) cannot be solved, as the relative growth rate does not slow down, while the growth coefficient r for logistic growth and the maximum value are not identified. The reason is the ongoing exponential growth.

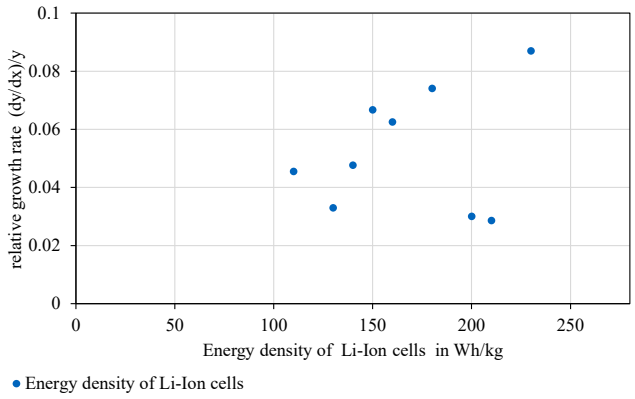


Figure 15: Relative growth rate over energy density of Li-Ion cells

In Fig 16, the exponential fit shows a correspondence of 0.976 for the coefficient of determination R^2. Data points were matched with the names of cells from Sanyo/Panasonic.

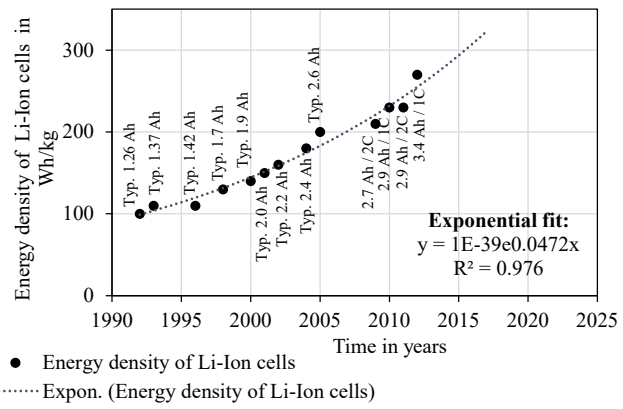


Fig 16: Exponential growth of energy density of Li-Ion cells.

Conclusion: The energy density of Li-Ion cells is still in its performance phase. The substitution of the material composition of anode, cathode or separator leads to the extension of the exponential performance. The ongoing exponential growth has an annual growth rate of 4.72 %. Using the 70 rule, a performance doubling of the energy density in Li-Ion cells can be expected in 14.83 years. The Fraunhofer Institute for System and Innovation Research predicts that post-Li-Ion technology can reach a performance of up to 500 Wh/kg on the level of individual cells [20].

4. Summary and Conclusion

In response to the rising demand of companies to make reliable decisions based on TM and technology road maps, this paper proposed a mathematical approach for technology forecasting and assessment. The benefit for academics and industry is a quantitative method for identifying the level of technology maturity. The mathematical description of achievement potential using the logistic growth function has been introduced and compared with the S-curve model. Table VI gives an overview of the analysed case studies.

Table 6: Overview of results

Overview of results		
Technology	***Phase in the S-curve model***	***Period of performance doubling in years***
Facebook user	Key phase	-
Tensile yield point of flat bar steel	Key phase	-
Number of transistors per unit area on integrated chips	Performance phase	2.04
Fuel efficiency per dimension of VW Golf series	Performance phase	48.95
Energy density of Li-Ion cells	Performance phase	14.83

The 70 rule for deriving the period of performance doubling, is a method for forecasting the technological development in its performance phase. Both rules are applied to a series of example cases. The results clearly show that the key technologies such as the number of transistors per unit area on integrated chips, the fuel efficiency per dimension and the energy density of Li-Ion cells are still growing exponentially.

5. Outlook

Further investigation and discussion is needed to clarify whether integrating other relevant aspects that influence the technology performance forecast is meaningful. Firstly, an analysis must clarify which aspects can be considered as relevant for technological growth besides improvements in process and device technology: For instance, it is generally assumed that interdependencies exist between the improvement of the growth rate and aspects such as the financial environment and its

characteristics such as resources, organisational structure and corporate culture. Secondly, the renewal time, during which the technology performance improves, must be measured precisely. It influences the shape of the S-curve model and can lead to transitions between the technological growth phases. Further possibilities, such as data mining and deriving key indicators for research and development topics, e.g. more precise online map data [25], lead to well-grounded decisions in strategic technology management. Technology road maps based on fundamental figures will be more successful in the long run. One of the next steps is the integration of automated search algorithms to gain a larger database for more accurate predictions of growth rates. An approach based on big data can help to gather more information over a shorter period. The future management of research and development will lead to a new dimension of decision-making tools.

Contributions

Special thanks go to Mr. Claudio Sutter, who helped to collect raw data for further investigation into technological growth during his master's thesis in 2015.

Acknowledgement

This work was presented at the IEEE 2016 International Conference on Industrial Engineering and Engineering Management (IEEM16)

References

[1] M. Fries, M. Lienkamp: "Technology Assessment Based on Growth Functions for Prediction of Future Development Trends and the Maximum Achievable Potential", in IEEE Proceedings of 2016 International Conference on Industrial Engineering and Engineering Management, Indonesia, 2016. https://doi.org/10.1109/IEEM.2016.7798140

[2] J. Höhmann, "New tools for strategists," Harvard Business Manager Germany, vol. 6/2014, June 2014.

[3] M. Steinert, L. Leifer, "Scrutinizing Gartner's Hype Cycle Approach", in IEEE Proccedings of Technology Management for Global Economic Growth, 2010.

[4] G. Schuh, D. Guo and M. Wellensiek, "Approach for the Measurement of Technology Management Performance and Value", in IEEE Proceedings of Portland International Conference on Management of Engineering and Technology, 2013.

[5] G. Schuh, S. Klappert, Technology Management – Handbook Production and Management 2. Berlin, Heidelberg: Springer Verlag, 2011, pp.43 ff.

[6] P. Blanchard, P. L. Devaney, G. R. Hall, Differential Equations. Boston, Cengage Learning, 2012.

[7] P. Blanchard, P. Cummings: Course Math226.1x Introduction to Differential Equations. Module 1, Video 6 & 7. Boston University, https://courses.edx.org, last call 26.04.2015.

[8] T. Wihler, Mathematics for Biology, Mathematical Institute, University of Bern, 2009, pp. 50-52.

[9] University of Leipzig, "Regression analysis", Informatics. http://www.stksachs.uni-leipzig.de/tl_files/media/pdf/lehrbuecher/informatik/Regressionsanalyse.pdf, last call 01.06.2016.

[10] Internet Live Stats, Facebook active users, http://www.internetlivestats.com/watch/facebook-users, statistics analysed from statista.com, last call 10.06.2016.

[11] M. Musgrove, "Facebook passes MySpace with global boost", The Washington Post - Technology, June 2008.

[12] Salzgitter Flachstahl GmbH, "Moderne Stahlvielfalt mit Blickpunkt Zukunft was die Automobilindustrie erwarten darf" (in German). Hannover Messe, 2013.

[13] T. Häberlein, Computer Engineering. Wiesbaden: Vieweg + Teubner Verlag, 2011, pp. 2-5.

[14] W. M. Holt, "Moore's law: A Path Going Forward", in IEEE Proceedings of International Solid-Stat Circuits Conference, Februrary 2016.

[15] European Parliament and Council of the European Union, "Setting emissions performance for standards for new passenger cars as part of the Community's integrated approach to reduce CO_2 emissions from light-duty vehicles", Regulation (EC) No 443/2009, April 2009.

[16] United States Environmental Protection Agency, "EPA and NHTSA finalize historic National Program to reduce Greenhouse Gases and Improve Fuel Economy for Cars Trucks", https://www3.epa.gov/otaq/climate/regulations/420f10014.pdf, April 2010.

[17] Autodata, "Volkswagen – Golf VII", http://www.auto-data.net/de/?f=showCar&car_id=17891, last call 01/06/2016.

[18] Autoscout24, "Golf – 2008 until today", http://autokatalog.autoscout24.de/vw/golf-vi-2008-heute/, last call 01/06/2016.

[19] Autobild, "The real Volkswagen – VW Golf II 1983 - 1991/92", "http://www.autobild.de/artikel/vw-golf-ii-1983-1991-92--44133.html, last call 01/06/2016

[20] Fraunhofer Institute of System and Innovation Research, Roadmap energy storage for electromobility 2030, December 2015.

[21] A. Kinoshita, "Development of Sanyo Li/Ion Batteries", Sanyo Electric Co., Ltd, 2009.

[22] SANYO: Cell Type UR18650ZTA, Specifications, June 2010

[23] PANASONIC: Cell Type NCR18650PD, Specifications, February 2010, pp. 5-7.

[24] PANASONIC: Cell Type NCR18650B, Specifications, January 2013, pp. 5-7.

[25] M. Fries, M. Kruttschnitt, M. Lienkamp: "Multi-Objective Optimization of a Long-Haul Truck Hybrid Operational Strategy And a Predictive Powertrain Control System", in IEEE Proceedings of the 12th International Conference on Ecological Vehicles and Renewable Energies, Monaco Monte Carlo, 2017. https://doi.org/10.1109/EVER.2017.7935872

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1050-1058 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Improving the Performance of Fair Scheduler in Hadoop

Shou-Chih Lo*, Ya-Wen Cheng

Department of Computer Science and Information Engineering, National Dong Hwa University, 974, Taiwan

ARTICLE INFO

Article history:
Received: 28 April, 2017
Accepted: 29 June, 2017
Online: 15 July, 2017

Keywords:
Cloud Computing
Hadoop
Fair Scheduling
MapReduce
Data Locality

ABSTRACT

Cloud computing is a power platform to deal with big data. Among several software frameworks used for the construction of cloud computing systems, Apache Hadoop, which is an open-source software, becomes a popular one. Hadoop supports for distributed data storage and the process of large data sets on computer clusters based on a MapReduce parallel processing framework. The performance of Hadoop in parallel data processing is depended on the efficiency of a job scheduling algorithm underworking. In this paper, we improve the performance of the well-known fair scheduling algorithm adopted in Hadoop by introducing several mechanisms. The modified scheduling algorithm can dynamically adjust resource allocation to user jobs and the precedence of user jobs to be executed. Our approach can properly adapt to the runtime environment's condition with the objective of achieving job fairness and reducing job turnaround time. Performance evaluations verify the superiority of the proposed scheduler over the original fair scheduler. The average turnaround time of jobs can be largely reduced in our experiments.

1. Introduction

This paper is an extension of work originally presented in PlatCon-17 [1], where we show the general ideas. Significant changes are added in this extended paper to explain more about background knowledge, related work, design philosophy, and performance evaluation. Nowadays, huge volumes of data are generated on the Internet every day due to the popularity of social media and portable devices, and this opens up a new era of big data [2]. How to extract interesting information from these data has become a hot topic in science and commercial fields. Big data analytics needs the support of scalable data storage and powerful data process. Cloud computing [3], which provides distributed storage and parallel data processing on commodity computer clusters, just meets this requirement on data analysis.

Cloud computing, as defined by National Institute of Standards and Technology (NIST) [4], should be composed of five essential characteristics: broad network access, rapid elasticity, measured service, on-demand self-service, and resource pooling. Moreover, cloud computing has four basic deployment models: public cloud, private cloud, community cloud, and hybrid cloud, and has three service models: software as a service (SaaS), platform as a service (PaaS), and infrastructure as a service (IaaS).

Many large on-line services are constructed by the technique of cloud computing. These services are usually developed on business-based cloud platforms such as Amazon's EC2, Google's GAE, and Microsoft's Azure. In academic research, Apache Hadoop is a popular cloud platform to test and verify research ideas. Hadoop is a Java-based and open-source software framework that supports a full set of cloud techniques: distributed file system, parallel data processing, and distributed database system.

To do parallel data processing in Hadoop, users need to write MapReduce programs [5]. These kinds of programs include two major processing codes: map and reduce to reflect a two-phase data processing flow. The map code mainly deals with data filtering and data transforming, and the reduce code mainly deals with data aggregating. Users submit these MapReduce programs into the Hadoop system as jobs. Each job contains several map tasks and reduce tasks. The map task typically reads an input data file in the form of key-value pairs, and then generates an intermediate result file in key-value pairs as well. All values with the same key will be grouped together, and then are processed by the reduce task that usually generates aggregated values.

The WordCount application that outputs the frequency of each distinct word occurred in a document file is usually used to explain the MapReduce framework. The map task will scan the input file and output the occurrence of each word in the format of (founded

*Corresponding Author: Shou-Chih Lo, Email: sclo@gms.ndhu.edu.tw

word, 1). Apparently, we only need to accumulate this value 1 for those outputs having the same founded word to be the word frequency. The reduce task just does this grouping and counting.

A Hadoop system can be built over clusters of computers that are typically named as nodes. All map and reduce tasks will be dispatched to these nodes. For parallel data processing, the input data file to a MapReduce job will be equally partitioned into several data chunks (or blocks) with a typical size of 64 MB. Each of these data chunks will be replicated and distributed into different nodes for fault tolerance. The same number of map tasks to the number of data chunks are cloned, and each copy of the map task handles one data chunk. Because there are redundant data chunks, the map task will select the closest data chunk for access efficiency. The reduce task can be cloned as well, and each copy handles the values corresponding to a certain set of keys. This MapReduce working model is outlined in Figure 1.

The Hadoop system can simultaneously run many map and reduce tasks from different users in a computer cluster environment. These tasks will contend system resources such as CPU, memory, and disk I/O for the completion of their jobs. Moreover, those tasks reading data chunks from far-away nodes will compete for network bandwidth. Therefore, a proper task scheduling algorithm is necessary for optimizing resource utilization and avoiding any resource starvation.

Hadoop provides three built-in scheduling modules: First-In-First-Out (FIFO) scheduler, fair scheduler, and capacity scheduler. These schedulers have their own features and have different influences on performance such as execution time and waiting time in different situations [6-10]. Fair scheduler and capacity scheduler generally perform better than FIFO scheduler. These built-in schedulers have a common drawback, which the scheduling policy is almost fixed and is not flexible to the change of working conditions. In this paper, we focus on the fair scheduler and propose some modifications to improve the scheduling throughput under the goal of resource fairness among users.

Our proposed scheduling algorithm is adaptive, because it can dynamically tune some working parameters such as the job priority and the waiting time for resource allocation. The runtime environment's conditions such as current workload and remaining resources are considered in the determination of job running order and the amount of resources allocated to each job. A modified fair scheduler is then coded into the Hadoop system and is examined in a real testing environment. Evaluation results show that the modified fair scheduler can significantly reduce the average turnaround time of a job by over 20 percent as compared to the original fair scheduler.

The remainder of this paper is organized as follows. Section 2 briefly introduces the basic job scheduling architecture in Hadoop and the related work on scheduling algorithms. Section 3 presents and discusses the proposed five mechanisms of improving the performance of fair scheduler. Performance evaluation is conducted in Section 4. Finally, some concluding remarks are given in Section 5.

2. Background Knowledge and Related Work

Hadoop, which is developed under the Apache projects, is an open-source software for reliable, scalable, and distributed computing. The Hadoop system is composed of two basic units: a distributed file system and a distributed data processing engine. The Hadoop distributed file system (HDFS), following the similar

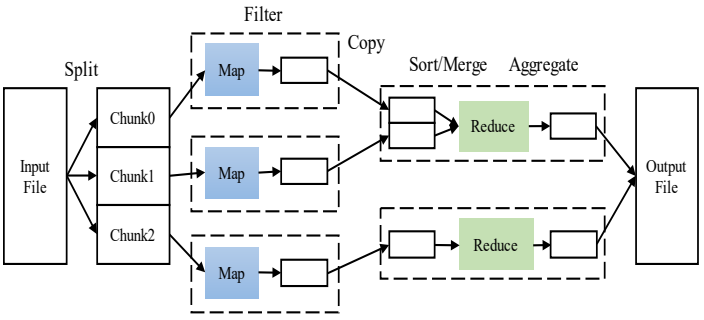


Figure 1. MapReduce working model

concept of Google file system (GFS) [11], can manage distributed data storage across computer clusters built from commodity hardware. A HDFS cluster consists of a single namenode, a master server that manages the namespace and access of files, and multiple datanodes, slave servers that store data chunks and are coordinated by the master server. A file is split into several data chunks which are then stored into datanodes. The namenode will record the mapping of each data chunk to each datanode.

The distributed processing engine in Hadoop is based on the MapReduce framework. A MapReduce program contains a map procedure that performs the data filtering and sorting operations, and a reduce procedure that performs the data summarizing operations. MapReduce programs are submitted by users to Hadoop as jobs. These jobs will be executed over a set of computing nodes where one master node is selected to schedule job executions on the other slave (or worker) nodes. The management functions of system resource and job execution are included in the MapReduce module for Hadoop version 1. These functions, however, are dedicated to anther software module for Hadoop version 2: Yet Another Resource Negotiator (YARN) [12]. No matter the different versions of Hadoop, the job scheduling process is relied on a job scheduler. In this paper, we follow the framework of Hadoop version 1 to illustrate the operation of different job schedulers.

Figure 2 shows the job execution flow in Hadoop. There is a JobTracker at the master node that plays the roles of monitoring/ allocating system resources and scheduling user jobs. The JobTracker gives each submitted job a job ID as an internal identification number and puts the job into a job queue (or pool). A certain sorting policy is applied to these jobs in the queue according to some comparing factors such as job ID, job priority, and job submitted time. When a job is ready to be executed, the associated map and reduce tasks are created. The number of map tasks is determined by the number of data chunks split from an input file. The number of reduce tasks can be configured by the user but is usually one by default.

The JobTracker will manage and monitor all system resources such as CPU, memory, and disk contributed from all slave nodes. These system resources for job running are configured as resource slots in Hadoop. The resource slot used for the running of a map (or reduce) task is called a map (or reduce) slot, respectively. Due to the limited system resource, map and reduce tasks will contend these resource slots. On each slave node, a TaskTracker will manage the actual running of tasks. The TaskTracker will request a task to run from the JobTracker whenever there is a free resource slot. The TaskTracker will also periodically report the execution status and the available resource to the JobTracker by sending a heartbeat message.

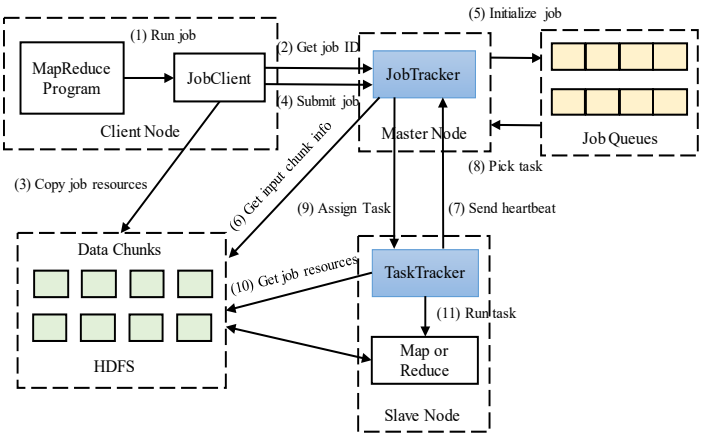


Figure 2. The job execution flow in Hadoop

A map task will read an input data chuck through the HDFS. To save network bandwidth, this data retrieval should happen locally or be done from a close node. This raises the data locality issue when assigning a task to a slave node [13]. There are three levels of data localities (see Figure 3) from good to bad:

- Node-locality: The node where data are stored is the same as the node where data are processed.
- Rack-locality: The node where data are stored is different from the node where data are processed, but these two nodes are located in the same rack.
- Non-locality: The node where data are stored is different from the node where data are processed, and these two nodes are located in different racks.

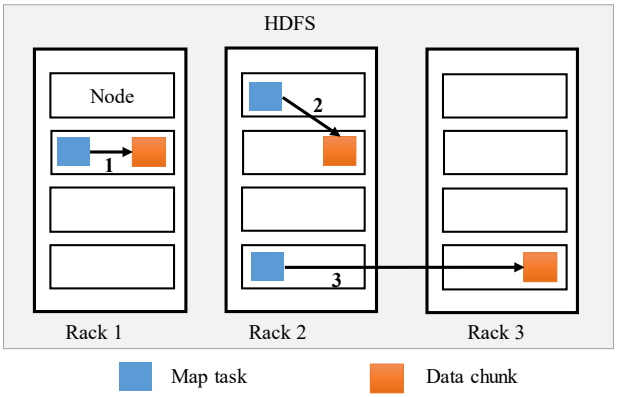


Figure 3. Data locality: (1) Node-locality, (2) Rack-locality, and (3) Non-locality

The Hadoop scheduling problem can be described below: Suppose that each node is configured with some map and reduce slots for the execution of map and reduce tasks, respectively. Given a set of jobs and nodes, we need to determine the order of running jobs and the slot assignment to tasks with the objective of high data locality and high execution throughput. This problem with multi-objective optimization is proven to be NP-hard [14].

In Hadoop, there are three well-known scheduling algorithms implemented as three schedulers: FIFO scheduler, fair scheduler, and capacity scheduler, respectively. The default FIFO scheduler uses one job queue to do slot allocation. Jobs in the queue are sorted in ascending order of their submitted times. To break the tie, the job ID assigned by the JobTracker is considered. All free map and reduce slots in the system can serve the job being scheduled.

In the example of Figure 4, suppose that there are three map slots and three reduce slots in the system, and each job (i.e., Job1~Job3) has three map tasks and three reduce tasks. The sequence to allocate slots to these jobs is Job1, Job2, and then Job3. For simplicity, we assume that each task uses one resource slot with the same occupation time in the figure. At first, all the map tasks of Job1 get map slots for 20 s execution time. Next, these map slots are released and re-allocated to the map tasks of Job2. Meanwhile, all the reduce tasks of Job1 get reduce slots for 20 s execution time. As can be seen, reduce tasks start executing just after the finish of all map tasks of the same job. The drawback of this scheduler is the possible long waiting time for a task to be executed when there are many jobs in the queue. This becomes unfair particularly when a user submits many jobs and the latter users need to wait.

The fair scheduler developed by Facebook equally allocates resources to job users. Each user has its own job queue (or pool), and all resource slots are fairly distributed to these user pools. Figure 5 shows the example slot allocation under the same condition with the previous example. Here three pools (P1, P2, and P3) are introduced with each pool having the resource of one map slot and one reduce slot. Each map task of these three jobs gets one map slot at the same time and runs for 20 s. The same case happens for reduce tasks. As compared to the previous example, each submitted job can start running immediately but the time to finish the job becomes longer. In a real situation, one pool without sufficient resources can borrow free (or idle) slots from another pool to increase resource utilization. To prevent resource unfairness, each pool is configured with the minimum and maximum resource capacity. Fair scheduler also supports preemptive mode by which a low-priority running job can be aborted for releasing resource to a high-priority job.

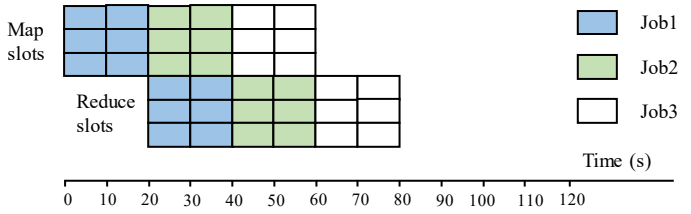


Figure 4. A scheduling example using FIFO scheduler

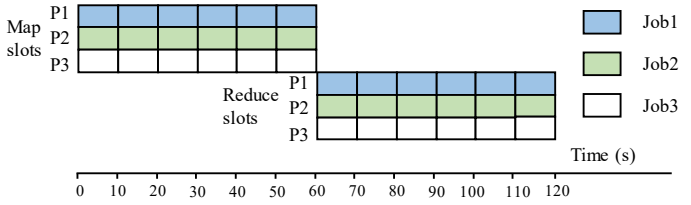


Figure 5. A scheduling example using fair scheduler

The capacity scheduler developed by Yahoo acts similar to fair scheduler, but uses queues instead of pools. Each queue has a defined resource capacity and is assigned to an organization or a group of users. Due to the nature of organization structure, queues can be constructed into a multi-level hierarchy. To increase running efficiency, capacity scheduler can allocate more than one resource slot to a heavy task. Moreover, several tasks can be assigned together to the same TaskTracker in batch mode to reduce the scheduling overhead. Overusing this batch mode, however, would cause load unbalancing among slave nodes. A summary table (see Table 1) is given to compare the characteristics of these three schedulers.

There are several scheduling algorithms proposed to improve the built-in schedulers. The reader is referred to [15-16] for a complete view. To increase data locality, the LATE (Longest Approximate Time to End) scheduler was proposed [17], where a delay timer is set during which a TaskTracker will wait for a suitable task that can fetch data locally. When a TaskTracker requests a task to do from the JobTracker, the JobTracker will scan the job queue to find a task with the node-locality data access level within a short time interval. If it is failed to find such a task, a task with the rack-locality data access level is then searched within a long time interval. If it is still failed, the JobTracker will randomly select one task to this TaskTracker.

The delay concept above is introduced into the naïve fair scheduler [18-19]. The setting of suitable delay time would be challenging in a dynamic changing environment. This delay time is configured as a fixed value in fair scheduler. To increase data locality, one research work implements two job queues for the FIFO scheduler [20]. Jobs having the node-locality potential are put into one queue, and jobs having the rack-locality potential are put into the other queue. Finally, these two queues are merged together and are fed into the FIFO scheduler. The process of data may have certain relationships. For example, some rare data are only processed by certain tasks. In this situation, the system can mark the locations of these rare data and assign irrelevant tasks to other locations for increasing data locality [21].

Table 1. Summary of different schedulers

Scheduler	Characteristics
FIFO Scheduler	• Single job queue • Resource allocation to jobs is considered one by one. • Scheduling overhead is low • There are resource starvation problems
Fair Scheduler	• Multiple job queues • Resource allocation to jobs is considered together. • System resources are fairly allocated to users • Support preemptive mode and resource borrowing
Capacity Scheduler	• Multiple job queues • Resource allocation to jobs is considered together. • System resources are allocated to users according to a certain organization policy • More than one resource slot can be allocated to one task • Support batch mode in job scheduling

The job priority will affect the job order in a job queue. If each user can freely assign the job priority, all users tend to set their jobs to the highest level, and this does not make sense. One auto-setting mechanism is proposed in [22] by considering several factors such as the job size, the average execution time and the scheduled time of a task in a job. A high priority is usually given to a small-size or fast running job. The job size can be simply estimated by the number of tasks involved in a job [23].

In Hadoop, map slots and reduce slots are separated and cannot be interchangeably used. Dynamic borrowing between them can improve slot utilization and system throughput [24]. The map or reduce slot in Hadoop represents a computing unit and the amount of available slots in a node is configured in advance. Dynamically changing the slot number according to the real computing power in speeds of CPU and I/O can also improve system performance [25-26].

The diversity of data, jobs, and computer nodes will also affect the scheduling performance, and this is called the skew problem [27]. Different map tasks from the same job may generate different amounts of data, causing data skew. Different jobs take different execution times depending on algorithms but not job sizes, causing computational skew. Resource slots from different computer nodes have different computing powers, causing machine skew. These skews cause unbalancing workload in a distributed computer cluster. There are many research efforts on designing load balancing scheduling algorithms [28-31]. For example, jobs are classified into CPU-bound and I/O-bound types and are put into different job queues [31].

3. Modified Fair Scheduler

Fair scheduler with the delay mechanism has good performance in general against the other schedulers in Hadoop. However, fair scheduler overemphasizes fair resource allocation to jobs and ignores the differences between jobs. These differences may even change over time when jobs are running. We hope to improve the performance of fair scheduler by considering some runtime conditions and further adjusting resource allocation to jobs. We provide some modifications to the fair scheduler by introducing the following mechanisms: job classification, pool resource allocation, resource-aware job sorting, delay time adjustment, and job priority adjustment. Their brief introductions are given in Table 2.

Table 2. The goal of each proposed mechanism

Mechanism	Goal
Job classification	Separate small-sized jobs from large-sized jobs
Pool resource allocation	Periodically reallocate resource slots to pools according to their remaining actual needs
Resource-aware job sorting	Determine the job or pool order for resource allocation based on more criteria including resource requirement and occupation
Delay time adjustment	Dynamically adjust the delay time for a suitable task with high data locality
Job priority adjustment	Dynamically adjust the job priority based on the change of its data locality level

3.1. Job Classification

In fair scheduler, system resources are equally allocated to user pools. All jobs belonging to the same pool will contend for the limited resource. In real cases, different jobs would have different resource requirements. A large-sized job consumes more resource than a small-sized job. Based on the principle of shortest job first, we separate small-sized jobs from large-sized jobs and allocate resources to them individually.

The job size is measured by the number of map and reduce tasks in a job. In general, the number of map tasks is greater than

that of reduce tasks, because most MapReduce jobs focus on data extraction from large input data sets. For simplicity, we concentrate on the map slot allocation to map tasks, so the job size is defined as the number of map tasks only. A job is recognized as a small-sized one if its job size is no greater than a threshold (minJobSize). This threshold value is not a fixed value but is updated periodically by the JobTracker (per 500 ms in our implementation) to be the currently smallest job size among all jobs in the system.

The basic concept of the proposed job classification is to direct all small-sized jobs into another system pool (called sharedPool) and all large-sized jobs into their originally belonging pools. In other words, there is a system pool shared by all users besides individual user pools. Jobs in the sharedPool are scheduled using FIFO for simplicity.

In an extreme case, if all jobs are small-sized, this job classification becomes useless. To prevent this condition, the ratio of the number of total small-sized jobs (SmallJobs) to the number of total jobs (TotalJobs) in the current system is considered. If this ratio is greater than a threshold, the sharedPool mechanism is disabled and all small-sized jobs remain in their own user pools. This threshold value is set to be the reciprocal of the average number of map slots configured in a salve node. For example, if each slave node has four map slots on average, the threshold value is 0.25. This means that we allow using sharePool when there is at most one small-sized job running on a slave node on average. Figure 6 shows the flow of job classification.

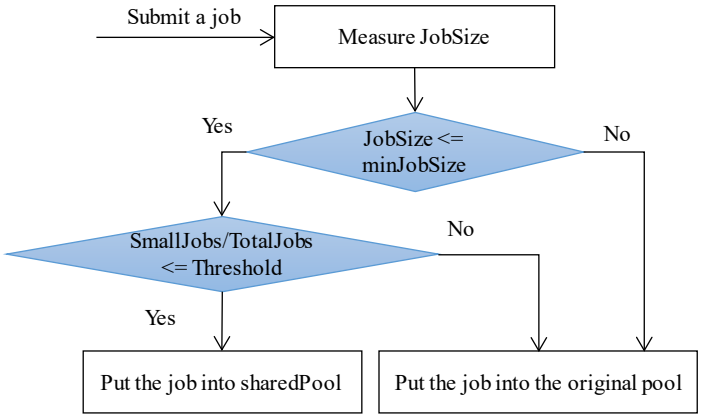


Figure 6. Job classification flow

3.2. Pool Resource Allocation

System resources are proportionally allocated to user pools according to weighted values assigned to user pools in fair scheduler. Without the pre-knowledge of traffic shape, it is hard to set a suitable weighted value to each user pool, and this weighted value is set to be equal most of the time. However, if this weighted value can automatically reflect the current resource requirement, resource allocation can always fit real situations and becomes more efficient. Based on this concept, a dynamic pool resource allocation is proposed.

The resource allocation to pools is periodically adjusted every 500 ms in our implementation. The allocation of map slots and reduce slots to each user pool is performed separately. We first check whether the sharedPool is enabled without any allocated resource. If that is the case, a certain amount of resource slots is allocated to the sharedPool. For the efficient schedule of this small amount of small-sized jobs in the sharePool, the FIFO scheduler is applied here. We allocate a half of resource slots needed by the currently smallest job in the system to the sharedPool. That is, the number of resource slots allocated to the sharedPool is $\lceil minJobSize/2 \rceil$. This decision is for the reason that large-sized jobs should get more resource and hence we allocate limited resource to small-sized jobs. For example, if minJobSize is four in map tasks and two in reduce tasks, this sharedPool gets two map slots and one reduce slots.

The remaining free resource slots (RemainingSlots) are proportionally allocated to the other user pools according to their actual needs. The portion of total resources allocated to a user pool is based on the ratio of the number of pending tasks in a user pool (PoolTasks) to the number of total pending tasks in all user pools (allPoolsTasks). A pending task is a task that is not allocated with any resource slots and is waiting for execution. The ratio above indicates the remaining resource requirements of a user pool against the total remaining resource requirements of all jobs. Remember that only map (or reduce) tasks are considered in the counting of the number of tasks when map (or reduce) slots are allocated. To prevent one user pool with heavy workload from getting too much resource, the actual number of allocated slots cannot exceed the maximum number (maxAllocated) configured in the system. Figure 7 shows the flow of this dynamic pool resource allocation.

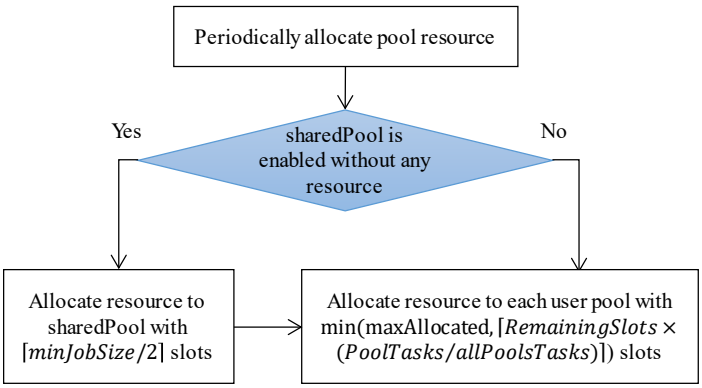


Figure 7. Pool resource allocation flow

3.3. Resource-Aware Job Sorting

The order to allocate resource to jobs is based on the order of jobs. All jobs are in the same system pool for FIFO scheduler, but are in different user pools for fair scheduler. Therefore, another pool sorting is necessary besides the job sorting for fair scheduler. The job or pool order is determined based on a sorting policy applied in the FIFO comparator or the FairShare comparator in Hadoop. These two comparators, which are involved in FIFO scheduler and fair scheduler, respectively, will return the order between two target jobs or two target pools based on some reference factors. Normally, these reference factors include the job priority, job ID, and submitted time. Here, we additionally consider the remaining resource requirements and the current resource occupation of a job or a user pool to do fine-grained sorting. The principle is to allocate resource first to those jobs or pools that have occupied less resource or desire for more resource.

At first, we describe the way to determine the job order in the same pool, which can be used both in FIFO scheduler and fair scheduler. We measure the remaining resource requirements of a

job by the number of pending map or reduce tasks in the job. Remember that only map tasks are counted when we consider the map slot allocation. We define the pending task ratio of a job (JobPendingRatio) in (1) as the number of pending tasks in a job to the number of total pending tasks in a pool. This ratio is high if the corresponding job has more unfinished tasks against other jobs, implying high resource requirements in the future.

JobPendingRatio = (#pending tasks in a job)/(#pending tasks in a pool) (1)

The current resource occupation of job is measured by the number of currently running tasks in the job. We define the occupied resource ratio of a job (JobOccupiedRatio) in (2) as the number of currently running tasks in a job to the number of totally allocated slots to a pool. If a task is assigned with at most one resource slot, the number of allocated slots is equal to the number of running tasks. A job with more occupied resource should have low precedence to contend new resource.

JobOccupiedRatio = (#running tasks in a job)/(#allocated slots to a pool) (2)

Our sorting policy for jobs in the same pool is designed as follows. Five comparing factors are examined in sequence if there is a tie as in Figure 8. The high order is given to the job with a high job priority (more detailed settings are discussed latter). The subsequent order is given to the job with a high JobPendingRatio and then with a low JobOccupiedRatio. To break the tie, the submitted time and the job ID are then examined.

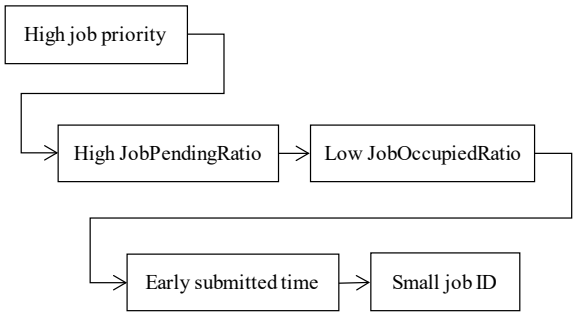


Figure 8. Sequence to examine the job order

In fair scheduler, there are several pools and the order of these pools should be determined before jobs are scheduled. Our sorting policy for pools is also from the perspective on resource requirements. First, the number of all pending tasks for each job in a pool is accumulated as the maximal resource demand of a pool (PoolDemand) in (3). Suppose that each user pool is configured with the minimum number of resource slots (minAllocated) in advance. If PoolDemand is less than minAllocated, minAllocated would be set to be PoolDemand to reflect the actual minimum resource need. This step is for tuning the minAllocated value.

PoolDemand = ∑#pending tasks, for all jobs in a pool (3)

Next, the number of currently running tasks for each job in a pool is accumulated as the current workload of a pool (PoolWorkload) in (4). If PoolWorkload is no greater than minAllocated, this means that the pool has sufficient resource to accommodate new tasks and hence a high pool order is given. That is, the pools are in the order of pools with sufficient resource followed by pools with insufficient resource.

PoolWorkload = ∑#running tasks, for all jobs in a pool (4)

Those pools having sufficient resource are further sorted by comparing the occupied resource ratio (PoolOccupiedRatio) of PoolWorkload to minAllocated in (5). The remaining resource is large if this ratio is low, and hence a high order is given to a pool with small PoolOccupiedRatio.

PoolOccupiedRatio = PoolWorkload / minAllocated (5)

Those pools having insufficient resource are further sorted by comparing the PoolWorkload and the weight of a pool that is given by the user or the system to indicate the pool priority. The ratio of these two values is computed in (6) and is denoted as WeightedPoolWorkload. A high order is given to the pool with small WeightedPoolWorkload. Here we suppose that the pool priority is high if the corresponding pool weight is large. Figure 9 shows the comparing sequence. If any tie happens, a random order is given.

WeightedPoolWorkload = PoolWorkload / (weight of a pool) (6)

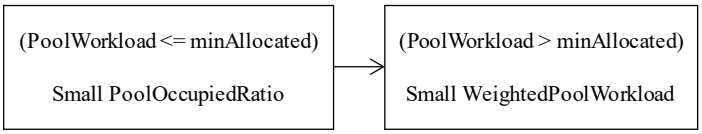


Figure 9. Sequence to examine the pool order

3.4. Delay Time Adjustment

To increase data locality during job scheduling, JobTracker will seek first for those tasks in pools having node-locality or rack-locality levels. The setting of delay time during task scheduling is proven to be helpful on data locality. The TaskTracker will wait for a suitable task satisfying node-locality in a pool until the time out of a short delay timer. If no such task is found, the TaskTracker continues waiting for a suitable task satisfying rack-locality in a pool until the time out of a long delay timer. Otherwise, one task is randomly selected from a pool and is fed to the TaskTracker.

The fixed setting of these delay timers in the current implementation of fair scheduler is not suitable in a dynamically changing environment. A long delay timer will increase the average turnaround time of jobs, and a short delay time will decrease the average data locality level. Here, a dynamic approach is applied by setting the timer according to the recently average waiting time as follows:

NodeLocalityWaitedTime = the average waiting time of tasks satisfying node locality (7)

RackLocalityWaitedTime = the average waiting time of tasks satisfying rack locality (8)

In our implementation, we use two pairs of counters (TimeCount and NumberCount) to maintain each waited timer as the value of TimeCount/NumberCount. If a task is scheduled with node-locality level, the scheduled waiting time is added into TimeCount and NumberCount is increased by one. Another pair of TimeCount and NumberCount is used when a task is scheduled with rack-locality level.

3.5. Job Priority Adjustment

In Hadoop, each job can be given with one of five priority values from high to low: VERY_HIGH, HIGH, NORMAL, LOW, and VERY_LOW. However, the job priority is not fully used and is set to a fixed value (NORMAL) without any specific specification. This priority value is one of the factors that are considered in job sorting. If this priority value can be dynamically adjusted to reflect the data locality level of a task for a job, the job with high data locality will be scheduled with high precedence. This strategy can increase overall data locality level and hence reduce the data access time.

We consider the change of the data locality level according to two recently scheduled tasks of the same job. There are nine possible changes as listed in the first column of Table 3. The job priority remains the same if there is no change on data locality. If the data locality is downgraded with one level (or two levels), the job priority is also downgraded with one level (or two levels). This job after downgrading will be sorted behind other jobs, and hence JobTracker has more opportunities to find other tasks with good data locality levels. Similarly, the job priority is upgraded with one level (or two levels) if the data locality is upgraded with one level (or two levels). If the job priority keeps in VERY_HIGH (or VERY_LOW) due to no change of data locality level, a minor adjustment to HIGH (or LOW) is applied for the balance of fairness. This prevents a job from keeping with very high or very low priority for a long period of time. The former case makes a job having high priority to seize system resource, and the latter case makes a job having the chance to starve for system resource.

Table 3. Change of job priorities

Data locality change	**Original Job Priority**				
	VERY_HIGH	HIGH	NORMAL	LOW	VERY_LOW
Node→Node	VERY_HIGH	HIGH	NORMAL	LOW	VERY_LOW
Node→Rack	HIGH	NORMAL	LOW	VERY_LOW	LOW
Node→None	NORMAL	LOW	VERY_LOW	VERY_LOW	LOW
Rack→Node	HIGH	VERY_HIGH	HIGH	NORMAL	LOW
Rack→Rack	VERY_HIGH	HIGH	NORMAL	LOW	VERY_LOW
Rack→None	HIGH	NORMAL	LOW	VERY_LOW	LOW
None→Node	HIGH	NORMAL	LOW	VERY_LOW	LOW
None→Rack	HIGH	VERY_HIGH	HIGH	NORMAL	LOW
None→None	VERY_HIGH	HIGH	NOMAL	LOW	VERY_LOW

4. Performance Evaluation

An experimental testing environment is established to evaluate the proposed fair-share-based algorithm against the original algorithm. The capacity scheduler is not examined, since it is little bit hard to configure a uniform testing environment to compare these schedulers together. Two physical computer servers are selected from two buildings on our campus. Each server is configured with four virtual machines (VMs) using the software of VMware ESXi. Each VM is installed with the operation system of CentOS 5.5 and Hadoop 1.2.1. These VMs are configured with physical IP addresses and are connected together to be one Hadoop cluster. That is, we have eight nodes in the cluster, and one node acts as both master and slave nodes and the other seven nodes act as slave nodes. Those nodes on the same physical computer are viewed as on the same rack. The hardware specification of the master node is: one dual-core CPU, 8GB memory, and 200 GB disk space. The specification of the slave node is: one dual-core CPU, 4 GB memory, and 200 GB disk space.

The replication factor of HDFS is set to three by default in Hadoop, which implies one data chunk is replicated to three nodes. However, we set the replication factor to one here and make the general node-locality level low. The reason is for the easy observation on the improvement of data locality. All the submitted jobs are based on the WordCount program which outputs the frequency of each distinct word occurred in an input document file. The job size is controlled by the size of input file. In other words, each job can contain a different number of map tasks but contain one single reduce task. The initial job priorities of all jobs are NORMAL. There are three user pools with equal weight in the system. All the submitted jobs will evenly be distributed into these pools. The minAllocated and maxAllocated values for each user pool is three and six in number of map slots, respectively. The default delay time in the original fair scheduler is 3000 ms. The above experimental settings are summarized in Table 4.

Table 4. Experimental settings

Physical server	2
Virtual machine	8
Hadoop system	1.2.1
HDFS replication	1
Default delay time	3000 ms
Job size	6~25 map tasks
User pool	3

Two sets of experiments are performed. All jobs in the first experiment have the same size, but have different sizes in the second experiment. For each experiment, two cost metrics are observed as follows:

- Average turnaround time: The average time taken between the submission of a job for execution and the completion of result output.
- Average node-locality ratio: The ratio of the number of task schedules satisfying node locality to the number of total task schedules.

4.1. Experiment A

In this experiment, each node is configured with four map slots and one reduce slot. There are totally 32 map slots in the system. The job size (in number of map tasks) is set to be 6, 10, and 18. For each case, 5, 10, and 15 jobs are respectively examined. That is, all submitted jobs in each testing round have the same size. For this reason, the following proposed mechanisms does not contribute too much on performance improvement. First, the job classification does not work here. Second, the pool resource allocation makes no significant difference on each user pool. Therefore, we mainly observe the proposed mechanisms on resource-aware job sorting, delay time and job priority adjustments. Experimental results are listed in Table 5 and Table

6. Each gain value shown in the tables, indicating the improvement percentage, is defined as follows:

Gain = (Modified FS – Original FS) / Original FS × 100% (9)

As the total number of tasks is increased, the average turnaround time is increased for both the original fair scheduler and our modified fair scheduler because of heavy workload. The modified fair scheduler can significantly reduce the average turnaround time by 21.4% to 54.5% against the original fair scheduler. This indicates that the modified fair scheduler performs proper job sorting and decreases unnecessary waiting delays. These make the utilization of resource slots high. In general, more reduction gains are achieved if the job size is getting large under the same number of jobs. This shows that the modified fair scheduler makes efficient schedule on many map tasks. The original fair scheduler will suffer from serious resource contentions under heavy workload and the lack of flexible scheduling makes the performance worse.

Table 5. Comparison of average turnaround time for jobs of same size

Job size	Number of jobs	Original FS	Modified FS	Gain (%)
6	5	289.6 s	188.2 s	-35%
	10	1083.2 s	761.3 s	-29.7%
	15	1791.7 s	1175.2 s	-34.4%
10	5	886.9 s	462.2 s	-47.8%
	10	1433.7 s	1103.3 s	-23%
	15	2008.1 s	1577.6 s	-21.4%
18	5	2212.3 s	1003 s	-54.6%
	10	3845 s	1894.1 s	-50.7%
	15	5681.4 s	2988.7 s	-47.3%

Table 6. Comparison of average node-locality ratio for jobs of same size

Job size	Number of jobs	Original FS	Modified FS	Gain (%)
6	5	0.09	0.13	+44.4%
	10	0.22	0.23	+4.5%
	15	0.28	0.28	+0%
10	5	0.29	0.32	+10.3%
	10	0.34	0.32	-5.9%
	15	0.33	0.35	+6.1%
18	5	0.21	0.25	+19%
	10	0.28	0.34	+21.4%
	15	0.3	0.36	+20%

Table 6 shows the performance comparison on data locality. First, the general node-locality ratio is low, because the replication factor is only one. Second, the node-locality ratio is improved in general by using our modified fair scheduler. Among our proposed mechanisms, only the mechanism of job priority adjustment considers the factor of data locality. Therefore, our scheduler can slightly increase the node-locality level but cannot guarantee this is always true as one counterexample in the table. Third, improvement gains become clearer when the job size is getting large. When we take a close look at the experimental result of each round, we found that data-locality levels in the original fair scheduler are variant over different testing rounds. For example, there may have several task schedules with node-locality level in the current testing round but with rack-locality level in the next testing round. However, our modified fair scheduler keeps data-locality level in a more stable state under the same input condition.

4.2. Experiment B

In this experiment, each node is configured with two map slots and one reduce slot. There are totally 16 map slots in the system. The job sizes are mixed with five different scales: 6, 10, 16, 21, and 25. For each job size, one, two, and three jobs are respectively generated such that there are totally 5, 10, and 15 jobs in the system. All the proposed mechanisms will function in this experiment. The experimental results are listed in Table 7 and Table 8. As can be seen, the modified fair scheduler largely decreases the average turnaround time by 57% in average, and slightly increases the average node-locality ratio by 3.87% in average.

As compared to Experiment A where only three proposed mechanisms function well, improvement gains become excellent in this experiment. This means that the proposed job classification works very well and the dynamic pool resource allocation based on real resource requirement makes slot allocation more efficient.

In this experiment, we also examine the FIFO scheduler. The FIFO scheduler runs worse in the first two testing suits of Table 7, but accidently has good performance in the last testing suit. This is because that any fair-based schedulers need to set an upper bound on the number of allocated resource slots to a user pool. The performance is affected when this upper bound is reached. This setting is for allowing more users to share system resources and preventing one user to occupy all resources. The FIFO scheduler however allocates all available resources to the current user and does not consider the future resource needs for other coming users to the system. It is expected that the FIFO scheduler would perform worse when there are many users and many large-size jobs in the system.

The average node-locality ratio is close for these three schedulers as shown in Table 8. This implies that data locality is not a dominant factor on the job turnaround time. Our modified fair scheduler performs more stable and efficient than other schedulers do due to multiple considerations.

Table 7. Comparison of average turnaround time for jobs of different sizes

Job size	Number of jobs	Original FS	Modified FS	FIFO	Gain (%)
6/10/16 /21/25	5	1242 s	364.6 s	513.6 s	-70.6%
	10	2343.1 s	1372.5 s	1532.2 s	-41.4%
	15	5585.3 s	2285.5 s	2131.933 s	-59.1%

Table 8. Comparison of average node-locality ratio for jobs of different sizes

Job size	Number of jobs	Original FS	Modified FS	FIFO	Gain (%)
6/10/16/ 21/25	5	0.25	0.26	0.27	+4%
	10	0.32	0.44	0.36	+3.8%
	15	0.32	0.44	0.36	+3.8%

5. Conclusions

Hadoop provides parallel data processing based on the MapReduce framework. The performance of MapReduce programs is dominated by a job scheduling algorithm. Fair scheduler, which is one of the built-in schedulers in Hadoop, provides good performance in general and is widely used. However, the scheduling policy of this naïve fair scheduler is almost fixed and is not flexible to the change of working conditions. In this paper, we propose five mechanisms: job classification, pool resource allocation, resource-aware job sorting,

delay time adjustment, and job priority adjustment to improve the fair scheduler by monitoring the actual resource usage and the runtime environment's condition. The conducted experiments on a real testing platform show that the modified fair scheduler can largely reduce the average turnaround time of a job and slightly increase data locality against the naïve fair scheduler.

In the future, more types of jobs such as CPU-intensive and I/O-intensive programs will be conducted in the experiment. Also, our current work only emphasizes the slot allocation and data locality for map tasks, more enhancements for reduce tasks can be furthered considered.

Acknowledgment

The authors would like to thank the anonymous referees for their helpful suggestions. This research was partially supported by Ministry of Science and Technology of the Republic of China under Contract No. MOST 104-2221-E-259-016.

References

[1] Y. W. Cheng and S. C. Lo, "Improving Fair Scheduling Performance on Hadoop," in Intl. Conf. on Platform Technology and Service, Feb. 2017.

[2] A. Labrinidis and H.V. Jagadish, "Challenges and opportunities with big data," in VLDB Endowment, **5**(12), 2032-2033, Aug. 2012.

[3] M. Armbrust, A. Fox, R. Griffith, and A. D. Joseph et al., "A view of cloud computing," Communications of the ACM, **53**(4), 50-58, Apr. 2010.

[4] The NIST Definition of Cloud Computing, Sept. 2011, http://dx.doi.org/10.6028/NIST.SP.800-145

[5] J. Dean and S. Ghemawat, "MapReduce: simplified data processing on large clusters," Communications of the ACM, **51**(1), 107-113, Jan. 2008.

[6] S. Liu, J. Xu, Z. Liu, and X. Liu, "Evaluating task scheduling in Hadoop-based cloud systems," in IEEE Intl. Conf. on Big Data, 47-53, Oct. 2013.

[7] A. Rasooli and D. G. Down, "A hybrid scheduling approach for scalable heterogeneous Hadoop systems," in SC Companion: High Performance Computing, Networking Storage and Analysis, 1284-1291, Nov. 2012.

[8] Y. Chen, A. Ganapathi, R. Griffith, and R. Katz, "The case for evaluating MapReduce performance using workload suites," in IEEE Intl. Symposium on Modeling, Analysis & Simulation of Computer and Telecommunication Systems (MASCOTS) , 390-399, Jul. 2011.

[9] L. Li, Z. Tang, R. Li, and L. Yang "New improvement of the Hadoop relevant data locality scheduling algorithm based on LATE," in Intl. Conf. on Mechatronic Science, Electric Engineering and Computer (MEC), 1419-1422, Aug. 2011.

[10] K. Yamazaki, R. Kawashima, S. Saito, and H. Matsuo, "Implementation and evaluation of the JobTracker initiative task scheduling on Hadoop," in Intl. on Symposium Computing and Networking (CANDAR), 622-626, Dec. 2013.

[11] S. Ghemawat, H. Gobioff, and S. T. Leung, "The Google file system," in ACM Symposium on Operating Systems Principles (SOSP), 29-43, Oct. 2003.

[12] V. K. Vavilapalli et al., "Apache Hadoop YARN: yet another resource negotiator," in Annual Symposium on Cloud Computing, Oct. 2013.

[13] J. Xie, S. Yin, X. Ruan, and Z. Ding et al., "Improving MapReduce performance through data placement in heterogeneous Hadoop clusters," in IEEE Intl. Symposium on Parallel & Distributed Processing, Workshops and Phd Forum (IPDPSW), 1-9, Apr. 2010.

[14] D. A. Heger, "Optimized resource allocation & task scheduling challenges in cloud computing environments," dheger@dhtusa.com, 2010.

[15] B. T. Rao and L. S. S. Reddy, "Survey on improved scheduling in Hadoop MapReduce in cloud environments," Intl. Journal of Computer Applications, **34**(9), 29-33, Nov. 2011.

[16] N. Tiwari, S. Sarkar, U. Bellur, and M. Indrawan, "Classification framework of MapReduce scheduling algorithms," ACM Computing Surveys, **47**(3), Apr. 2015.

[17] M. Zaharia, A. Konwinski, A. D. Joseph, R. Katz, and I. Stoica, "Improving MapReduce performance in heterogeneous environment, " in USENIX Conf. on Operating Systems Design and Implementation, 29-42, Aug. 2008.

[18] M. Zaharia, D. Borthakur, J.S. Sarma, and K. Elmeleegy et al., "Delay scheduling: a simple technique for achieving locality and fairness in cluster scheduling," in Proc. European Conf. on Computer Systems, 265-278, Apr. 2010.

[19] M. Zaharia, D. Borthakur, J. S. Sarma, and K. Elmeleegy et al., "Job scheduling for multi-user map reduce clusters," ECS Department, University of California, Berkeley, Apr. 2009.

[20] H. Zhao, S. Yang, Z. Chen, H. Yin, and S. Jin, "A locality-aware scheduling based on a novel scheduling model to improve system throughput of MapReduce cluster, " in Proc. Intl. Conf. on Computer Science and Network Technology (ICCSNT), 111-115, Dec. 2012.

[21] T. Y. Chen, H. W. Wei, M. F. Wei, and Y. J. Chen et al., "LaSA: A locality-aware scheduling algorithm for Hadoop-MapReduce resource assignment," in Proc. Intl. Conf. on Collaboration Technologies and Systems (CTS), 342-346, May 2013.

[22] P. Nguyen, T. Simon, M. Halem, D. Chapman, and Q. Le, "A hybrid scheduling algorithm for data intensive workloads in a MapReduce environment," in IEEE Intl. Conf. on Utility and Cloud Computing (UCC), 161-167, Nov 2012.

[23] M. Pastorelli, A. Barbuzzi, D. Carra, M. Dell'Amico, and P. Michiardi, "HFSP: size-based scheduling for Hadoop," in IEEE Intl. Conf. on Big Data, 51-59, Oct. 2013.

[24] S. Tang, B. S. Lee, and B. He, "Dynamic slot allocation technique for MapReduce clusters," in IEEE Intl. Conf. on Cluster Computing (CLUSTER), 1-8, Sept. 2013.

[25] S. Tang, B. S. Lee, and B. He, "DynamicMR: A dynamic slot allocation optimization framework for MapReduce clusters," IEEE Trans. on Cloud Computing, **2**(3), 333-347, Jul.-Sept. 2014.

[26] S. Kurazumi, T. Tsumura, S. Saito, and H. Matsuo, "Dynamic processing slots scheduling for I/O intensive jobs of Hadoop MapReduce," in Intl. Conf. on Networking and Computing (ICNC), 288-292, Dec. 2012.

[27] Y. Kwon, M. Balazinska, B. Howe, and J. Rolia, "Skew-resistant parallel processing of feature-extracting scientific user-defined functions," in ACM Symposium on Cloud computing (SoCC), Jun. 2010.

[28] B. S. Kapil and S. S. Kamath, "Resource aware scheduling in Hadoop for heterogeneous workloads based on load estimation," in Intl. Conf. on Computing, Communications and Networking Technologies (ICCCNT), 1-5, Jul. 2013.

[29] Y. Le, J. Liu, F. Ergun, and D. Wang, "Online load balancing for MapReduce with skewed data input," in IEEE INFOCOM Conf., Apr. 2014.

[30] X. Xu, L. Cao, and X. Wang, "Adaptive task scheduling strategy based on dynamic workload adjustment for heterogeneous Hadoop cluster," IEEE Systems Journal, **10**(2), 471-482, Jun. 2016.

[31] S. Y. Hsieh, C. T. Chen, C. H. Chen, H. C. Hsiao, and R. Buyya "Novel scheduling algorithms for efficient deployment of MapReduce applications in heterogeneous computing environments," IEEE Trans. on Cloud Computing, http://doi.ieeecomputersociety.org/10.1109/TCC.2016.2552518

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1059-1063 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Ranking of Two Multi Criteria Decision Making Cases with Evidential Reasoning under Uncertainty

Farzaneh Ahmadzadeh*

Post-Doctoral, Department of Innovation, Design and Engineering, University of Mälardalen, 631 05, Eskilstuna, Sweden

ARTICLE INFO

Article history:
Received: 21 April, 2017
Accepted: 15 June, 2017
Online: 15 July, 2017

Keywords:
Evidential reasoning
Uncertainty
Decision making
Prioritization

ABSTRACT

Many decision problems have more than one objective that need to be dealt with simultaneously. Moreover, because of the qualitative nature of the most of real world problem it is an inevitable activity and very important to interpret and present the uncertain information for making effective decision. The Evidential Reasoning (ER) approach which is one of the latest development within multi criteria decision making (MCDM) seems to be the best fit to synthesize both qualitative and quantitative data under uncertainty. To support this claim, two case studies were tested to illustrate the application of ER for prioritization and ranking of decision alternative to support decision process even with uncertain information. The overall goal of the first case study is to identify and prioritize factors that can be considered maintenance-related waste within the automotive manufacturing industry. The result after applying ER shows "inadequate resources" and "weather /indoor climate," respectively, are the highest and lowest average scores for creating maintenance-related waste. This prioritization methodology can be used as a tool to create awareness for managers seeking to reduce or eliminate maintenance-related waste. The aim of the second case study is to look at the possibility of having a new approach for sustainable design. So through a literature review six design strategies were taken into consideration in order to develop a new approach based on all advantages (sustainable factors) of the six approaches. For ranking and finding out about the most important factors the evidential reasoning (ER) approach is used. Based on ER all the important factors, apart from the one collected from interviews are a part of eco-design. So it means among all strategies eco-design is the most dominant strategy in term of environment. However two of the important factors are not found in any strategy but in interviews. These factors can be used as the building blocks for a new approach. The importance of having a better structured decision process is essential for the success of any organization, so it can be applied widely in most of real world problem dealing with making effective decision.

1. Introduction

It has become more and more difficult to see the world around us in a uni-dimensional way and to use only a single criterion when judging what we see [1]. The decision making process for any organization may be key factor for its success. Decision maker's wishes to evaluate the performance of the alternative with different criteria simultaneously. In many situations these objectives/ criteria may be conflicting. These objectives are associated with the possible consequences (or outcomes) that results from choosing an alternative [2]. The branch of decision analysis which deals with this kind of problem is called multi-criteria decision making (MCDM). Many MCDM methods have been developed, such as multiple attribute utility theory (MAUT) and analytical hierarchy process (AHP) [3, 4]. Most of these methods are suitable for solving small scale MCDM problems without uncertainty. In uncertain situations, the Fuzzy Multi-Criteria Decision Making (FMCDM) approach provides an ideal option; it has been tested by a number of researchers to rank alternatives in different situations [5]. However, the fuzzy approach is used only when uncertainty is predominant. In other words, when a particular parameter is quantifiable with fair degree of accuracy, or there are a missing or incomplete data this approach need not be used. Most real-life decisions use a mixture of qualitative and quantitative attributes with varying degrees of uncertainties, increasing the need for the development of scientific methods and tools that are rational, reliable, repeatable, and transparent. Since, it is essential to properly represent and use uncertain information for making effective decision, it is

* Corresponding Author: Farzaneh Ahmadzadeh, Mälardalens högskola, 631 05 Eskilstuna | Email: Farzaneh.ahmadzadeh@mdh.se

compulsory to use the multi-level evaluation framework for assessing different type of uncertainty inherent in data like missing data, incomplete data which is one of the many research limitation when it comes to qualitative data. Therefore in this paper an evidential reasoning (ER) approach has been introduced to address this problem. Two case studies is examined to emphasize the effectiveness of this approach. The rest of the paper is organized as follows. Section 2 briefly outlines the Evidential Reasoning (ER) approach. Section 3 explains the first case study for prioritization of maintenance related waste. Section 4 provides the second case study when it has been applied for developing a sustainable product design, while Section 5 offers a conclusion.

2. Evidential Reasoning Approach

The Evidential Reasoning (ER) advocates a general, multi-level evaluation process for dealing with MCDM problems. The process can model various types of qualitative and quantitative uncertainties and is developed on the basis of Dempster-Shafer evidence theory [6] and evaluation analysis model and decision theory. In ER, A complex general property which is usually difficult to assess directly is broken down and operationalized by using well-defined, measurable concepts that together constitute the general property. The result of such a breakdown is a multiple attribute framework taking the shape of a tree (hierarchy) structure, with assessable basic attributes at the lowest level. The assessment of these basic attributes can be aggregated to an assessment of the upper level of the tree. The Dempster-Shafer mathematics are designed to aggregate the uncertainties in the basic attributes to a total uncertainty of the total assessment. Steps for the overall assessment of the complex general property are suggested in [6, 7] and summarized in [8] are as following:

2.1. Definition and representation of a multiple attribute decision problem

Define a set of L basic attributes include all the factors influencing the assessment of the upper level attribute as follows:

$E = \{\varepsilon 1, \varepsilon 2, \dots, \varepsilon L\}$

Now estimate the relative weights of the attributes where ω_i is the relative weight for basic attribute ε_i and is normalized so that $\sum \omega_{i=1}$ and $0 \leq \omega_i \leq 1$. Moreover define N distinctive evaluation grades H_n, $n=1,\dots,N$ as a complete set of standards to assess each option on all attributes.

For example:

$H=\{H_1=worst, H_2=poor, .., H_{N-1} = Good, H_N=Excellent\}$

For each attribute ε_i and evaluation grade H_n a degree of belief β_n is assigned. The degree of belief denotes the source's level of confidence when assessing the level of fulfillment of a certain property.

2.2. Basic probability assignments for each basic attribute

Let $m_{n,i}$ be a basic probability mass, representing the degree to which the ith basic attribute ε_i supports a hypothesis that the general attribute is assessed to the nth evaluation grade H_n. Then, $m_{n,i}$ is calculated as follows:

$$m_{n,i} = \omega_i \beta_{n,1} \quad (1)$$

Let $m_{H,i}$ be the remaining probability mass unassigned to each basic attribute $\varepsilon_{i,}$ so $m_{H,i}$ is calculated as follows :

$$m_{H,i} = 1 - \sum_{n=1}^{N} m_{n,i} = 1 - \omega i \sum_{n=1}^{N} \beta_{n,i} \quad (2)$$

Decompose $m_{H,i}$ into $\bar{m}_{H,i}$ and $\tilde{m}_{H,i}$ as follows:

$$\bar{m}_{H,i} = 1-\omega_i \text{ and } \tilde{m}_{H,i} = \omega_i (1 - \sum_{n=1}^{N} \beta_{n,i}) \quad (3)$$

$$m_{H,i} = \bar{m}_{H,i} + \tilde{m}_{H,i} \quad (4)$$

2.3. Combined probability assignments for a general attribute

The assessments of the basic attributes constituting the general property are aggregated to form a single assessment of the general property. The probability masses assigned to the various assessment grades, as well as the probability mass left unassigned, are denoted by $m_{n,I(L)}$, $\bar{m}_{H,I(L)}$, $\tilde{m}_{H,I(L)}$ and $m_{H,I(L)}$. Let $I(1)=1$. This gives us $m_{n,I(1)} = m_{n,1} (n=1,\dots,N)$, $\bar{m}_{H,I(1)} = \bar{m}_{H,1}$, $\tilde{m}_{H,I(1)} = \tilde{m}_{H,1}$ and $m_{H,I(1)} = m_{H,1}$. The combined probability masses can be generated by aggregating all the basic probability assignments using the following recursive ER algorithms:

$\{H_n\}$:

$$m_{n,I(i+1)} = K_{I(i+1)}[m_{n,I(i)} \times m_{n,i+1} + m_{H,I(i)} \times m_{n,i+1} + m_{n,I(i)} \times m_{H,i+1}]$$
$$n = \{1,2,\dots,N\} \quad (5)$$

In equation (5), we continue to let $i=1$. The term $m_{n,1}$, $m_{n,2}$ measures the degree of attributes ε_1 and ε_2 supporting the general attribute y to be assessed to Hn, the term $m_{n,1}$, $m_{H,2}$ measures the degree of only ε_1 supporting y to be assessed to H_n, and the term $m_{H,1}$, $m_{n,2}$ measures the degree of only ε_2 supporting y to be assessed to H_n.

$\{H\}$:

$$m_{H,I(i)} = \bar{m}_{H,I(i)} + \tilde{m}_{H,I(i)} \quad (6)$$

$$\tilde{m}_{H,I(i+1)} = K_{I(i+1)}[\tilde{m}_{H,I(i)} \times \tilde{m}_{H,i+1} + \bar{m}_{H,I(i)} \times \tilde{m}_{H,i+1} + \tilde{m}_{H,I(i)} \times \bar{m}_{H,i+1}] \quad (7)$$

$$\bar{m}_{H,I(i+1)} = K_{I(i+1)}[\bar{m}_{H,I(i)} \times \bar{m}_{H,i+1}] \quad (8)$$

$$K_{I(i+1)} = \left[1 - \sum_{t=1}^{N} \sum_{\substack{j=1 \\ j \neq t}}^{N} m_{t,I(i)} \cdot m_{j,i+1}\right]^{-1} \quad i = \{1,2,\dots,L-1\} \quad (9)$$

In equation (7), the term $\tilde{m}_{H,1}$, $\tilde{m}_{H,2}$ measures the degree to which y cannot be assessed to any individual grades due to the incomplete assessments for both ε_1 and ε_2. The term $\bar{m}_{H,1}$, $\tilde{m}_{H,2}$ measures the degree to which y cannot be assessed due to incomplete assessments for ε_2 only. The term $\tilde{m}_{H,1}$, $\bar{m}_{H,2}$ measures the degree to which y cannot be assessed due to incomplete assessments for ε_1 only. The term $\bar{m}_{H,1}$, $\bar{m}_{H,2}$ in equation (8) measures the degree to which y has not yet been assessed to individual grades due to the relative importance of ε_1 and ε_2 after ε_1 and ε_2 have been aggregated. $K_{I(2)}$ as calculated by equation (9) is used to normalize $m_{n,I(2)}$ and $m_{H,I(2)}$ so that :

$$\sum_{n=1}^{N} m_{n,I(2)} + m_{H,I(2)} = 1 \quad (10)$$

2.4. Calculation of the combined degrees of belief for a general property

Calculating the combined degrees of belief for a higher level property. Let β_n denote the combined degree of belief that the higher level property assessed to the grade H_n, generated by combining the assessments for all the associated basic attributes ε_i. β_n is then calculated by:

$$\{H_n\}: \beta n = \frac{m_{n,I(L)}}{1-\bar{m}_{H,I(L)}} \quad n = \{1,2,\dots,N\} \quad (11)$$

$$\{H\}: \beta n = \frac{\tilde{m}_{H,I(L)}}{1-\bar{m}_{H,I(L)}} \quad (12)$$

Steps 1-4 can now be employed for the other sub-trees, to obtain combined degree of belief for the higher level of the hierarchy model.

2.5. Using linear utility function

In this step, the utilities of the respective assessment grades $H_{1...n}$ are estimated via utility functions ($u(H_n)$). This estimation can be accomplished for instance by means of a range of methods and techniques that can be utilized for this purpose. In this paper however we will not dwell on the subject of utility estimations, rather we assume that the utilities of the respective assessments grade can be appreciated in a linear fashion. Therefore top level score of the hierarchy model can be obtained by $\sum \beta n\, u(H_n)$, n=1…N.

3. First Case Study: Prioritization Of Maintenance-Related Waste

The reduction and elimination of maintenance-related waste is receiving increasing attention because of the negative effect of such waste on production costs. The overall goal of this research is to identify and prioritize factors that can be considered maintenance-related waste within the automotive manufacturing industry [9].

3.1. Identification of Waste

To identify maintenance-related waste in the manufacturing industry, we held six workshops at five manufacturing companies. Brain writing and brainstorming were the main data collection tools. In total 465 maintenance-related wastes were discussed during the workshops. The classification into categories was performed by three researchers and through discussions, 16 final categories were decided upon. It was visible from the workshop analysis that the origin and cause of the maintenance-related waste could be linked to human factors. Therefore, in order for classification and model provision of maintenance-related waste linked to human activities, different literature in the area of human errors in maintenance field have been studied. the most efficient and relevant classification was related to a study about maintainer error by the Naval Safety Center's Human Factors Analysis and Classification System-Maintenance Extension (HFACS-ME) which was adapted for maintenance mishaps in aviation [10]. So, HFACS-ME is accepted as the basic framework and the 16 categories are incorporated into this model based on their similarity. The mentioned model is revised when no suitable category were found.

3.2. Constructing Survey

A survey was developed based on the identified maintenance-related wastes on the lowest level of the hierarchy model. It contains 28 questions; because of having no informative knowledge about different type of the waste it is assumed that all the waste attributes have equal relative weight (importance). Five distinctive evaluation grades are used to assess each question: H= {Very low, Low, Average, High, Very high}. The respondents were asked to assess each waste by assigning their belief degree to these five grades. A belief degree represents the strength to which the grade is believed to be appropriate for describing the opinion on the criterion. For example subjective judgement of an expert for the first question about "how much "inadequate process" are responsible for waste was: (Very high=0%, High= 10%, Average=20%, Low= "no idea", Very low=40%).

3.3. Data Analysis and Discussion

The main purpose in prioritization the human factors responsible for maintenance-related waste was to identify strengths and weaknesses which could form a basis for subsequent detailed assessments and help create action plans to address the weaknesses. This means management teams can focus on different factors to reduce or eliminate waste based on their importance for creating waste. A Windows-based Intelligent Decision System (IDS) is applied to implement the ER approach. IDS is a general-purpose multiple criteria decision analysis tool; it provides graphical interfaces to build a decision. The group belief degrees entered for each evaluation grades and for 28 questions (which were designed based on the lowest level of MWC-HF model) into IDS. As result of IDS for rankings of maintenance-related waste at the lowest level shows, "inadequate resources" and "weather /indoor climate," with average scores of 54% and 22% respectively, are the highest and lowest average scores for creating maintenance-related waste; see Table 1. This prioritization methodology can be used as a tool to create awareness for managers seeking to reduce or eliminate maintenance-related waste.

Table 1. Ranking of the maintenance related waste created by human factors

Maintenance related waste based on human factors	Score (%)	Rank
Inadequate Resources	54	1
Inadequate Supervision	52	2
Mental State	50	3
Poor EEM (Early Equipment Management)	48	4
Inadequate Process	47	5
Inadequate Documentation	46	6
Poor Spare Part Handling	45	7
Adaptability/ Flexibility	43	8
Inadequate Design	42	9
Inappropriate Operation	42	10
Judgment / Decision Making	40	11
Assertiveness	38	12
Communication	37	13
Training Preparation	37	14
Physical State	35	15
Unavailable/ Inappropriate	35	16
Inadequate Customer Demand	31	17
Certification Qualification	30	18
Lack of Employee Engagement	30	19
Inaccessible	29	20
Supervisory Misconduct	29	21
Limitation	28	22
Infringement	27	23
Uncorrected Problem	27	24
Environmental Hazards	26	25
Confining	24	26
Error and Violation	23	27
Weather /Indoor Climate	22	28

4. Second Case Study: Developing Sustainable Product Development Strategy

It has become increasingly important for producing companies to reduce their environmental impact. Companies are focusing more on preventing environmental issues by taking sustainability into the product development process, and not just reducing emissions from manufacturing the product [11].

Product development needs to be done with considering sustainability and without compromising future generation's ability to satisfy their needs. There are several strategies and methods developed to guide companies towards sustainability. The aim of this case study is to look at the possibility of having a new approach for sustainable design. So through a literature review six design strategies were taken into consideration in order to develop a new approach based on all advantages (sustainable factors) of the six approaches. Those six strategies are: eco-design, green design, cradle-to-cradle, and design for environment, zero waste and life cycle approaches. Together with literature review an interviews were conducted with managers from companies working with product development in Sweden to identify as many sustainable factors as possible. For ranking and finding out about the most important factors the evidential reasoning (ER) approach is used. The reason for application of ER is the qualitative nature of the data (factors) which add more uncertainty. Based on the literature several advantages and disadvantages are defined, both in regard of the environment and in a business perspective [12].

4.1. Result of Literature Review and Interview

Results shows, Eco design is a tool with most advantages, and green design has most disadvantages. By looking at the advantages, patterns emerge in the different approaches. By grouping the 38 advantages below similar advantages are merged. The disadvantages that were found are fewer than the advantages, most likely because the research focus on the benefits of the strategies. Several of the advantages can be seen as factors of sustainable design and by defining them there is a possibility of finding which factors are important to a new approach to sustainable design. The factors that were found is presented, in Table 2 with the design strategies related to each factor. To support the literature review and find other factors than the ones conducted from the literature review, three semi structured interviews were conducted with managers from companies working with product development in Sweden. Factors that were drawn from the interviews are: material selection, reduce energy usage, reduce emissions, minimize use of toxic substances, increased competitiveness and economic benefits. Some of these factors correspond directly to factors drawn from the literature, but two factors are added: "material selection" and "reduce emissions".

Table 2 - Factors of sustainable design and the corresponding strategies

Factors	Design strategy
Reduce energy usage	Eco-design
Reduce material usage	Eco-design, Life-cycle approaches
Reduce use of non-renewable resources	Green design
Reduce waste	Design for Environment
Eliminate waste	Cradle-to-cradle, Zero waste
Eliminate emission	Zero waste
Minimize use of toxic substances	Eco-design, Zero waste
Minimize waste	Green design
Recycle materials/component	Cradle-to-cradle, Design for environment, Zero waste, Life-cycle approaches, Eco-design
Reuse material/components	Zero waste, Life-cycle approaches, Eco-design, Cradle-to-cradle
Increase product functionality	Eco-design
Increase product lifespan	Eco-design
Increase use of renewable energy	Green design, Cradle-to-cradle
Increase use of renewable materials	Green design, Life-cycle approaches, Cradle-to-cradle
Increase use of biodegradable materials	Cradle-to-cradle
Closed loop material flow	Cradle-to-cradle
Holistic Approach	Life-cycle approaches, Cradle-to-cradle
Social standards	Green design, Cradle-to-cradle
Economic benefits	Eco-design, Cradle-to-cradle, Zero waste
Increased competitiveness	Eco-design

4.2. *Constructing Survey*

Based on the 20 factors collected from the literature review and additional 2 factors collected from interviews a survey was designed. The survey was sent together with instructions to people working with product development. The respondents were asked to answer the importance of each factors in sustainable product development based on five grades of H= {un-important, Not very important, Quite important, Important, Very important}. They were given the opportunity to answer the questions by assigning their degree of belief, from 0 to 100%, in different grades and for different answers. If they weren't sure of the importance of a factor, they could give the answer "don't know". The surveys were answered by 10 respondents with an average of 8 years of experience in product development.

4.3. Data Analysis and Discussion

The mean value for each grade and factor based on the results from the survey was calculated by adding up the respondents' degree of belief in each grade and entered into the IDS. The factors of sustainability are not arranged by hierarchy, it is assumed that all factors are top-level criteria.

The result of applying ER through IDS shows that all factors are important but the most important factors, with a percentage score of over 65%, which is the mean value of all factors, are: "Minimize use of toxics substances" (82%), "Increased competitiveness" (76%), "Economic benefits" (75%), "Reduce material usage" (74%), "Material selection" (72%), "Reduce emissions" (69%), "Increase product functionality" (69%), see Table 3.

By looking at the factors from Table 2 it is clear that most of the important factors are part of the eco-design strategy. Material selection" and "reducing emission" are factors that were obtained from interviews with companies. In other words all the important

factors, apart from the one collected from interviews are a part of eco-design. So it means among all strategies eco-design is the most dominant strategy in term of environment.

Table 3 – Important design factors and relevant score

Factors	Score (%)	Rank
Minimize use of toxic substances	82	1
Increased competitiveness	76	2
Economic benefits	75	3
Reduce material usage	74	4
Material selection	72	5
Reduce emissions	69	6
Increase product functionality	69	7
Reduce waste	64	8
Increase use of renewable energy	64	9
Social standards	64	10
Increase use of renewable materials	63	11
Holistic view	62	12
Recycling components/materials	61	13
Reduce use of non-renewable resources	60	14
Minimize waste	59	15
Reusing components/materials	58	16
Increase use of biodegradable materials	58	17
Increase product lifespan	57	18
Eliminate emissions	56	19
Reduce energy usage	55	20
Circular material flow	54	21
Eliminate waste	53	22

5. Conclusion

Many of the real life problems need making decision under uncertainty that is, choosing action among a set of actions considering different criteria based on often imperfect observations, with unknown outcomes. The Evidential Reasoning (ER) is one of the latest developments within MCDM literature and appears to be the best fit to handle uncertain information. ER can model multiple attribute decision problems which have both quantitative and qualitative attributes. In this paper ER is introduced and it is applied in two different case studies for prioritization and ranking of different factors. In the first case study it is applied to rank different maintenance related waste linked to human factors. The result showed, among all 28 factors identified in the workshop studies, “Inadequate Resources”, “Inadequate Supervision”, “Mental State of the workers” are the most important factors for creating waste by human in maintenance context at considered automotive manufacturing industry. Second case study look at the possibility of having a new approach for sustainable design. So through a literature review six design strategies were taken into consideration in order to develop a new approach based on all advantages (sustainable factors) of the six approaches. For ranking and finding out about the most important factors the evidential reasoning (ER) approach is used. After applying ER for the second case study the result showed among the sex sustainable design strategies most of the important factors were found in the eco-design strategy, however that strategy also contains factors that are not as important, and two of the important factors are not found in any strategy but in interviews. These factors represent the building blocks for a new approach. As a future research extension modelling of other type of uncertainty, such as interval uncertainties, uncertainties in other parameters of a decision problem such as criterion weights and belief degrees is recommended.

Acknowledgment

The research work is part of the initiative for Excellence in Production Research (XPRES) which is a cooperation between Mälardalen University, the Royal Institute of Technology, and Swerea. XPRES is one of two governmentally funded Swedish strategic initiatives for research excellence within Production Engineering.

References

[1] M. Zeleny, *Multiple Criteria Decision Making.* California: McGraw-Hill, 1982.
[2] L. H. Alencar, A. T. Almeida, D. C. A. Morais, “ Multicriteria group decision model aggregating the preferences of decision-makers based on ELECTRE methods, Pesqui Oper, 3**0**, 687-702, 2010.
[3] R. R Lekurwale, M. M Akarte, D. N. Raut, “Framework to evaluate manufacturing capability using analytical hierarchy process”. Int J Adv Manuf Tech, **76**(1), 565-576, 2015.
[4] T.L. Saaty, *The Analytic Hierarchy Process*. University of Pittsburgh. 1988.
[5] C.T. Chen, “A fuzzy approach to select the location of the distribution center”. Fuzzy Set Syst, **118**, 65-73, 2001.
[6] G.A Shafer,. *Mathematical Theory of Evidence*. Princeton University Press. 1976.
[7] J.B. Yang , B.G. Dale., C.H.R. Siow, “Self-assessment of excellence: an application of the evidential reasoning approach”. Int J Prod Res, **39**(16), 3789–3812, 2001.
[8] J. Pontus, L. Nordström, R. Lagerström, “Formalizing analysis of enterprise architecture”. Lect Notes Bus Inf P, 35-44, 2007.
[9] F. Ahmadzadeh, M. Bengtsson, “Using evidential reasoning approach for prioritization of maintenance-related waste caused by human factors—a case study”, Int J Adv Manuf Tech,, https://doi.org/10.1007/s00170-016-9377-7, 1–15, 2016.
[10] J. Schmidt, D. Lawson, R. Figlock. Human factors analysis and classification system maintenance extension (HFACS-Me), review of select NTSB maintenance mishaps, an update, 2003. http://www.faa.gov/about/initiatives/intenance_hf/library/documents/ media/hfacs/ntsb_hfacs-me_updated_study_report.pdf
[11] G. Johansson, “Succes factors for integration of eco design in product development”. Environmental Management and Health, **13**(1), 98-107, 2002.
[12] F. Ahmadzadeh, K Jederström M. Plahna, A. Olssona, I. Foyera “An investigation of the most important factors for sustainable product development using evidential reasoning”, Numerical Algebra, Control and Optimization., In press.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1064-1069 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Applied Horizontal and Vertical Geothermal Heat Exchanger with Heat Pump System to Provide Air Conditioning for an Academic Facility in Mexico

Daniel Alcantar Martínez, Oskar Javier González Pedraza*, Crisanto Mendoza Cavarrubias, J. Jesus Pacheco Ibarra, Jorge Alberto Rangel Arista

Faculty of Mechanical Engineering, Universidad Michoacana de San Nicolás de Hidalgo, UMSNH, Morelia, Michoacán, México.

ARTICLE INFO

Article history:
Received: 28 April, 2017
Accepted: 11 June, 2017
Online: 15 July, 2017

Keywords:
heat pumps
geothermal energy
geothermal heat exchanger

ABSTRACT

At present in Mexico, the renewable energy has become more important due to the great dependence of the country for fossil fuels. Within the several applications of renewable energy, there are the geothermal applications for the air conditioning of spaces. This technology employs heat pumps that interexchange heat with the ground. This technology is relatively young in Mexico, leaving a large field for study and application throughout the country. In this way, to calculate the correct sizing of geothermal heat exchangers, it is necessary to calculate the thermal loads of the complex in which this technology of geothermal heat pumps using vertical heat exchangers type U will be installed, to perform the calculation of thermal loads Autodesk Revit® software was used, with which was possible to make a virtual model in detail of the botanical center that is located in Morelia, Michoacán, Mexico and belongs to the Universidad Michoacana de San Nicolás de Hidalgo (UMNSH).

This study shows the results of the analysis of the installations and determination of the thermal loads of the complex due to this type of infrastructure. By obtaining the values of the thermal loads, the dimensioning of the heat exchanger was archived, which will have to be installed to cover the thermal requirement of this system and his installation, in addition to the selection of the heat pump. This complex of 2 levels, where, on the first floor there are cubicles and laboratories and on the second floor, several common areas.

The design was developed in detail in Autodesk Revit 2015. After obtaining the thermal loads, the GLHEPro software was used for dimensioning the Vertical heat exchangers with the number and depth of the exchangers was obtained. the GLD 2014 software was used for dimensioning the Horizontal heat exchangers with the number and depth of the exchangers was obtained.

1. Introduction

Climate change is, according to experts, one of the greatest challenges facing humanity. This means that all public and private actors in global society must work together to reduce greenhouse gas emissions and adapt them territorially and sectoral to the potential effects so that we build a prosperous, sustainable and peaceful future [1].

An alternative to collaborate with the environment and the application of sources of renewable origin:

Mexico presents in most of the territory medium and low enthalpy resources that are not exploited for the generation of electricity, and to which they are not used for direct uses. In Mexico, geothermal energy is solely employed to power generation. Direct use is still in a developing phase and currently it is restricted to supply bathing and swimming facilities. Geothermal heat pumps usage is minimal and underdeveloped without available data.[2] Geothermal energy is an alternative that can be used to produce electricity, store energy for heating and air conditioning devices in various facilities, and provide heat for a variety of commercial and industrial processes. Due to the constant temperature in the subsoil, a geothermal heat pump

*Corresponding Author: M.C. Oskar Javier González Pedraza, Santiago Tapia 403 Morelia, Michoacan México | Email: oniblueoskar@gmail.com

system typically has better performance than conventional systems. Geothermal heat pumps are used around the world to heating and cooling of buildings. Heat pump efficiency relies on the temperature of the fluid that provides the heat, in the side of ground, which is affected by the annual temperature profile and the pits distribution [3]. Low Temperature Geothermal Energy (LTGE) is the energy accumulated in the field generated from the heat exchange with the atmosphere at low temperature, it is an alternative energy source capable of satisfying the energy demand for domestic heating and cooling.

Horizontal Geothermal heat exchangers (HGHEs) with vertical springs have been increasingly used in Geothermal Sytem of Heat Pumps (GSHP). The current studies have been carried out by utilizing field trials and simulations focused in reveal the heat transfer mechanisms of HGHEs to procure an adequate performance of heat exchenge [4]. In [5] it is analized a solution of the interaction between a Photovoltaic generator (PV) and a GSHP to be used in a residential building. The main goal is to review the operative functioning of a self-sufficient system that enable the maximal self-consumption in the building itself: these systems can provoke a major penetration of Renewable energy sources in a complex energetic context.

Vertical geothermal heat exchangers are more common to be installed in areas where there is not enough space to place horizontal type heat exchangers. For the installation of the type of vertical exchanger configuration, drilling is performed in a very small space of land; The advantage of having a vertical type configuration is that it has a lower temperature fluctuation, which gives higher values of the operating coefficient.

Figure 1 shows schematically the operation of a heat pump with geothermal heat exchangers. In Mexico, the construction of an air conditioning system by means of heat pumps by means of geothermal heat exchangers has not been reported. Therefore, the present study studies the technical viability of the installation of this type of systems in national territory.

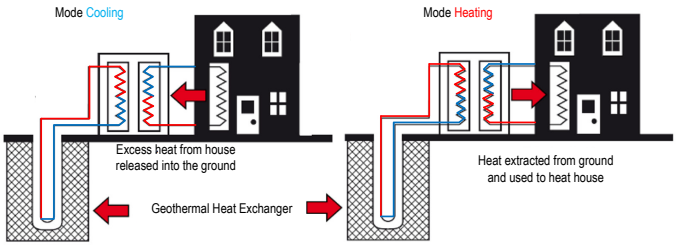


Figure 1. Operating of a conditioning system using a Heat pump with geothermal heat exchangers.

2. Case Study

Revit® software was used to calculate the thermal loads, which is contained within Autodesk® CAD software. This software allows to detail the whole and its spaces, for the calculation of the thermal loads of the complex. The case study refers to a building consisting of 2 levels, in the first level are the laboratories of sinecology, mycology, vascular and non-vascular plants, Palynology and sanitary. On the second level is a cafeteria, herbarium area, a silver reception area, quarantine area, multipurpose room and library. The calculation of thermal loads is essential for the selection of the different heat pumps to be installed in the system, as well as the dimensioning of the air ducts and the geothermal heat exchanger system. In **Figure 2,** show different views of the resort on its facade and plan views of the distribution of spaces.

The construction of the model on the Autodesk Revit® platform is specifically done using the Building Information Model (BIM) technique, and includes architectural design, structural engineering, MEP and construction. BIM is the process of generating and managing building data during its lifecycle, using dynamic real-time, three-dimensional building modeling software to reduce time and resources in design, construction and calculation. This process produces the building information model (also abbreviated BIM), which covers building geometry, spatial relationships, geographic information, as well as the quantities and properties of its components. The software has a basic of international climatic data, which allows, in the software, to locate the exact region of the building, as well as its orientation. When detailing each internal component of the model the software is able to determine the thermal loads through a year and with this determine the loads of summer and winter. The software is based on the methodology ASHRAE 2005[6].

Figure 2. 3D view of the outer facade of the complex.

In this platform, it is possible to calculate the thermal loads for both summer and winter, to locate in detail the location of the complex to increase the accuracy of the calculation of thermal loads due to solar incidences in walls, ceilings and windows, as well as infiltrations due to the Air currents of the area in question, in addition to allowing the dimensioning of the air distribution ducts. Table 1 summarizes the overall calculation of the maximum and peak values shown and flux densities of the entire complex, can be noted that the maximum cooling load is 65.6 kW and maximum load Heating is 1.28 kW. With this we can appreciate that the requirement of these educational facilities will be primarily in cooling mode.

Table 1. Herbarium Summary

Entries	
Types of construction	Educative Center
Area (m²)	756.16
Volumen (m³)	2,080.20
Calculation Results	
Maximum value of total cooling load (W)	**73,906**
Maximum cooling value (month and hour)	May 02:00 p. m.
Maximum value of sensitive cooling load (W)	65,602
Maximum latent cooling load value (W)	8,304
Maximum cooling capacity (W)	73,906
Maximum value of cooling airflow (L / s)	5,821.20
Maximum heating load value (W)	**1,286**
Maximum value of heating air flow (L / s)	5,572.30
Checking Sums	
Density of cooling load (W / m²)	97.74
Density of the cooling flow (L / (s · m²))	7.7
Cooling flow / load (L / (s · kW))	78.77
Cooling area / load (m² / kW)	10.23
Density of heating load (W / m²)	1.7
Density of heating flow (L / (s · m²))	7.37

3. Analysis and Results to the Virtual Model of the Facility

It was conducted a loads analysis to cooling and heating, in order to select the proper zones in the complex to install the Air conditioning System. Once review, the value of load in the two working modes comes from the herbarium, followed by the library and then all the laboratories at the bottom and finally the multipurpose room and the editing room, this can be seen in Figure 3. Figure 4 shows the thermal loads of heating and appreciated that spaces having larger load are the library, the multipurpose room and assembly, and Zone of reception of plants, this is because they are zones with great incidence solar by the windows and of conduction by ceiling and walls, which will be the zones of priority for the system of conditioning.

Figure 3. Distribution of thermal loads of the school of refrigeration loads (W)

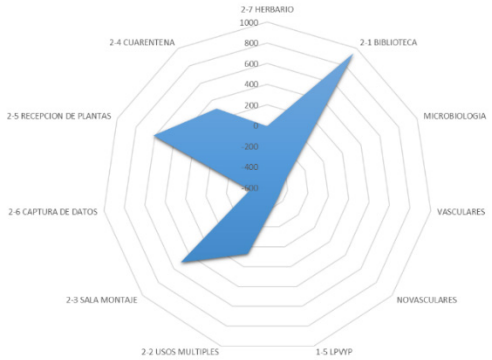


Figure 4. Distribution of heating loads of the school of heating loads (W)

In Figure 5, the monthly average thermal loads shown in cold mode, these loads are due to the transfer of heat generated by the occupants, electrical equipment and laboratory equipment, they are also considered heat gains due to windows and the incidences of the sun, and the transfer of heat through the walls of the enclosure, ceilings and floors. As you can notice the months in which it will require more use of the cold mode of the heat pump will be from the month of May until October of each year.

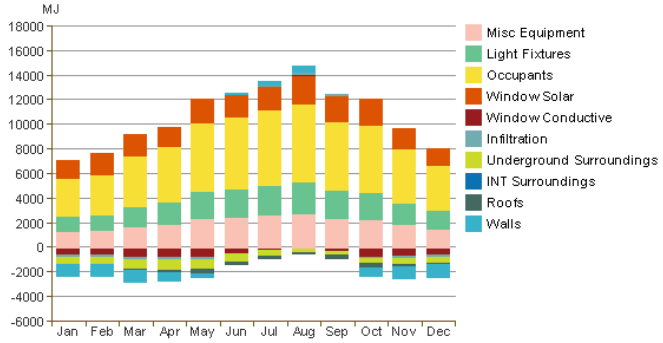


Figure 5. Average monthly heat load in cold mode

In Figure 6, the monthly average thermal loads shown in heat mode, these loads are due to the transfer of heat emitted by the occupants, electrical equipment and laboratory equipment, they are also considered heat gains due to windows and the incidences of the sun, and the transfer of heat through the walls of the enclosure, ceilings and floors. It can be noted in the figure that the months in which the system will be used in heat mode will be the months from November to March of the year.

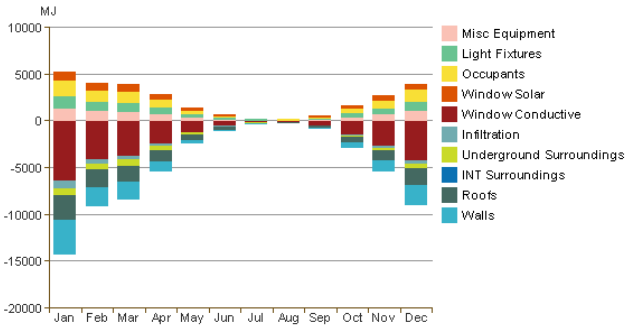


Figure 6. Monthly average heat load in heat mode

In Figure 7, the consumption of electricity over a year shows, you can note that the maximum consumption occurs in the months of December to March, in these months do not have daylight saving and lighting Solar is for few hours so the electricity consumption is higher while the other months have a lower consumption due to the little use of lighting.

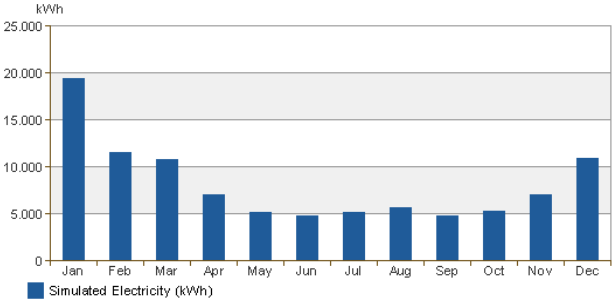


Figure 7. Monthly consumption of electrical energy of the complex

4. Results and Discussion

After obtaining the value of the thermal load under which the whole complex will be operating, the dimensioning of the heat exchangers was started, for which two commercial software's were used. In the first instance, we will perform the dimensioning of the vertical heat exchangers and then the horizontal ones, it is worth mentioning that, between the two systems, the thermal load of the complex will be divided.

In order to calculate the length of the heat exchanger, we will use specialized software tools. In this case, the GLHEPro® numerical tool was used, which specifies the type of heat exchanger to be installed, physical characteristics of the vertical heat exchanger, Specifications of the floor type, in addition to the type of coolant to be used in the heat exchangers which in this case would be water. After all the values required for the simulation of the well are loaded, the thermal loads are introduced to the system which were calculated by the Revit® software.

The U-type heat exchanger, proposed for the storage and recovery of geothermal energy from the well and subject of the present analysis, converting the soil into a thermal sink, the pipes were considered to be high density polyethylene (PAD), the grout is of a mixture of concrete, silica sand and water, the material around the terrain, are those found in the field in which these exchangers are being installed, the type of soil in the area will provide the values of the thermal properties. The average thermal conductivity of the soil obtained by the rheological drilling column is 0.8657 W / (m ° K) and the thermal resistance of the exchanger is 0.3048 ° K / (W / m), the volumetric flow of the system is recommended by IGSHPA [7] of 1.9962 L / s, the thermal conductivity of water 0593 W / (m ° K) and heat capacity of 4173.36 kJ / (° Km $^{3)}$. In Figure 8 the scheme proposed heat exchanger is shown.

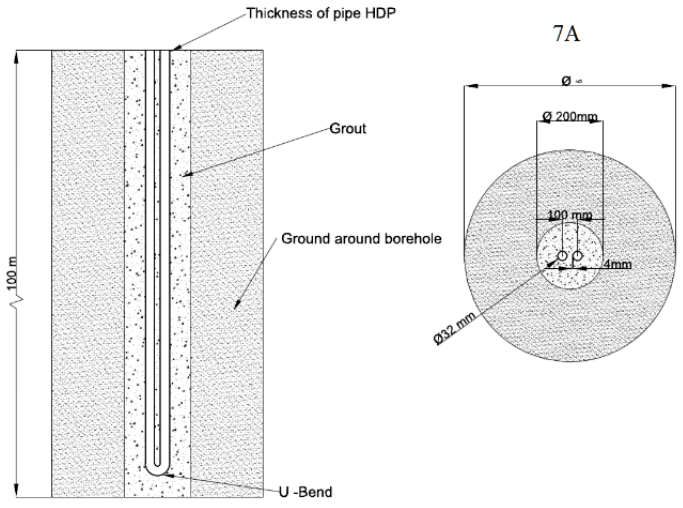
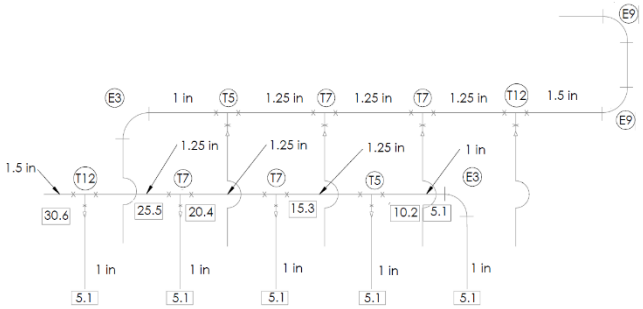


Figure 8. Schematic of the geothermal heat exchanger.

Once the calculation was carried out in the software, the results obtained are from the need of 4 vertical exchangers with a depth of 88.29 m and a total drilling depth of 353.16 m, and a distance between interchangers of 4.57 m. The maximum temperature value will be 25.51 ° C and with a minimum of 6.23 ° C and with temperature peaks ranging from 31.53 ° C and -6.67 ° C. With these temperatures. It can be concluded that it may be necessary to add some antifreeze to the system for its operation and with that the system works properly. The software has a tool to perform the optimize of the appropriate size, using this tool the software tells us that the length of the exchangers will be 81.15 m with a total drilling length of 324.6 m, and a separation distance between exchangers of 4,572 m. The maximum temperature will be 25.93 ° C and with a 6.5 ° C and the temperature peaks at 32.22 ° C and -6.67 ° C. With these temperatures, the need to add some antifreeze to the system is also confirmed for its operation.

Figure 9 shows the proposal for the interconnection of the heat exchangers, the figure shows the dimensioning of pipe exchangers, like the inkjet heads and refrigerant recovering also the dimensioning of the accessories for installation, As well as the mass flow in gal/min in each of the sections of the heat exchanger. For sizing, the horizontal heat exchanger shown in Figure 11 was performed using the software GLD 2014, this tool allows for vertical and horizontal dimensioning exchangers with a large number of configurations. This software uses an iterative method for its solution by feeding parameters divided into specifications of fluid, soil, piping, configuration and auxiliary equipment.



7B

CODE	DESCRIPTION	QUANTITY
T26	2" X 2" X 1"	6
T22	2" X 1-1 /2" X 1"	2
T15	1-1 /2" X 1-1 /2" X	2
T12	1-1 /2" X 1-1 /4" X	2
T7	1-1/4" x 1-1/4 " x	4
T5	1-1/4" x 1 " x 1"	2
E3	1" x 1"	2
E9	2" x 2"	2
O2	Pipeline 1" RD 9	1010 m
O3	Pipeline 1-1/4 RD 9	20 m
O4	Pipeline 1-1/2 RD 9	20 m
O5	Pipeline1-3/4"	20 m
O6	Pipeline 2"	60 m
U2	U 1"	10

Figure 9. Schematic of the air conditioning system using a geothermal heat pump

Figure 10 shows the behavior of the maximum average temperature of the exchangers during operation over 10 years of operation. It can be observed on the temperature profile curve the effect provoked to the underground due the operation of heat exchangers on this soil. From the graph, it is clear that variations are minimal as well as soil perturbances. The variance of 1°C demonstrate that along 10 years of usage the system will not affect the surroundings or the environment where installed.

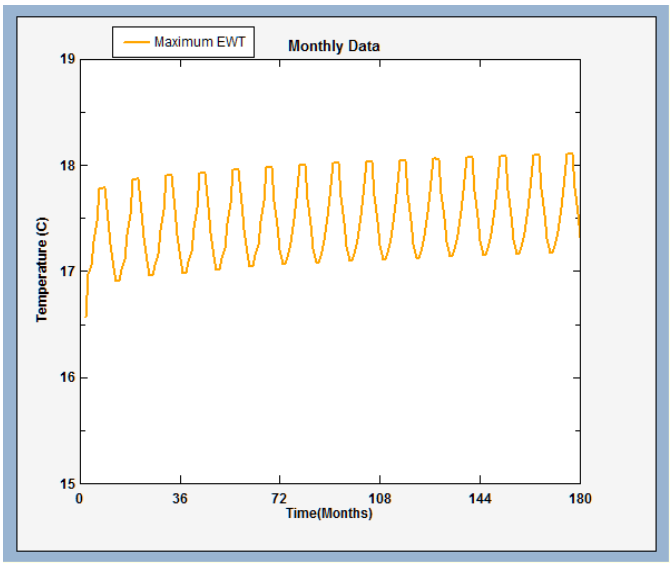


Figure 10. Behavior of the set of heat exchangers with maximum temperature

For the dimensioning of the horizontal heat exchanger shown in Figure 11 was performed with the help of GLD 2014 software, this tool allows the design of vertical and horizontal exchangers with a large number of configurations. This software uses an iterative method for its solution by feeding proper parameters: such as specifications of fluid, soil, piping, configuration and auxiliary equipment.

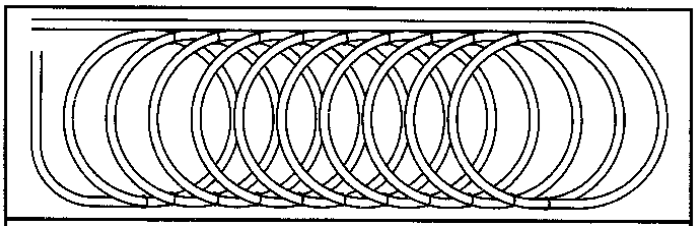


Figure 11. Slinky horizontal geothermal heat exchanger for trenches

The parameters of the pipe comprise the thermal conductivity of the pipe, and its dimensions as outer and inner diameters and the die, plus the type of internal flow of the pipe. In the configuration of the heat exchangers the software has the different types of trenches that can be applied, in this study we used the trench slinky type, this is a pipe coil, this technique requires less space than other techniques, using only 4 lines with a separation of 6.1 m between them to a depth of 1.8 m and a width of 20 cm, the particular measures of the slinky type heat exchanger is a diameter of 91 cm and a separation between them of 52 cm. The next thing is to quantify the heat gains of the system by use of pumps and auxiliary equipment, in this case only a 2.5 HP heat pump is taken into account to be able to move all the fluid through the heat exchangers to the heat pumps.

After entering the parameters required by the software as in the previous calculation, the iteration is performed to be able to quantify the dimensions of the system, in this case the result obtained is 4 trenches with a total length of 101 m, whereby each Trench will have a length of 25.3 m, the total length of the system will be 1347.6 m of pipe

In Table 2**,** the selection of the heat pump for space with increased demand and usage shown. The criterion of the selection of heat pumps is made by means of the air flow of the system for each space and these must also meet the requirements of thermal load in both cold and heat mode [7]. These values are also obtained by the 3D model space subject of analysis which are shown in Table 2. In Mexico it has a large number of companies that provide conventional air conditioning systems, in this case it was decided to make the selection of heat pumps in the catalog of the company Daikin.

Table 2. Selection of heat pump for Herbarium

Space	Maximum cooling load value (W)	Maximum cooling load value (BTU/hr)	Cooling airflow (L/s)	Maximum heating load value (W)	Maximum cooling load value (BTU/hr)	Heating air flow (L/s)	Daikin Unit
HERBARIUM	18,308	62,469	1,480	17,295	59,013	1,359.60	48
LIBRARY	5,463	18,641	441.8	947	3,231	531.1	30
LBM	8,047	27,458	654.7	897	3,061	623.3	42
MULTIPLE USES	5,287	18,040	427.6	70	239	401.8	42
ASSEMBLY ROOM	2,920	9,963	236.2	512	1,747	285.1	

5. Conclusion

The purpose of this work was to design the system for air conditioning using heat pumps with U-type vertical heat exchangers through the use of different software for a student laboratory complex located in Morelia, Michoacán, Mexico. With this dimension, it makes clear the feasibility of installing these systems in Mexican territory and the learning of technical and field capacities in a multidisciplinary way. In addition to the development of a methodology for the integration of different software such as Autodesk Revit® and GLHEPro®, in which it was possible to model the building in detail for the calculation of thermal loads and the dimensioning of vertical heat exchangers. The length of the exchangers will be 81.15 m with a total of 4 heat exchangers, and a separation distance between exchangers of 4,572 m In addition to the horizontal heat exchanger system which are formed of 4 trenches with a total length of 101 m, Whereby each trench will have a length of 25.3 m, the total length of the system will be 1347.6 m of pipe, the heat pumps selected for the highest priority spaces for its Daikin models 30, 42 and 48. Which meet the need for mass air flow and thermal load in both heat and cold.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The authors wish to acknowledge the SENER and CONACYT institutions for their support in CeMIE_Geo projects, particularly the P-30 project.

References

[1] S. General, "GEOTHERMAL Outlook Limited for Some Uses but Promising for Geothermal Heat Pumps," 1994.

[2] N. Lenhardt and A. E. Götz, "Geothermal reservoir potential of volcaniclastic settings: The Valley of Mexico, Central Mexico," *Renew. Energy*, vol. 77, pp. 423–429, May 2015.

[3] A. Capozza, A. Zarrella, and M. De Carli, "Long-term analysis of two GSHP systems using validated numerical models and proposals to optimize the operating parameters," *Energy Build.*, vol. 93, pp. 50–64, Apr. 2015.

[4] D. Wang, L. Lu, and P. Cui, "A new analytical solution for horizontal geothermal heat exchangers with vertical spiral coils," *Int. J. Heat Mass Transf.*, vol. 100, pp. 111–120, Sep. 2016.

[5] A. Franco and F. Fantozzi, "Experimental analysis of a self consumption strategy for residential building: The integration of PV system and geothermal heat pump," *Renew. Energy*, vol. 86, pp. 1075–1085, Feb. 2016.

[6] 2005 ASHRAE, *ASHRAE. 2005. Capítulo 30: Nonresidential Cooling and Heating Load Calculations (Cálculos de cargas de calefacción y refrigeración en zonas no residenciales), 2005 ASHRAE Handbook - Fundamentals Roth, Stephen. 2006.* 2005.

[7] C. Remund, R. Carda, P. Rawlings, and J. Bose, "GROUND SOURCE HEAT PUMP RESIDENTIAL AND LIGHT CEMMERCIAL DESIGN AND INSTALLATION GUIDE." Oklahoma state University, Stillwater, Oklahoma, 2011.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1070-1079 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Space Time - Track Circuits with Trellis Code Modulation

Marius Enulescu

Engineering & Design, Obrascon Huarte Lain, 1400- SKI, Norway

ARTICLE INFO

Article history:
Received: 29 April, 2017
Accepted: 14 June, 2017
Online: 15 July, 2017

Keywords:
Track Circuit
Trellis coding
Space Time diversity

ABSTRACT

The track circuits are very important equipments used in the railway transportation system. Today these are used to send vital information, to the running train, in the same time with the integrity checking of the rail. The actual track circuits have a small problem due to the use of the same transmission medium by the signals containing vital information and the return traction current, the running track rails. But this small problem can produce big disturbances in the train circulation, especially in the rush hours.

To improve the data transmission to the train on-board equipment, the implementation of new track circuits using new communication technology were studied. This technology is used by the mobile and satellite communications and applies the principle of diversity encoding both time and space through the use of multiple transmission points of the track circuit signal for telegram which is sent to the train. Since this implementation does not satisfy the intended purpose, other modern communication principles such as 8PSK signals modulation and encoding using Trellis Coded Modulation were developed. This new track circuit aims to solve the problems which appeared in the current operation of track circuits and theoretically manages to transmit vital information to the train on board equipment without being affected by disturbances in electric traction transport systems.

1. Introduction

In the paper [1] "Track circuit with space-time coding" it has been shown that the present track circuits used in metro lines have some problems, connected with the train transmission telegrams due the electrical perturbation produced by the train return current. To have one conclusion about this problem, at the Bucharest Metro line, measurements were taken and oscillograms picked-up, for more than 100 track circuits. These measurements were conducted to find the real cause of the sudden emergency brakes applied by the running metro trains. The trains lose the command telegram for a while and apply the brakes for safety reasons. The telegram contains the permissive orders to run regarding the permissive speed and distance to go. These emergency brakes lead to disturbances in the train time schedule and jeopardize the passengers' safety. To solve this problem, several solutions have been proposed but none of them have been able to fully meet the proposed goal.

In paper [1] "Track circuit with space-time coding" an embodiment was proposed to be utilized, using the existing signalling equipment, to ensure the transmission of the telegrams to the train using two transmitter points at a fixed position from one another. The two telegrams emitted by tuning unitsTU1-a and TU1-b arrive to the train at a time interval δ one after the other. This interval enables the rapid verification of the telegrams if they contain errors and if they do, the receiver can correct them.

The disadvantage of the embodiment is that on the distance between the two transmitters TU1-a and TU1-b, the train receives only the second telegram, leading to the need of a geographical map at the on board receiver and the verification of the position of the train if it is found between the TU units. But the main disadvantage is the use of additional equipment which has a high cost, namely one TU unit for each track circuit.

The train receives orders that authorize its movement entire in full safety conditions. The orders are embedded in a telegram that is transmitted through the track rails. The telegram addressed to the train can be disrupted and the train cannot recognize the transmitted order. And this fact leads on the train to apply the emergency brake [2].

The telegram is affected while it passes through the communication channel via rails [3]. This channel is not a special channel built for data transmission. Because of this, the telegrams

*Corresponding Author: Marius Enulescu, Engineering & Design, Obrascon Huarte Lain, 1400- SKI, Norway | Email: mariusenulescu@yahoo.com

are affected by the return traction current and electrical noise produced by the electrical motors of the train. In this paper such issues are analyzed and a new solution is proposed to solve the telegram loss by the train, when the telegram is affected by disturbances similar to Dirac impulses.

The QPSK-MIMO track circuit proposed in previous works cannot provide telegram transmission during the window time requested by the receiver. The two telegrams required for MIMO diversity are composed of 8 bits needed to identify the track circuit, 29 bits of information, 1 utility bit, and 33 bits allocated for the verification of the telegram by CRC code totaling 72 bits per telegram [4]

For the transmission of the two telegrams in the window time allocated by the receiver for the reception of a telegram, namely 630 ms [5], one should be reduced the bit time by way of changing the oscillators in the existing receivers, which does not correspond with the scope to improve the current system with minimum changes. The second method of shortening the transmission time of the telegram is to decrease the length of the transmitted telegram. In order to reduce the transmitted telegram, the only option is to drop the CRC code and replace it with another encryption and error protection system. In order to transmit the 38-bit utility telegram it is needed to look for a coding to allow the telegram to be checked and the transmission to be carried out in less than 315 ms even below 300 ms, because 15 ms should be allocated to the time between the two telegrams.

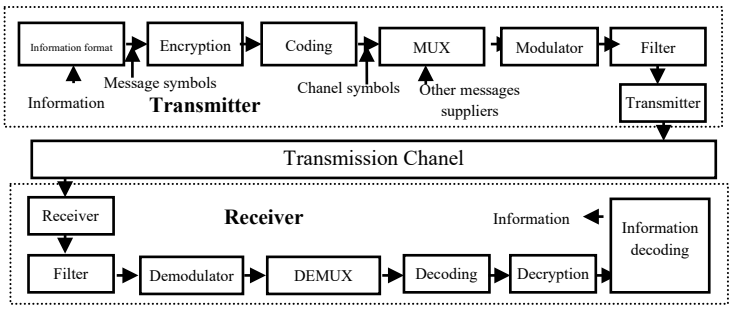


Figure 1. Transmission of information in communication systems

In the previous works have been implemented the transmission technologies used by today's WIFI communications systems so far, in this paper is proposed to try to use Trellis encoding. Trellis encoding is a coding commonly used in current communication systems. The transmission of the telegram in the current system is performed by transmitting the information in 30-bit information blocks which are accompanied by a 33-bit CRC block. If an informational bit is affected at the moment of reception, when performing the polynomial division modulo two, the result is not identical to that of the CRC code, and the telegram is declared disrupted. If a convolution code is used then information bits and control bits are combined into a set structure, practically decoding is bitwise with a continuous bit and if a bit is disturbed then it is not considered to be the entire lost word. In the classical data transmission systems functional blocks appear as in Figure 1. In a communication system two main functional blocks can be distinguished, namely the transmitter and the receiver. The main functionality blocks are: „information formatting block" - it collects the information that is the subject of the transmission and processes it by obtaining the message symbols and then these messages enter into the „encryption block" to ensure data protection.

The message is coded [6] to be transmitted through the communication channel in order to obtain the channel-specific symbols, the „multiplexing block" which allows the access to the transmission channel of several message sources. The received messages are modulated and passed through a „band pass filter" that only allows the passage of useful frequencies to the „broadcasting block". The „broadcast block" amplifies and does the adaptations with the „data transmission channel" towards the reception. At the „reception", the information is extracted by the reverse execution of the operations performed in the „broadcasting block". If a more efficient system is desired, the complexity of the system must be increased, with the advantages and disadvantages of the solution. Increasing the system performance can be done by using Trellis Coded Modulation (TCM). The use of Trellis coding performs coding and modulation of the signal in the same block, which reduces the time required for data processing, but at the same time requires specialized electronic components made by state-of-the-art techniques, allowing complicated operations. Trellis encoded data transmission systems have very good noise protection and are used for high-speed data transmissions. In the case of a limited power communication channel, increasing energy efficiency can be achieved by using error correction codes. Corrector error codes supplement additional bits to the transmitted encoded information sequence. Adding supplementary bits to the useful information results in the need for longer words to be transmitted, and if they need to be received in a given window time, it is necessary for the modulator to work at a higher frequency requiring a wider frequency band. The maximum data transmission speed through an ideal communication channel (stable, homogeneous, invariant in time, without distortion) is limited. This limit, called the capacity of the communication channel C [bit / s], was established by Shannon for an ideal channel with add-on Gaussian white noise, with a mean power of the useful signal and assuming the infinite coding, as to be:

$$C = B \log_2 (1 + P/N) \text{ [bit/s]} \quad (1)$$

Where the capacity of the communication channel is denoted by C and B is the bandwidth available for communication, the average signal power is P and the average noise power is N. In the case of a limited frequency band communication channel, increasing the spectral efficiency can be achieved by using modulators that maintain a large distance between the signals. These modulator processes are based on Shannon's 2nd theorem: a source with the information flow R (bit / s) and a channel with capacity C (bit / s) and if R <C then there is a code with words of length „n" such that the probability of a decoding error P_e to be $P_e \leq 2^{-nE(R)}$ where E (R) is a non-negative function called the exponent errors - the theorem allege that indifferent how much the channel transmission is disturbed, the broadcast can be made with a minimal error probability. The sequence error probability (as well as the probability of symbol error) is limited up by a decreasing function ratio d_{min}^2 / N_0 and is well approximated by this expression when the signal / noise ratio is high. In the above expression d_{min}^2 is the square of the minimum Euclidean distance between two possible signal vector sequences. The Euclidean distance is a straight line between two points "s_1" and "s_2", each point is defined by the Cartesian coordinates (2). The distance between the two symbols is given by the relationship.

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \quad (2)$$

For digital transmission, the equivalent of the Euclidean distance is the Hamming distance given by the relation (3), this is

zero if the binary words are the same. These two distances are useful in signals processing.

$$d_H(x,y)=\sum_{i=1}^{n} d_H(x_i,y_i); \quad d_H(x,y)=\begin{cases} 1 \text{ daca } xi \neq yi \\ 0 \text{ daca } xi = yi \end{cases} \qquad (3)$$

For an optimal communication, a long sequence of signals with maximum separation must be transmitted, and the receiver must decide on these long signals rather than individual bits or symbols.

If the designing process of these long signal sequences is properly performed, the error probability of the message „v_e" decreases exponentially with the length n.

$$P_e < k e^{-(R_0-R)n} \qquad (4)$$

Where the transmission rate R is less than the maximum rate R_0 as given by the Shannon capacity of the channel. In conventional transmission, encoding is separate from the signal modulation operation. Coding is done at the digital level, this in generally adds bits to the streams to be transmitted and requires an increased bandwidth. At the receiver the decoding occurs after the signal demodulation. The demodulation can be done by electronic circuits for decision, with voltage thresholds, and is called „hard demodulation" or can be done by sampling the received signal and comparing these samples with memorized patterns of the signal accompanied by Gaussian noise, this is a „soft demodulation". "Soft demodulation" brings a 2 dB gain compared to „hard demodulation". The decoding operation is executed for the entire data stream that feeds the decoder, whether it contains an error or not. Decoding can be done by sampling the received analogue signals and comparing them with memorized patterns, this is a comparative decoding. The theoretical loss of the decoder by electronic decision over the comparative decision is about 2 dB. Modulation of coded signals can be a 2D modulation, which uses the dependence between the initial phase and the phase of the quadratic symbols. It is manageable to have a 4D modulation; where it introduces the dependence of the symbols of two successive time intervals. A Trellis modulation can be used, that introduces the dependence of each successive symbol. Trellis and multidimensional codes are designed to maximize the Euclidean distance between the possible sequences of the transmitted symbols. The distance between the nearest possible symbols transmitted in the signal space determines the performance of the code $Pc \sim e^{-d_{min}^2/2N^2}$ where d_{min} is the minimum distance between the signal sequences and N is the noise power in the channel. Trellis Coded Modulation combines modulation and encoding in the same block, while the bandwidth is not increased. It has the same symbol rate, but redundancy is introduced by using to encode a constellation with more points, than it would be needed without encoding. For Trellis Coded Modulation the number of point doubles but the symbol rate is the same and the spectral power remains the same. Since there are several possible points per symbol, it can be considered that for a given noise signal ratio the probability of error increases. As with the conventional coding, there is a dependency between different symbols. Only certain secure sequences for successive points of the signal constellation are allowed for Trellis encoding. By applying these constraints optimally at the reception, the probability of error decreases, as a measure of the transmission performance while growth the coding gain. This is the difference between the transmission through a channel with the same noise to signal ratio of the same information in a coded system and a non-coded system. The key element of Trellis Coded Modulation, to improve the power efficiency of a coded channel, is to insert a memory circuit. To compensate the redundant bits added at the signal encoding, a constellation with more than 2^n symbols is used. The design of the encoder and the constellation map is collaborated to reduce the complexity of the decoder. Using an 8-layer constellation for signal encoding, theoretically 2 bits per symbol can be transmitted with an error below 13dB. Using a coded modulation it is possible get a 6dB gain at the same symbol rate without increasing bandwidth [7]

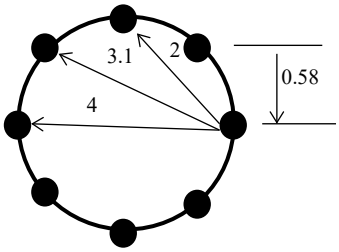


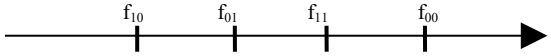
Figure 2. Euclidean distances for 8PSK

Trellis encoding, for a given limited channel bandwidth, at first, will be determined the rate of symbols that can be transmitted. The size of the 2L alphabet, which is required to produce the required bit rate, is determined. The constellation size is doubled and inserted into the encoder, which will generate an extra bit. It is not necessary that all the bits from the encoder input to be passed through the encoder. There are several ways to encode the bits in the symbols. The selected mapping can dramatically improve or dramatically reduce the code performance. Ungerboeck proposed a heuristic technique called mapping symbols, by partitioning symbol sets. The coding philosophy is to position the large constellation of 2^{2L+1} signals in smaller sets. The Euclidean distance between the signal point sequences in the different subsets is substantially increased (can be of the same order for all points in the same subset). Performance can be determined by the distance between symbol sequences in different subsets. Trellis encoding produces a substantial increase in the Euclidean distance between the signal sequence points so that at the receiver, the Viterbi algorithm can be used to detect the signal.

Ungeboeck in 1976 showed that for channels with bandwidth restricted by using convolution signal encoding, the gain of coding by comparison with binary encoding of the same signal can be substantially improved [8]. He proposed the use of modulation and coding together, to increase the Euclidean distance between the signal sequences. This was called Trellis encoding of the channel, because the sequence states in the finite state automaton that encode the signal levels describe a trajectory in the form of trellis for the possible passages. The largest Euclidean distance between the signal sequences, the lowest error rate for the noise signal ratio is $P_e = N_{free} Q(d_{free}/2Z)$ where N_{free} is the mean of the sequence numbers at the distance d_{free} and $Q(x) \sim e^{-x^2/2}$.

The code words consist of level sequences modulated. Trellis encoding uses dense signal sets but restricts what sequences can be used. These produce a minimum distance gain and the code requires a time dependence of the signal sequences that allow the receiver to pass over the noise bursts like the transmitted sequence estimates. Since the shape of the symbol and the transmission rate are unchanged, coding does not require any bandwidth increases.

This theory can be applied in track circuits for telegram transmissions to the train in FSK system, for which two symbols are used. To use a TCM for the order telegram, the FSK modulation signal can be used and, according to Ungerboeck's rules, a dual-signal constellation with 4 frequencies is needed:



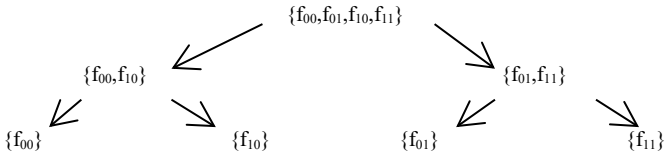
The frequency band is kept by setting the conditions:

$$f_{00}-f_{10}= f_0-f_1$$

$$\text{and} \quad f_{00}-f_{01}=f_{01}-f_{10}=f_{01}-f_{11}$$

This requires a slightly more complicated receiver, but not so much. Indeed, the demodulator is a frequency-to-voltage converter: for example, for the frequency f_0 at the output will have -U (volts) and for the frequency f_1 the output will be + U (volts).

If there are 4 frequencies, the demodulator gives to the output for each frequency a preset voltage value as in the following example: For f_{00} the output will be -3U, for f_{11} is + 3U, for f_{01} is -U and for f_{10} output will be + U. The four-level signal ± U, ± 3U is easy the convert into di-bits (two-bit group). The four-frequency set is divided so that the minimum distance between two characteristic frequencies increase:



The new aspect of coded modulation Trellis is that the encoding and modulation operations are not treated as independent operations, but as a single operation. The signal received, instead of being first demodulated and then decoded, is processed by a receiver that combines demodulation and decoding in a single operation. The detection process involves more soft decisions than hard decisions (the incoming signal is processed before making the decision with which data source symbol it matches). To exemplify the operation, a convolution Trellis coder is used as in Figure 3

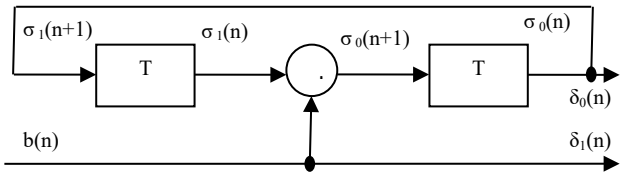


Figure 3. Transmission of information in convolution systems

The input bit b_n goes unmodified at the exit and is renamed $\delta_1(n)$, $\delta_1(n)=bn$

The convolution encoder has 4 states, given by bits $\delta_0(n)$ and $\delta_1(n)$. These states are the outputs of some delay elements (bi-stable, tipping circuits) that use a tact T equal to the duration of one bit. For the coder shown in Figure 3:

$$\sigma_0(n+1) = bn \oplus \sigma_1(n)$$

$$\delta_0(n) = \sigma_0(n)$$

$$\sigma_1(n+1) = \sigma_0(n) \qquad (5)$$

Table 1 Frequencies of signal issued to Trellis encoding

$\delta_1(n)$	$\delta_0(n)$	frequency
0	0	f_{00}
1	0	f_{10}
0	1	f_{01}
1	1	f_{11}

The two output bits δ_0 (n) and δ_1 (n) select the frequency of the transmitted signal as shown in Table 1:

The four-state Trellis graph which represents the operation of the convolution encoder is shown in Figure 4:

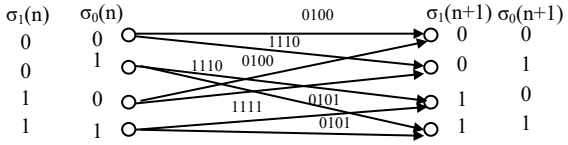


Figure 4. Four-State Trellis Graph

The convolution encoder state table, according to the previous states and the input data described above, is as follows:

Table 2 Trellis convolution coder states

$\sigma_1(n)$	$\sigma_0(n)$	b_n	$\sigma_1(n+1)$	$\sigma_0(n+1)$	$\delta_1(n)$	$\delta_0(n)$	frequency
0	0	0	0	0	0	0	f_{00}
		1	0	1	1	0	f_{10}
0	1	0	1	0	0	1	f_{01}
		1	1	1	1	1	f_{11}
1	0	0	0	1	0	0	f_{00}
		1	0	0	1	0	f_{10}
1	1	0	1	1	0	1	f_{01}
		1	1	0	1	1	f_{11}

If the encoder state is "00" and the input bit is b_n = "1", the encoder switches to the state "01" and the transmitted frequency is f_{01}. One advantage of the FSK modulation is that it does not need to equalize the transmission channel. For demodulation, the receiver uses only the input signal frequency, ignoring the amplitude and phase. For the above Trellis encoded scheme, the Viterbi algorithm can be used at reception. The algorithm was first proposed in 1967 and is an asymptomatic optimal decoding technique for convolution codes. It can be used to determine the encoded signal sequence close to the received signal sequence with the condition that the signal sequences transmitted are finite. The decoder determines the nearest signal sequence coded, directly from the information channel output. The most probable errors made by the decoder's optimal soft decision are related to the Euclidean minimum distance of the transmitted symbols.

2. Trellis Track Circuit - functionality analysis

For the TI21M track circuit an informational telegram, consisting of 29 informational bits, shall be transmitted to the train. They are divided according to the table 3 [9, 10].

From the table 3, it can be seen that the new Trellis Track Circuit must send the following information: [11, 12]

- The target distance - bits 0-7, contains 8 informational bits to transmit the allowed distance whereon the train can travel.

- The target restriction - bits 8-10, contains 3 informational bits to convey the target restriction. This tells to the train if a speed restriction or a change in the type of traffic follows.

- Door left- bit 11, an informational bit to transmit to the train if left-handed doors are allowed.

- Door right - bit 12, an informational bit to transmit to the train if the right door is allowed to be opened.

- Entry into the controlled area- bit 13, an informational bit that informs the train if it enters or exits in the system-controlled area.

- Passing over- bit 17, an informational bit informing the train that an area, where the train can lose the telegram for a short distance, is coming.

- Current speed and target speed- bit 18-24, 7 informational bits informing about the train speed and target speed.

Table 3 Meaning of bits in the telegram transmitted by the TI21M track circuit

bits	information
0-7	Target distance
8-10	Target restriction
11	Left doors
12	Right doors
13	Entrance in an controlled area
14-16	Track circuit markers
17	Over pass
18-24	Target and current speed
25-28	Target Gradient
29	Utility
30-63	CRC Cod

In total there are 22 bits of information to be transmitted to the train. The rest of the information being transmitted, by the current track circuits, is linked to track circuit's markers, which can be assimilated by the track circuit labelling. The CRC code bits can be removed, if another message verification method is used at the reception. The telegram that needs to be transmitted by the track to the train now has 30 informational bits.

If a Trellis encoding scheme is used, like the one shown in Figure 5.30, with a simple systematic feedback response coder and parity check, at two information bits from the encoder input it will have three bits at its output. This type of coding, with a ratio of 2/3, brings at the encoder output a 45-bit encoded trellis with parity check, where "h^i " are the parity coefficients.

For this scheme, Ungerboeck calculated in the article [13] the following characteristics: the number of states is eight for two information symbols per modulation cycle. The parity coefficients are h^2= 04, h^1= 02, h^0=11 and d^2_{free}/Δ_0^2 = 4.586 and coding gain is 3.6dB, all this for 8PSK modulation with

$\Delta i = 2\sin(\pi/8)$, $\sqrt{2}$,, 2 with i = 0,1,2

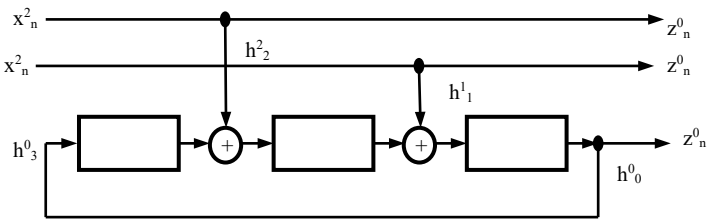


Figure 5. Trellis Systemic with Feedback Coder (3, 2, 3)

For systematic encoding, the input bits appear unchanged at the encoder output; this type of encoder cannot generate a catastrophic code. A catastrophic encoder is a coder for which at least one code word with the finite Hamming distance corresponds to an input with an infinite Hamming distance.

For the track circuit the important thing is that a 2 bits / sec / Hz transfer rate is achieved with a coding gain at the same 3.6 dB noise signal ratio versus a no-coded transmission.

For track circuits, the optimum transmission frequencies are in the audio frequency band due to the limitations imposed by the broadcast channel.

For 200 Hz signal bandwidth and Trellis encoding spectral efficiency 2½ bits / sec / Hz, 252 bits to the reception in the window time of the current system can be sent. This leads the train to be able to receive 4 telegrams in a timely manner with the information required for the system. If the allocated band is increased at 300 Hz, four information telegrams can be sent to the train, each of 64 bits, without reaching the maximum spectral efficiency limit for Trellis encoding. These four telegrams will be transmitted from two different points of the track to the moving train. Using a systematic parity convolution coder, a 45-bit telegram will be obtained; the remaining bits up to 64 bits can be used to send additional information to the train or even to redundantly transmit vital information within the same telegram. The speed and distance information can be transmitted twice in the same telegram.

Practically the new telegram that is sent to the train will contain the following information:

• target distance - bits 0-7 and 33-41, contains 2 * 8 informational bits to transmit the distance where on the train can travel

• target restriction - bits 8-10, contains 3 bits of information to convey the target restriction this tells to the train if there is a reduction in speed or a change in the type of traffic

• Left doors - bit 11, an informational bit to transmit to the train if it is allowed to open the doors on the left side

• Right door - bit 12, an informational bit to send to the train if it is allowed to open the doors on the right side

• Entry into the controlled area - bit 13, an informational bit that informs the train if it enters or exits the system-controlled area

• Pass over - bit 14, an informational bit informing the train that there is an area where the train can lose the telegram over a short distance

• Current speed and target speed - bits 15-21 and 41-47, 2 * 7 informational bits informing about the train speed and target speed.

After applying the Trellis 2/3 modulated encoding with 8PSK and parity checking, the following distribution of information inside the telegram will be obtained:

Target distance-bits 0-12 and 43-55, vital double information
Target Restriction - Bits 13-16
Left - Right Doors - Bits 17-19
Entry into Controlled Area and Over - Bits Crossing 20-22
Current speed and target speed - bits 23-32 and 33-43, vital information doubled
Track circuit label bit 56-64.

Performing the optimal Trellis encoding is done by mapping the symbols to ensure maximization of the free distance between them. This mapping is done by using special software programs to obtain the maximum efficiency of the encoder.

The information telegram set out above contains 37 bits required for the on-board system to meet the safety requirements for automatic train movement. These 37-bit information bits are inserted into a systematic convolution coder with feed-back and parity verification.

The encoder will have at the output 3 bits for every 2 bits at the input. For the transmission of the symbols it is necessary to use an 8-level modulation and for the beginning the use of the 8PSK modulation is to be considered.

The most probable errors made by the decoder's optimal soft decision interfere between the $\{e_n\}$ and the $\{d_n\}$ transmitted and the decoded signals, if they are close in terms of Euclidean square spacing. The free quadratic free distance is called free distance and is given by the relation

$$d^2_{free} = \min \sum_{e_n \neq d_n} |e_n - d_n| \text{ where} \{e_n\}, \{d_n\} \in C. \quad (6)$$

For this, the codes must be designed to have the maximum free Euclidian distance and not for Hamming distance. The redundancy required for coding comes from expanding the signal set without increasing the bandwidth [13]. The mapping should be based on the grouping of signals in symbols with a large distance between the signal subsets. To maximize the free distance the map of the symbol set is used as shown in the example represented in Figure 6.

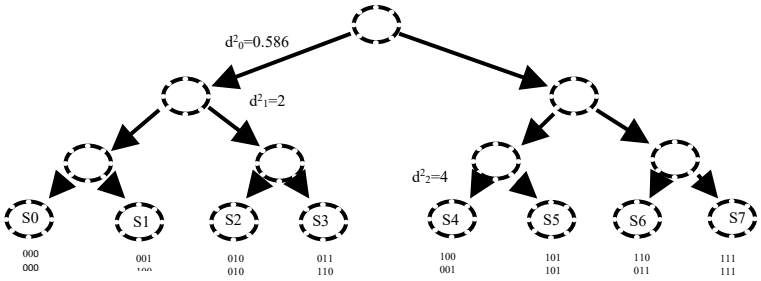


Figure 6. Symbol mapping by partitioning

3. Trellis track circuit – on-board receiver

The on-board receiver must receive the telegrams transmitted in the track from two distinct transmitters located in front of the train. The on-board receiver distinguishes these telegrams by differentiating the Trellis encoded modulated signals. The on-board receiver to process the received telegrams will have the block diagram shown in Figure 7.

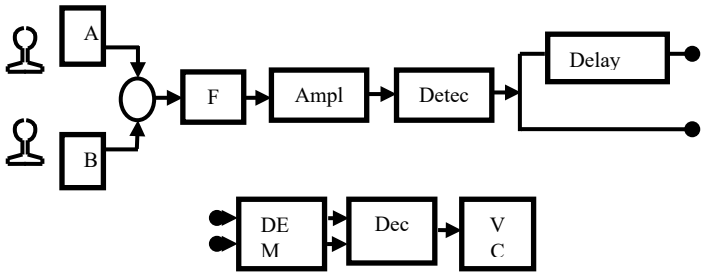


Figure 7. Trellis Track Circuit on-board receiver

The signal captured by the pick-up coils is summed by a differential summing circuit and then passed through the FTB bandwidth filter to keep only the useful signal. The detector block must separate the two telegrams received from the two separate transmitters. The telegrams arrive at a time difference dependent on the distance between the two transmitters and the gap between the telegrams introduced by the interlocking device. The first telegram is entered into a delay line until the second telegram arrives, then the telegrams are inserted into a demodulation and decoding block. The telegrams are inputted into a decision block that extracts the train data from the two telegrams, to transmit them to the train vital computer for execution.

The "Differential Summing Circuit" sums up the useful signals received by the pick-up coils, installed in front of the first axle of the train. The useful signal is passed through a band pass filter which only serves to extract the signals from the useful frequency bands.

The output from the "Band Pass Filter" is amplified and inputted into a delay line, until the second telegram is received. When the second telegram is received, the two telegrams are inserted into the "Demodulation and Decoding Block".

The "Decoding and Demodulation Block" consists of two separate demodulators and decoders for the two telegrams. The signal decoding can be made "hard" by demodulation and then decoding, but this demodulation causes an irreversible loss of information received before decoding. At the receiver a sequence of sampled voltages corresponding to the transmitted bits are created. The received signal is sampled and forms a bit sequence, to decode this sequence which can apply a hard decision process that decodes the samples by assigning them pre-set voltage values. This process can make an erroneous decision especially in the comparison threshold area. In order to avoid these inconveniences of hard decoding, a soft demodulation-decoding is used, which means that the decoder operates directly on the non-quantified "soft" samples of the channel [14]. The patterns are expressed as $v_n = a_n + w_n$, where a_n is the discreet signal transmitted by the modulator and w_n is the representation of the samples of the Gaussian noise addition process. The decoding rule of the decoded optimal sequence is to determine from the set "C" of the encoded signal sequences that can be produced by the coder-modulator system the sequence $\{a_n\}$ with the Euclidian square minimal distance that fulfils the relation:

$$|r_n - a_n|^2 = \text{Min} \sum |r_n - a_n|^2 \text{ with } \{\underline{a}_n\} \in C \quad (7)$$

The Viterbi algorithm is used to perform this decoding-soft demodulation. This algorithm was proposed in 1967 and is an asymptomatic optimal decoding technique for convolution codes. It can be used to determine the sequence $\{a_n\}$ of the encoded signal near the received non-quantized signal sequence $\{r_n\}$ provided that

the sequences of the encoded signal generated $\{a_n\} \in C$. The coder determines the nearest encoded signal sequence, directly from the non-quantized channel output. The most probable errors made by the decoder's soft decision-soft interfere between the $\{a_n\}$ transmitted and $\{b_n\}$ decoded signals, if they are close one to another within the Euclidean quadratic distance, this distance is the free distance and is given by the relationship:

$$d^2_{free}=\min\sum_{e_n\neq b_n}|a_n - b_n| \text{where}\{a_n\},\{b_n\} \in C \quad (8)$$

For this reason, system codes must be designed to have the maximum free Euclidean distance. For Trellis encoding, Viterbi decoder uses two metrics: branch metric and track metric. The branch metric is the measurement of the distance between what was transmitted and what was received and is defined for each path in the signal braiding. For "hard" decoding this branch metric is the Hamming distance between the expected bits and received bits. The track metric is the value associated with the trellis state, the value associated with each node, for the "hard" decision it corresponds to the Hamming distance of the most probable path from the initial state to the current state in the trellis. The most probable path is that with the smallest Hamming distance, measured on all possible paths between two states Hamming's smallest distance minimizes the total number of errors and is the probable path with the smallest bit error ratio [15]. The Viterbi algorithm calculates metric paths for pairs of states using the branch and track metrics calculated for previous states [14]. The key points are as follows: the optimal path of the infinite past signal to all trellis states at time "n" is known. The algorithm extends these interactive paths from state "n" to state at time "n + 1" Choosing the best way, the algorithm applies a "good" tag for each new path or "cancelled" for all other paths that cannot be extended. The good paths tend to unite in the same historical path after a time "n_d" with an effective delay decoding "D" (so the random change of the values of "d" is very likely to be less than "D"). The information associated with a transition is the historical "n_D" path can be selected at the output. The received signals are considered to be disturbed by non-corrected Gaussian noise samples with variable σ^2 in each signal size (in the subway system the noise is not Gaussian and patterns with system-specific disturbances must be used). The probability that at the given time the decoder makes an erroneous decision among the values of the associated signals for parallel transitions or to start making a series of mistaken decisions along the divergent paths for more than one transition from the correct path is called event probability error. At high noise signal ratios this probability is generally approximated by:

$$Pr(e) \approx N_{free}\, Q\left|{}^{d_{free}}/_{2\sigma}\right| \quad (9)$$

Where Q (x) is the Gaussian error integrity (10), and N_{free} is the average of the numbers of neighbouring sequences with the d_{free} distance that go to any state in the transmitted signal sequence and come together after one or more transitions.

$$Q(x) = \frac{1}{\sqrt{2\pi}} \int e^{-y^2/2} dy \quad (10)$$

From the relationship (9) it can be approximated that at high noise signal ratios, the probability of event error associated with the target distance and then d_{free} becomes negligible.

For the embarked system the receiver uses the Viterbi algorithm and it implements two sets of patterns for comparison. The pattern sets for the disturbed signal must be correlated with the communication channel used and the disturbances that occur in it. Depending on the results obtained, the handset can automatically switch between sets of patterns to reduce the probability of error.

Because the 8PSK modulation is used on the receiver, a symbol rotation may occur due to interference or even oscillator instability. If in the Trellis diagram there is a large free square-space between symbols, then a small rotation of the diagram does not affect the received symbol, but if there is a small free square space between the symbols, then a small rotation of the diagram can change the decision boundaries of the symbols to make an erroneous decision. In most of the digital frequency carrier systems, the phase of the loop-specific carrier system has to be tracked. These loops estimate the phase shift of the received signal and decoder decisions, the estimated phase shift controls the phase of the demodulator carrier. For the new system, two symbols are used for each transmitted telegram to synchronise the demodulator with the modulator phase.

The output of the Trellis "Demodulation-Decoding block" is connected to the "Decision block" input. This block receives the two sets of decoded signals, each for a received telegram, compares them and checks the parity bits. Having two telegrams issued by two transmitters at different points of the runway, the train can extract the correct information even if the telegrams have been disturbed during the passage of the communication channel. The "Decision block" maps the bits from the Trellis decoder demodulator output to the telegram pattern and extracts the information about the received orders. The bits received for each telegram contain redundant information on speed and distance (maximum safety information for train running). The "Decision block" has four sets of information that regulate the train speed over a reserved and secured traffic distance. This information is received by the way of two telegrams sent through two transmitters separated from each other by the length of a track circuit.

The verified information is transmitted to the train computer for execution.

4. Trellis track circuit – transmitter

Taking into account the above mentioned functionality, the block diagram of the track circuit transmitter becomes that of Figure 8.

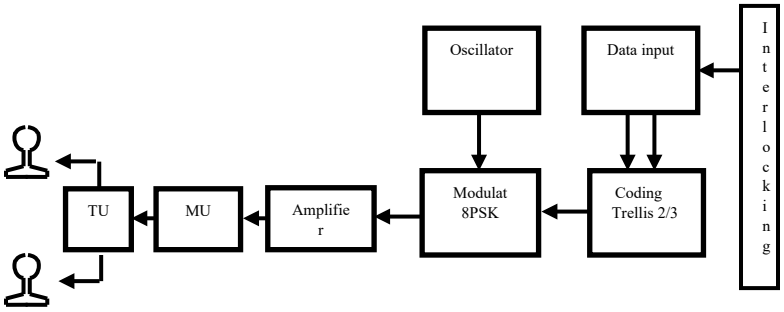


Figure 8. Trellis circuit block diagram scheme

The electronic interlocking system sends the orders to be executed and also the data which will be transmitted to the train, to the track circuit transmitter. The orders received contain instructions on how the track circuit transmitter should work, these modes being "detection mode" and "telegram mode".

The information received from the interlocking is processed in the “Input block”. This block upon receipt of the "detection mode" order extracts from its own memory track circuit label and maps it to obtain the optimal Trellis encoding symbols. This mapping is required to optimize the encoding. The information is cyclically sent by two bits to the “Trellis Encoding block”. This operation is performed by the transmitter until another order is received.

"Telegram mode" is executed upon receipt of the order received from the interlocking. On receipt of this order the transmitter receives the information, which must be included in the telegram to be sent to the train, maps this information to obtain the optimal symbols and sends these symbols to the “Trellis Encoding block”.

The “Input block” will also prepare the synchronization symbols needed to synchronize the receivers; these symbols are included in the train telegram or in the detection message.

The “Trellis Encoding block”, can be that shown in Figure 5, it is a block that encodes the two information bits received from the “Input block”. The “Encoding block” performs Trellis' systemic with feed-back coding and parity verification. The block encodes the received information and sends it by three bits to the “Modulation block”.

The “Modulation block” is designed to modulate a carrier frequency with the three-bit symbol received from the “Encoding block”. The internal diagram of this block is given in Figure 10 and performs the modulation using the symbol diagram shown in Figure 9

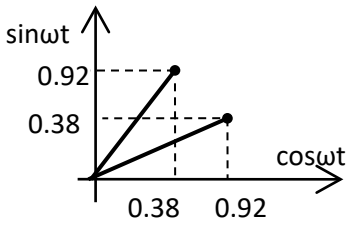


Figure 9. Symbols diagram 8PSK

From the chart shown in Figure 9 it can be noticed that: for QPSK modulation where there are only two coordinates “I” and “Q”, for 8PSK it also need to define two amplitudes ± 0.38 and ± 0.92 to obtain the 8PSK modulation angles in the four quadrants.

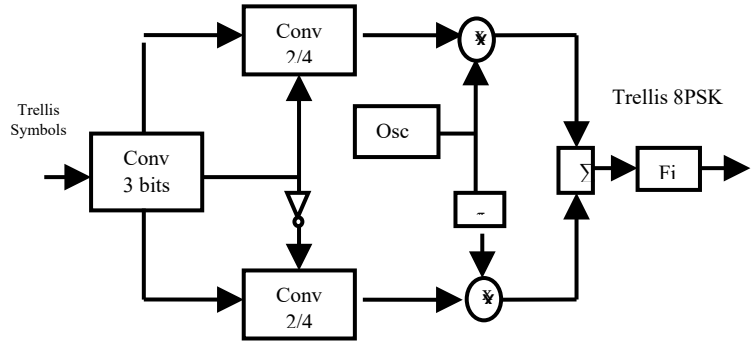


Figure 10. 8PSK Modulator block diagram

In the block diagram shown in Figure 10, the three-bit symbols from the output of the “Trellis Encoding block” are inserted into a serial / parallel converter, which separates the three bits on the three paths that define the symbol in the signal diagram. The bit 1 is used to control channel “I” , bit 2 to control channel “Q”, and bit 3 is used to define the two amplitudes. The amplitudes of the “I” and “Q” signals must be different in order to obtain the spatiality between the 8 phase states needed for 8PSK encoding states. The three bits command the 2-4 amplitude converters located on each channel “I” and “Q”, the bit 3 controls what amplitude is applied to each channel ± 0.38 or ± 0.92.

Outputs from the converters 2-4 are multiplied by the cosine and sine functions having the required carrier frequency, f_1 or f_2 for the line 1 and f_3 or f_4 for the line 2. These two signals are rotated continuously with $3\pi / 8$ before forming the signal pulse to avoid that the signal envelope instantly passes near zero. For an optimal transmission, where must always be a difference between two consecutive symbols of 112.5^0

The two waveforms obtained after multiplication will be summed and applied to the input of a “Band Pass Filter” that only selects the signal useful for transmitting through the communication channel.

The signal rotated at this point is described by the relationship:

$$R_n=S_n\, e^{\frac{j3\pi n}{8}} = e^{\frac{j2\pi C_n}{8}} * e^{\frac{j3\pi n}{8}} = e^{\frac{j\pi(2C_n+3n)}{8}} = a_n + jb_n \quad (11)$$

Where the “S_n “signal is

$$S_n = A_c e^{\frac{j2nC_n}{8}} = A_c \cos\frac{2\pi C_n}{8} + j\, A_c \sin\frac{2\pi C_n}{8} = I + j\, Q;\ \ n = 0 \div 7 \quad (12)$$

At the output of the 8PSK modulator block, it is a Trellis-encoded-modulated signal that, in order to be transmitted through the communication channel, in our case the rail tracks, requires amplification and adaptation to achieve the maximum power transfer to the load. These preparatory operations are performed by the “Amplifier block” and the “Mu” and “Tu” blocks that make the adaptation with the cable or railway tracks.

5. Trellis track circuit – track receiver

The track circuit receiver has the function to make the decision whether the track circuit is free or occupied. In order to do this, the receiver has a schematic block as shown in Figure 11.

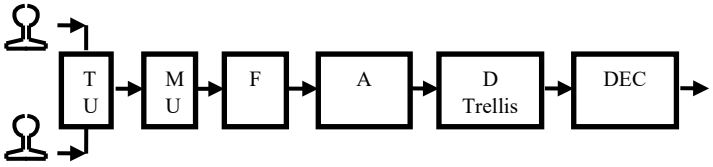


Figure 11. Trellis-8PSK track circuit block diagram

“Tu” and “Mu” units have the role of adapting impedances to get a maximum power transfer. The received signal is passed through a band pass filter that only keeps the useful frequency band

The amplifier “A” takes the signal from the output of the filter “F” and increases its power so it can be applied to the Trellis demodulator-decoder input “D”.

The demodulator - decoder Trellis “D Trellis” applies the Viterbi algorithm to the signal and extracts the received data. The data is extracted by “soft” decisions taken on signal samples that are compared to the noise-signal patterns. The patterns are made

for the type of communication channel specific interference. For the track circuit, the receiver has two types of signal patterns, one with moderate disturbances and another for high disturbances. The receiver estimates the received symbols by applying these patterns and if the estimate does not match, then the receiver will change automatically the pattern sets to obtain optimal decoding and demodulation. The “demodulation-decoding block” also includes a tracking loop for the carrier phase. This block directly controls the frequency of the demodulator. Trellis demodulation and decoding at the track receiver is made identically to the on-board receiver. The symbols on the output of the Trellis demodulator-decoder block are introduced into a “Decision block”.

The “Decision block” is designed to rebuild and check the track circuit label. If the label is correct, the receiver will make the free or occupied track circuit decision. The track circuit receiver also collects other signals transmitted by the neighbouring transmitters, retrieves the track circuit labels and sends the information to the interlocking system. The interlocking system will receive the message with all the labels collected by the track circuit receiver and will have an overview of all the status of the track circuits in that area.

6. Connection of Trellis track circuit equipment

In the track circuit connection scheme, the transmitters and receivers are connected to the rail at the same point via an adapter transformer that has the role of adapting the impedances so as to obtain a maximum power transfer for the 8PSK modulated signal with the carrier frequencies f1 and f2. Frequencies f1 and f2 are chosen to achieve the best transmission performance through the railroad communication channel. This choice of carrier frequencies can be done using the communication channel models described in [16].

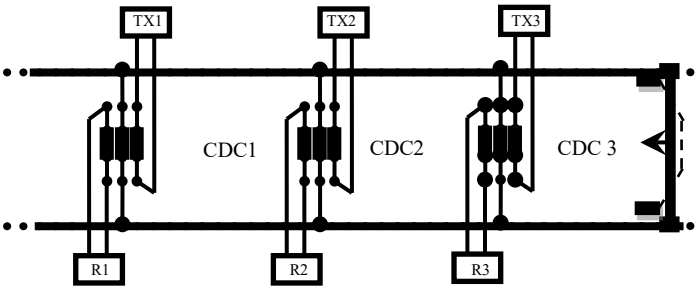


Figure 12. Connection of Trellis track circuit equipment

The carrier frequency is chosen accordingly to the electrical characteristics of the respective section of the line, these characteristics being dependent on humidity, ballast, line, etc.

The operation of the CDC2 track circuit in the two operating modes is as follows:

"Detection" mode - the Tx_2 transmitter conveys a 64-bit telegram made by Trellis 2/3 coding of the 8-bit circuit-specific identification tag and a 30-bit "0" and "1" string modulated at a frequency of 20 Hz. The telegram is received by the R_1, R_2 and R_3 receivers, making the decision that the track portions between them and the transmitter Tx_2 are free, and the receiver R_2 checks the correct operation of the transmitter chain Tx_2.

When the CDC3 track circuit is occupied by the train having the direction of going to the CDC2, the interlocking device decides to switch the Tx_3 transmitter as well as the Tx_2 transmitter in the "telegram" mode.

"Telegram" mode - the Tx_3 transmitter conveys a 64-bit telegram composed by Trellis 2/3 coding of the 8-bit transmitter label Tx_3 and a 30-bit string containing the information required for the ATP system. This telegram is received by the on-board receiver as well as the R_3 receiver that check the transmission chain as well as the receiver R_2 which recognizes the transmitter identification tag Tx_3, making the decision that the CDC2 is free.

"Telegram" mode - the Tx_2 transmitter transmits a 64-bit telegram composed by Trellis 2/3 coding of the 8-bit transmitter label Tx_2 and a 30-bit string containing the information required for the ATP system. This telegram is received by the on-board receiver which has doubled the information received, benefiting from the diversity principle, as well as the receivers R_3 and R_1 which decide that the CDC1 and CDC2 are free and the receiver R_2 is checking the transmitter chain Tx_2.

When the train pass on the CDC2 track circuit, the transmitter Tx_2 and Tx_1 are switched to "telegram" and the Tx_3 transmitter is switched to "detection" mode. The command of the transmitter operations in both modes is performed by the interlocking system according to the train movement authority, and this system also provides the 30-bit string required for the ATP system

It is noted that for the operation of this track circuit there is no need for alternating the carrier frequencies to separate the signals of the neighbouring track circuits. This separation is done digitally by recognizing the labels transmitted. Alternating the carrier frequencies is only to help the receiver to easily separate the telegrams conveyed by two consecutive transmitters.

Identification tags may also be received at greater distances after two or more track circuits and their receivers extract the respective label and send it to the interlocking device, which sees an overview and takes the decision that the distance between the receiver and the transmitter whose label has been received as "free".

Therewith the interlocking system can transmit to the train many telegrams using more than two transmitters, in this case the train has several telegrams that contain the same ATP information with multiple track circuit labels, which leads to a more efficient extraction of the ATP data system even in the case of Dirac disturbances. The limitation is given only by the window time available for telegram reception and the time required to extract the useful information, this being for an ideal line without loss and without attenuation.

In the case of real railway lines, there are losses due to ballast resistance and attenuations due to the constructive impedances of the railway. And in order to transmit the modulated signals on the carrier frequency "f" by the transmitters at a great distance, compensation capacitors can be installed, at each 50m or more. Compensation capacitors have the role of creating resonance points on the carrier frequency used, so it can pass on distances greater than the length of a track circuit. The length of the subway track circuit is about 250 meters, making it ideal for subway trains used.

7. Conclusions for the Trellis track circuit

The new track circuit with TCM aims to solve two of the problems encountered in the operation of the current track circuits, namely the loss of the telegram by the train and the false

occupation of the track circuits in front of the train. These two problems emanate due to the Dirac impulses which arise in the return traction current path.

The new track circuit uses the transmission diversity principle to solve the loss of the telegram by the train.

Due to the limitation imposed by the length of the telegram, it was necessary to use Trellis 2/3 convolution code with a 8PSK signal modulation

The new telegram contains a tag for each transmitter, which it is used to identify free-circuits but also help the train to distinguish between received messages. By using these labels, the second problem encountered in the current track circuits, namely the false occupation of the track circuit in front of the train due to the Dirac disturbances occurring on the traction return side, is solved. The receiver will periodically receive telegrams transmitted from several transmitters and these telegrams are spaced from each other with the time required to travel on the distances from the transmitters to the receiver.

The realization of these circuits requires specialized FPGA circuits that can perform complex functions necessary for the coding-decoding and modulation-demodulation operations of the signals used. Due to these complicated systems the reliability is decreasing but this can be avoided by using good quality components.

Data loss caused by electromagnetic perturbation, especially due to the electrical arch in the return current path, is decreased drastically by using multiple transmitters. On the other hand, the data loss can increase due to 8PSK modulation but this is compensated by using multiple identical telegrams transmitted from different points and a good Trellis code.

The incident response time of the on-board equipment is improved by using the Trellis code and multiple similar telegrams.

The new track circuit uses a new rail connection concept, namely a magnetically coupled return coil with two coils that are tuned to deliver maximum power transfer from the transmitter to the track as well as from the track to the track circuit receiver.

Future works will be done with the scope of improving this new track circuit by finding a suitable Trellis code for metro line conditions and which fits the patterns used by the Viterbi algorithm

Conflict of Interest

The authors declare no conflict of interest.

References

[1] M. Enulescu, "Track circuit with space-time coding" International Conference on Applied and Theoretical Electricity (ICATE), 2016. DOI 10.1109/ICATE 2016.7754690

[2] Bombardier Transportation - Safety Analysis of OC Functions

[3] Bombardier Transportation - Track circuit TI-21M technical manual M2

[4] Data description for track circuit M2 - Bombardier Transportation

[5] On-board equipment description M2 - Bombardier Transportation

[6] http://webspace.ulbsibiu.ro/macarie.breazu/ACM/Codare Predictiva.pdf

[7] www.complextoreal.com Trellis Coded Modulation tutorial

[8] Trellis-Coded Modulation with redundant signal sets part 1 introduction Gottfried Ungerboeck 1987 IEEE communication magazine

[9] Telegram Decoder Software Block Specification - Tigris ATP/ATO Vehicle Equipment Platform Coded Track Circuit Interface, CTI. (GBSIG document) – Bombardier Transportation

[10] EBICAB 800 Telegram Display Utility, Software Description - Bombardier Transportation

[11] Functional description, Supervision of contact positions - Bombardier Transportation

[12] Functional description M2 - Bombardier Transportation

[13] Trellis-Coded Modulation with redundant signal sets part 2 State of the art Gottfried Ungerboeck 1987 IEEE communication magazine

[14] Viterbi Decoding of Convolutional Codes MIT 6.02 draft Lecture notes October 2010

[15] www.complextoreal.com Coding and decoding with convolutional codes tutorial 12

[16] M. Enulescu "Communication Architecture for Public Metropolitan Subway Transportation System" PhD. Thesis, "Politechnica" University Bucharest 2016

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1080-1085 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Support Vector Machine based Vehicle Make and Model Recognition System

Muhammad Asif Manzoor*, Yasser Morgan

Faculty of Engineering and Applied Sciences, University of Regina, SK, Canada

ARTICLE INFO

Article history:
Received: 01 May, 2017
Accepted: 15 June, 2017
Online: 16 July, 2017

Keywords:
Vehicle Classification
Support Vector Machine (SVM)
Scale Invariant Feature Transform (SIFT)
Speed-Up Robust Transform (SURF)
Bag-of-Features (BoF)

ABSTRACT

Vehicle analysis is a very useful component in various real world applications. In this paper, we have developed a Vehicle Make and Model Recognition (VMMR) system using Support Vector Machine (SVM). Scale Invariant Feature Transform (SIFT) and Speed-Up Robust Transform (SURF) are used to extract local features from an image. Bag-of-Features (BoF) model is used to create visual dictionaries and convert the local image features into global image feature representation. Multiple dictionaries of different sizes are created for both features; SIFT and SURF and the dataset is coded using these dictionaries to determine the best size for the visual dictionary. NTOU-MMR is a publicly available vehicle dataset which we have used to evaluate the performance of proposed VMMR system. 92% recognition rate is achieved by using the proposed VMMR system.

1. Introduction

This paper is an extension of work originally presented in 2017 IEEE 7th Annual Computing and Communication Workshop and Conference (CCWC) [1]. In original work, we had developed a Vehicle Make and Model Recognition (VMMR) system using Bag of SIFT features and used Support Vector Machine (SVM) for the classification task. In this extended work, we have investigated Speed-Up Robust Features (SURF) along with SIFT features. We have investigated the effect of dictionary size over the recognition rate. We also have analyzed the margin size between for binary SVMs and its relation with the recognition rate.

Machine vision based vehicular analysis is suitable for many scenarios. Machine vision based techniques require installation of cameras to capture the videos and/or images and it also requires computing power to process these captured videos/images. Traffic cameras care widely available at present time; which can be used to capture real-time videos/images. These techniques do not require any installation of devices/sensor in the vehicles; which makes vehicular analysis simple. However, machine vision based applications have their own challenges.

Vehicular analysis applications may include License Plate Recognition, Vehicle Detection, and categorization of vehicles into bus, cars, trucks etc.; the majority of research is done for these mentioned scenarios. The focus of this work is to classify the vehicles according to their Make, Model and Manufacturing year. We are using machine vision-based approach to recognizing a specific instance of a vehicle. We are using Machine learning algorithm to classify input image/video according to the vehicle present. Machine learning algorithm provides the mean to train the system using a training dataset and then predict the outcome for new and unseen images. A lot many challenges are associated with this problem; some of these challenges are:

1. Image acquisition in outdoor environment.
2. Varying and uncontrolled illumination conditions.
3. Varying and uncontrolled weather conditions.
4. Wide variety of available vehicle appearances.
5. Similarities between different vehicle models.

* Corresponding Author: Muhammad Asif Manzoor, Faculty of Engineering and Applied Sciences, University of Regina, SK, Canada, AsifManzoor@uregina.ca

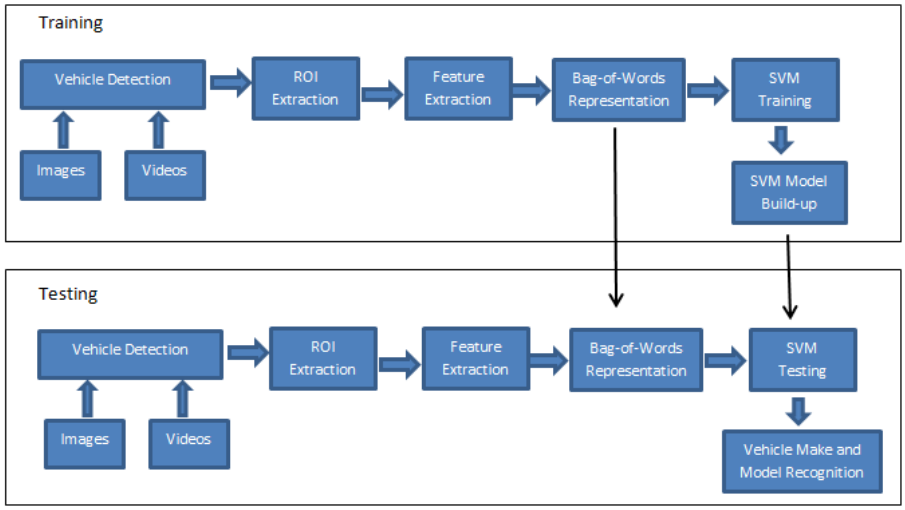


Figure 1: VMMR system Architecture

The architecture of VMMR system developed in this work is given in Figure 1. The system is designed to classify the vehicles using the frontal images. The input to the system can be either images or videos. If the videos are used as input sources, individual frames (images) must be extracted and will be processed separately. First, we must train the VMMR system based on input training dataset. The first step is to detect whether a vehicle exists in the input image. If it does not then we cannot use this non-vehicle image for the training. The vehicular images contain background along with vehicle and some parts of the vehicle (like a windshield) are almost identical for various vehicles' models. Hence, we define Region of Interest (ROI); which can be easily distinguishable for various models. Images features are extracted from this ROI and converted into image feature vector to represent a specific vehicle model. Lastly, machine learning algorithm is trained using these image feature vectors. Scale Invariant Feature Transform (SIFT) [2] and Speed-Up Robust Features (SURF) [3] are used to extract local image features. Bag-of-features (BoF) model is employed here to build the visual dictionaries and to transform the local image features into global image feature vector. We have used Support vector machine (SVM) [4], [5] as classifier in this work; which is supervised learning algorithm.

We have reviewed some related articles in section 2 and presented the dataset used in section 3. Section 4 discusses the VMMR system methodology in detail. Experimental results are provided in section 5 and section 6 conclude this work.

2. Literature Review

VMMR is a challenging task; some of the challenges are presented in Section 1. We will provide an overview of some of the related research work.

In [6], authors proposed VMMR system using the rear-view images. Authors defined the shape and geographical features with respect to taillights and license plate and used these features for vehicle recognition. The proposed system is developed to recognize the vehicles during the night under limited lighting conditions. The initial step is to recognize the location of license plate and authors used license plate location to calculate the features. Genetic Algorithm is also used in this work for feature selection. Genetic algorithm improves the recognition rate by 0.4%. However, the effect of Genetic Algorithm over the computational time is not discussed.

In [7], authors designed a system to identify vehicles' type by using deep convolutional neural network. The proposed system identifies vehicles without vehicle detection. Authors studied VGGNet, GoogLeNet, and CaffeNet (three well known convolutional neural network method). They had suggested including of vehicle/non-vehicle classification as pre-training for three convolutional neural network methods. Data enhancement techniques are used for performance enhancement. The proposed method is tested against cars dataset [8] and 79.5% accuracy is achieved in this work.

In [9], authors used side profile images in their work to recognize the vehicle's Make and Model. Authors have used five different classification techniques in their research to develop VMMR system; Random Forest, Evolutionary Forest, HoG based Random Forest, HoG based Linear SVM and HoG based RBF SVM classifier. A pole-mounted camera is used to capture the images and a dataset is created containing more than 10,000 images. These images contain 86 different make and model and are divided into 9 categories.

In [10], the oriented contour points are extracted from frontal image to recognize the vehicles in their research. Authors have used three voting and distance error to determine the make and model. The dataset tested in this work contains partially occluded images as well and have 830 vehicle images. Authors have reported recognition rate of 90.6%.

In [11], Contourlet transform is applied to extract features and applied localized directional feature selection criterion in their work. They also have used Two-Dimensional Linear Discriminant Analysis to reduce the feature dimensionality. Authors have used support vector machine classifier in their proposed framework.

Table 1: Details of Dataset

Vehicle	Year	Train	Test	Vehicle	Year	Train	Test
Toyota Altis	2008-10	260	504	Nissan Tiida	2009	108	139
Honda CRV	2003-09	224	258	Toyota Altis	2005-06	216	227
Toyota Camry	2008-10	117	169	Mits. Zinger	2010	12	13
Honda Civic	2010	79	243	Mits. Outlander	2010	26	50
Honda Fit	2012	34	35	Toyota Wish	2010	68	45
Honda Fit	2009	19	13	Mits. Savrin	2008	36	12
Toyota Camry	2005-06	109	122	Toyota Wish	2005-09	107	77
Toyota Camry	1999	21	5	Mits. Lancer	2007	16	61
Nissan March	2007-08	96	92	Toyota Yaris	2008	155	147
Suzuki Solio	2008	34	84	Ford Liata	2003	9	10
Toyota Vios	2008-10	242	292	Toyota RAV4	2009	79	0
Nissan Livna	2010	119	128	Ford Escape	2009	50	45
Nissan Teana	2010	66	29	Toyota Innova	2008	15	29
Nissan Sentra	2003	22	20	Ford Mondeo		38	10
Nissan Sentra	2005	24	15	Toyota Surf	2008	40	30
Nissan Cefiro	1997	67	22	Ford Tierra	2006	34	16
Nissan Cefiro	1990	36	9	Toyota Tarcel	2005	110	70
Nissan X-rail	2007	37	89	Total		2725	3110

3. Dataset

[12] is a publicly available dataset which contains vehicular images. We have used this dataset in order to analyze our VMMR system. The dataset contains complete images (vehicle and background) as well as the region of interest images. We have used images in which ROI is already extracted. For a real world scenario, images are captured during different time and different weather conditions. Images are also partially occluded with irrelevant objects like pedestrians. The viewing angle for these images is -20 degrees to 20 degrees. The detailed description of each category and training and testing images available for each category is given in Table 1. The images are divided into classes on the basis of make, model, and shape (manufacturing year). The dataset contains thirty six different types (shapes) of vehicles. The dataset contains 2725 images for training purposes and 3110 images for the testing process. Number of available images for testing and training are given in Test and Train columns respectively in Table 1.

4. Methodology

The architecture of VMMR system is given in figure 1. The dataset provides the images in which Region of Interest is already extracted. Hence we are focusing on following three tasks of the VMMR system.

1. Feature Extraction: we have applied Scale Invariant Feature Transform (SIFT) and Speed-Up Robust Features (SURF) in this work.
2. Bag-of-Features (BoF) model: we have used BoF model to create visual dictionaries and encode the local image features into a global image feature vector.
3. Support Vector Machine (SVM): SVM is a supervised learning algorithm and it is used here as a classifier.

4.1. Feature Extraction

The first step for many machine vision applications is feature extraction. The input image is processed and local prominent interest points are located. A robust descriptor is used to encode these interest points which are invariant to many different kinds of noise.

Scale Invariant Feature Transform (SIFT)

Scale Invariant Feature Transform (SIFT) is introduced by David Lowe [2]. We have computed SIFT interest points in every image (training and testing dataset). Every SIFT descriptor construct a descriptor based on the histogram of gradient direction and magnitude around the interest point. The features, extracted using SIFT algorithm, are invariant to rotation and scaling. These features are also not affected by slight changes in view point, noise, and illumination. The operational flow chart of SIFT feature detection is given is figure 2.

SIFT algorithm detects interest point using scale-space maxima detection. Once the robust and invariant interest points are selected; these are encoded into 128-dimensional feature vector based on their appearance in a 4 x 4 patch. We have used standard SIFT algorithm in our work to detect the local image features.

Speed-Up Robust Features (SURF)

Speed-Up Robust Features (SURF) descriptor works similarly to SIFT descriptor but SURF is quicker as compared to the SIFT [3]. Herbert Bay et al. presented the idea of SURF feature detector at the 2006 European Conference on Computer Vision. SURF is partly inspired by SIFT feature detector algorithm. Hessian Matrix is used to detect the interest points in SURF algorithm. The first step is to apply Hessian Matrix on the integral image. The next step is to locate the Extrema and unstable Extremes points are discarded. Finally, orientations are assigned to construct the SURF descriptor. SURF descriptor is a 64-dimensional feature vector. Figure 3 shows the operational flow chart of SURF feature detection.

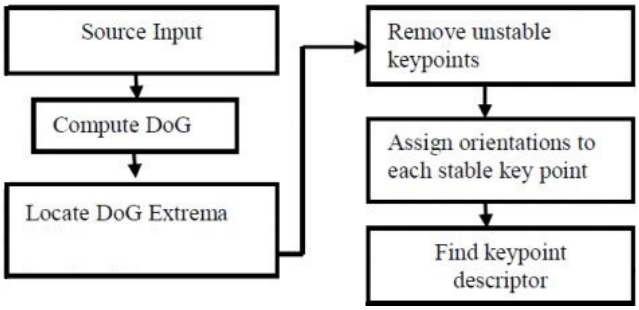


Figure 2: Flow Chart for SIFT feature detection

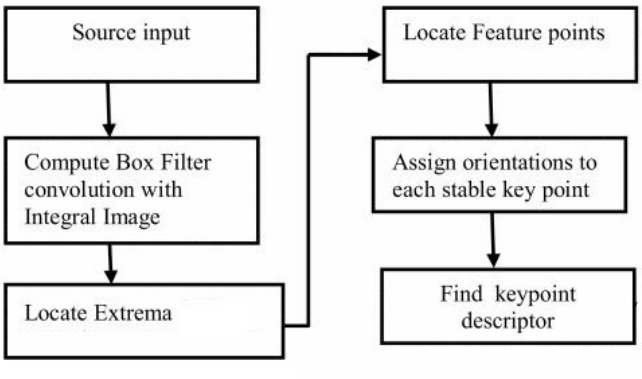


Figure 3: Flow Chart for SURF Feature Detection

4.2. Bag-of-Feature (BoF) Model

Bag-of-features (BoF) model [13] used in this work is based on bag-of-Words (BoW) model. BoW is originally used for document classification; a document is represented using histogram based on the occurrence of words. We must have to create visual dictionaries for an image that can be later used to construct a histogram. Using this histogram of the visual word, an image feature vector is created to represent the image. Whether we are using SIFT or SURF features; both are local image feature and require global feature representation. We will describe the BoF process in terms of SURF descriptor. Same BoF process is applied for SIFT algorithm.

We have used K-means clustering over all SURF features and the centers of each cluster represent a visual word. The collection of these visual words is used to create the dictionary. After creating the dictionary, each SURF feature is mapped into specific visual word and an image is represented by the histogram of the visual words. The dictionary is created by using only training dataset and the same dictionary is used to encode the testing dataset images as shown in figure 1.

Support Vector Machine

Support Vector Machine [4], [5] is binary classifier; multiple binary SVM can be combined to create multi-class SVM. SVM is used for both regression and classification problems. Support vector machine is successfully applied to many real world problems. SVM creates wide decision boundary (as wide as possible) to partition two categories as opposed to other machine learning algorithm; where single line decision boundary is used to separate categories. So SVM is also named as maximum margin classifier. Kernel functions are applied on training dataset to convert linearly non-separable categories into linearly separable classes.

Let assume a training dataset S= $\{(x_1, y_1), (x_2, y_2),\ldots, (x_n, y_n)\}$, where $x_i \in R^n$ and is input feature vector and $y_i \in \{-1, 1\}$ is the category. Let weight and bias of hyper plane is given by w and b. $\varphi(x)$ represents transformed dataset (transformed using kernel function). The hyper plane between the two categories can be defined as:

$$w.\varphi(x) + b = 0 \quad (1)$$

The optimization problem for calculation of w and b is:

Minimize

$$\varphi(w) = \tfrac{1}{2} \|w\|^2 \quad (2)$$

Subject to

$$y_i(w.\varphi(x) + b) \geq 1$$

New variables ζ_i (slack variable) and C (regularization constant) are included in above optimization problem:

Minimize

$$\varphi(w,\zeta) = \tfrac{1}{2} \|w\|^2 + C \sum_{i=1}^{N} \zeta_i \quad (3)$$

Subject to

$$y_i(w.\varphi(x) + b) \geq 1 - \zeta_i, \zeta_i \geq 0$$

Where ζ is used to relax the hard margin constraint and C is used to manage the trade-off between classification error and maximal margin of separation.

One-vs-All [14] and One-vs-One [15] are two popular methods to construct multiclass SVM using multiple binary SVMs to deal with real world scenarios.

5. Experimental Results and Discussion

We have proposed Support Vector Machine based VMMR system and we have used SURF and SIFT features in this work. We have used different sizes and SVM configuration to attain better recognition rate. We have tested our proposed method using publicly available dataset. Figure 4 and Table 2 provides the details about SURF based recognition rate. SIFT based recognition rates are provided in figure 5 and Table 3.

The recognition rate is given on vertical axis and vocabulary size is given on horizontal axis in figure 4. Each graph line is generated for different SVM configuration. The best recognition rate among all the results is 90.90% with using SURF features. As evident from the figure 4 and table 3, better results as compared to others are generated with larger dictionary size and higher value of C. The recognition rate for all the variations is given in tabular format as well. The best recognition rate is achieved at C=8, 10, 12 and the vocabulary size, in this case, is 3000 visual words.

The recognition rate for SIFT image features is given in figure 5. Table 3 shows the recognition rate in a tabular format. The best recognition rate achieved in this case is 92.13%. As compared to SURF, the good recognition rate is achieved at 600 and 1800 vocabulary size; which is less as compared to SURF results. The recognition rate of 92.13% is achieved by using a vocabulary size of 600 visual words.

With these results, we can conclude that the SIFT can achieve better recognition rate as compared to SURF based recognition. However, considering other factors during selection process between SIFT and SURG may change the result.

Figure 4: Recognition rate for SURF features for various SVM configuration and vocabulary size

Table 2: Recognition rate for SURF features for various SVM configuration and vocabulary size

Size	C=2	C=4	C=6	C=8	C=10	C=12
800	88.23	71.93	86.79	86.43	86.30	86.30
1000	89.52	87.78	89.26	89.10	88.88	88.84
1200	86.95	87.33	89.65	89.55	89.36	89.36
1400	87.11	89.49	87.75	87.81	87.81	87.81
1600	88.88	89.78	87.52	87.49	87.56	87.62
1800	87.11	87.78	89.29	89.33	89.16	89.10
2000	89.45	87.59	89.26	89.26	89.23	89.23
2200	89.90	89.33	88.62	88.59	88.46	88.46
2400	77.36	89.16	90.39	90.16	89.97	90.06
2600	86.98	89.07	76.82	76.62	76.27	76.43
2800	90.55	90.16	88.36	88.23	88.17	88.14
3000	89.58	77.07	90.80	90.90	90.90	90.90
3200	90.39	87.91	90.32	90.23	90.23	90.13
3400	89.00	90.84	90.58	90.61	90.58	90.58

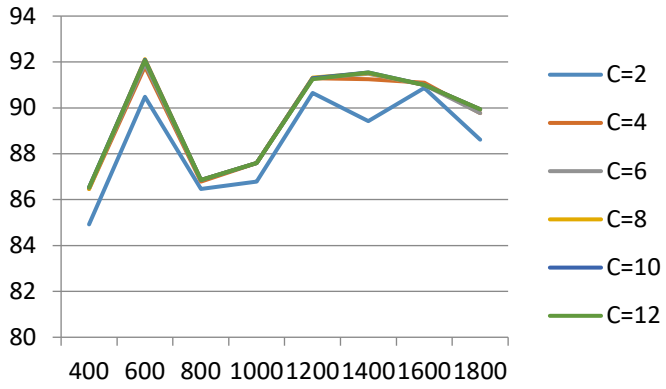


Figure 5: Recognition rate for SIFT features for various SVM configurations and vocabulary sizes

Table 3: Recognition rate for SIFT features for various SVM configurations and vocabulary sizes

Size	C = 2	C = 4	C = 6	C = 8	C = 10	C = 12
400	84.92	86.46	86.53	86.46	86.53	86.53
600	90.48	91.80	91.93	92.12	92.09	92.06
800	86.46	86.79	86.85	86.85	86.85	86.85
1000	86.79	87.59	87.59	87.59	87.59	87.59
1200	90.64	91.29	91.29	91.32	91.29	91.25
1400	89.42	91.25	91.48	91.51	91.54	91.54
1600	90.87	91.09	91.00	91.00	91.00	91.00
1800	88.62	89.78	89.78	89.94	89.94	89.94

6. Conclusion

VMMR is an important component of the intelligent transportation system and smart cities. We have presented support vector machine based VMMR system in this work. This VMMR system takes images as input and extracts the SIFT and SURF features; which are later used during the classification process. Bag-of-features (BoF) model is another component of VMMR system; which provides image feature vector. The proposed VMMR system is evaluated with the publicly available dataset and 92.13% recognition rate is achieved.

References

[1] M. A. Manzoor and Y. Morgan, "Vehicle Make and Model classification system using bag of SIFT features," 2017 IEEE 7th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, 2017, pp. 1-5.

[2] Lowe, David G. "Distinctive image features from scale-invariant keypoints." International journal of computer vision 60.2 (2004): 91-110.

[3] Bay, Herbert, et al. "Speeded-up robust features (SURF)." Computer vision and image understanding 110.3 (2008): 346-359.

[4] C. J. C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," Data Mining and Knowledge Discovery, vol. 2, pp. 121–167, June 1998.

[5] V. N. Vapnik, The Nature of Statistical Learning Theory. New York, NY, USA: Springer-Verlag New York, Inc., 1995.

[6] Boonsim, Noppakun, and Simant Prakoonwit. "Car make and model recognition under limited lighting conditions at night." Pattern Analysis and Applications (2016): 1-13.

[7] Zhang, Feiyun, Xiao Xu, and Yu Qiao. "Deep classification of vehicle makers and models: The effectiveness of pre-training and data enhancement." 2015 IEEE International Conference on Robotics and Biomimetics (ROBIO). IEEE, 2015.

[8] Krause, Jonathan, et al. "3d object representations for fine-grained categorization." Proceedings of the IEEE International Conference on Computer Vision Workshops. 2013.

[9] Boyle, Jonathan, and James Ferryman. "Vehicle subtype, make and model classification from side profile video." Advanced Video and Signal Based Surveillance (AVSS), 2015 12th IEEE International Conference on. IEEE, 2015.

[10] Clady, Xavier, et al. "Multi-class vehicle type recognition system." IAPR Workshop on Artificial Neural Networks in Pattern Recognition. Springer Berlin Heidelberg, 2008.

[11] Zafar, Iffat, Eran A. Edirisinghe, and B. Serpil Acar. "Localized contourlet features in vehicle make and model recognition." IS&T/SPIE Electronic Imaging. International Society for Optics and Photonics, 2009.

[12] [Online]. Available: http://mmplab.cs.ntou.edu.tw/MMR/

[13] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray, Visual categorization with bags of keypoints, in Proc. ECCV Workshop Stat. Learn. Comput. Vis., 2004, pp. 122.

[14] Vapnik, V.: Statistical Learning Theory. Wiley, New York (1998)

[15] Kreßel, U.: Pairwise classification and support vector machines. In: Schölkopf, B., Burges, C., Smola, A. (eds.) Advances in Kernel Methods: Support Vector Learning, pp. 255–268. MIT Press, Cambridge (1999)

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1086-1091 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Estimation of voltage unbalance at an AC traction substation with different train operational scenarios

Thanatchai Kulworawanichpong*, Nattapong Mingpruk, Tosaphol Ratniyomchai

School of Electrical Engineering, Institute of Engineering, Suranaree University of Technology, Nakhon Ratchasima, 30000, Thailand

ARTICLE INFO

Article history:
Received: 30 April, 2017
Accepted: 24 June, 2017
Online: 16 July, 2017

Keywords:
Voltage unbalance factor
AC electric railway
Cyclic permutation of phases connection

ABSTRACT

This paper presents the modeling and simulation of the voltage unbalance in the power supply system of an AC electrified railway. The operating voltage at the traction substation is obtained by MATLAB/SIMULINK program and the voltage unbalance is evaluated by considering the voltage unbalance factor (VUF). Suvarnabhumi Airport Rail Link (ARL), in Bangkok, Thailand, is chosen to be a case study. The ARL is supplied by two single-phase traction transformers connected as a cyclic permutation of phases. According to the extension plan of the ARL, four service scenarios with different headways, auxiliary loads and total weights of the trains were simulated to study how the different scenarios influence the VUF at the three-phase side. It can be summarized that at the same headway, the VUF was increased when the train weight and auxiliary power load were increased. In case of different headways, the longer the headway, the more voltage unbalance. The extension plan of the ARL in the near future can be achieved with the voltage unbalance level under the maximum allowable value.

1. Introduction

This paper is an extension of work originally presented in IEEE/SICE International Symposium on System Integration (2016) [1]. However, the additional experimental pieces proposed in this paper is that the voltage unbalance at an AC traction substation is effected by the three factors of the train operation, including train headway, number of train and additional auxiliary power. This consideration is evaluated for the extension plan of the ARL in the near future. The results would be verified and discussed as follows:

There are presently various forms of ground transportation that help the passengers, particularly the international tourists, to make a decision. The mass rapid transit system is one of the most favorite ground transportation systems and currently vital to developing countries, especially regarding the fundamental economy and society. In Thailand, the tourism industry grows quickly. Therefore, the number of the international tourists at the Suvarnabhumi airport also increased. The cheapest cost and shortest time to travel from the Suvarnabhumi airport to Bangkok city center and other mass rapid transit lines is traveling by the Suvarnabhumi airport rail link (ARL).

*Corresponding Author: Thanatchai Kulworawanichpong, 111 University Avenue Muang, Nakhon Ratchasima 30000 Thailand | Email: thanatchai@gmail.com

The power feeding system of the ARL is the 25 kVrms overhead catenary system receiving electric power from the three-phase AC transmission system. The traction substation transformers are connected as a cyclic permutation of phases, i.e. an alternate phase connection such as A-B, B-C, and C-A. In practical, the voltage unbalance and voltage drop are frequently occurred at the point of common coupling depending on the train operations. In additions, the State Railway of Thailand has planned to extend the service from Suvarnabhumi airport to Don Muang airport in the near future. Therefore, the energy consumption by trains at the traction substation will definitely increase. Obviously, the voltage unbalance at the three-phase side of the traction substation is likely to increase too. To assess the feasibility of the extension plan, four expected train operational scenarios have been simulated by using MATLAB/SIMULINK to evaluate the voltage unbalance and to study the effects of the headways, train weight, and auxiliary power on the voltage unbalance.

This paper presents the study of the voltage unbalance at the traction substation of the ARL in four different scenarios. The contents are divided into six sections. Section 1 is an introduction. Section 2 describes a train movement, power consumption, and voltage unbalance factor (VUF) calculation. Section 3 describes the ARL power supply system and parameters calculation. Section 4 presents the modelling in MATAB/SIMULINK and simulation

procedure. Simulation results and discussion are described in section 5. Lastly, a conclusion is in section 6.

2. Train Movement and Voltage Unbalance Calculation

A train motion is calculated by the Newton's second law of motion taking account of gradients, speed limitation, modes of operation, etc. In order to obtain the required power, speed, and position of the train, it is indispensable to simulate a train movement. In this section, the basic modules within the train simulation are briefly reviewed [1]. In addition, the definition of the voltage unbalance factor is proposed.

2.1. Train movement calculation

A train motion is resisted by several resistive forces, e.g. train resistance, track gradient force, etc. The free body diagram of all forces acted upon the train on an inclined plane is shown in Figure 1. By applying Newton's second law, the train movement equation is expressed in (1):

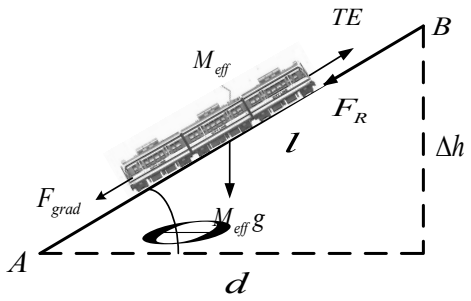


Figure 1. Free body diagram of a train on an inclined plane [1].

$$F = TE - F_{grad} - F_R = M_{eff}\alpha \tag{1}$$

where TE denotes the tractive effort (N), F_{grad} denotes the gradient force (N), F_R denotes the train resistive force (N), M_{eff} denotes the effective mass (kg) and α denotes the train acceleration (m/s^2)

- Tractive effort (TE): the force generated by a locomotive to pull a train [2].
- Gradient force (F_{grad}): the weight of a train opposes the train movement when a train is on a positive slope and helps accelerate when it is on a negative slope. The gradient force is usually expressed in the form of Δh/l, where Δh is the vertical distance, l is the slope length or horizontal distance from point A to B, see Figure 1. With a small arbitrary angle θ the gradient force can be approximated by using (2), where g is the acceleration rate due to gravity of the earth.

$$F_{grad} = M_{eff}\, g \sin\theta = \frac{M_{eff}\, g \Delta h}{l} \tag{2}$$

- Train resistive force (F_R): the motion of the train is opposed by a number of resistive forces. The overall resistance can be formulated as follows [3]:

$$F_R = A + Bv + Cv^2 \tag{3}$$

where v is the vehicle speed (km/h) and the coefficients A (kN), B (kNh/km) and C (kNh^2/km^2) are all constants called the Davis coefficients [4].

- Effective mass (M_{eff}): it is the sum of tare mass (mass of train without loads or passengers) and rotational inertia of the rotating components on the train which is expressed as follows [3]:

$$M_{eff} = M_t\left(1+\lambda_w\right) + M_l \tag{4}$$

where M_t is the tare mass, λ_w is the rotary allowance and M_l is the freight or passenger load.

The modes of train motion considered in this study include accelerating, constant speed, coasting, and braking modes. The speed-distance curve describing each mode of motion is illustrated in Figure 2.

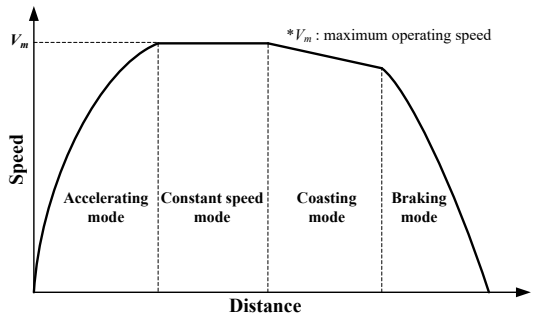


Figure 2. Train operating modes [1].

2.2. Train Power Consumption

The power required by a train P is given by the following expression [5]:

$$P = \frac{TE \times v}{\eta} + P_{aux} \tag{5}$$

where η denotes the efficiency of conversion of the electrical power to the mechanical power, TE denotes the tractive effort, v denotes the instantaneous speed and P_{aux} denotes auxiliary power.

2.3. Voltage Unbalance Factor

The voltages of a three-phase power supply are ideally equal in magnitudes and phase angles of 120º apart from each other. The three-phase system with those conditions are called a balanced three-phase system and its phasor diagram is shown in Figure 3-A. If the three-phase voltage unbalance occurs, both magnitudes and phases are altered, see Figure 3-B. It can be seen that the neutral point (n) will move to the position of n_2. The effects of voltage unbalance, for example, are: the reduction of three-phase induction motor operating lifetime, performance or even failure, etc.

According to IEEE 1159 – 2009, unbalance in a three-phase system is defined as the ratio of the magnitude of the negative sequence component to the magnitude of the positive sequence component, expressed as a percentage. This definition can be applied for either voltage or current. Typically, the voltage

unbalance of a three-phase service is less than 2%, especially when a single-phase load is presented. Mathematically, the voltage imbalance is represented as follows [7]:

$$\%VUF = \frac{|V_{neg}|}{|V_{pos}|} \times 100\% \quad (6)$$

where $\%VUF$ is the percentage of the voltage unbalance factor, V_{pos} is the positive sequence component and V_{neg} is the negative sequence component.

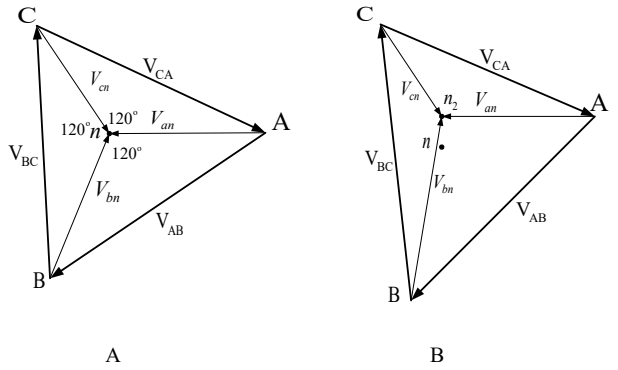


Figure 3. A) Phasor diagram of a three-phase voltage balance and B) Phasor diagram of a three-phase voltage unbalance (modified from [6]).

3. ARL System Description and Parameters Calculation

To obtain the voltage solutions of the three-phase power supply and the voltage unbalance factor at the point of common coupling of the ARL traction substation .The parameters calculation and structure of an AC traction power supply, a transformer connection, and an overhead contact system are briefly proposed as follows:

3.1. Cyclic permutation of phases connection

The connection by cyclic connection of phases as shown in Figure 4 can be considered for the advantages in disposition and easier operation of substations, whereas the three-phase short-circuit power in the primary network is remarkably high if compared with single-phase load power. In this arrangement, the single transformer is fed with two phases and the primary three-phase network can be balanced by systematically distributing the phase connection based on the traction loads [8].

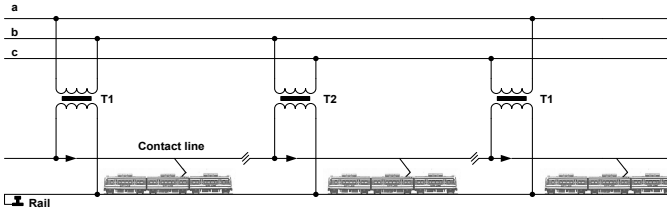


Figure 4. AC traction substation with cyclic permutation of phase

3.2. Structure of the ARL Traction Power Supply

The entire line of the ARL is supplied by a traction power substation (TPS) with 2×20 MVA transformers, located at 8.078 km near Ramkham Haeng station. A minimum short circuit capacity of the utility grid is 2.7 GVA. Transformer's characteristics data are: Ur = 69/27.5 kV, Sr = 20 MVA, Oil Natural Air Natural, Ukr.= 12 %. The single-line diagram of the substation is illustrated in Figure 5.

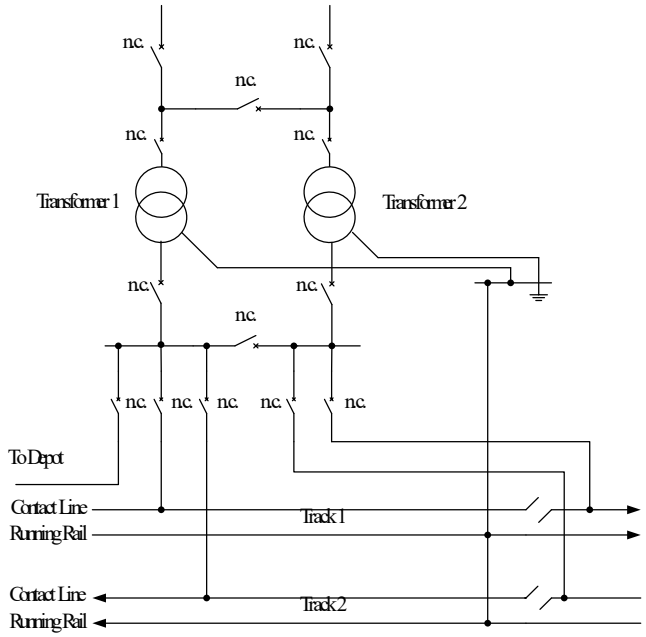


Figure 5. The substation infeed with two transformers [1].

3.3. Overhead Contact System

The overhead contact system in Figure 6 consists of catenary (C), rails (R), and earth (E) described as follows: the catenary (C) is formed by a contact wire and one or two messenger wires. They share the current and are always at the same potential. To reduce the system complexity, their impedances are combined all together to form a single conductor. There are two rails (R) in the same track, which are treated as a single cylindrical conductor. There is a wire called "earth wire" that is used for safety reasons and it normally carries no current. It is assumed that the leakage current from the rails flows underneath the track with a depth of D_E through a virtual conductor called the return earth by (7) and an earth resistivity can be calculated by (8):

$$D_E = 655\sqrt{\frac{\rho_E}{f}} \quad (7)$$

$$\rho_E = 547 f^{-0.434} \quad (8)$$

where f is an operating frequency (Hz) and ρ_E is the earth resistivity ($\Omega \cdot m$).

A reactance caused by electrical induction play an important role in a transmission system for an AC railway power supply. The value of conductor's reactance in the system is dependent on the conductor arrangement and the distance between conductors, see Figure 6. Impedance per unit length of the catenary and running rail is expressed in (9) and (10).[9]:

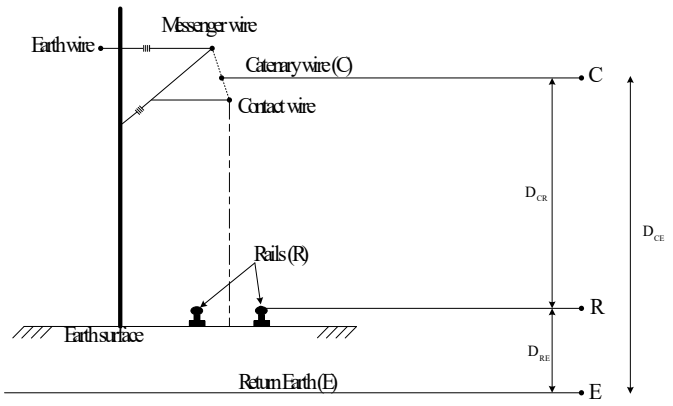


Figure 6. Overhead contact system configuration [9].

$$Z_C = R'_C + j2\pi f \mu_0 \left(\frac{1}{8\pi} + \frac{1}{2\pi} \ln \frac{D_{CE}}{GMR_C} \right) \quad (9)$$

$$Z_R = R'_R + j2\pi f \mu_0 \left(\frac{\mu_R}{8\pi} + \frac{1}{\pi} \ln \frac{D^2_{RE}}{RG \times GMR_R} \right) \quad (10)$$

where

Z_C is catenary's impedance per unit length (Ω/km)

Z_R is rail's impedance per unit length (Ω/km)

R'_C is catenary's resistance per unit length

R'_R is rail's resistance per unit length

μ_0 is magnetic permeability of air

μ_R is magnetic permeability of rails

GMR_C is catenary's geometric mean radius

GMR_R is rail's geometric mean radius

RG is a distance between rails in the same track

D_{CE} is a distance between catenary and earth

D_{RE} is a distance between rail and earth

4. Modelling and Simulation Procedure

The model of the traction transformer, catenary, rails and train of the ARL are modelled by using the built-in blocks in MATLAB/SIMULINK. The configuration of the ARL power supply system is briefly mentioned. Lastly, the evaluation of the VUF at the traction substation of the ARL with different scenarios of the train operations is discussed as follows:

4.1. The model of the transformer in the traction substation

The two traction transformers in the ARL traction substation are alternately connected to different pairs of phases. To estimate the operational voltage and voltage unbalance at the traction substation, the transformer model is formed by using the multi-winding transformer built-in block in MATLAB/SIMULINK program described as follows:

This connection has the same configuration as the single-phase connection except it connects to different pairs of phases. For example, the first transformer is connected to phase AB and the second transformer is connected to phase BC as shown in Figure 7.

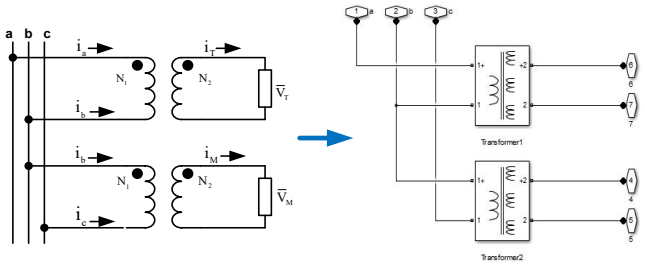


Figure 7. Cyclic permutation of phase connection in MATLAB/SIMULINK.

4.2. ARL traction power supply modelling in MATLAB/ SIMULINK

The AC traction power supply modelling is developed in the block models in MATLAB/SIMULINK. The model consists of 4 main parts: (i) AC traction substation, (ii) transformers, (iii) overhead catenary system, and (iv) electric trains as shown in Figure 8. The 69 kV three-phase AC transmission system is represented by the three-phase source built-in block. The transformers will step down the voltage from 69 kV to 27.5 kV using the multi-winding built-in block connected as a cyclic permutation of phases. The overhead catenary and running rails have the impedances obtained from the textbook [9] and BS EN 50149:2012 standard. The series RLC branch built-in block is used to model this transmission and return circuit system. Lastly, the electric trains are modelled as a series RLC load.

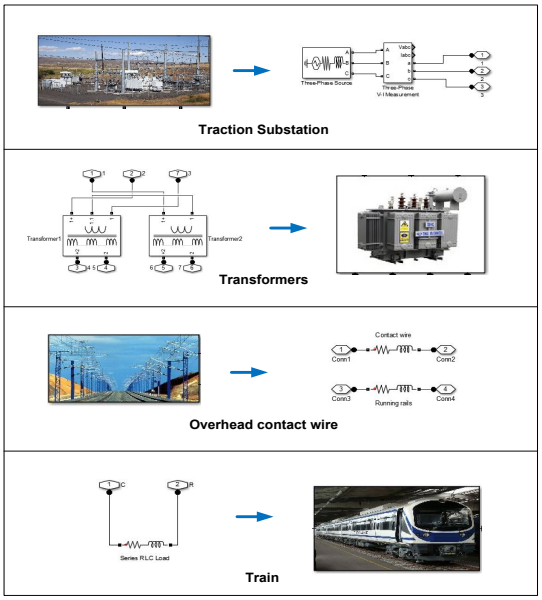


Figure 8. The power feeding components' models in MATLAB/SIMULINK.

The ARL system was modelled for the simulation tests in this paper. The ARL was the first electric mass transit railway operated with overhead catenary system or overhead contact system to commence operations in Thailand. The ARL is composed of 8 passenger stations and one traction substation as shown in Figure 9. Parameters' setting for this simulation can be found in Appendix.

Figure 9. Suvarnabhumi Airport Rail Link map [1].

The AC traction substation of the ARL draws the electrical energy with the voltage level of 69 kV from the metropolitan electricity authority (MEA). The traction substation transformers with an alternate phase connection transform the grid voltage into the train's nominal voltage of 27.5 kV. The total service distance is 28.298 km and the traction substation's position is at 8.078 km. In the simulation, the trains depart from both tracks at the same time with the same headway throughout the service, e.g. the first and fifth trains start the operation at the same time, also the second and the sixth trains and so on as shown in Figure 10.

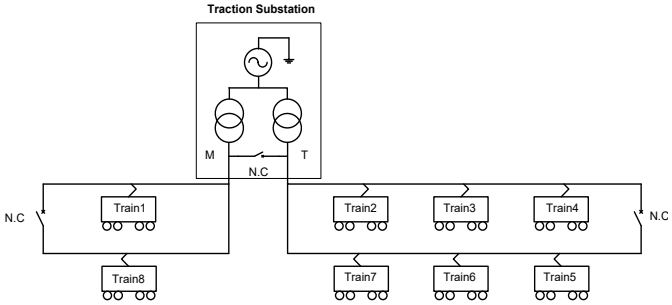


Figure 10. The schematic diagram of the ARL traction power supply system [1].

4.3. Simulator structure

The simulator has 2 major parts: main program (script M-file) and the power network solver (SIMULINK). The main program performs the multi-train movement, performance calculation, and the network capture, i.e. sorting trains on service, setting the distances between the components in the system and the train data. Then, the trains' positions and powers together with the distances recently mentioned are transferred into the SIMULINK block model to solve the voltage solution (primary and secondary voltages of the substation) and VUF. Finally, those values are returned to the main program to be stored and display the graphical illustration. The operating diagram is shown in Figure 11.

5. Simulation Results and Discussion

The results showed the VUFs at the point of common coupling of the ARL traction substation with four different scenarios as shown in Figure 12. The operational conditions, i.e. headways and the train configurations, are given in Table 1. Each train has the maximum speed of 70 km/h and the speed profile is shown in Figure 13.

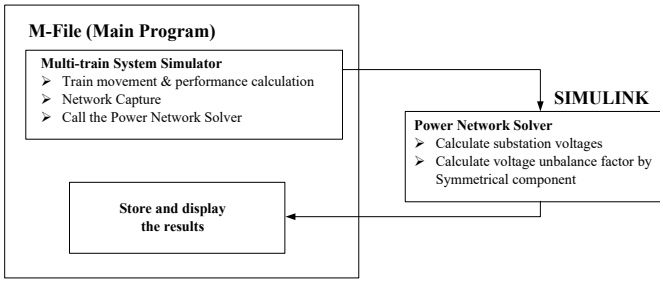


Figure 11. Diagram of the simulation program.

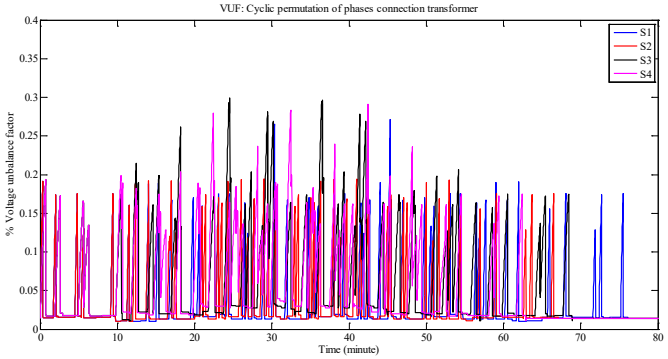


Figure 12. Percentage of VUFs in each scenario

Table 1: Operational conditions of Suvarnabhumi Airport Rail Link.

Scenario	Headway	Train configuration	Total weight	Auxiliary power
S1	15 min	3 car train	183.67 t	200 kVA
S2	12 min	3 car train	183.67 t	200 kVA
S3	12 min	4+3 car train	422.64 t	450 kVA
S4	10 min	4+3 car train	422.64 t	450 kVA

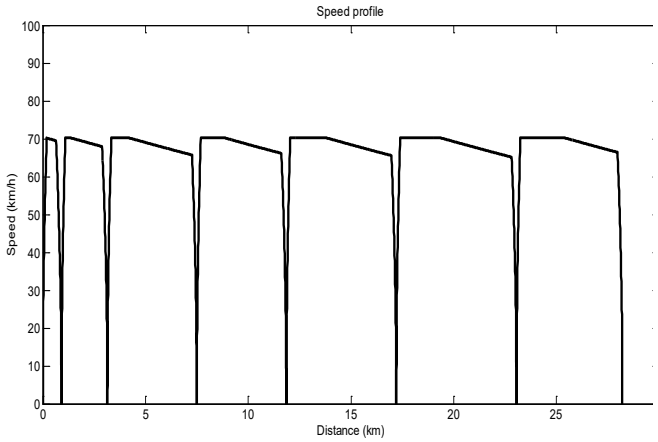


Figure 13. Speed profile.

Considering scenario S1 and S2 in Figure 12, the VUF obtained from S2 was found to be lower than that from S1 because the total weight and auxiliary power of S1 and S2 are equal to each other but the headway of S1 is longer than that of S2. It can be noted that when the headway is shorter, the number of trains appear

in the system, so the total loads on the left and the right of the substation were more likely to have the same values leading to less voltage unbalance. To study the effect of different total weight and auxiliary power of the train on the voltage unbalance, scenario S2 and S3 were compared. As the total weight and auxiliary power load of S3 were greater than those of S2, more power would be consumed in S3, and there was more possibility of different amount of consumed power between the left and right side of the traction substation. Thus, the voltage unbalance of S3 was more than that of S2. In addition, the comparison between S3 and S4 gave the same result as the S1-S2 comparative result. However, due to more weight and auxiliary power load in S3 and S4 compared to S1 and S2, the voltage unbalance in S3 and S4 was more than that of S1 and S2, respectively and the voltage unbalance of S3 was more than that of S4 because the headway of S3 was longer. In the viewpoint of the maximum %VUF for each scenario as shown in Table 2, it helped confirm the above explanation since the results seen from Figure 12 were consistent with the results from Table 2.

From the results, it could be concluded that the longer the headway, the more voltage unbalance. That meant the voltage unbalance was more reduced when the trains were released more frequently. The total weight and auxiliary power also affected the voltage unbalance. According to IEEE 1159 –2009 standard, the %VUF must not exceed 2%. Therefore, the service extension of the ARL can be carried out without any violation of voltage unbalance based on this simulation results.

Table 2: The Maximum %VUF

Scenario	%VUF Max	Comparison with IEEE 1159 -2009 standard (2%)
S1	0.2717	Less than 2% (passed)
S2	0.1946	Less than 2% (passed)
S3	0.2993	Less than 2% (passed)
S4	0.2906	Less than 2% (passed)

6. Conclusion

This paper presents the study of voltage unbalance caused by the AC railway power supply system using MATLAB/Simulink . The VUF was the key factor in this consideration. The test system was the ARL linking Suvarnabhumi airport to the Bangkok city center .The traction transformer connection type was the cyclic permutation of phases .The VUF was estimated in four scenarios of the train operations. The results showed that at the same headway, the VUF was increased when the train weight and auxiliary power load were increased due to the increase in energy drawn by the trains .For different headways, the longer the headway, the more voltage unbalance. However, the increased VUF was still within the limit of 2% imposed by IEEE 1159 – 2009 in every scenario .Accordingly, the extension plan of the ARL by increasing routes, service distance, and the number of trains in the future will slightly affect the voltage unbalance. However the VUF is limited within the acceptable value.

Acknowledgment

We would like to express our sincere gratitude to Suranaree University of Technology for financially supporting this research.

Appendix

Overhead catenary system for ARL:

- Contact wire: Cu 107mm² (20% abrasion)
- Messenger wire: Bz 70mm²
- Return Conductor: E-Al 120mm²
- Rails: UIC 60, 15%

Traction power substation for ARL:

- 2x20 MVA transformers (cyclic connection)
- 69/27.5 kV, 12% impedance, ONAN
- Short-circuit power of the substation is 2.7 GVA

Train resistance parameters:

3-car train	4+3-car train
Tare weight: 131 tons	Tare weight: 299 tons
Rotary portion of tare weight: 6%	Rotary portion of tare weight: 6%
Maximum payload: 44.9 tons	Maximum payload: 105.7 tons
Maximum design speed: 160 km/h	Maximum design speed: 160 km/h
Maximum acceleration: 1 m/s2	Maximum acceleration: 1 m/s2
Maximum deceleration: 0.88 m/s2	Maximum deceleration: 0.88 m/s2
Gear/Motor/converter efficiency: 82%	Gear/Motor/converter efficiency: 82%
Maximum auxiliary power: 200 kVA	Maximum auxiliary power: 450 kVA
Maximum tractive effort: 200 kN	Maximum tractive effort: 400 kN
Train resistance formula: A = 2.1222 B = 0.00618 C = 0.3197	Train resistance formula: A = 4.8438 B = 0.01412 C = 0.53287

References

[1] N. Mingpruk, U. Leetonand, T. Kulworawanichpong, “Modeling and simulation of voltage unbalance in AC electric railway systems using MATLAB/Simulink”. IEEE/SICE International Symposium on System Integration, 2016.

[2] Applications of Mathematics, “Tractive effort, acceleration and braking”, The Mathematical Association, 2004.

[3] S. Lu, “Optimising power management strategies for railway traction”. PhD Thesis, University of Birmingham, UK, October 2011.

[4] BP. Rochard, F. Schmid, “A review of methods to measure and calculate train resistances”. Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit, Vol. 214, No. 4, p. 185–199, 2000.

[5] T. Kulworawanichpong, “Optimising AC electric railway power flows with power electronic control”. PhD Thesis, University of Birmingham, UK, November 2003.

[6] M. Kalantari, M. J. Sadeghi, S. Farshad, S. S. Fazel, “Modeling and Comparison of Traction Transformers Based on the Utilization Factor Definitions”, Internation Review on Modelling and Simulation (I.RE.MO.S.), Vol. 4, N. 1, pp. 342 – 351, February 2011.

[7] IEEE recommended practice for monitoring electric power quality, IEEE std 1159 – 2009 (Revision of IEEE std 1159 - 1995), pp. c1 – 81, June 2009.

[8] F. Ciccarelli, M, Fantauzzi, D. Lauria, R. Rizzo, “Special Transformers Arrangement for AC Railway Systems,” Department of Electrical Engineering – University of Naples Federico II, Naples (Italy), 2014.

[9] F. Kiessling, R. Puschmann, A. Schmieder, E. Schneider, Contact Lines for Electric Railways, 2nd revised and enlarged edition, pp. 664 – 666, 2009.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1092-1106 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

A Microwave Reflectometer Prototype for Detection of Body Through the Rubble of Collapsed Buildings

Syed Daniyal Ali Shah*,

Department of Electrical Engineering, COMSATS Institute of Information Technology (CIIT), 54000, Lahore, Pakistan

ARTICLE INFO

Article history:
Received: 29 April, 2017
Accepted: 24 June, 2017
Online: 16 July, 2017

Keywords:
FMCW
ISM Band
MATLAB
Microwave
Radar
Reflectometer
Signal Processing
Transceiver

ABSTRACT

There is a growing interest in using radio frequency in imaging for problem solving in situational awareness, surveillance, search and rescue operations, and screening for security. This has been fetched from recent development in addition to innovative technologies in microwave radar imaging applications. Despite the fact that no new phenomenology is exposed in this field, the prospects to eventually realize systems that operate under adverse weather conditions, interferences and obscuration has prompted newer prospects. However, the existing systems for RF imaging are not easy to afford and are not easily available and there is a need to make affordable imaging available to support people's lives. The effort provided an economic, compact and easy solution for human body detection applications using Radio Frequency (RF) by designing and developing a microwave transceiver that can give data in a suitable form for further signal processing and data acquisition. In this project, different types of RF imaging techniques and the most feasible techniques that may be employed for under debris search and rescue operations were studied. Link budget analysis and software simulations were conducted. Individual components along with the transmitting and receiving chains were tested on VNA. Gain and radiation pattern of antennas was determined in anechoic chamber. During this project the performance of transceiver, regarding its detection ability, range resolution, signal to noise ratio (SNR) etc., were also investigated. At the end a prototype FMCW transceiver is given that is able to detect human body under debris or behind a walls whether living or dead, which is effective at transmitting as well as receiving the reflections and giving data in an audio format for further signal processing.

1. Introduction

This paper is an extension of work originally presented in 19th IEEE International Multi Topic Conference (INMIC 2016) [1]. The initial scope of the project was to build antennas and conduct software simulations to define the parameters on which a transceiver could be built like type of transceiver design that could be employed, operating frequency, transmit power and bandwidth etc. Furthermore, the work was extended to develop and test an economical and simple transceiver system to demonstrate the concept and feasibility of the prototype. Not all of the initial proposals could come to reality but a transceiver reflectometer was developed to demonstrate the feasibility of proposal. An Analog to Digital Converter (ADC) and a Data Acquisition Device (DAQ) used for conducting experiments and collecting data. There was no dedicated electronic hardware used for data collection and generating image, which reduced the complexity of the technical requirements, but imposed certain constraints on the capability of the system. This project was to provide foundation research and work for more complex system to be built in the future. Figure-1 has been presented for thorough understanding of the project and an idea of what is being conceived through the project. The project outcome may be transformed for commercial use.

The data given by radar contains the bearing and also range to the Transceiver and the body which is buried under debris. Hence, it is employed in a number of areas in which the requirement for this kind of locating is vital. Pakistan is located at the north far western region of the Indian subcontinent which also overlaps with the Indian as well as the Eurasian tectonic plates. The Sindh and also some of the regions in Punjab province rest on the north

*Corresponding Author: Syed Daniyal Ali Shah, House No 135, Street 3, Block C, DHA Phase 8, Lahore Cantt, Pakistan. +92-321-3141555, 9daniyal@gmail.com

western nook of the Indian tectonic plate, whereas, Baluchistan province as well as the vast majority of the Khyber Pakhtunkhwa exists within the Eurasian tectonic plate which primarily makes up the Iranian plateau. The Northern Regions as well as Azad Kashmir rests mostly in Central Asia besides the edge of the Indian tectonic plate, therefore, are vulnerable to ferocious earthquakes exactly where both the tectonic plates collide. Due to the lack of advanced equipment a lot of lives are lost during these natural disasters. If rescue response teams are able to equip themselves with devices that are able to detect life under debris, the amount of lives that can be saved by this device are unimaginable. The objective is to make a microwave transceiver that is able to detect living beings under debris and is man pack handheld.

The capability to discover moving objects inside a structure or behind the wall [2] is also one of the noteworthy applications of the RF transceiver. The transceiver delivers users with situational knowledge for an operational benefit. The through the wall detection capabilities are priceless for both equally the Armed forces and Law Enforcement agencies associated with tactical operations in Urban Terrain. Military Operations in Urban Terrain (MOUT) necessitates successful attainment of military aims. These capabilities significantly improve the pre-planning and also execution of military operations.

An imaging transceiver model comprises a wireless transmitter that produces radio waves in specific direction. Whenever, these waves come in contact with a physical entity they are generally echoed and also dispersed in lots of different directions. The objective is to make some of these radio waves echo back form a living body the place where channel do not favor optical visibility. A Figure 1 of such a transceiver system is shown below which is being conceived for the project.

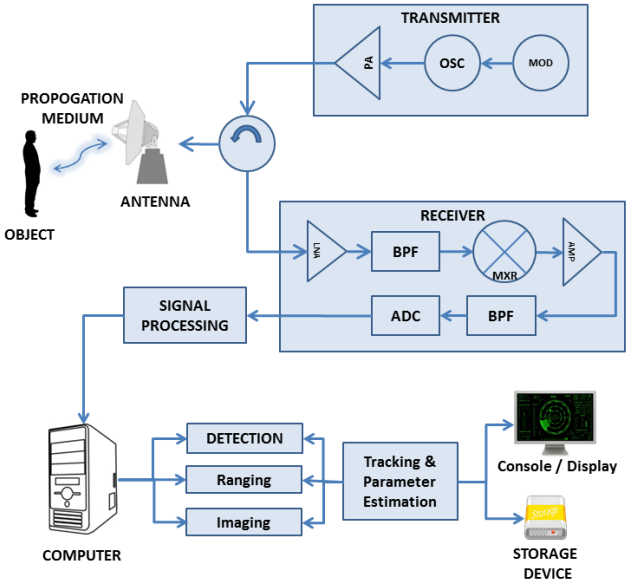


Figure 1 - Overview of Generic Transceiver System.

A link budget calculation of all of the gains and losses of the components and parts of transceiver will be presented and also analyzed. Additionally, the losses inculcated as a result of propagation through the channel would certainly be taken into consideration. Albersheim's equation [3] is used to figure out the needed Signal to Noise Ratio (SNR) for the particular detection and also false alarm probabilities as well as the maximum detectable range for monostatic radar [4] is going to be determined and investigated for the desired range of operation. Comparing between different types of radar techniques, specifically Pulse Doppler (PD) as well as Continuous Wave (CW) [5], the most appropriate technique is going to be implemented for transceiver design. The true to life differences between them may very well rely more upon the compromises than the hypothetical attainable. Antenna selection as well as design is among the most crucial areas of a wireless system for dependable Radio Frequency (RF) communication. One grave constraint using this elucidation is that the Electromagnetic (EM) waves are emitted in an Omni direction pattern, so the quantum energy in the bearing being studied is a very trivial portion of that radiated energy from the antenna. To obtain an acceptable amount of received power, the transmitting antenna should be directional.

1.1. Project overview

To identify the issues in the creating of RF system the properties of such system were investigated along with the environment. Receiver as well as transmitter architectures were studied and also their viability in existing Integrated Circuit (IC) solutions for instance superheterodyne (superhet) technology [6], direct detection and power meters. The operating frequency selected for the project is in industrial, scientific, and medical radio band (ISM band) exactly 2.4 GHz range [7], and follows maximum transmitting output power limit of 30 dBm or simply 1 watt in accordance with the government regulations [8]. In addition to mere component accessibility numerous other issues were also well thought-out while developing and planning this system. Field testing of the prototype was carried out and outcomes are reported. It was summarized by looking at upcoming future directions in RF microwave transceiver concepts and improvements that can be made to the existing system.

2. Design Objectives

There is certain required criterion that need to be met in order to work in conditions the device is intended to operate and function. There are also certain government regulations that need to be met so that the equipment in utility is within the legal boundaries.

2.1. Low cost

The transceiver needs to be low cost, it needs to be easily affordable for governmental and non-governmental organizations, for a poor country with low economic conditions and meagre resources.

2.2. Compact size

The transceiver needs to be compact in size so that it can be easily moved and operated in affected areas like a region struck with earthquake.

2.3. Easy to operate

The transceiver needs to be user-friendly and easy to learn, as complex equipment requires specialized training which is not easy to resource and afford.

2.4. Low maintenance

Transceiver should be low on maintenance as higher maintenance costs needs higher affordability.

2.5. Low operation cost

Transceiver need to have lower operating costs as it matters to buyers why they are operating in large quantities.

3. Regulations requirements

The PTA (Pakistan Telecommunication Authority) guidelines for unrestricted wireless apparatus functioning in the ISM frequency bands [9] is restricted. The ISM RF frequency bands were initially reserved for electromagnetic frequency that is generated by industrial, scientific as well as medical (ISM) apparatus. In order to abide by these rules, the equipment bandwidth will remain within the ISM Band of operation particularly 2.4735 to 2.4835 GHz which also gives us the bandwidth availability of 10 MHz only. Max wireless radiated power to be fed to the antennas should be less than 30 dBm or 1 watt. The obligation for remaining within all these radiated power restrictions falls on the manufacturer. Even if the equipment is professionally built by specialists, or in case if it is self-designed. In order to remain within these regulations, the maximum wireless transmitter output power to be fed to the antennas should not exceed 30 dBm and Effective Isotropic Radiated Power (EIRP) should not exceed 30 dBm. Keeping this in mind it was decided to keep the maximum transmit power at 29 dBm or 794 mW.

4. Zero IF Design

A Direct Conversion Receiver also referred to as Zero Intermediate Frequency (IF) Receiver, synchrodyne or even homodyne is a RF receiver model that down converts the arriving RF signal making use of synchronous demodulation detection which is driven with the help of oscillator, the frequency of which is equal to or close to the transmitted frequency of the expected signal [10]. Whereas, in the case of regular superheterodyne receiver in which this really is achieved after transformation to an IF (Intermediate Frequency).

The simplicity of carrying out merely a single conversion of frequency decreases the elementary RF circuit complication, however, other concerns relating to dynamic range do ensue. In its unaltered form it turned out unsuitable to obtain Amplitude Modulation (AM) as well as Frequency Modulation (FM) radio signals without employing a Phase Locked Loop (PLL). Even though these kinds of engineering problems realized this approach rather unrealistic when the technique was in its initial development stages in 1930's. Present technological innovations and software defined radio have particularly re-established its use in various areas such as certain consumer based equipment [11].

4.1. Zero IF Design

Even though the superhet method happens to be effective receiver architecture, it really is nowadays finding competition from Direct Conversion Receiver (DCR) technique also known as the Zero IF receiver (Figure 1) [6]. Here, the Local Oscillator (LO) frequency is placed very near to RF transmitted frequency of the acquired signal. The mixed output is directly at baseband, and also there is absolutely no IF stage required.

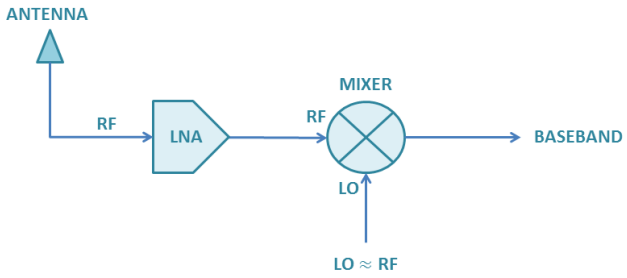


Figure 2 – Block Diagram of Zero If Design.

The RF signal is fed into the mixer in Direct Conversion Receiver the same way as in superheterodyne receiver [6]. In contrast to the superhet, the frequency of LO does not have an offset from RF frequency. The outcome is a demodulated baseband output signal just like the one obtained from a superheterodyne technique employing synchronous detection following an IF stage [10]. To put it differently, the down conversion is performed in just one frequency transformation. This prevents the complication of superhet like image rejection problems, IF stages and a couple of frequency conversions.

Despite the fact that this technique decreases the basic circuit complication in principle, it places serious requirements on almost all stages, as well as noise, stability, dynamic range, tuning range, and distortion. For certain very carefully chosen as well as developed purposes, ICs could make Zero IF receivers competitive or even better than the superhet with its IF stages.

4.2. Technical Problems

To come at par with the efficiency of the superheterodyne down conversion, several of the features typically sorted out by the IF phase needs to be achieved at baseband. Since there are no IF amplifiers with high gain, therefore, by making use of Automatic Gain Control (AGC) the power level at baseband output contrasts over a very wide dynamic range. This really is major engineering challenge which restricted the feasibility of the model. An additional concern is the incapability of the model to employ envelope detection of amplitude modulated signals. Hence, reception of AM or FM signals directly requires locking of the phase of the local oscillator to the transmitted frequency, an even more challenging job as compared with the more widely used envelope detector at the output port of an IF implementation stage in superheterodyne model. Nevertheless, this may be prevented when it comes to a Direct Conversion model making use of quadrature detection and after that Digital Signal Processor (DSP) is employed [12]. Making use of software defined radio approach, the two quadrature outputs could be processed so that any kind of demodulation can be carried out as well as filtering on baseband signals from frequencies around the LO frequency [12]. The significant increase of digital equipment, together with improvements in the analogue components associated with the frequency transformation to demodulated signal, has consequently created this more straightforward topology effective in several applications.

4.3. Advantages

The beat frequency signals which are unwanted by products from the frequency mixing stage do not require any additional signal processing because they are rejected by using a Low Pass

Filter (LPF) at the output. The demodulator model carries the added benefit of higher selectivity, which is for this reason a precision demodulator. The model concepts could be expanded to allow splitting up of adjoining channel transmitted signals in which sidebands might overlay the desired transmitted signal. The model furthermore enhances the detection ability of pulse modulation transmitted signals.

5. Modulation Requirement

Continuous Wave Frequency Modulated (CWFM) also called Frequency Modulated Continuous Wave Radar (FMCW) radar [13] is a small distance calculating radar efficient at figuring out range. This enhanced dependability by delivering range measurement together with velocity measurement, that is certainly important if multiple sources of reflection are arriving at the radar antenna. Additionally, it is put to use as wave radar, proximity sensors and early warning radar. Doppler effect is not necessarily required for identification when frequency modulation is employed.

In such a transceiver the Transmitted (Tx) signal of a well-defined steady continuous wave frequency changes back and forth in regularity over a set time frame by a controlling modulated signal. The distinction of frequency between the receiver and the transmitted signals raises with delay, together with range. This smudges away the Doppler frequency. Reflections from an object are after that mixed with the Tx signal to generate a beat frequency which will lead us to the range of the object after demodulation.

There are lots of feasible modulation forms [14] that can be employed for various measurements & requirement needs but saw tooth or triangular modulation scheme is of the interest. The saw tooth wave modulation pattern is employed in comparatively large distance measurement applications along with a minimal effect of Doppler shift in frequency for instance, a maritime radar for navigation [15]. Saw tooth modulation is the single most widely employed in FMCW radars when distance measurement is favorite for targets that do not have moving elements. Range data and the Doppler speed are mixed by means of this approach. Modulation could be switched off on alternative scans to recognize speed making use of carrier frequency shifts of unmodulated wave. This enables distance and also speed to be found with a single reflectometer set. Triangular waveform modulation is often used to accomplish the identical objectives. In a triangular shaped frequency changing, a range calculation is usually carried out on both the increasing and also on the dropping edge simultaneously, the precise Doppler shift in frequency could be figured out from two measurements. The shift between the two different frequencies is two times the Doppler frequency [15]. Due to the fact the two differential frequencies are not at the same time available this comparison demands digital signal processing, with saving the measured results and processing them later. This modulation enables simple splitting up of the difference frequency Δf of the Doppler shift frequency f_D.

Distance is restricted to one fourth of the wavelength of the transmitted modulation. Measurable distance for 100 Hz FM will be 500,000 meters. That restriction is dependent upon the kind of modulation as well as demodulation. The succeeding normally is applicable [15].

$$IR = f_r - f_t = \frac{c}{(4 \times \lambda)} \quad (1)$$

Where:

$IR = Instrumentrd\ Range$
$c = speed\ of\ light$
$\lambda = Modulated\ Frequency$

The radar will certainly record inaccurate ranges for echoes from ranges beyond the distance resolution, for example from the moon. Range determination for FMCW are merely dependable around 60% of the instrumented distance, or even for about 100 Hz FM is 300 km. In order to maintain versatility of the transceiver system the majority of the modulation patterns such as sinusoidal, triangular and Frequency Shift Keying (FSK) are integrated in the transceiver design but triangular modulation scheme for the intended purpose was used.

6. Dynamic Range

The transceiver needs to be effective at dealing with RF signal reflections from surface of debris and also shorter range objects as well as sense minute RF signals in close proximity to the noise floor form the human skin. The ratio of the highest receivable RF signal to a bare minimum of detectable reception is known as the dynamic range which is calculated for the system as below [4]:

$$Dynamic\ Range = 20 \log \left(\frac{Vrms_{max}}{Vrms_{min}}\right) \quad (2)$$

$$Dynamic\ Range = 20 \log \left(\frac{0.071\ v}{0.018\ v}\right) \quad (3)$$

$$Dynamic\ Range = 11.9\ dB \quad (4)$$

It is actually 11.9 dB for a particular transmit bandwidth. The highest energy of receiver RF signal in terms of maximum Root Mean Square (RMS) voltage at receiver should not saturate the transceiver front end as well as considering a 14 dB Low Noise Amplifier (LNA) is also given to the acquired signal, it is also the highest sampling rate of the ADC. Furthermore, for a FMCW transceiver, the greatest RF signal in the acquired bandwidth is determined by a compression point of 1 dB the down converter mixer in which the gain turns from linear behavior this shoulders virtually no saturation once the mixed signal is achieved. A bare minimum of detectable RF signal power in terms of minimum RMS voltage at receiver needs to be above the received background noise level as well as have a minimum detectable SNR. In the majority of reflectometer in which detection of target is the primary application, SNR of 8 dB is mostly required [16]. In the majority of applications in Ground Penetration Radar (GPR), a minimum voltage also needs to have a nominal amount of signal to clutter ratio (SCR) so that the received signal is capable of determining and also identifying the target profile.

The FMCW transceiver's dynamic range will certainly have an effect on the greatest distance at which an object could be identified. Usually, radars will have a greater equipment dynamic range as compared to sampling dynamic range. The ADCs dynamic range in dB is as comprehensive as $20\ log\ (2^N)$, where N is the amount of bits or even roughly N times 6 dB [17]. Consequently, a 16-Bit ADC will have a 96 dB of calculated dynamic range. However, additional elements impact the authentic attainable dynamic range [18]. There are techniques for

instance like stacking which might be employed to enhance the SNR and also AGC to get over the issue of dynamic range variances [16].

7. Project Architecture

7.1. Methodology

An FMCW radar comprises fundamentally of a transmitting chain and a receiving chain collectively known as transceiver along with a data acquisition device. The transceiver is a small in size, portable and a light weight module which typically consists of the antennas integrated as standalone transmitter as well as receiver antennas. An RF wave typically in microwave region is produced by a Voltage Controlled Oscillator (VCO) which can be a modulated waveform which directly feeds the transmitting antenna, or even its strength may also be amplified by employing a power amplifier. A typical form of transmitting chain is shown below.

Figure 3 - Block Diagram Transmit Chain.

If a single antenna is being used, then a ferrite circulator then to differentiate the receiving and transmitting signals is essentially required, since simultaneously transmitting as well as receiving signals are detected by the FMCW radar. When the Tx and Rx antennas are close to one another there is a direct positive feedback between the two antennas which is like crosstalk in terms of communication. In order to mitigate this typically, the antennas employed are polarized, the polarization direction is rotated by 180° against one another. In order to further decreases crosstalk a shielding plate is also used between the antennas. Considering that the measurement is conducted as a frequency variance between the transmitting and receiving signal, the RF signal that is generated by this direct coupling could be suppressed or filtered because the frequency output is very less as compared to the coupling frequency.

Depending on the application of CW radars solely the Doppler frequency has to be processed, when transmitting a constant frequency or alternatively in the event that very low modulation frequency is applied as compared to the transmitting frequency. This consists of frequencies just approximately 16.5 KHz by applying an FMCW transceiver working in K–Band (24 GHz) and the estimated speeds for recording are around 360 Km/h. Similarly, for the case in which detection of the object is the sole purpose of the transceiver the system does not require very high received signal bandwidth. For that reason, as microprocessor there may be a regular stereo audio processor, that is manufactured in large quantities as well as being used for instance in audio cards for home laptops or computers. Even just in the rectangular shape modulation (FSK) this kind of a processor can be employed conditionally.

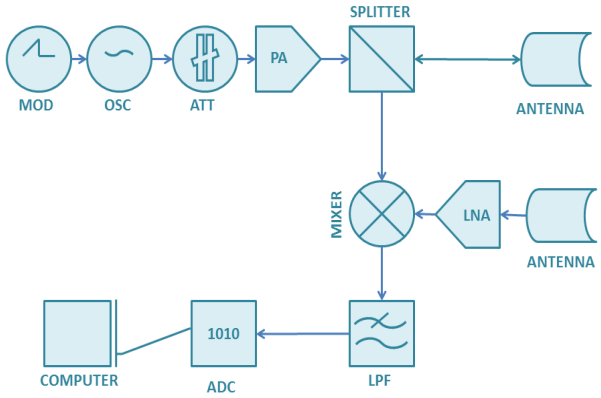


Figure 4 - Block Diagram Generic Transceiver Design Without Circulator.

One of the difficulties of this system design is that the baseband signal also varies over the bandwidth if the transmitted signal. This issue is resolved in signal processing of the received signal in which the reference signal is parsed in to data, wave by wave, with respect to a reference signal which is a square waveform.

7.2. Link budget calculations

Keeping in view all the design criteria, constraints, component losses and previous analysis of the losses and absorption of building materials following designs are visualized. A separate transmitter and receiver design is given below in which approximately 90% signal is lost due to reflection, absorption or attenuation. Signal propagation or path loss models [19] were not calculated although a rough estimation of the propagation losses by carrying out literature review on attenuation in propagation was considered. The transmitted signal bandwidth is kept at 300 MHz which is the maximum bandwidth of the system this allows to achieve a decent range resolution.

Since VCO source that could produce such a high power signal was not available so an amplifier was used to increase the power

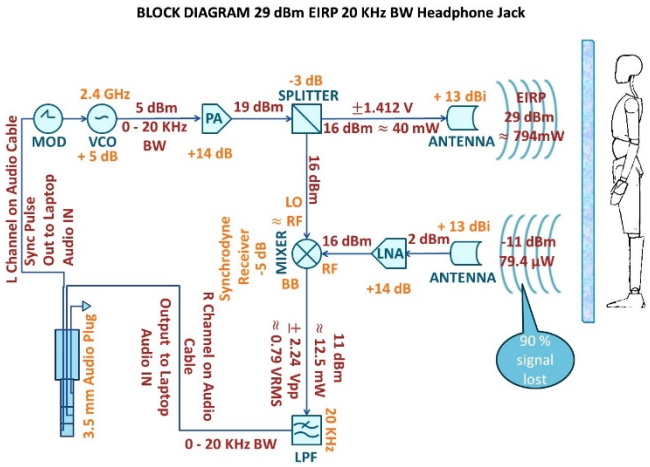
as desired. This lead to an increase in components employing an attenuator in addition to an amplifier lead us to the desired power of 16 dBm. Although this produced the desired results but also increased the noise floor unnecessarily.

An innovative way was devised to decrease the noise floor as well as to reduce the quantity of components required. This ultimately lead to the decreased cost of transceiver which was also one of the main goal to deliver a low cost solution. After transceiver design research and looking deep into radar operational fundamentals [20] a design which is simplistic yet effective for the detection application was conceived.

If the transmit bandwidth of 300 MHz is maintained which is within the frequency range of Data Acquisition Devices (DAQ's) and if the modulation frequency to the VCO is known in real time, then it can be directly down converted from 2.4GHz using stolen carrier from the transmitter. But it is coupled with a problem that the received signal varies over the frequency range of 300 MHz. This seems like there is high rate frequency hopping in the receiver system, which can be tackled in signal processing, when received data is parsed taking reference from the synchronization pulse from the modulator. This design is given below assuming similar characteristics of propagation loosing 90% of the received signal.

Now a data acquisition device had to be acquired it was found out that data acquisition devices that are available nowadays are quite big to be portable enough to use in the application so it required to figure out how to acquire signal using some handheld portable device. It was later suggested that if the transmitted bandwidth is kept at or below 20KHz which is within the frequency range of most audio processors and at such low frequencies (20 KHz) can be directly down converted from 2.4GHz using stolen carrier from the transmitter. The output can then be directly fed to a sound card of a computer or audio processor but the bandwidth of the system is greatly reduced which resulted in decreased range resolution, this technique could be useful in detection but this was not supportive in range measurement or object profiling.

If the transmitted bandwidth is kept in audio range (20 KHz) which can be processed by audio processors and at this frequency range the direct down conversion from 2.4GHz using stolen carrier from the transmitter varies over the frequency range of 20 KHz this indicates there is frequency hopping in the system. This issue was dealt by taking a signal from the modulator which is in Synchronization with the input frequency of VCO this way the data can be parsed wave by wave and hop the frequency at the same rate. The synchronization frequency is given to the left channel of audio jack and the received signal is given to the right channel of the audio jack. The acquired results can be directly fed in an audio jack input of any computer instead of using a DAQ or any other ADC. This greatly reduces the cost of the transceiver as now days a lot of high speed precision DAQ's are quite expensive. The design is given below.



This is the final design of the Microwave transceiver prototype that was implemented.

7.3. Modulator

In order to provide input to VCO's V-Tune XR-2206 IC was utilized. The XR-2206 is a function generator incorporated monolithic circuit effective at generating good quality sine, rectangular, triangle, ramp, and also impulse wave-forms of great stability as well as precision. The generated wave-forms could be both frequency as well as amplitude modulated by an external voltage source. Frequency could be chosen by an external input over a frequency range of 1 Hz to in excess of 1 MHz. The circuit is a good choice for telecommunications, instrumentation applications, and also function generator uses demanding sinusoidal signal FM, AM, or even FSK production requirement. The oscillator's frequency could be swept linearly over a 1:2000 frequency band with an external voltage control, while maintaining minimal signal distortion level.

The circuit designed has a voltage supply of 9-12 V_{DC}. The waveforms that can be generated are Square, Sine & Triangle. The impedance is 600 Ohm + 10% and the frequency range of operation is 1Hz - 2MHz.

The characteristics of the waveforms are given below at the input voltage of 9 volts Direct Current (DC) that can be employed to achieve results.

- Sine Waveform

Amplitude:	0 – 8 V
Distortion:	> 1% (at 1 kHz)
Flatness:	± 0.05 dB (1 Hz – 100 kHz)

- Square Waveform

Amplitude:	0 – 8 V
Rise Time:	> 50 ns at 1 kHz
Fall Time:	> 30 ns at 1 kHz
Symmetry:	> 5% at 1 kHz

- Triangular Waveform

Amplitude:	0 – 8 V
Linearity:	> 1% up to 100 kHz

The square wave is in frequency synchronization with triangle and sine wave forms so it can be used as sync pulse for Data processing [3]. Frequency chirp was carried out in order to cover the bandwidth using triangular wave so demodulated waveform is also scattered over the bandwidth as stolen carrier from the transmitter is being taken.

7.4. Audio Amplifier / Low Pass Filter

A LPF permits signals with a frequency less than a particular cut-off frequency as well as attenuates the frequencies more than the cut-off frequency. The degree of attenuation for every single frequency is dependent upon the LPF model. A Sallen Key Low Pass filter was used to get rid of electronic noise in a circuit, however, it may also be employed to avoid aliasing by restraining the frequencies of signals to be dealt with. The filter response could be transformed by replacing an individual resistor or even capacitor value.

In the case, LPF is utilized to decrease signal noise. Noise is a group of undesirable, random electronic signals that can be found in RF electronic circuits. Almost all RF devices are prone to the unwanted effects of noise, especially the ones that operate at comparatively lower energy signals. One more use for filtering is to avoid aliasing at the time of signal processing. Additionally, an LPF was employed to make sure that only lower frequency signals are allowed to pass through and high frequency noise is going to be discarded.

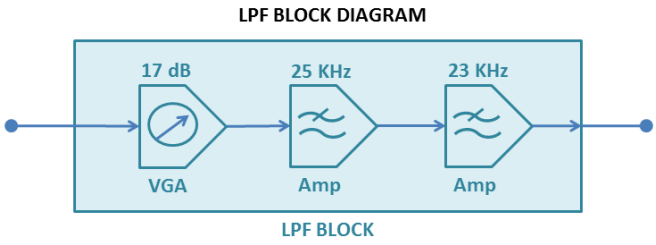


Figure 6 - LPF Block Diagram Representation.

The type of filter employed was Sallen Key Active filter topology simply because it provides a straightforward to execute filtering topology. The cut-off frequency can be adjusted by altering an individual resistor or a capacitor value. The filter constructed in Lab was tested on an oscilloscope at multiple frequencies and a bode plot was produced for the analysis of the constructed LPF. One such reading is shown. The yellow waveform is the input frequency signal and the blue waveform is the output signal that has been passed through the Sallen Key Low Pass Filter.

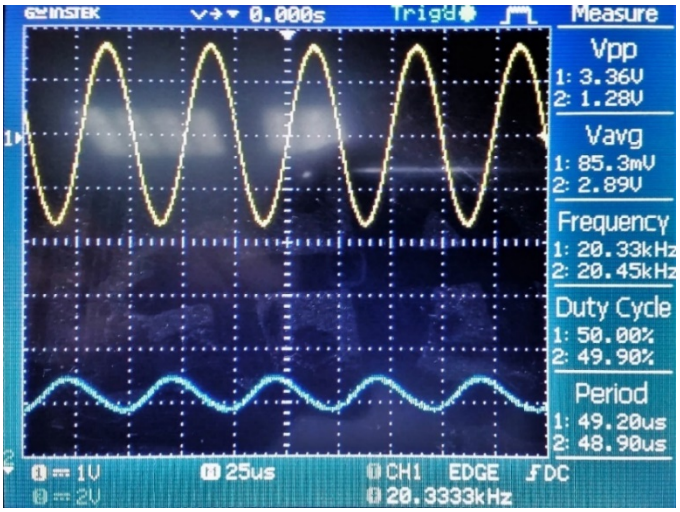


Figure 7 - Sallen Key LPF Result at 20 KHz.

The filter is functioning as desired. A -3dB gain factor was desired at 20 kHz and the filter produced slightly higher attenuation than -3dB. Calculations have been given below:

$$V_{in} = 3.36\,V \tag{5}$$

$$V_{out(IDEAL)} = \frac{V_{in}}{\sqrt{2}} = \frac{3.36\,mV}{\sqrt{2}} = 2.37\ V \tag{7}$$

$$V_{out(actual)} < V_{out(IDEAL)} \tag{8}$$

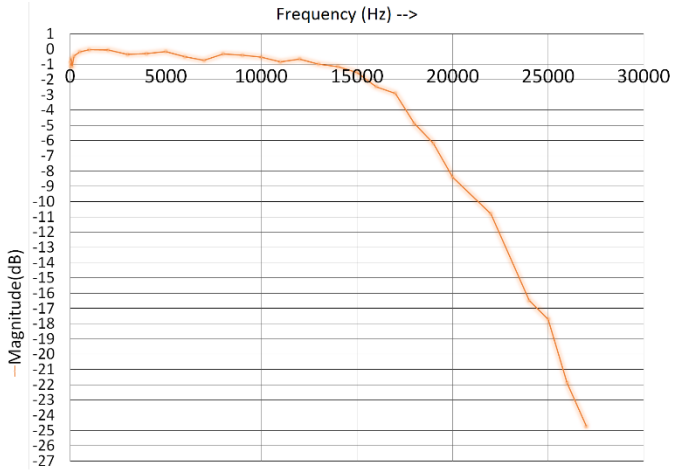


Figure 8 - Bode Plot 20 KHz LPF.

Since the filter is producing better than -3dB attenuation at 20 kHz, the Filter is working better than expected. This is likely due to parts tolerances that varied between 1 % to 5 %. Due to the ease of design and implementation, a Sallen Key filter is the perfect candidate for electronic filtering. It is ideal for eliminating electrical noise and since Operational Amplifiers (OP AMPs)

draw a negligible amount of current, the filter will not load the circuit to which it is attached.

7.5. Understanding Microphone Input Sensitivity

As the analogue to digital microphone input includes different units, comparing a single type to another can be difficult to understand; however, they share a very common unit of measurement in the audio frequency signal domain, input power level in dBV [21]. One may have an analogue voltage input or a PDM modulated input, however, their maximum acoustic input power as well as SNR can be instantly analyzed. The difference between the 94 dB Sound Pressure Level (SPL) reference and the noise floor of microphone input of 65dB, altering the input relative to the 94 dB SPL, would result in changing dBV output at the microphone jack input.

7.6. Choosing Sensitivity and Setting Gain

A balance between the microphone's input port noise floor, distortion, clipping point as well as level of sensitivity decides whether or not a microphone input is an effective match for a specific practical application. A microphone input with significantly higher level of responsiveness may require significantly less pre-amplification gain before the ADC, however, it might have much less headroom before clipping, compared to a microphone with reduced level of responsiveness.

In signal processing applications, similar to the this, in which a microphone with increased level of responsiveness is much more likely to achieve the higher audio input, distortion, as well as result in clipping. Alternatively, a higher level of sensitivity might be suitable in the application forms, in which the audio signal is attenuated as the length from the source to the microphone input port raises.

Nominal line level for consumer equipment is $-10\,dBV$, which is equal to $10^{-10V/20} = 316\,mV_{(RMS)}$. For simple sinusoidal input waves $\sqrt{2}\,.316\,mV_{(RMS)} = 447\,mV_{(peak)}$ or in other words $2\sqrt{2}\,.316mV_{(RMS)} = 894\,mV_{(peak\ to\ peak)}$ theoretical maximum input of microphone jack [22].

However, audio equipment will often include as much as $20\,dB$ of headroom above this value before clipping, so a full scale signal could be $10\,dBV$, or $3.16\,V_{(RMS)}$ or $8.94\,V_{(peak\ to\ peak)}$ [22]. Of course, every manufacturer wants to be louder than everyone else in competition, so devices tend to have higher inputs and outputs than they should be producing, consequently inputs have less headroom than they should have.

The aim would be to remain within the $894\,mV_{(peak\ to\ peak)}$ limit in all circumstances so that problems during signal processing are not experienced.

7.7. Pyramidal horn antenna

Following are the dimensions and pictorial representation of fabricated antenna:

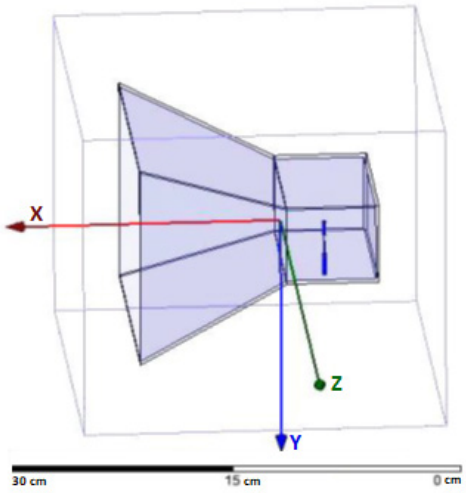


Figure 9 - HFSS Model Visualization Of Ism Band Antenna.

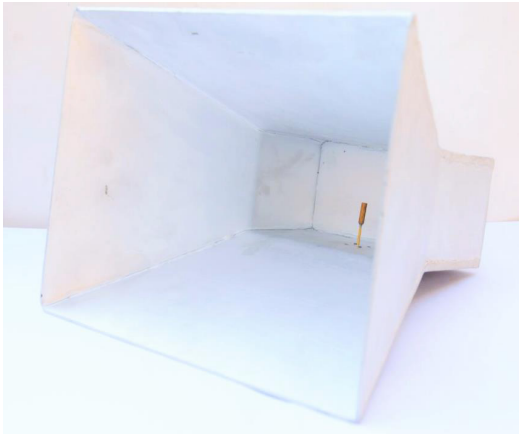


Figure 10 - Fabricated Model of ISM Band Antenna.

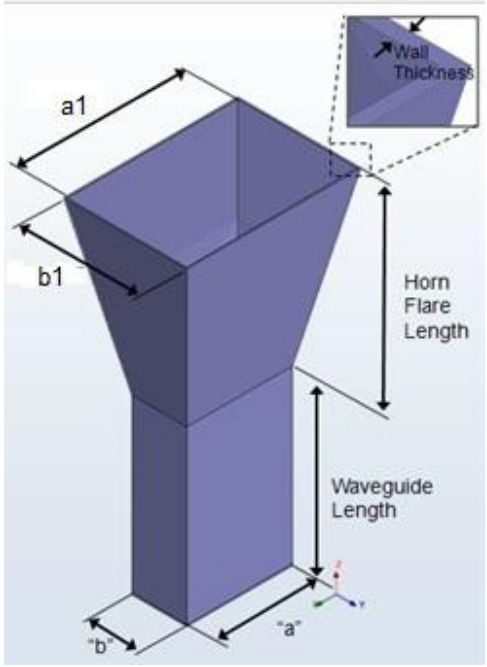


Figure 11 – Dimensions for Horn Antenna .

Table 1 - Dimensions of Horn Antenna

Units	Cm
Material	Aluminum
Connector type	SMA
Waveguide	WR-430
Waveguide “a“ dimension	10.92
Waveguide “b“ dimension	5.46
Waveguide length	10.37
Horn size in “a1“ dimension	18.66
Horn size in “b1“ dimension	14.51
Horn flare length	31.1
Wall thickness	0.203
Gain	Approx. 13dB
Lambda	12.244

7.8. Results of Antenna

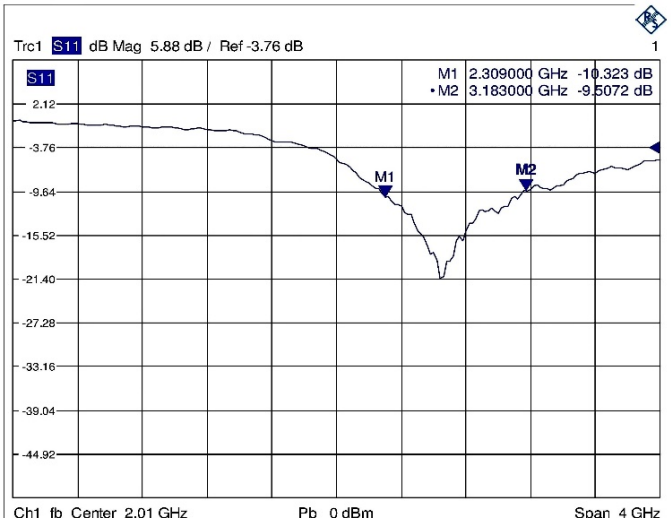


Figure 12 - Return Loss of Fabricated Tx Antenna.

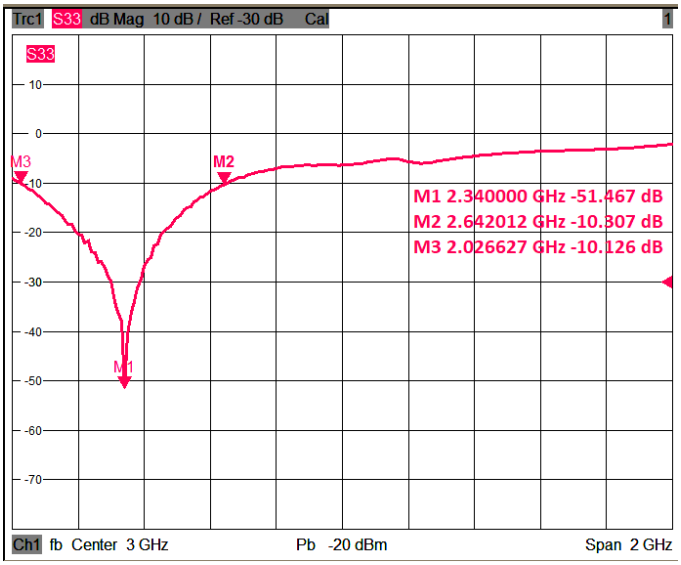


Figure 13 - Return Loss of Fabricated Rx Antenna.

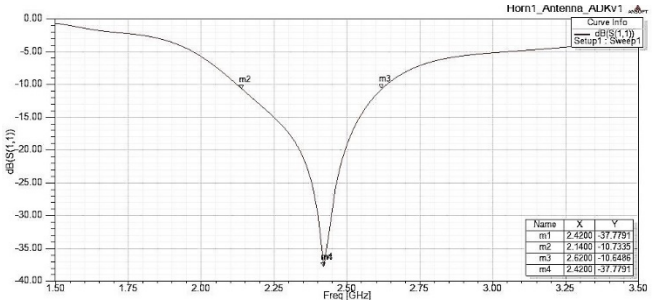


Figure 14 - Return Loss Of HFSS Simulated Antenna.

The return loss that is calculated by High Frequency Structure Simulator (HFSS) software simulation is -37.7791 dB and the return loss measured by the network analyzer is at a minimum of -21.40 dB. The -10 dB bandwidth of the Tx antenna is calculated from simulation as 480 MHz (2.62 GHz - 2.14 GHz) and the one measured from network analyzer is 883 MHz (3.183 GHz - 2.3 GHz) it can be seen that the centre frequency is shifted a little to 2.55 GHz but it still remains within the desired bandwidth of operation. The -10 dB bandwidth of the Rx antenna measured from network analyzer is 622 MHz (2.642 GHz - 2.02 GHz) it can be seen that the centre frequency is shifted a little to 2.34 GHz but it still remains within the desired bandwidth of operation.

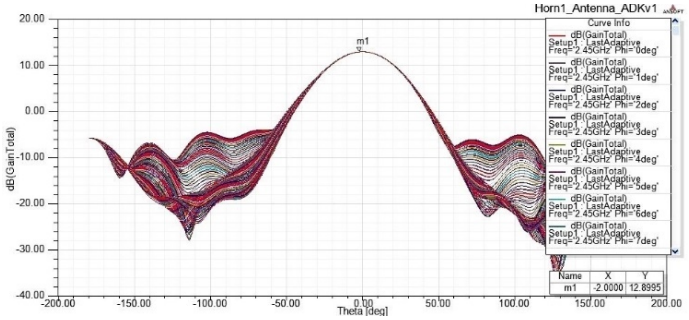


Figure 15 - Gain of HFSS Simulated Antenna.

Antennas Radiation pattern was also tested in Antenna Test Range (ATR) at Space and Upper Atmosphere Research Commission (SUPARCO) Laboratories. The plots of the results obtained are shown below.

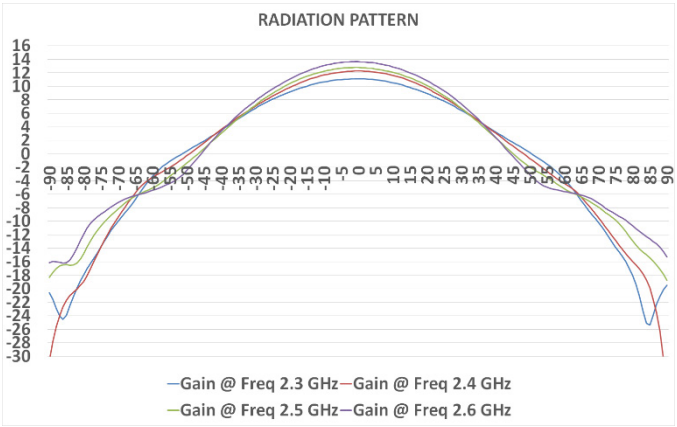


Figure 16 - Gain Measurement at ATR (Tx).

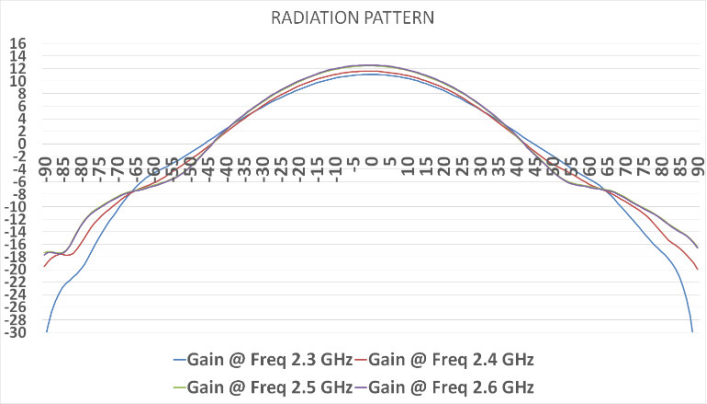


Figure 17 - Gain Measurement at ATR (Rx).

The actual gain measurement at 2.4 GHz of Tx antenna comes out to be 12.2744 dB, whereas, in simulated antenna results it can be seen that it comes out to be 12.8995 dB this is an acceptable loss that may be experienced in practical measurements. Similarly, for Rx antenna it comes out to be 11.5885 dB. Complete detailed results are mentioned in appendix.

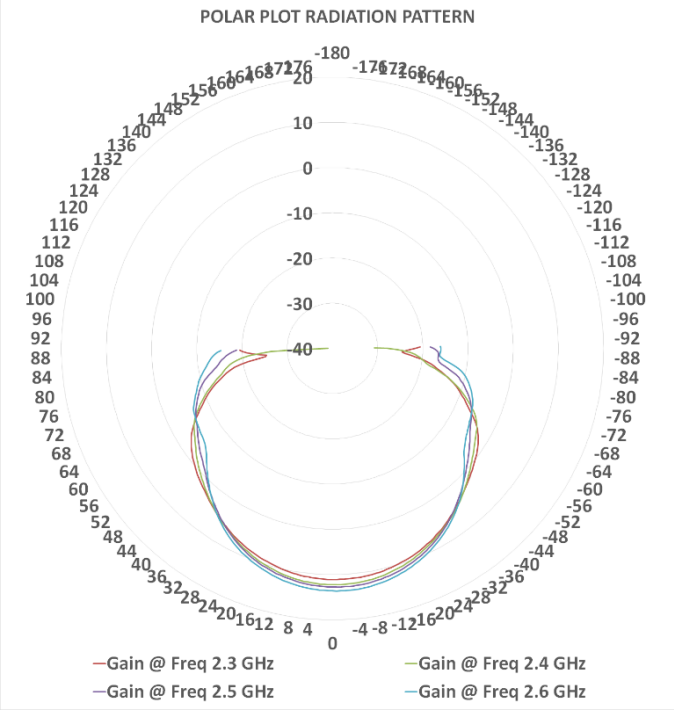


Figure 18 - Polar Radiation Pattern Plot (Tx) in ATR.

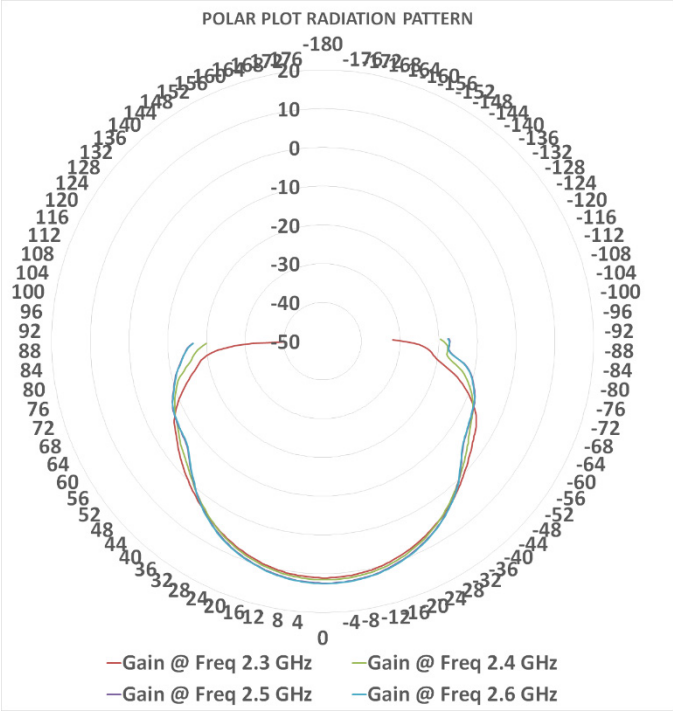


Figure 19 - Polar Radiation Pattern Plot (Rx) at ATR.

7.9. Power supply

In order to power the transceiver, a straightforward yet effective way by making use of widely available voltage regulator IC's to power the system was employed. 12 V DC from electric batteries was employed and required voltage was regulated by the components employing 78xx linear voltage regulator IC's. The 78xx or for certain producers L78xx, LM78xx, MC78xx and so forth are a series of linear static voltage regulators which are self-contained and have a built-in circuit. These are widely used in electric circuits demanding a regulated and managed power source because of their simplicity and inexpensive IC's. They generate a voltage which is positive relative to ground.

Light Emitting Diode (LED) lights were also used for indication of functioning of each separate IC input voltage and used comparatively higher values for capacitance for better ripple free and fixed voltages. The complete circuit diagram of power supply is given on next page. This circuit was implemented in the design. A Printed Circuit Board (PCB) was fabricated which was used to supply power to components.

Schematics of transceiver power supply are given below:

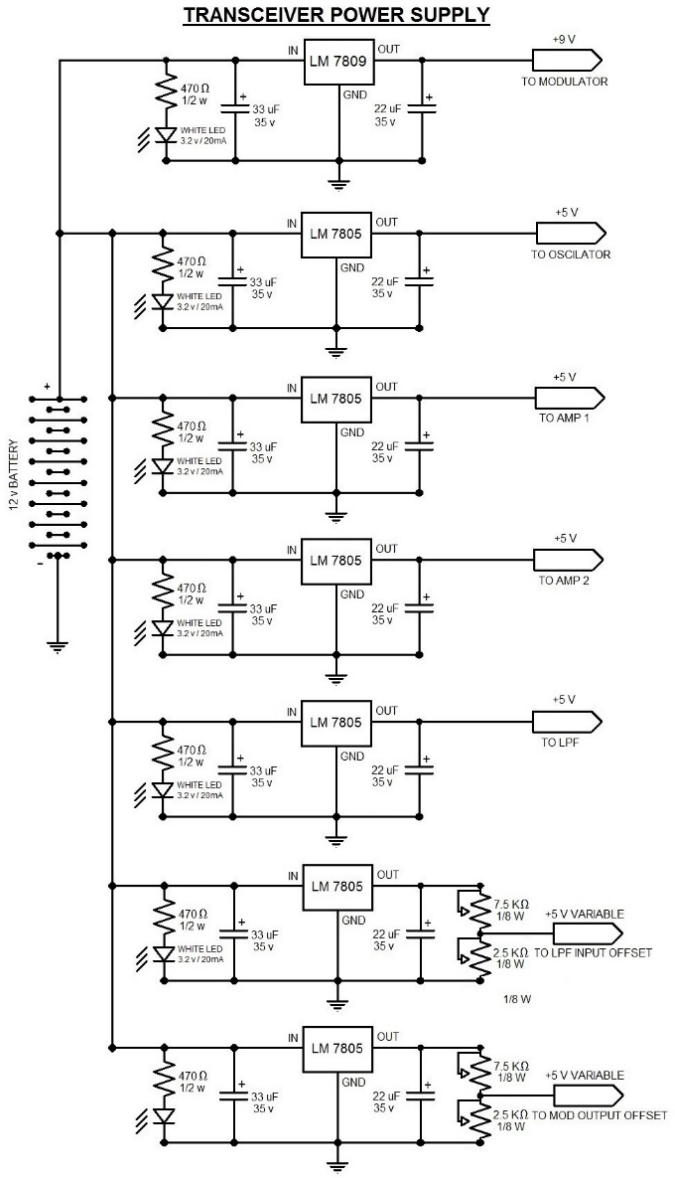


Figure 20 - Transceiver Power Supply Schematic.

8. Design implementation

8.1. Transmitter development stage

Each and every component of transmitter was connected and results were tested with spectrum analyzer.

First of all, VCO was given a constant voltage in order to generate a fixed frequency of 2.4 GHz then amplifier was connected in series and the results were visualized on spectrum analyzer. Similarly, the splitter was added to the chain and a 50 Ω termination was applied to the other port of splitter in order to avoid reflection from it. Then a total gain of 14.65 dBm was achieved and when the cable loss was included which was calculated prior to the testing (approximately 1.5 dB calculated before testing) then the total power comes out to be 16.15 dBm. This is very close to the calculated value of 16 dBm during link budget analysis at 2.4 GHz.

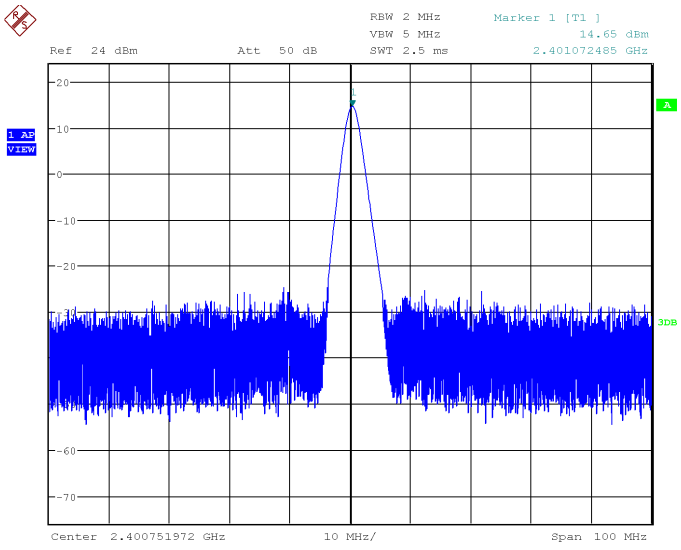


Figure 21 - VCO + PA + Splitter Output Response.

The block diagram of transmitter chain is given below this will be the total power feed into the antennas at 2.4 GHz.

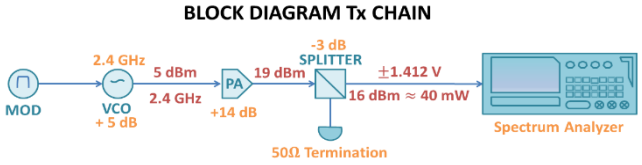


Figure 22 - Block Diagram VCO + PA + Splitter Testing.

8.2. EIRP measurement at anechoic chamber test

The gain of the antenna is measured by using the two antennas with known gain one as a transmitter and other as a receiver. Subsequently replaced the Tx antenna with the fabricated antenna and again measured the gain then with some calculations, the gain of the fabricated antenna can be found. The transmitted power fed into the antennas is measured previously although their values are given in table. The EIRP is calculated at different frequencies in order to cover the entire bandwidth.

Table 2 - EIRP Calculations Table

SR NO.	FREQ (F_C)	TX GAIN (G_T)	TX POWER (P_T)	EIRP (P_T+ G_T)
1	2.3 GHz	11.11 dBi	16.75 dBm	27.86 dBm
2	2.4 GHz	12.27 dBi	16.66 dBm	28.98 dBm
3	2.5 GHz	12.79 dBi	16.68 dBm	29.47 dBm
4	2.6 GHz	13.64 dBi	16.68 dBm	30.32 dBm

All the values are given excluding the cable losses that were calculated prior to the testing for gain measurement which was calculated to be 1.2 dB and for transmitting power fed into the antennas the cable loss was measured to be approximately 1.5 dB.

The calculated EIRP of the transmitter was 29 dBm but it can be seen that the EIRP of the transceiver increases with the increase

in frequency. There is almost 3 dB increase in EIRP when frequency is increased from 2.3 GHz to 2.6 GHz. This is mainly due to the increase in transmitter antenna gain increase, as the frequency increases the beam width of the transmitting antenna increases, which leads to the increase in overall transmitted power.

8.3. Receiver development stage

For any equipment comprising a receiver and a transmitter the loopback test [23] is a smart method to test the receiver. The essential loopback assessment for the integrated Microwave transceiver is conducted. The baseband signal visualization is conducted by oscilloscope which helps as an output analyzer while the microwave front end is in test with the help of a test attenuator. By means of this method in Transceiver (TRX) development, the microwave examination equipment can largely be evaded eventually leading to the reduced cost of test. The existing limitations and trade-offs are also discussed in the next chapter in terms of test observability controllability, and feasibility vs performance also the influence of production tolerances and the detectability threshold test was also conducted. This loopback test is conducted to provide a comprehensive picture of the transceiver functionality also in order to test Tx and Rx front ends before deploying antennas and actually start transmitting and receiving. The actual received signal of applied communication transceiver under test is also included. The receiver was tested using loop back technique where an attenuator is applied at the transmitter feed and is given to the received signal feed. The block diagram of the system under test is given below.

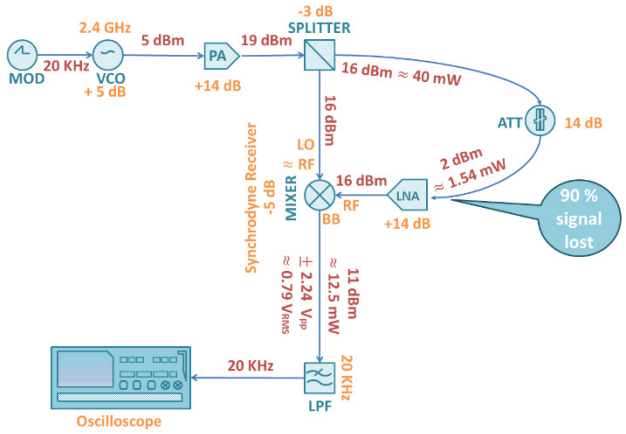


Figure 23 - Block Diagram Receiver Test (Loop Back Test).

Since the RF built-in circuits function at gigahertz frequencies and as they are growing to be more and more sophisticated, in apparatus assessment the typical testing aimed toward comprehensive requirements of the engaged RF components seems unrealistic and also expensive when it comes to the instrumentation as well as evaluation time. Direct RF measurement results on components demand additional contact elements that are likely to degrade the microchip overall performance and efficiency while using the needle probes obstruct the measurement results. The received signal result shown below tells that the receiver is functioning as desired and the frequency in use (20 KHz) would be the maximum frequency that is going to be used for the system.

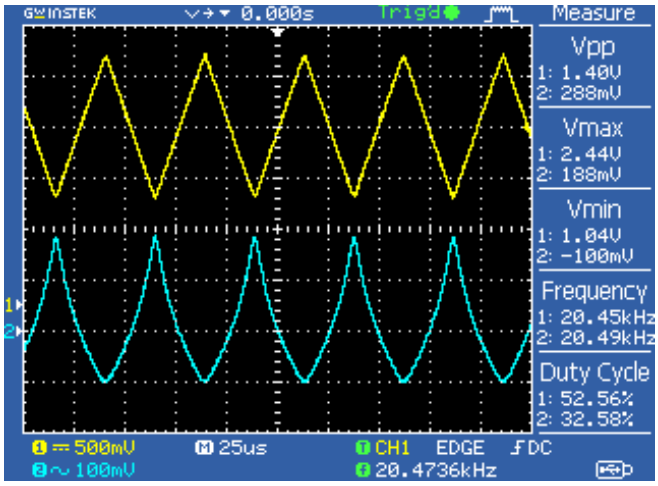


Figure 24 - Receiver Loop Back Result (20 KHz).

Using this approach as carried out in the front end test the detectability thresholds for different errors could be determined as well as various loop back results can be analyzed for the transceivers detectability. As the limitation and tolerances of the RF circuitry are likely to drive the RF components away from specifications the Design for Correction (DFC) is practically essential and also it needs to be merged with link budget analysis for an in depth approach. Through DFC the effect of variation due to tolerances on assessment is mostly alleviated, however, the demand for repeated tests is not reduced.

In spite of the problem of not having the schematics of the microwave components, a simple yet effective way was used to calculate the phase difference that will be produced at the mixer. It can be seen that there is a phase shift of approximately 180° in the transmitted and the received signals.

8.4. Data acquisition stage

A signal processing code block diagram has been given below.

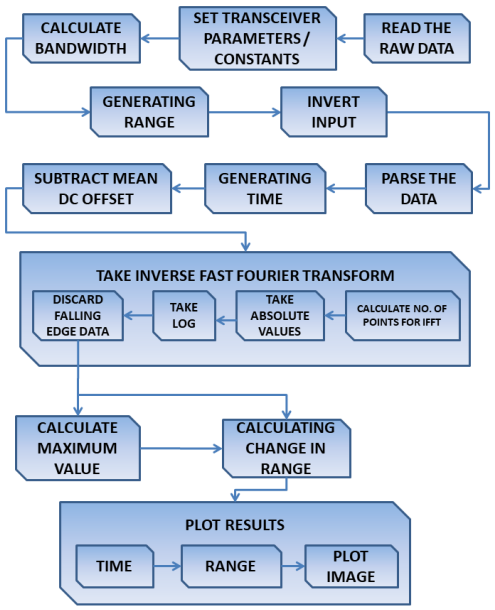


Figure 25 - Block Diagram of DSP MatLab Code.

8.5. System integration

The modules which make the core of the microwave Tx and Rx chain were assembled. The picture below shows the assembly.

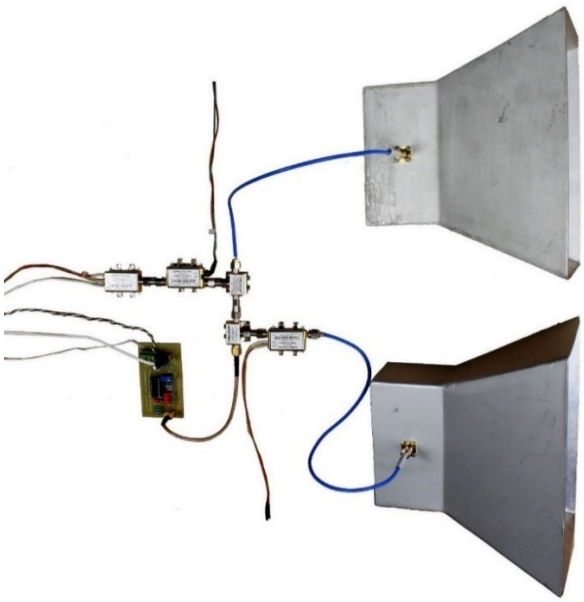


Figure 26 - Complete TRX Assembly with Antennas.

9. Evaluation

9.1. Oscilloscope Rest Results

First of all, the test results are conducted without any material in front of the body. By placing a body directly in front of the transceiver. The results are taken on oscilloscope and also a recording is made which was processed by the MatLab code in order to get a better picture. The test with a 9-inch (22.86 cm) thick wall in between the body and transceiver. The body was placed at a distance of 5 ft. from the wall. The test results in time domain are given below as visualized on oscilloscope.

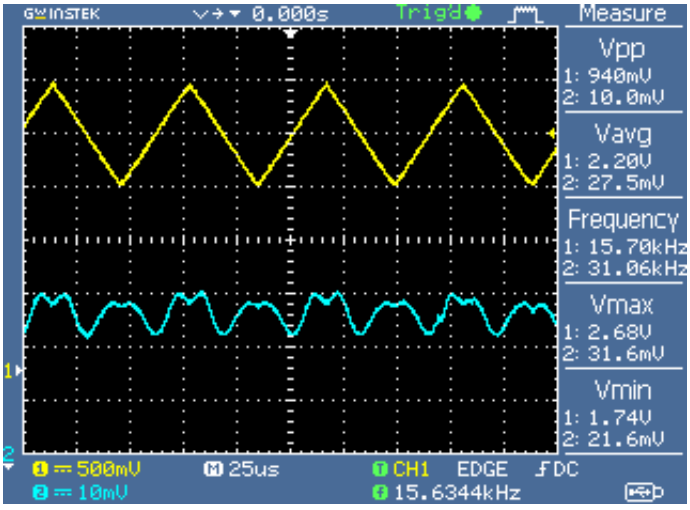


Figure 27 - Far Field (Wall) Test without Body.

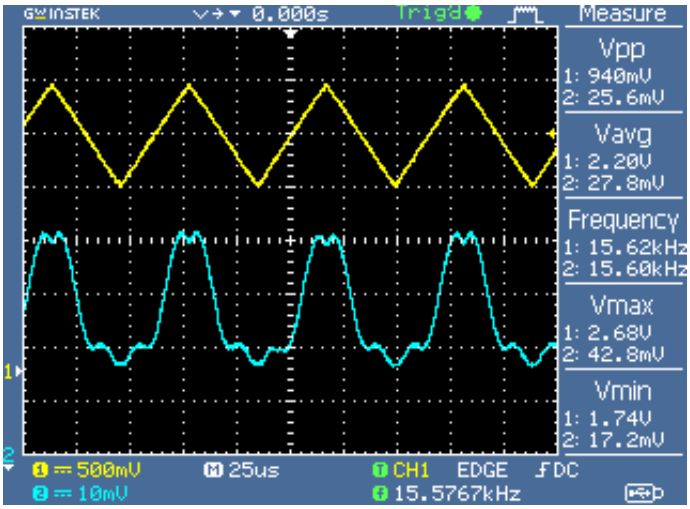


Figure 28 - Far Field (Wall) Test with Body 1 Foot (30.48 Cm) from the Wall.

9.2. Computer Audio Card Results

Audacity software was used to record the test results. First of all, the transceiver results are acquired by placing the transceiver with the wall and body was only placed once behind the wall these results are captured at multiple frequencies.

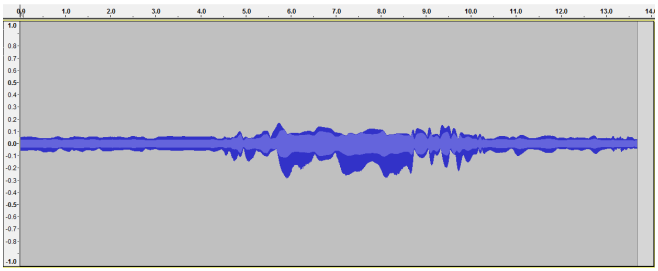


Figure 29 - Audio Pictogram at 5 KHz (Body was Placed Once).

Now the transceiver results are acquired by placing the transceiver with the wall and body was only placed multiple times behind the wall these results are captured at multiple frequencies. The multiple detection was conducted by setting threshold at the noise floor. The resulted pictograms are given below.

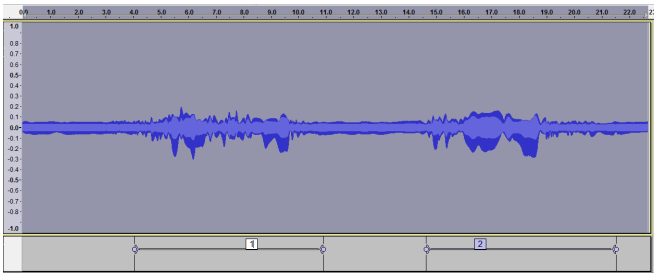


Figure 30 - Audio Pictogram at 5 KHz (Body was Placed 2 Times).

9.3. MATLAB 2D Images Results

Recordings from different frequencies is processed using MATLAB and RTI plot is developed.

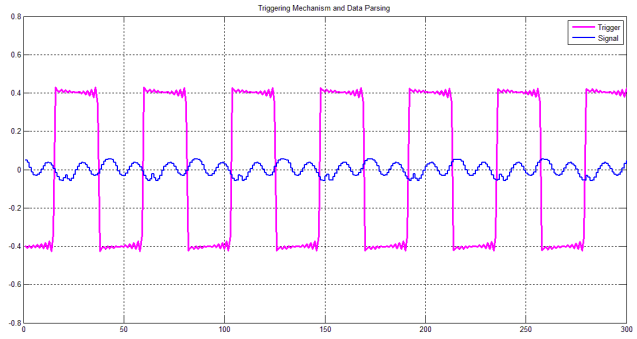


Figure 31 - MatLab Data Parsing 1 KHz.

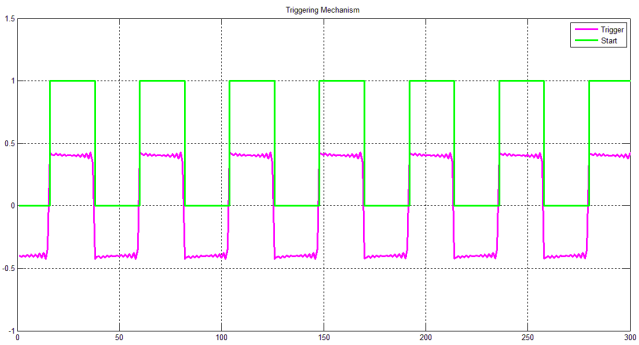


Figure 32 - MatLab Triggering Mechanism 1 KHz.

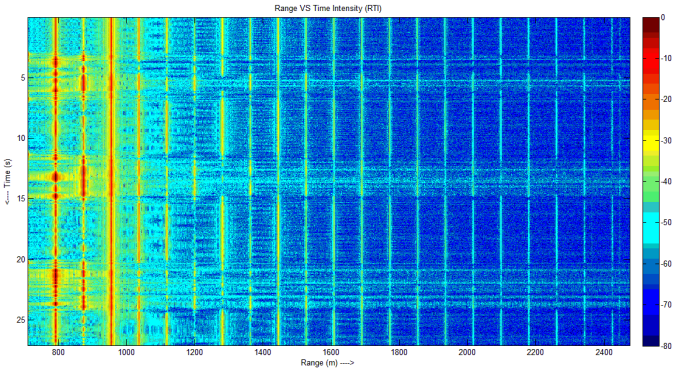


Figure 33 - MatLab Range vs Time Intensity 1 KHz (Body was Placed 3 Times).

The transceiver results are acquired by placing the transceiver with the wall and body was only placed 3 times behind the wall with specific time intervals these results are captured at 11 KHz. The multiple detections can be observed in RTI plot given above.

10. Conclusion

A completely built-in 2.4 GHz FMCW transceiver is basically designed and produced in this particular report. Employing XR 2206 IC for the purpose of function generator IC as an FMCW engine and using module based components with SMA connectors can significantly decrease the complication of the circuit as well as hardware designs. Which helped accomplish the project goals and targets realized that were set within initial project proposal. FMCW within the system was analyzed as well as simulated, using various modulation techniques. Microwave front end came to the realization in MMIC technology is exhibited as well. With baseband processors as well as high-gain antennas incorporated, this project offers an entire transceiver realization instance, which demonstrates encouraging probability of future under debris search and rescue or through wall imaging applications. After designing the transceiver and horn antennas, several experiments were conducted on human skin and other materials (concrete, wood, glass and sand), to understand and distinguish between if a living or nonliving body is present behind the material or otherwise. From the experiments that were conducted, it was observed that the body was easily detectable and materials are easily distinguishable when the body was close to the material but when the body moved away from the material it was difficult to detect the body beyond a distance of 7 ft (2.1 m).

Developing of this kind of a transceiver experiences a lot of challenging problems. Nevertheless, the energy requirements of this kind of a transceiver which is effective at detecting body of a human under rubble is dependent on operating environments. To be able to model a time domain relying microwave transceiver, antenna gain, wireless transmitter, energy consumption, highest possible input power and the highest possible voltage levels as well as form of modulation is must. Frequency generator, system dynamic range together with noise figure of Rx are the variables that depend on the attenuation along with the operating environments. Consequently, Link Budget Analysis own fundamental share when planning these kinds of transceivers. Within this report, a Simulation of Link Budget Analysis, that is in MatLab environment, is presented to be able to evaluate the functional bandwidth of the System. Moreover, these simulation outcomes are reconfirmed by actual measurements.

One of the most interesting observation was a presence of beat pulse in the form of amplitude variance of the received signal when living being is placed in front to the transceiver. Which was not received during presence of nonliving body. Similarly, there was a repeated pattern in received frequency components when living being is placed in front of the transceiver. This could be due to heart beat or breathing of a person if so then this could also be employed to detect living and nonliving beings under debris. More research is needed to look into this observation.

This range of the system is greatly increased by using high power amplifiers and employing horn antennas. During the tests performed, the antenna was directly interfaced with the transceiver that enabled us to transmit 794 mW of power or 29 dBm EIRP. That much power would greatly increase the penetrating and distinguishing ability of the antenna. Further experimentation, research and analysis will determine exactly how much power is to be used in real life scenarios.

This project provided foundation for further research and work on more complex system to be built in the future by upcoming students. This will also help students gain understanding and interest in subjects such as electromagnetic theory, antenna & radio wave propagation and radar systems.

Conflict of Interest

I certify that there are no affiliations with or involvement in any organization or entity with any financial interest, or non-financial interest in the subject matter or materials discussed in this manuscript.

Acknowledgment

First of all, I would like to thank Almighty Allah and then my parent's, relatives & friends who kept supporting me in all the phases of my study, both economically and morally.

Here, I would explicitly like to acknowledge my teachers and professors, for their guidance and encouragement to work hard, being my mentors. I have found them enormously supportive, caring and reassuring throughout my course of studies at CIIT. Their analytical, systematic and investigative comments on my work yielded positive results by stimulating my mind and inspired me to think of novel ideas, approaches as well as procedures in the field of electrical telecommunications.

I must here acknowledge support given to me by Space & Upper Atmosphere Research Commission (SUPARCO) Pakistan in test and trials of project work at SUPARCO LABS. Without their kindness proofing of equipment would not have been possible. Staff at Satellite Research and Development Centre SUPARCO had been extremely forthcoming in providing the training and test equipment for proofing.

Once again all praises are for Allah Almighty who provides us with all the resources and knowledge, making sure that we are able to make their adequate use for the benefit of humanity.

References

[1] D. A. A. K. Syed Daniyal Ali Shah, "Microwave Transceiver Prototype for Under Debris Search and Rescue Operations," in *19th IEEE International Multi Topic Conference (INMIC 2016)*, Sahiwal, 2016.

[2] J. E. P. Jr., G. L. Charvat, J. Goodwin and M. Tobias, "Through-Wall Imaging Radar," *Lincoln Laboratory Journal ,* Vols. Volume 19, , no. 1, 2012.

[3] M. A. Richards, Fundamentals of Radar Signal Processing, vol. Second Edition, McGraw-Hill Professional Engineering, 2014.

[4] M. Skolnik, Introduction to Radar Systems, vol. 3rd Ed, New York: McGraw-Hill, 2001, p. 49.

[5] B. Mullarkey, " The differences between pulse radars and FMCW ones," [Online]. Available: Navigate-us.com.

[6] C. R. Leutz, Notes on a Super-Heterodyne, QST, Hartford, CT: American Radio Relay League, p. 11–14.

[7] S. I. –. T. o. F. Allocations, CHAPTER II – Frequencies, ITU Radio Regulations, ARTICLE 5 Frequency allocations.

[8] I. T. Union, ARTICLE 1 - Terms and Definitions, 19 October 2009.

[9] G. G. W. &. W. Henry Goldberg, The Genesis of Unlicensed Wireless Policy, George Mason University, April 4, 2008.

[10] W. Hayward and D. Bingham, Direct Conversion - A Neglected Technique, 1968.

[11] "The History of the Homodyne and Synchrodyne," *Journal of the British Institution of Radio Engineers,* April 1954.

[12] "Quad Demodulators Arm Direct-Conversion Receivers," *IEEE Microwaves & RF,* 9 February 2011.

[13] J. Lesurf, "Improved forms of radar," [Online]. [Accessed 24 01 2016].

[14] "Frequency-Modulated Continuous Wave Radar," [Online]. Available: Radartutorial.eu. [Accessed 7 8 2016].

[15] J. M. Ridenour, "Radar System Engineering," MIT Radiation Lab, vol. 1, p. 629, 1947.

[16] S. Erst, Receiving Systems Design, The Receiver, Dedham, Ed., Artech House Inc., 1985.

[17] W. Chang, L. Huan and L. Yubai, "A Practical FMCW Radar Signal Processing Method and Its System Implementation," 6th International Conference on ITS Telecommunications Proceedings, p. 1195 – 1199, 2006.

[18] D. Smith, Digital Signal Processing, The ARRL Handbook for Radio Communications., 81 ed., D. Reed, Ed., 2004.

[19] Mahafza and B. R., "Radar system analysis and design using MATLAB".

[20] M. I. Skolnik, Radar Handbook, vol. 3, New York: McGraw Hill, 2008.

[21] J. Eargle, Handbook of Recording Engineering, Springer. p. 4. ISBN 0-387-28470-2., 2005.

[22] B. Fries and M. Fries, Digital Audio Essentials, O'Reilly Media. p. 147, ISBN 0-596-00856-2.

[23] R. Rudersdorfer, Radio Receiver Technology: Principles, Architectures and Applications, John Wiley and Sons. pp. 1–2., 2013.

[24] E. H. Armstrong, A new system of short wave amplification, New York: Proceedings of the IRE - Institute of Radio Engineers, 1921, p. 3–11.

[25] T. K. Sarkar, R. J. Mailloux, A. A. Oliner, M. Salazar-Palma and D. L. Sengupta, History of Wireless, John Wiley and Sons, 2006, p. 110.

[26] J. B. Hagen, Radio-frequency electronics: circuits and applications, Cambridge University Press - Technology & Engineering, 1996, pp. 12-58.

[27] Improvements in or relating to superheterodyne radio receivers, GB 426802, 12 October 1933.

[28] F. E. Terman, Radio Engineers, McGraw-Hill, 1943.

[29] J. J. Carr, RF Components and Circuits, 2002.

[30] G. Malanowski, The Race for Wireless: How Radio Was Invented or Discovered?, 2011, p. 69.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1107-1113 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Rotary Cutting of Electrical Steel Laminations - A Contrast to Traditional Stamping

Markus Hubert*, Simon Kutter, Marco Ziegler, Michael Schneider, Joerg Franke

Friedrich-Alexander-University Erlangen-Nuremberg, Department for Mechanical Engineering, Institute for Factory Automation and Production Systems, 90429, Germany

ARTICLE INFO

Article history:
Received: 30 April, 2017
Accepted: 12 June, 2017
Online: 16 July, 2017

Keywords:
Rotary Cutting
Electrical Steel Processing
Toroidal Core Tests

ABSTRACT

Rotary cutting of electrical steel laminations is a novel, promising alternative in the processing of electrical steel strip with enormous economical and technical potentials. This paper is an extension of work originally presented on the 6th International Electric Drives Production Conference (E|DPC) in Nuremberg, Germany 2016. The scope of this paper includes a theoretical description of the rotary cutting process, an explanation of the machine and tooling setup, the basics of the Toroidal Core Tests and a discussion of the theoretical and experimental results. The methodology applied is basing on the Toroidal Core Test per DIN 60404-6, which is elementary described in section 3. Toroidal core tests allow a characterization in regards of technical, process and functional aspects, what is also mainly extended in this work. Finally, a classification of the rotary cutting process into existing processing technologies of electrical steel strip is feasible with respect to its limits and possibilities.

1. Introduction

Due to increasing demand of high-performance electric drives for the automotive market, German manufacturers are under pressure towards global competitors. The lamination stack forms the skeleton of the electric drive and guides the magnetic field in application. [1, 2]

In mass production of laminations, stamping is the technology with the highest application rate. Furthermore, the stacking process can be fully integrated into the stamping tool with e.g. the interlocking technology. However, complex machine tools are very expensive and thus, small volumes and frequent product changes are not profitable. Instead laser cutting can also be used for small lot sizes as no expensive tools are necessary. Despite high flexibility, laser cutting is not competitive due to long cycle times and high energy costs. Moreover, the unloading and stacking of laminations is currently only possible manually. [1, 2]

In the production of very thin laminations (smaller than 0.30 mm) for electric drive systems to reduce eddy current losses at high frequencies, the conventional stamping process reaches its limits due to several reasons: The cutting gap widths between punch and die tools is very small (2%-5% of the sheet thickness). Thus, the production of the complex cutting tools is no longer profitable. To absorb the high stamping forces, the kinematics structure of the stamping machine and machine´s foundation are very massive. Consequently, high investment costs occur. Moreover, the stamping machine cuts sheets at intervals. Thereby large masses must be accelerated and decelerated, what leads to high energy costs and noise pollutions. Due to these reasons, the enormous demand to thin sheets cannot be handled with the conventional stamping process. [2]

This is the reason why a new approach is required. Rotary cutting of thin electrical steel laminations is very promising. Rotary cutting is a continuous cutting process which is currently used in paper industry or for metal profiles. There are several advantages due to its compact structure and special cutting kinematics: [2]

- The distribution of the cutting force over the entire cutting profile is more uniform and the maximum force is lower.
- The continuous cutting process without accelerations and decelerations in strip feed and tools increases the productivity and the quality of the cutting edges.
- The cutting gap can be set to zero as the stamp does not immerse into the die.
- The tool structure is simpler and the tooling costs are lower.
- Fast tool introduction for new tools, as the development and the production is less complex.

*Corresponding Author: Markus Hubert, Fuerther Straße 246b, 90429 Nuernberg, Germany | Email: markus.hubert@faps.fau.de

- The machine structure is more compact thus the energy costs are lower.

The rotary cutting process is characterized by a counter movement of two parallel driven shafts. These tool rollers consist of a punch and a die cylinder. Due to the movement of the cylinders, the electric steel sheet is carried frictionally through the shaft´s gap. Hence the laminations are cut out by the mesh of the rollers. The tool rollers may not lift off during the cutting process, thus a radial preload is necessary.

In contrast to conventional stamping, the kinematic motion of the rotary cutting process is described by a cycloid movement. Basically, each point on a rotating circle describes a cycloid curve on a straight line or a circle wheel. Thus, the punch also describes cycloid curves on the steel strip. [4, 5]

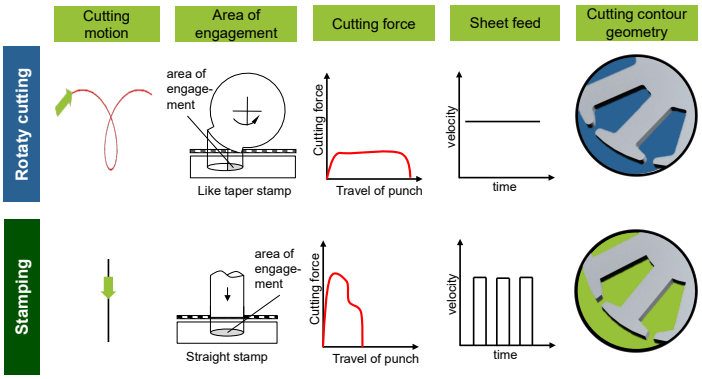


Figure 1 Comparison of Rotary Cutting with Stamping [3]

According to Figure 1 and [3], the conventional stamping process differs significantly from rotary cutting process through the following points:

- The cutting movement at stamping is linear with a permanent acceleration, deceleration and change of direction. In contrast, the rotary cutting cylinder moves continuous and describes a cycloid motion.
- During the rotary cutting process, the area of engagement is very small. At conventional stamping, the entire outline is cut at once.
- The maximum cutting force at stamping is very high and the travel of punch is very low.
- The speed of the sheet feed is constant during the rotary cutting process. Instead, the sheet feed speed is cyclical and clocked at stamping.
- Regarding the cutting contour geometry, laminations show sharp edged corners instead of a curved outline.

2. Comprehensive Analysis of the Rotary Cutting Method using Toroidal Core Tests

For a methodical examination of the cutting process, a suitable geometry is necessary. This is secured by toroidal core tests. With this simple geometry, the process behavior can be investigated during the separation process. Especially, cutting parameters, process stability and wear behavior can be evaluated. Thereby, it is possible to focus the tests on the cutting process without additional influences by complex contour elements, such as small radii. Secondly, it is possible to determine important characteristic values like power losses or magnetic and electric properties. In addition, an assessment of the sheet geometry is possible. Thus, process influences on the geometry can be investigated. To enable this, the toroidal core tests are comprehensively analyzed according to DIN IEC 60404-6. [4]

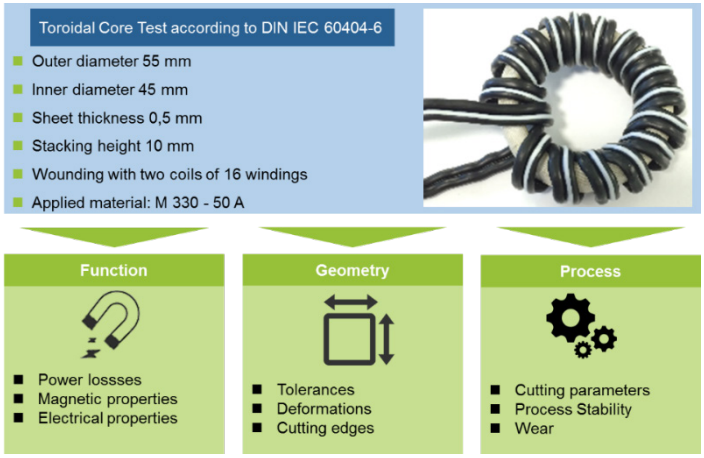


Figure 2 Toroidal core test per DIN 60404-6

The toroidal test specimens' inner diameter D_i is 45 mm and outer diameter D_o is 55 mm. The lamination stack length l_{stack} is 10 mm consisting of 20 single toroidal laminations with sheet thickness s = 0,5 mm. The single laminations are stacked without additional insulation layers between the laminations and wrapped with tape to a toroidal core. Finally, for investigation of electrical and magnetic properties, the toroidal core is wound with two coils of 16 windings each. [4]

Table 1 provides an overview of the manufacturers specific mechanical and magnetic properties for the applied material M 330 - 50 A. [4]

Table 1 Material Card for M 330 – 50 A

Material designation	EN 10027-1	M330-50A
Producer	Langhagel	
Nominal thickness	[mm]	0,50
Density	[kg/dm³]	7,65
Max. hysteresis and eddy-current losses	1,0 T / 50 Hz [W/kg]	1,35
	1,5 T / 50 Hz [W/kg]	3,30
Anisotropy of the losses	[%]	±14
Min. magnetic polarisation	2500 A/m [T]	1,49
	5000 A/m [T]	1,60
	10000 A/m [T]	1,70
Yield strength	$R_{p0,2}$ [N/mm²]	340 - 380
Tensile strength	R_m [N/mm²]	480 - 520
Hardness	HV5	170 - 200
Bending coefficient	-	3

3. Machine Setup and Process Description for Rotary Cutting of Toroidal laminations

3.1. Machine and Tooling Setup

The prototype rotary cutting machine for toroidal laminations consists of four main components: Cutting unit, electric drive,

control and decoiler. The drive unit is an *SIEMENS* asynchronous servomotor of type *"SIMOTICS M-1PH8"*. With a higher-level control, important motor parameters such as number of revolutions or torque can be set, controlled and monitored. [4]

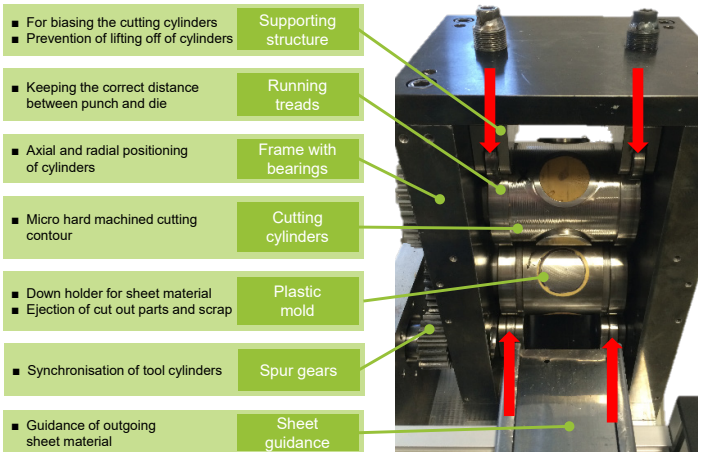


Figure 3 Prototype Rotary Cutting Device for Toroidal Core Samples [4]

The cutting unit itself essentially consists of five components: Tool set, gear box, pressure rollers, bearing unit and housing. The housing consists of four plates which are detachably connected by screws and pins. Thus, the cutting cylinders can be changed or subsequently adjusted. Two pressure rollers, which press against the running treads, relate to the housing to transfer the preload force on the cutting cylinders. Therefore, the cutting cylinder immerses in the material without lifting. [4]

The bearing unit is in the side walls of the housing. One side consists of two double-row angular contact ball bearings to absorb radial as well as axial forces. On the other side two cylindrical roller bearings are mounted as floating bearings. [4]

Moreover, at the drive side of the tool set, the spur gear is located. To ensure a synchronous and opposite rotation two gears of the same size are mounted on the shaft ends of the cylinders. To ensure the required dimensional accuracy and perfect cutting edges, one gear is being secured by a clamping sleeve in order to align both tool cylinders in radial direction. The pinion transmits the torque from the drive shaft to the spur gears. [4]

The cutting cylinders are core components of the rotary cutting machine for electrical steel. The tool set is arranged on top of each other. For rotary cutting processes in general, three tooling concepts are possible: wedge-action cutting, resilient cutting and stamp-die-cutting. However, for rotary cutting of electrical steel, only stamp-die-cutting is viable. [2, 4]

With this tooling concept, one cylinder acts as stamp and the other cylinder as die. For simple stamped holes with low accuracy requirements, the stamp may immerse into the die. For electrical steel processing with complex contour geometries and complete outlines, the stamp does not immerse into the die. Instead, the cutting unit must be designed in a way, that the stamp is located on the outer diameter of the die, the so called "zero to zero cutting" [5]. As a result, the lamination is cut by the stamp without immersion. Thus, a small die clearance is sufficient. To obtain reproducible cutting results, all process parameters have to harmonize with each other. [4]

In the conducted experiment, the geometry of the stamp is given by the toroidal geometry. The die has the corresponding negative form. In pre-trials the die clearance of 0.03 mm was determined as optimum for different sheet thicknesses. Thus, the outer diameter of the die D_{oD} is manufactured 0.06 mm larger and the inner diameter D_{iD} is manufactured 0.06 mm smaller compared to the stamp. The circumference of each cylinder is equipped with four cavities in total. The cutting circle diameter d_c of the stamp cylinder as well as of the die cylinder is 108 mm. [4]

All cavities of the stamp and the die cylinders´ are filled with an elastic polyurethane casting compound with hardness of 90 Shore A. During the cutting process, the plastic performs as sheet downholder stamp-side and counterholder die-side. Moreover, the plastic ejects the lamination from the cutting area. This is necessary as otherwise the sheet remains stuck in the cavities after cutting. [5]

Since the sheet does not automatically exit the stamp completely after cutting, an external tensile force must act in sheet feed direction. Consequently, the sheet leaves the machining zone completely without jamming up. [4]

3.2. Description of the Rotary Cutting Process

The rotary cutting process is divided into 4 phases (according Figure 4). Following, the rotary cutting phases are described for stamp-die-cutting with closed cutting outline. [4, 6]

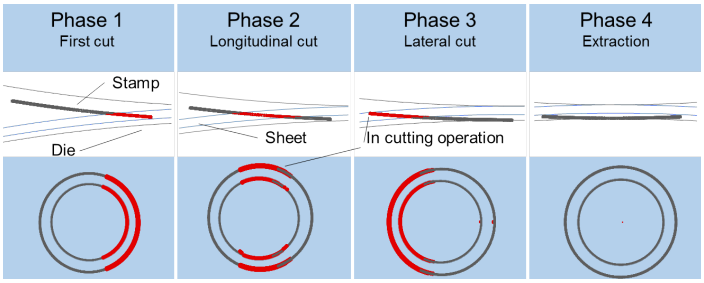


Figure 4 Four phases of rotary cutting referring to [6]

When the front edge of the stamp touches the sheet surface, the 1st phase of the cutting process starts. By further rotation, the cutting force is constantly increasing by forming the first crack until the entire sheet thickness is severed. [4, 6]

The cylinders continue to rotate, whereby the lateral cutting edges of the stamp cuts in the 2nd phase. The force level remains almost constant. [4, 6]

During phase 3, the back cutting edge impinges on the sheet metal and completes the cut. When the sheet is completely divided, the force level decreases abruptly. [4, 6]

After the rotary cutting process, the cutting sheet is pressed out of the die in the 4th phase. [4, 6]

For the description of the rotary cutting process, many different parameters are given. A decisively parameter is the angle of impact α, which defines starting- and endpoint of the line of action [7]. The angle of impact α depends on the circle diameter d, the vertical center distance a and the sheet thickness s and is derived from law of cosines. For good cutting results, the angle α should be very small. Thus, the length of line in action l_s and length of the contour in action l_c are getting small. [4]

According to [3], the angle of impact for equal diameters is defined by (1).

$$\alpha = \cos^{-1}\left(1 - \frac{s}{d} + \frac{s^2}{d^2}\right) \quad (1)$$

With a sheet thickness s of 0.5 mm and a cutting circle diameter d of 108 mm, the angle of impact α is located at 5.5°. [4]

In contrast to the angle of impact α, the force progression during cutting is difficult to determine. By means of a calculating software, the length of contour in action l_c for each contour can be calculated. This parameter behaves similarly to the force progression and can be used as a reference value. For the toroidal core test, the following diagram results (according Figure 5): [4, 8]

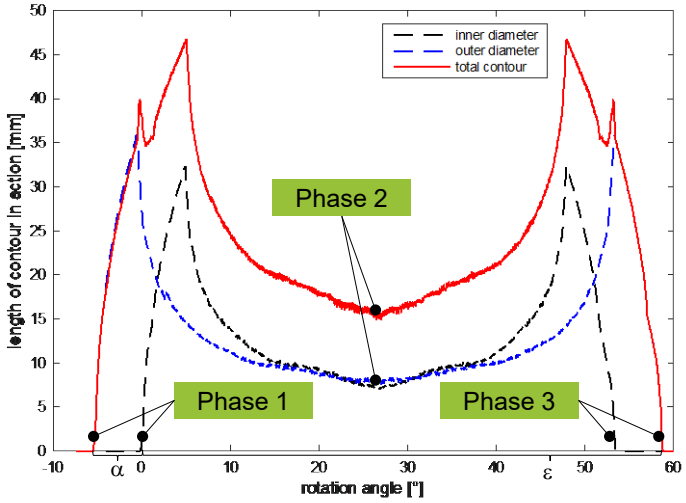


Figure 5 Length of contour in action per angle [8]

The total contour (red line) is explained as the summed curve of inner and outer diameter (black and blue line). [4]

A further parameter is the immersion depth of the stamp into the sheet Δy. This value is varying during the cutting process and depends on the circle diameter d, the sheet thickness s, the angle of impact α and the entry angle β. According to [3], the immersion depth Δy is: [4]

$$\Delta y = \left(\frac{d}{2} + s\right) \cdot \cos\beta + \frac{d}{2} \cdot \cos\alpha - d \quad (2)$$

with:

$$\beta = \cos^{-1}\left(\frac{d^2 + d \cdot s + s^2}{d^2 + 2 \cdot d \cdot s}\right) \approx \alpha \quad (3)$$

During the cutting process several velocities arise at the rotary cutting machine. The most important parameter is the circumferential speed v_c, which results from the circumference $d \cdot \pi$ and the number of revolutions n.[4]

$$v_c = d \cdot \pi \cdot n \quad (4)$$

By means of (4), the resulting speed in x- and y-direction can be calculated. All equations for the different speed components can be found in [3]. [4]

Since the speed of the lower sheet surface is faster than the upper one, there is a sheet speed difference Δv_{sheet} [4]. Table 2 gives an overview of applied process parameters.

Table 2 Applied Parameters of the test bench

Cutting circle diameter	d	[mm]	108
Diameter of lamination	d_c	[mm]	55
Sheet thickness	s	[mm]	0.5
Angular range	ε	[°]	58.4
Max. length of line of action	$l_{s,max}$	[mm]	5.18
Number of cavities	N_{Cav}	-	4
Number of revolutions	n	[min^{-1}]	10
Preload torque	M_P	[Nm]	60
Output per minute	X	-	20
Sheet speed difference	Δv_{sheet}	%	≈1
Transmission ratio motor/ cutting unit	I	-	1:2
Cutting speed in X-direction	v_x	[m/min]	3.4
Max. cutting speed in Y-direction	v_y	[m/min]	0.65

3.3. Simulation model

Within the research project, a simulation model is created in ANSYS Workbench 15.0 and explicit finite element calculations are done by the LS-DYNA solver. Thereby, the tool cylinders are defined as not deformable rigid bodies and the electrical steel sheet as a flexible solid body. The simulation model is assumed as quasi static. [8]

The material model of the electrical steel is defined by Johnson and Cook. This approach is very popular in metal cutting process simulations [9]. The three elementary material behaviors included in the model are material hardening, strain rate effects and thermal softening [10]. However, in this FEM model strain rate effects and thermal softening are neglected. [8]

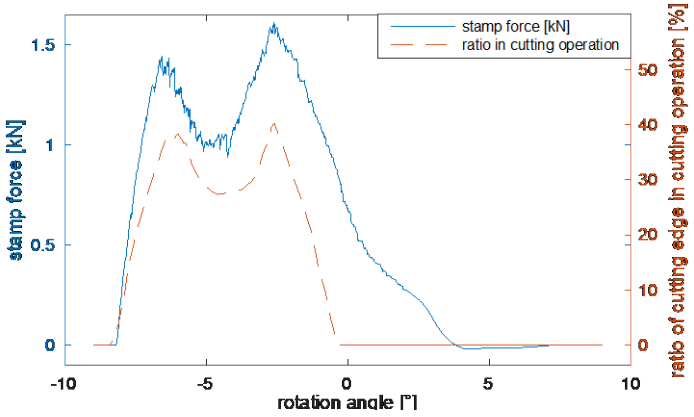


Figure 6 Simulated force-rotation path [8]

The CAD model of the geometry only includes important components. In addition, the cylinders are limited to 7° large shell elements. These simplifications reduce the required calculation time significantly. [8]

Figure 6 shows a comparison of the stamp force engaged cutting length over the rotation angel for a simplified geometry of a round stamp with a diameter of 5.00 mm and a steel strip thickness of 0.50 mm. At a rotation angle of -2.5°, the maximum stamp force of about 1.6 kN is reached. At this point, the entire cutting edge is in contact with the sheet. The pulling out of the sheet causes a negative stamp force. [8]

4. Results and Discussion of the Rotary Cutting Experiments

To analyze the influence of the cutting process, several electric sheet strips are cut with the test bench. The preload torque for all tests is 60 Nm and the sheet thickness is 0.5 mm. First, the process behavior is evaluated. Second, the geometry and the functionality of the laminations are analyzed. Micro sections of the different phases are made to investigate the cutting edges of the toroidal laminations. The contour of the geometry is analyzed with a multi-sensor coordinate measuring machine *VideoCheck IP 400 HA* from *Werth*. And third, the electrical and magnetic behavior of the toroidal core is evaluated on an electrical steel tester *MPG 200* from *Brockhaus Measurements*. [4]

4.1. Characteristics of process behavior

The toroidal core samples are cut completely from the steel strip. However, the samples adhere at the inner contour due to slight wedging. Thus, the separation occurs manually. The toroidal samples are slightly deformed and bent in sheet feed direction, as high forces occur. Thus, the laminations must be arranged after cutting. A further reduction of the bending arises with an adjustment of the machine settings. Larger cylinders or an additional recoiler with a further tensile force decrease the deformations. Since the design of the test bench is very simple, the cutting speed is very low. However, a higher cutting speed improves the results significantly. The cutting results at a sheet thickness of 0.5 mm and 60 Nm preload torque is shown in Figure 7. [4]

Figure 7 Rotary cut toroidal laminations

A high process stability is mandatory for a high cutting quality at the rotary cutting process. To fulfil perfect process settings, several important parameters must be optimized. A precise alignment of the tool set is essential to provide a perfect cutting clearance and avoid squeezing the sheet. Another important factor of the rotary cutting process is the setup of the correct preload torque. The cutting cylinders may not take-off, when the stamp immerses the sheet. However, deformation gets stronger with increasing preload torque. [4]

To implement the downholder and ejector function, the cavities of stamp and die are filled with plastics. During the cutting process, the plastic performs as a downholder and establishes a counterforce to avoid deformation. After cutting it also applies the force to eject the lamination from the die. In absence of the plastics component, the sheet strongly deforms and is pulled into the die by tensile stresses after the cutting process. Consequently, strong deformation occurs at the cutting edges and in the lamination. [4]

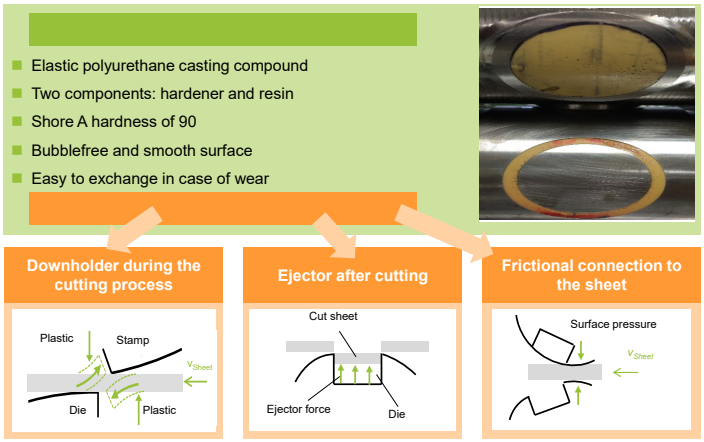
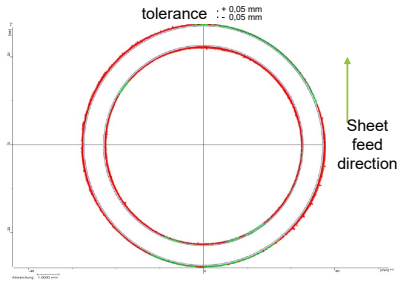


Figure 8 Characteristics of the plastics matrix

In contrast to the rotary cutting process, the process stability of the conventional stamping is very high. This is due to long experiences in this traditional process. The cutting speed and the application rate is thereby very high. In addition, the flat tooling structure and further downholder generate flat blanks without any deformations. However, the high cutting forces of up to 2,500 kN are very inefficient. To reach the required accuracy, massive and rigid stamps are necessary, which cause high investment costs. The High production costs of tools and large conversion effort are the main challenges of stamping with a wide variety. [1]

4.2. Toroidal lamination geometric properties

For the geometry measurement of the toroidal laminations, the Werth VideoCheck IP is used. In Figure 9 the results of the toroidal core sample are shown with a tolerance of ± 0.05 mm. The areas marked in red represent an exceedance of the tolerance limit. Consequently, the green areas are inside the tolerance. A total of approximately 25% of the measurement points are within the tolerances of ± 0.05 mm. [4]



	Lower tolerance	Upper tolerance
Outer diameter of stamp	55.02 mm	55.13 mm
Outer diameter of die	55.04 mm	55.06 mm
Inner diameter of stamp	44.87 mm	44.99 mm
Inner diameter of die	44.94 mm	44.96 mm

Figure 9 Matching of actual and target toroidal contour

In the tolerance field of ± 0.05 mm, the measured contour of the inner diameter exceeds the lower tolerance limit and the measured contour of the outer diameter exceeds the upper tolerance limit. Further investigations on the outer diameter show, that all measured values for the outer diameter are between 55.02 mm and 55.13 mm. Therefore, the average outer diameter is 55.07 mm. At the inner diameter, the measured values are between 44.87 mm and 44.99 mm. The average inner diameter is 44.93 mm. [4]

When considering the results of measurement, two factors are particularly of relevance. First, a general deviation of the dimensions of inner and outer diameter. The outer diameter is at most of the measuring points larger and the inner diameter smaller than the target contour. This is due to the tolerances of the tools. The cutting clearance of 0.03 mm is applied die sided and tolerated from 55.04 mm to 55.06 mm on the outer diameter and from 44.96 mm to 44.94 mm on the inner diameter. As a consequence, for future design of tools, the die is to be tolerated with the nominal dimension and the stamp must be adapted for the cutting clearance. The second deviation is at the inner diameter. It deviates most from the nominal dimension right of the inlet. The corresponding range on the outer diameter is close to the target dimension. This context implies a slight general deformation of the sheet. This can be explained by bumpy placement of the toroidal laminations on the measuring table and as a result the inner and outer diameter appear smaller in evaluation. [4]

An advantage of stamping is the very high accuracy of the manufactured parts. Stamped laminations do not have any deviations or failures and reach tolerance classes up to IT6/7. The separated laminations also can be stacked or packaged in the stamping machine. However, stamping of very thin sheets (lower than 0.30 mm) is process-related very difficult. Furthermore, the production of thin sheets with a high accuracy is cost-intensive and the processing speed must be reduced. This is due to the exact positioning of the thin and limp sheets in the stamping tools. [1]

For investigation of the cutting edge and burr, micro sections of the rotary cut laminations are made. The shown sectors in Figure 10 are the different phases of the rotary cutting process. The top of each micro section shows the stamp side and the bottom the die side of the lamination. [4]

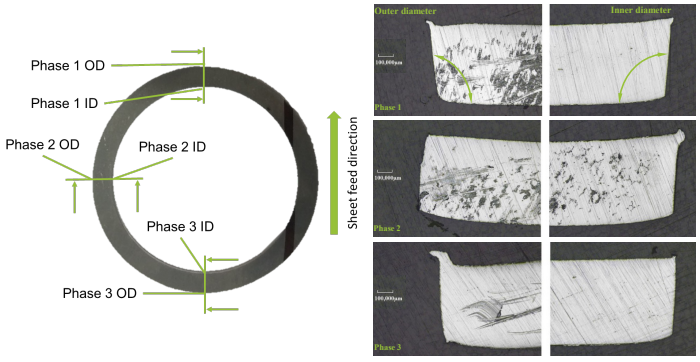


Figure 10 Microsections of rotary cut toroidal lamninations

The cutting edges of phase 1 show a typical cutting surface of shear cutting with the following areas. Top sided, there is a fractured face with burr. In the middle of the cutting edge, a smooth cutting zone is visible. Lastly, dented edges are at the bottom. Resulting from the inaccurate orientation of the cylinders, the largest burr formation is at the outer diameter of phase 3. In contrast, the inner diameter of phase 3 is nearly burr-free. Consequently, the cutting clearances of inner and outer diameter are different. The test results show that the cutting clearance of the outer diameter is too small. Moreover, it is possible that the stamp pushes against the sheet. As the cutting clearance is perfect at the inner diameter, no burr appears. The angels of all cutting edges are between 94.0° and 98.0°. These deviations arise from tilting between stamp and die. Due to the rotary movement, the angle changes constantly during cutting process. [4]

4.3. Toroidal lamination magnetic and electric properties

The electric and magnetic properties of the rotary cut toroidal cores are determined with an *Brockhaus Measurements* electrical steel tester. The Toroidal cores specimens consist of 20 single laminations, wrapped in tape and wound with two coils of 16 windings. [4]

To determine the hysteresis loop and the hysteresis- and eddy-current losses, measurements are carried out at polarizations of up to 1.50 T and frequencies between 50 Hz and 200 Hz. [4]

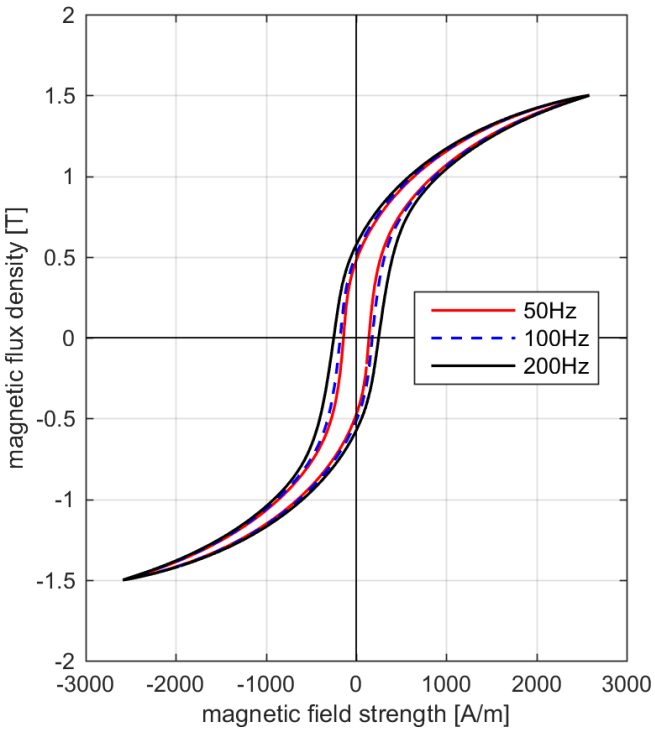


Figure 11 Hysteresis loops for rotrary cut electrical steel at 50, 100 and 200 Hz

At a frequency of 50 Hz and a field strength of 2,500 A/m, the toroidal core shows a polarization of 1.48 T. However, the hysteresis curve is very flat and does not have the desired Z-shape. Thus, the remanence is not very high after magnetizing the toroidal core. In a magnetization of 200 Hz, the remanence of the toroidal core is 0.57 T. With optimal magnetic properties, this would be as close as possible to the applied polarization of 1.5 T. [4]

According to Table 1, the total losses of the starting material at a polarization of 1.50 T and a frequency of 50 Hz are 3.30 W/kg. Although hysteresis and eddy-current losses, which are influenced by unevenness, burr formation and air gaps between the individual sheets, are significantly higher in case of the rotary cut toroidal cores, the result is typical for electric steel laminations. Due to deformation during the cutting process, cold forming occurs in the metal sheets, what is negatively effecting electrical and magnetic properties. Figure 12 shows the hysteresis and eddy-current losses of the toroidal core at different frequencies. [4]

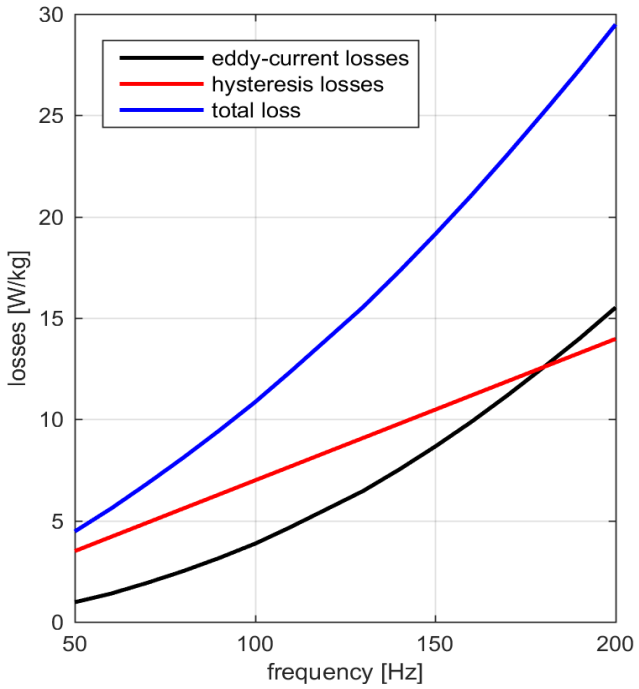


Figure 12 Hysteresis and eddy current losses of the toroidal core

The total of hysteresis and eddy current losses in the toroidal core at 50 Hz is 4.46 W/kg. This value is factor 1.35 greater than the starting value. After conventional stamping, the increasing of total losses is similar to this value. In order to achieve improvements in electrical and magnetic properties, deformation during rotary cutting has to be reduced by further development of the process. [4]

5. Conclusion

With the help of the rotary cut toroidal core samples, it has been shown that the process is generally suitable for the processing of electrical steel strip. However, the results also indicate that a great deal of development work still needs to be done. In particular, a detailed adjustment of the process parameters is necessary to achieve a stable level of high quality. To further intensify process understanding, a long-term experiment with a real application is recommended. Process stability, reproducibility and tool life can then be demonstrated.

6. Summary

This publication constitutes the state of research of rotary cutting of electrical steel strip using a toroidal core sample. Based on this analysis, a distinct statement on the technically feasible parameters is achieved and it allows a comparison with applied industrial processes.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This research activity was kindly supported by the research project “Entwicklung eines Rotationsschneidverfahrens fuer duenne Elektroblechfolien (E|RoCut)” funded by the Central Innovation Programme for SMEs (ZIM) of the Federal Ministry for Economic Affairs and Energy (BMWi) in Germany.

References

[1] J. Franke *et al.,* “Laserschneiden in der Elektroblechfertigung: Flexible und produktive Produktion dünner Elektrobleche für elektrische Traktionsmotoren in Elektrofahrzeugen,” *ZWF*, vol. 107, no. 9, pp. 642–646, 2012.

[2] M. Hubert, J. Franke, and J. Hackert, “Continuous rotational cutting of laminations for electric drives: Cutting process and tool setups for rotational cutting of laminations,” in *2014 4th International Electric Drives Production Conference (E|DPC): Proceedings: Sept. 30, 2014 - Oct. 1, 2014, Nuremberg, Germany*, Piscataway, NJ: IEEE, 2014, pp. 1–5.

[3] M. Hubert, H. Rosenow, J. Franke, and J. Hackert, “Rotary cutting as an alternative method in the processing of electrical steel strip: Geometric and kinematic determination of the process setup,” in *2015 5th International Electric Drives Production Conference (E|DPC): Proceedings: September 15th - 16th 2015, Nuremberg, Germany*, [Piscataway, New Jersey]: IEEE, 2015, pp. 1–5.

[4] M. Hubert *et al.,* “Analysis of the Rotary Cutting Process of Electrical Steel on Basis of a Toroidal Core Test,” in *2016 6th International Electric Drives Production Conference (EDPC)*: IEEE, 2016, pp. 230–236.

[5] J. Hackert, “Anlage zum Rotationsschneiden von Elektroblechen,” DE 10 2013 018 995, Germany DE 10 2013 018 995 B4.

[6] M. Schweitzer, “Prozessspezifische Merkmale des Rotationsschneidens,” PhD Thesis, Lehrstuhl für Umformtechnik und Gießereiwesen, Technische Universität München, München, 2001.

[7] G. Blumauer and E. Koehler, “Beitrag zur Untersuchung des Rotationsstanzens von Karton,” PhD Thesis, Fakultaet für Maschineningenieurwesen, Technische Hochschule Karl-Marx-Stadt, Karl-Marx-Stadt, 1974.

[8] M. Hubert *et al.,* “Explicit Finite Element Analysis for Rotary Cutting of Electrical Steel Sheet,” in *InAES 2016 – The 6th International Annual Engineering Seminar: Conference Proceedings; August 1st - 3rd, 2016*, Yogyakarta, Indonesia: IEEE, 2016.

[9] L. Bohdal and L. Kukielka, “The effect of selected material parameters on the stress and strain states in the process of cutting a sheet plate with circular cutters,” in vol. 4, *Task Quarterley: Scientific Bulletin of Academic Computer Centre in Gdańsk*, Politechnika Gdańska, Ed., 2006, pp. 391–400.

[10] G. R. Johnson and W. H. Cook, “A constitutive model and data for metals subjected to large strains, high strain rates and high temperatures,” *Proceedings of the 7th International symposium on ballistics*, pp. 541–547, 1983.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1114-1128 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Design for Smaller, Lighter and Faster ICT Products: Technical Expertise, Infrastructures and Processes

Omer Aydin*, Orhan Uçar

Research and Development Department, Netaş, 34912, Turkey

ARTICLE INFO

Article history:
Received: 12 May, 2017
Accepted: 14 June, 2017
Online: 16 July, 2017

Keywords:
Information and Communications Technology
Product Development

ABSTRACT

For the design of Information and Communications Technology (ICT) related products, basic level of technical expertise and simple development tools were enough in the early 1980s. As time goes by, new generation products designed by engineers are always expected to be smaller, lighter and faster, due to the increasing global competition and the application of hardware and software in every area of our lives. In order to be able to respond to these needs, engineers have always had to produce new technologies and solutions. In this sense, hardware related technologies have been growing much faster such as System on Chips (SoC), displays, batteries and other materials used in the ICT products. Designers have been diversifying software development processes and tools remarkably for new applications and constantly developing new communication standards. In this article, all the main design phases are presented starting from the system design to the product validation and test. As can be seen from the relevant sections, in order to design an ICT product, very deep technical expertise, expensive development and verification infrastructures are needed in the relevant fields.

1. Introduction

Before the 1980s, in the design process of Information and Communication Technology (ICT) products, it was sufficient to have technical expertise at the basic level such as entry-level material knowledge, basic skills in using measurement equipment, and the ability to use simple compilers in related fields. In those years, functionality was the main concern for the product design. Furthermore, the development processes and infrastructures were much simpler than today. In the modern era, due to the advents in technology, the emergence of new composite materials, the introduction of effective production methods and the rise and on-going diversification of consumer expectations, product design process is getting more complex. In order to meet new product development requirements, a multidisciplinary approach is required to synthesize respective knowledge from various disciplines, such as anthropology, sociology, physics, chemistry, environment, etc.. In the rise of today's consumer culture, functional satisfaction is no longer a mere reason of purchasing goods and services, while social factors substantially affect consumer decisions. Especially for mobile and wearable ICT products, the owned products are tools through which consumers express their identities. Thus, besides the proper functionality of the product, different factors, such as emotional response of the customers should be taken into account in the product design. Therefore, today's development processes are becoming too complex and multifaceted.

Until mid 1970s, the design and manufacturing processes of ICT products were mostly based on nonautomated or semi-automated processes, where technical hand-skill of the operator usually plays a crucial role. PCB (Printed Circuit Board) design and production were made by hand work, with Letraset and paper [1]. The use of the computer in the design process of ICT products was a revolution. Computer-aided design not only accelerated the design and manufacturing processes, but also resulted in a significant reduction in the related human errors. Moreover, the decrease in the duration of the design allows the production of more complex ICT products. The automation of the manufacturing of ICT products greatly increased the manufacturing volume and reduced the cost of the products that could provide the functionality never achieved before.

*Corresponding Author: Omer Aydin, Research and Development Department, Netaş, 34912, Turkey, +902165222682, aydin@netas.com.tr

The first commercial microprocessor was Intel 4004 (4-bit), which was intended for use for use in a calculator in 1971 [2]. Today, multi-core processors with high processing power are required in the design of broadband wired or wireless products or high-capacity IP (Internet Protocol) switching devices. These processors include dedicated coprocessors, special data transfer paths and data processing units, along with a large number of ARM (Advanced RISC-Reduced Instruction Set Computing-Machine) and DSP (Digital Signal Processor) processor cores running at gigahertz speeds [3]. In order to meet wide range of the required features, it is necessary to develop high performance software spread across tens of thousands of source files and, composed of millions of lines, which is designed to exploit the potential of the available hardware resources [4]. On the other hand, the hardware provides multi-core CPUs capable of performing hundreds of billions of operations per second and very high-speed and high-capacity data transmission interfaces. In addition, for the design of an ICT product, technical expertise and dedicated infrastructure is required related to mechanical design, thermal analysis, electromagnetic compatibility, product integrity and reliability, product testing and verification. As a result, the complex technological structure of ICT products entails the use of various advanced engineering skills in the design process.

In this article, various technical expertise, infrastructure and processes necessary to design an ICT product will be described. These technical expertises are related to system design, hardware design, mechanical design and software design, verification and production.

After a brief and formal definition of system design given in Section 2, issues in today's hardware design is discussed in details in Section 3. In this section, PCB design, signal and power integrity, recent developments in IC (Integrated Circuit) technologies, developments in manufacturing technologies, System on Chip (SoC) technologies, recent developments in RF (Radio Frequency) design and EMC/EMI, and concerns related to energy efficiency are considered. In this section, the requirements of a specific hardware design in today's technology are considered as an example.

Section 4 is concerned with mechanical design and discusses design considerations in the choice of materials to meet the requirements of the used process.

Section 5 discusses issues in software design. In this section, the development steps and troubles from the past to the present are given, followed by the related development processes, which are formed because of these progresses. In this section, trends emerged in the field of software development, especially in virtualization technologies, open source code development and effective use of frameworks are considered in order to create cost effective short-term software development. In addition, virtualization and IoT (Internet of Things) software emerged in this field are discussed. In Section 6, software validation tests and test methodologies are considered. In Section 7, the state-of-the-art ICT production processes are described and discussed. Finally, in Section 8, a practical implementation for a cutting edge ICT product design is described.

2. Design for Requirements: System Design

System engineering manages the product development process by taking into consideration the technical requirements, time and cost elements of the product design, production, delivery and maintenance activities [5]. System engineers may suggest either a new product on a new product or adding a new feature to an existing product, based on the feasibility study. Prior to a detailed system design, the general system-level design needs to be prepared by the system engineers. After performing a detailed study of product description, the functional behaviour and the basic system architecture are developed. This is the most critical step in the design of complex products. The ability to meet the additional requirements for the designed system depends on the flexibility and extensibility of the system design. For a system engineer, expertise and skills in all areas described in this paper are required in order to achieve the necessary engineering activities.

3. Design for Software Platform: Hardware Design

Hardware design consists of establishing system and/or subsystem hardware and mechanical infrastructure in order to achieve defined system, software and platform requirements. These requirements are concerned with the platform, mechanical dimensions, power consumption and environmental conditions, such as high/low temperature, humidity, shock, vibration, radiation, air pressure, Electromagnetic Compatibility (EMC) and Electromagnetic Interference (EMI).

In this Section, first PCB design with signal and power integrity issues will be introduced. After that, the evolution of hardware IC design and developments in manufacturing techniques will be explained. Later, SoC IC concept; application specific and high capacity ICs with multiple functions and interfaces will be pointed out. On the last part, 3D(3 Dimension) IC design technique for smaller and less power consuming ICs, developments in RF design, EMI/EMC and energy efficiency subjects will be covered.

3.1. PCBs (Printed Circuit Boards)

The first PCBs were used as early as 1900's. In those years, various materials like Bakelite and Masonite, plain old thin wooden pieces were used as PCB materials [6]. These PCB materials were drilled into the holes and the brass wires were riveting on the material. Over the following years, circuit sizes were reduced with the use of double-sided and multilayer PCBs and hot air soldering methods.

In the 1980s, surface mount components became the preferred option rather than through-hole components. In the 1990s and 2000s, the complexity of modern circuit boards increased and "micro-via" technology emerged in PCB manufacturing. In those years, the use of 5-6 mils of trace and space on PCBs was getting more widespread. On the other hand, Flex and Rigid-Flex PCBs have become an economical option and widely used. Today, ELIC (Every Layer Interconnect) technology allows for dynamic connection between any two layers within the circuit layout.

Nowadays the number of layers in a typical high-density PCB has increased up to 40. There are more than 350 kinds of dielectric materials in aerospace, communications, high-speed digital design, power, and wearable device areas. The operating frequency for epoxy / glass (FR4) material has been increased to 40GHz. In recent years, flex and rigid-flex cards have gained popularity. Since rigid-flex cards require fewer connections, they are more reliable and their long-term maintenance cost is reduced.

Printed Circuit Design was made manually before the introduction of Computer Aided Design (CAD) tools. Placement

of the components is done on transparent paper with ICs plotted as templates. Using this paper, a film that corresponds to the circuit layout is prepared and by the use of dry transfer technique, PCB is produced. PCB design is finalized by adding copper lines formed using Letraset papers.

Today there are several advanced CAD tools for the hardware design development. These tools enable multi-layer PCB design with high-speed signals and allows the diagnosis of layout problems related to built-in signal integrity, power integrity, cross talk, reflection add-on's before PCB production. Using 3D modelling features of these tools, mechanical mismatches can be detected on 3D PCB layout view. With the help of these tools, multi-layer, densely populated PCB design including high-speed data paths with low latency providing hundred billions per second processing power can be accomplished [7].

Figure 1 shows the layout of the baseband card of a 4G base station. On this card, there are both high-speed data paths (>50 Gbps) and high-speed transmission lines for inter-card communication (>10 Gbps). The design of the PCB enables the reference clock signals (312.5 MHz, 156.25 MHz and 122.88 MHz) to have very low jitter (4 ps RMS) and very small 50 ps rise/fall time.

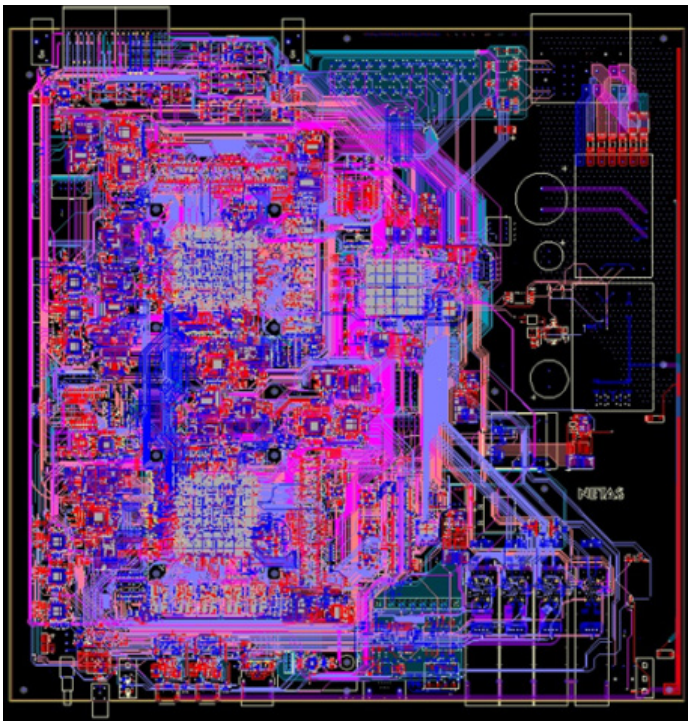


Figure 1 PCB layout for a 4G baseband card.

3.2. Solving Transmission Issues for High Speed Digital Design: Signal & Power Integrity

Signal Integrity deals primarily with the signal transmission performance within an electronic product. This issue is due to the electromagnetic behaviour of the system and the tools to study the related effects remain more or less the same since the beginning of electronic transmission. First transatlantic telegraph cable was exposed to serious signal integrity problems, and this led to invention of the mathematical tools for signal integrity analysis, such as telegraph equations, which we still use today [8].

Typical ideal and real waveforms of a signal are shown in Figure 2.

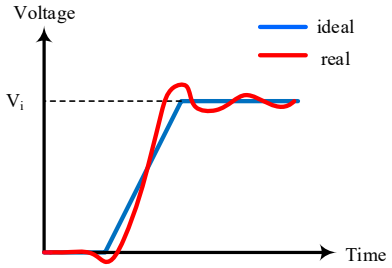


Figure 2 Ideal and real waveforms at the receiving gate.

On PCBs, when propagation delay of a signal approaches signal transition times (rise/fall time) and the system speeds exceed a few tens of megahertz, signal integrity analysis becomes a necessity. Today signal speeds reach several GHz.

Since 1990's, the requirements for signal integrity analysis are getting more and more important. Initially, PCB trace speeds have become an important issue in the PCB design. Designers approached this problem by considering analog characteristics of digital data paths. The main objective of signal integrity study is to ensure reliable transmission of digital signals between transmitter and receiver. While initially the main focus in signal integrity study was the connection between transmitter and receiver, in modern designs, signal integrity models are used in the design of transmitters, receivers and every part in between.

By the years of 2000s, power supply for digital circuits has become a big problem and a need for Power Integrity (PI) analysis has emerged. Signal integrity (SI) analysis shows signal's behaviour through PCB traces, conductors and passive elements. However, PI analysis examines power distribution lines by evaluating power plane impedance and decoupling capacitors. While PI is related to the system reliability analysis, its main goal is to ensure that drivers and receiver have enough current to send and receive 1's and 0's [9].

In the early phases of electronic circuit industry, available ICs came with one power and one ground pin. A bypass or decoupling capacitor between the power rail and the ground pin was more than enough to filter out the noise. Today, a complex Ball Grid Array (BGA) features hundreds of power and ground pins and more than 10 different voltages under 1V. These complex ICs typically draw currents of tens of amperes. In order to supply clean and noiseless power at these levels, a large capacitor network is required. The effective design of this capacitive network requires the accurate analysis of the propagation of energy in a power distribution network, which includes capacitors, vias and power/ground planes. For instance, , since 1990s the operating voltages drops from 5V to sub 1V's, where operating currents rises from 1A to several 10A's. These show that power consumption rises from 5W to several 10W's and impedance target drops from 500mΩ to1mΩ.

Today, EMC and Signal Integrity (SI) are the most important issues in electronic product design, which should be considered through all the design process. The main reason is the high data bandwidth demand of new generation high performance products and servers, which operate over hundreds of Gbps. EMC and SI Technologies on PCB and IC packages are one of the bottlenecks at these bandwidth requirements.

In Figure 3, the variation of signal rates used in serial links in the last 40 years is shown.

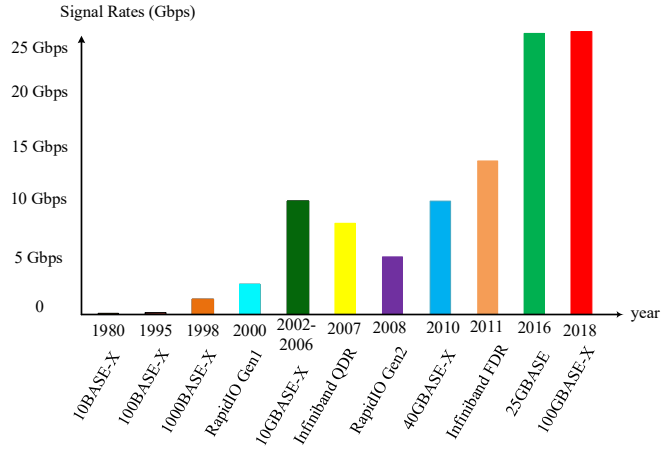


Figure 3 Signal rates of serial links in the last 40 years.

As seen in this figure, signal rates that were around 10 Mbps for 10 Base-X Ethernet in 1980, have reached 25 Gbps today. An interesting feature in this graph is the decrease in signal rates, which appear in years 2007 and 2008 [10]. In these years, implementation difficulties of 10 Gbps per line on long backplanes due to signal integrity issues led way to interconnect standards consist of multiple port high-speed serial lines such as RapidIO or Infiniband QDR. In 2007, Infiniband QDR drops line rate to 8 Gbps, however supports operation on four parallel lines to achieve 32 Gbps data throughput. In 2008, Serial Rapid IO standard defines the maximum line rate as 5 Gbps and supports 2, 8 and 16 parallel lines, which in turn supports total throughput of 10, 40 and 80 Gbps respectively.

One possible approach to increase data rates is to increase the Non-Return-to-Zero (NRZ) data rate with very high-speed serial links. Another possible way is to encode the data using PAM4 (Pulse Amplitude Modulation). Using PAM4 modulation, it is theoretically possible to reach a collective data transfer rate of 800 Gbps.

Nowadays, for PAM-4 coding, 13 GHz clock frequency is used and 400 Gbps Ethernet speed is reached with 8 lines, each featuring 53 Gbps speed [11].

As mentioned above, the ultimate goal is still smaller, lighter and faster products. In order to achieve these goals, additional issues such as the implementation of appropriate inter-system connection interfaces should be taken into account. In order to meet the design requirements, robust, miniature and high-speed connectivity solutions should be considered.

3.3. Progresses on Integrated Circuit and Manufacturing Technologies

The fastest change in the semiconductor technology era emerges in the field of processor and memory oriented semiconductor manufacturing. The key drive factors in this evolution are higher speed (executed commands per second, data transfer rate) and higher capacity (memory, number and different types of interfaces). Transistor size in IC's reaches from several micrometers to tens of nanometers. The prediction of Gordon Moore, one of Intel's co-founders, about transistor number in a dense integrated circuit will double every two year in 1965, proved to be right in the following 50 years as can be seen in below Figure 4 [12].

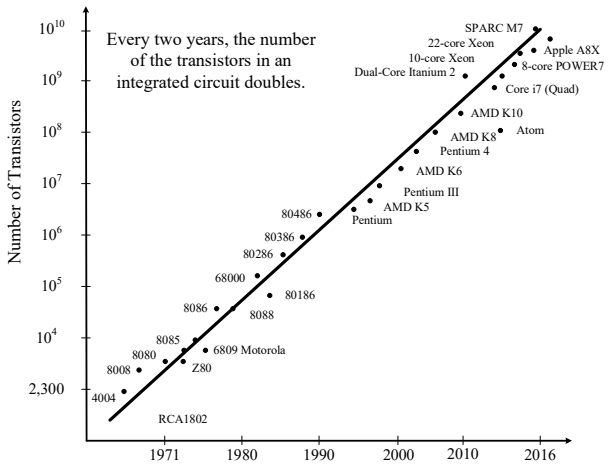


Figure 4 Number of transistors for CPUs and Moore law.

Due to the high integration capability, programmable semiconductors for different features in the same package can be developed for digital circuits, interfaces, power electronics, RF (Radio Frequency) and analog circuits as well as processors. Every new integrated circuit has more rule set and documentation for usage. Design, manufacturing and test processes become more and more complex and need different methods and tools than the previous ICs.

In recent years, performance and success metrics in integrated circuit design and manufacturing has changed dramatically. Up today, the main success metric for personal computer ICs is clock speed. Therefore, the manufacturing technology has developed according to these criteria. Nowadays, form-factor, cost and power consumption are the critical driving factors, especially in mobile market; that in turn has increased the importance of on-chip integration of functional hardware (e.g. power management, computing, audio/video, graphics and radio). In order to integrate these hardware on-chip, specifically designed for purpose design and simulation techniques are used. Today, there are around 20 fabs that can manufacture analog and basic digital ICs with 130nm ~ 28 nm process where there are only 5 fabs which can use 14/10nm process to manufacture market dominator SoCs [13].

The need in the present decade drives technology markets to use more and more integration such as mobile communication, data centers, automotive and IoT platforms. This demand has increased the importance of lower nanometer process and more complex manufacturing technologies. 7 nm processes is expected to be realized within the next two years, where studies and early trials continue for 5 nm [14]. As a result, more analog and digital functional block will be integrated within a single chip. Analog functional blocks will be manufactured with 14 nm process and very low power analog/RF circuits targeting IoT platforms can be designed. FinFET (Fin Field-Effect Transistor) technology may be one of the viable methods to achieve this goal.

The distinctive feature of the FinFET is that the conducting channel is wrapped by a slender semiconductor "fin" that forms the body of the device and the fin thickness determines the effective channel length of the device. FinFET technology will be the basic infrastructure for the complex semiconductors produced within next decade. Because of its features like the ability to be fabricated with its channel along different orientations in a single

die due to its nonplanar structure, the reduction of "Short Channel Effect according to its fin thickness and fin length, ability to control its threshold voltage with multiple voltage sources and reduction of off-state current. For example, a mobile chipset produced with 14 nm FinFET technology in August 2016 is performing % 30 faster, increased core speed by % 20 and consumes % 35 less power compared to its predecessor [15].

3D IC comes forward as an alternative to improve manufacturing technology and managing processes under 14 nm. 3D ICs are manufactured by placing silicon layers vertically on top of each other and connecting layers with TSVs (Through-silicon vias). Using this technology, a cost and performance effective development can be accomplished. $16 billion investment is forecast to build a foundry for a more advanced technology level of 7 nm process node [16]. 3D IC design is discussed in detail in the section 3.6 for size reduction and low power consumption purposes.

3.4. Processors: CPU, GPU, NPU, APU, PPU and TPU

CPU (Central Processing Unit) is the central processing unit of computers. After the invention of transistor, Intel 4004 CPU with 2300 transistors was produced in 1971 using a 10 µm technology operating at 740 kHz. Today, processors can be manufactured with 14 nm technology and contain 10 billion transistors. These processors are capable of running at 4.2 GHz speeds and support many types of instruction sets. General-purpose processors perform general purpose operations and are used in many areas, especially in personal computers and [17].

In GPU (Graphical Processing Unit) processors, it is important to have parallel command processing capabilities, as well as sequential processing capabilities that conventional CPUs have. This feature is also used for accelerating computation in areas such as 3D rendering in computer games, advanced scientific research, oil and gas exploration and self-controlled driving technology [18].

In terms of architecture, general-purpose processors can be made up of several CPU cores and can do several jobs simultaneously. Similarly, GPU processors have hundreds of cores to process parallel jobs efficiently. Today, Multi-GPU processors with many GPU processors are produced [19].

A NPU (Network Processing Unit) is a dedicated processor specially designed to perform network related functions. While these units are often intended to be used in embedded systems, they are also employed in the personal computers. For example, Mellanox's Indigo NPS-400 unit has an internal processor unit, which provide all services including traffic management at 400 Gbps capacity. This processor has development support in C programming language [20].

Large processor vendors such as Intel and AMD (Advanced Micro Devices) have developed APU (Accelerated Processing Unit) technology by combining the superior features of CPUs and GPUs. APU processors are used in areas where power consumption and performance are important. For example, a large GPU manufacturer Nvidia has collaborated with the ARM Company, to produce the APU named Tegra, and has entered the mobile and smartphone markets [21].

A PPU (Physics Processing Unit) is an interface, which has been developed for processing physics rules in graphics processing. These processors compute special operations such as the effect of gravity and the effects of collision of objects in accordance with the laws of physics. It was originally developed by Ageia in 2006. However, after merging with Nvidia, it is developed continuously as a middleware running on GPUs under the name PhysX [22].

A TPU (Tensor Processing Unit) is developed by Google for TensorFlow open source library. It is application-specific ASIC technology specially developed for machine learning. These units are widely used in the data centers of Google. An important asset of these units is reduced computational accuracy compared to GPUs. However, since fewer transistors are used in these units, they offer better performance than their counter parts in terms of unit power consumption [23].

3.5. Design for High Capacity: SoC (System on Chips)

In the ICT products developed today, processors, which include several different functional blocks, are used for high capacity and special function needs that cannot be met with generic processors on the market. In these processors, there are several CPU cores, general-purpose DSP cores and application-specific co-processors, which provide low latency required in applications such as error coding/decoding and Fast Fourier Transform (FFT) algorithms. These processors provides performance, easy development cycle, flexibility and product cost advantages. Today, programmable SoC, FPGA based SoC and RF specific SoCs are used for different applications.

Today, SoC manufacturers are purchasing the IP Core license from ARM to integrate it into the chip design. In this way, fabless companies can develop designed, tested and production ready processor cores. The new chips designed by these companies are produced in semiconductor fabrication plants (fabs) to meet high performance and low-cost requirements [24].

Figure 5 shows functional block diagram of a general SoC used in the design of ICT products.

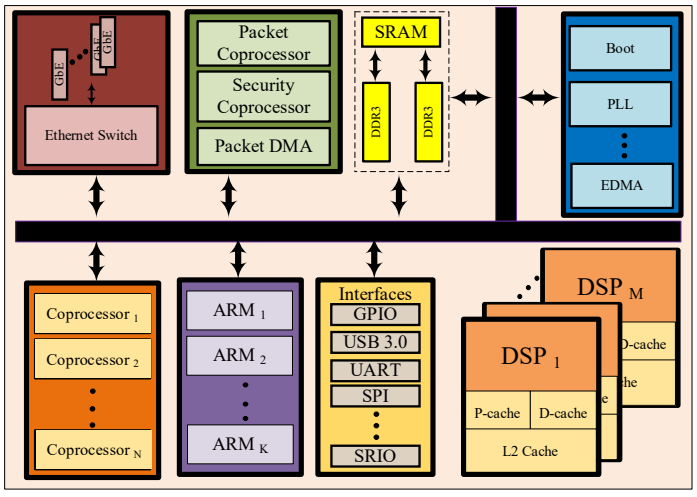


Figure 5 Functional block diagram of a SoC.

3.5.1. PSoC (Programmable System on Chip)

A cost effective solution to general purpose SoC ICs are programmable SoCs, which includes only specific functional blocks and interfaces required in the designed system. Using this architecture, unused functions and interfaces are not integrated into the specifically designed ICs and this provides the main advantages of ASIC (Application Specific Integrated Circuit), such as performance, form-factor and product cost. The

availability of the optimized use design of these functional blocks reduces design time and eliminates nonrecurring engineering cost of programmable SoCs. The flexibility of this architecture enables adding or removal of functional blocks in a very short time. Currently, these PSoCs can provide low capacity microcontroller, memory and interfaces and therefore they are being used for products and solutions targeting IOT platforms [25]. Cypress announced that an IOT solution using PSoC 6 series will be provided soon.

3.5.2. FPGA Based SoC (System on Chip)

To provide an alternative to general purpose processors used in product design, FPGA firms developed "soft" processors that can be put into standard FPGA components as object or source code. "Microblaze" and "Picoblaze" processors are good examples for this approach. The next step is placing "hard" "Power PC" processor core into "Virtex" FPGA. The latest FPGA family ZYNQ includes 2 ARM cores as standard product. ZYNQ Ultra scale+ SoCs are produced with 6 ARM core and 1 GPU (Graphic Processing Unit) core [26].

Central Processing Units (CPU) work sequential by design, increasing their performance can only be accomplished by increasing operating frequency or number of CPU cores. FPGA based processors have the advantage of parallel processing architecture, low latency and gigabit high-speed interfaces.

3.5.3. RF SoC (System on Chip)

In the future of the communication, transport, energy and automotive sectors as well as vertical IOT market, integrated circuits produced with low nanometer technologies are expected to be dominant. AMS (Analog-Mixed Signal) technologies and nm based analog block manufacturing processes used during production of these ICs are the main differentiating factor. Because of increased device model complexity of FinFET, complex layout effects impacting design specs, increased parasitic complexity, low-voltage driving intensive noise analysis issues during transition into nanometer production process reveals long and lots of simulation work and huge net lists. Lacking of fast, nm-accurate mixed signal verification limits design and manufacturing success, which leads to 50% first silicon failures. Research continues to integrate analog, digital and RF technologies to provide ultra-low power requirement of IOT applications. Main focus is developing new mixed-signal circuit architecture. In addition, the need and usage of RF MEMS (Micro Electro Mechanical Systems) in satellite communications, radar systems, and adjustable antennas increases.

3.6. IC Design for Size Reduction & Low Power: 3D Integrated Circuits

As mentioned in section 3.3, 3D integrated circuit emerged out as an alternative to improve manufacturing technology. Ever since the combination of analog and digital integration becomes a necessity; 3D ICs, which enables manufacturing of functional blocks with different technologies and rule-sets separately and integrate them in the last stage of the production, offer significant improvements over two-dimensional circuits. Average connections inside 3D ICs are much shorter and interfaces inside the IC can be work faster consuming less power. An application processor designed as 3D IC is reported to have 43% less space, 13% less power and 14% less connection length in average compared to conventional designs [27].

Figure 6 shows 3D IC structure including many different circuits with 3D interconnect.

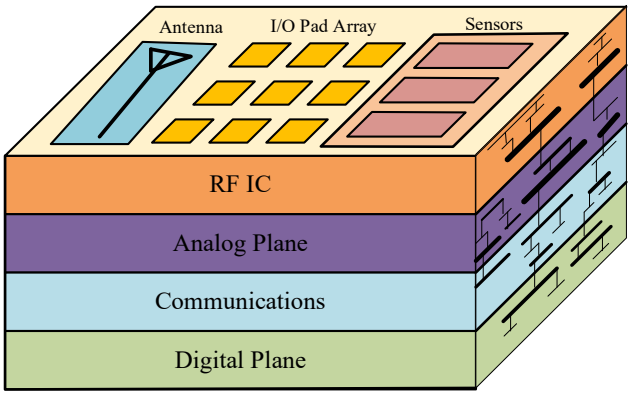


Figure 6 Different circuits on the same die, tied together with different 3D interconnect.

3.7. Batteries

When choosing a battery for a product, there are many criteria to take into account, such as capacity/volume, charging duration, voltage match, polarization/rate capability, weight, and cycle life/shelf life, operating temperature range, safety, toxicity and cost. For instance at the beginning of 1900's nickel cadmium batteries helped pave the way for modern technology, but because of cadmium's toxicity, alkaline batteries spread in the market after 1950's. Alkaline batteries were safer and cheaper but their energy density was not enough to supply high performance devices.

While electronic devices are getting more and more complex, more energy density and form factor are required. Nickel-metal-hydride and lithium-ion, emerged in the early 1990s, provides high energy density. For example, the energy density of lithium-ion is typically twice that of the standard nickel-cadmium. Today, lithium-ion is the fastest growing and most promising battery chemistry [28].

The high cell voltage (3.6V) capacity of li-ion's allow battery pack designs with only one cell and most of today's mobile phones run on a single cell. A nickel-based pack would require three 1.2 V cells connected in series. This allowed smart phones to have relatively larger display size and longer lifetime [29]. Lithium-ion batteries suit well to low temperature and warmer environments compared to NiMH cells. Li-ion can operate to a temperature of as low as –40°C that increases the data security at harsh environments.

There are specific cathode types for each application for instance laptops and smartphones use lithium cobalt dioxide (LiCoO2) and electrical vehicles use lithium nickel cobalt aluminium oxide (LiNiCoAlO2). Nearly half of the global cobalt supply is being used by in lithium-ion battery industry. Therefore in order to overcome from the lack of cobalt and achieve better battery performance researchers work on different alternative technologies such as calcium ion batteries, Na-Ion batteries, metal-air based rechargeable batteries and sulphur based rechargeable batteries.

3.8. Sensors

3.8.1. Magnetic Sensors

Essentially there are two types of magnetic sensor: which measures total magnetic field and magnetic field vector components [30]. There are various ways to sense magnetic fields and there are also many factors such as size, sensitivity, linearity, range, frequency bandwidth, dimensions, cost and power, which affects what sensor is best suited for an application.

Magnetic sensors contributed lots of field in ICT and industry. For instance, computers now have storage disks with very big capacity, higher safety standards have been achieved because of the high reliability of without contact switching in aircrafts [30].

Magnetic sensor application may be listed in four categories as given below.

- Low-Sensitivity Application: Hall Effect sensors are most commonly used ones to detect fields greater than the Earth's field.
- Medium-Sensitivity Applications: Anisotropic Magneto Resistance sensors have become standard off-the-shelf devices for use in medium-accuracy applications, such as compasses for mobile devices [31].
- High-Sensitivity Applications: Flux-gate sensors are widely used in military industry to detect ferromagnetic objects such as a mine, ship, tank, or aircraft [31].
- Medical/Biological Applications: The SQUID sensors are mainly used in bio magnetic applications, which helped to measure extremely weak magnetic fields generated by the brain [32].

3.8.2. Accelerometers

An accelerometer measures linear or angular acceleration. There are many types of accelerometers developed and reported in the literature. The vast majority is based on piezo-electric crystals, but they are too big to use in small products. Therefore, engineers developed smaller accelerometers using MEMS (Micro Electromechanical Systems) technology [33].

With their robust sensing with little power consumption, MEMS accelerometers have an important commercial potential [33]. Accelerometers are also used in medical equipment such as prosthetic arms. Another important application field is inertial navigation.

Inertial navigation is a navigation technique, which incorporates accelerometers and gyroscopes to track the position and orientation of an object relative to a known starting point, orientation and velocity [34]. For example, PIGA (Pendulous Integrating Gyroscopic Accelerometer) type of accelerometer is used in inertial navigation systems. PIGA sensors are used in guidance system of missile systems today, such as German Rocket V2 [35].

3.9. Displays

Until the beginning of this century CRT displays, which have low picture quality had been widely used in computers and televisions [36]. In the last 20 years, various display technologies have been developed by researchers. Generally brightness, contrast ratio, resolution, size, number of colours, viewing angle, power consumption, response time and life time are key points when selecting the right display for your application. LCD, LED and OLED displays have been most commonly used in consumer products and with the development of rugged touch displays defence industry has started to use them in many applications. Sun light & night vision readable, glove friendly touch screens are the most popular ones in defence industry. Additionally, helmet mounted wearable displays provides augmented data and reality for the next generation soldiers.

Progress in LCD technology has enabled the development of thin, high definition, large display screens. The development of curved screens has been provided with OLED technology. On the other hand, with the development of LAFi (Luminous Array Film) technology curved super-large screens are possible [37].

There are also other display technology such as DLPs, PDPs, FEDs, and electronic paper (electronic ink) competing for market share [38].

In recent years, considerable attention has been paid to 4K, or 8K TV displays, that makes the viewer feel immersed in the image scene [39]. 3D holography image, bendable flexible display with plasma technologies are also hot topics in this field.

3.10. RF Design for 5G

5G (5th Generation) communication technology and IoT are the main driving applications shaping future RF technologies. 5G technology is being developed as a solution for high data speed, low-latency and low power consumption requirements of Ultra HD, 360 degree video transfer, VR (Virtual Reality), smart home, smart city, autonomous vehicles, M2M (Machine-to-machine) communication applications. All these applications require high bandwidth to answer high data throughput needs. Existing bandwidths of frequency bands used for cellular communication cannot support requirements of 5G, research is being done to use 10 GHz, 28 GHz, 32 GHz, 43 GHz, 46-50 GHz, 56-76 GHz and 81-86 GHz bands [40]. Radio signal on these candidate frequencies is deeply affected by atmospheric fading, rain, buildings, human beings, flora and reflection. To overcome these difficulties in RF design, signal processing studies focus on adaptive beamforming. With adaptive beam forming, high data speeds according to the receiver locations can be achieved [41]. Complex circuit structures are used to solve the problems of adaptive beam forming and massive MIMO (Multiple-Input Multiple-Output) technologies. RF Insertion Loss should be lowered to reduce circuit complexity, dimensions and power consumption, which are usually very costly in microwave frequencies. Zero-IF architecture is used for this purpose. The resulting local oscillator leakage issue in Zero-IF architecture is solved with signal processing techniques in digital domain [42].

From semiconductor technology perspective, GaAs (Gallium Arsenide) is the dominating technology of microwave industry for years. SiGe (Silicon-Germanium) technology has major benefits over GaAs in high-speed applications. SiGe BiCMOS (Bipolar Complementary Metal-Oxide Semiconductor) technology also fulfils high-level integration requirements of beamforming systems [43].

Mechanical design is also an important part of RF IC design. Substrate and antenna should be taken into account in order to minimize the losses in integrated circuit design. Less than 50 GHz, antenna and several passive components can be designed as part of the substrate. Studies on SIW (Substrate Integrated Waveguide) continue to be used in this type IC designs. Antenna elements above 50 GHz are small enough to fit into integrated circuit [44].

3.11. Design for EMC/EMI

EMC and EMI technologies become a critical design criteria due to technological developments [45,46]. When necessary precautions against EMC and EMI are not taken during design process, design time and hardware design, mechanical design and test cost can be much higher. To verify if EMC requirements of the designed system are met, very expensive test infrastructure (anechoic chamber, signal generators with several signal types and output power, spectrum analysers, antenna, etc.) should be established. If required by regulations, these tests should be carried by certificated labs.

Today, the voltages in the circuits are low and the currents are high. For this reason, noise and interference problems in digital circuits are mostly due to magnetic field and inductance. For this reason, engineers should take precautions for the magnetic field instead of the electric field for the products they develop [47].

3.12. Energy Efficiency

Total electric consumption of the world has reached 20567 TWh (Tera Watt hour) in 2015 [48]. ICT sector share is %4.6 of total power consumption [49]. Although the figure is relatively low, studies to reduce electric consumption continue due to environmental concerns and cost. Sector participates in initiatives such as EARTH (Energy Aware Radio and network TecHnologies). In addition, operators with the collaboration of universities and research institutes conduct studies about future network infrastructures to reduce CO2 emission.

Figure 7 shows global energy network infrastructure and energy consumption of user terminals [50].

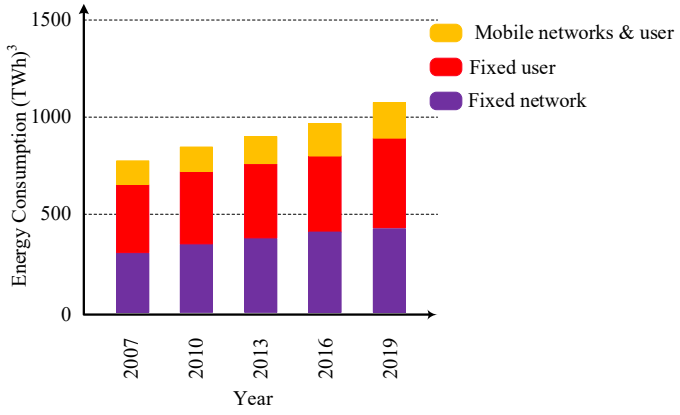


Figure 7 Global energy consumption for network infrastructure and user equipment during operational phase.

The first studies in mobile communication aimed to reduce energy consumption, targeted reducing current draw from battery in order to extend battery life. Advanced RF design techniques like linear and efficient PA (Power Amplifier) design with envelope tracking, adaptive transmit power control, radio aware software management, and development in semiconductor production, extend the battery life of smart phones from 310 minutes to 430 minutes in 5 years [51].

Increasing data requirements of smart phones, laptops and smart TV devices raises energy consumption of mobile communication systems. According to ITU (International Telecommunications Union), energy consumed in the network is %10 of the energy consumed by ICT eco-system and this value is sufficient to illuminate the entire world in 1985 [52].

In mobile communication sector, %80 of the energy is consumed in base stations [53]. Most of this energy is spent for cooling equipment, transmission losses, power amplifiers, transmitter/receiver devices, baseband processing and AC/DC and DC/DC converters.

Base stations do not operate at maximum capacity most of the day. Reducing power consumed at this low traffic intervals, reduces CO2 emission and operator costs dramatically. For this purpose, antenna muting, small cell usage and power-save mode implementation on radios are widely used.

Passive cooling design, their ability to provide service at network edge and low RF power of Small Cell devices helps to reduce operational costs. Small Cell usage improves battery life and coverage area for the users as it helps operators by optimizing energy consumption and easy-to-deploy characteristics in crowded areas.

Figure 8 shows power consumption in cellular networks [54].

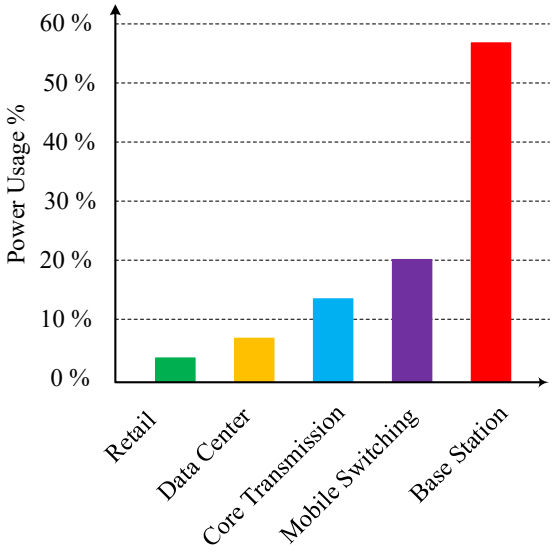


Figure 8 Power consumption in cellular networks.

Macro base stations are becoming more efficient over time. Thanks to improvements in active RF component design, it was possible to use RRH (Remote Radio Head) units. RRH units contribute to energy efficiency because they are designed with passive cooling and are close to the antenna, reducing RF transmission losses. For example, because of the 3 dB reduction in transmission losses to the antenna using RRH, the output power of the power amplifier is reduced by half.

As the use of massive MIMO and adaptive beam forming technologies in base stations, the power used per transmitted bit is reduced. In these technologies, passive cooling is possible because of the use of many small amplifiers in the base station and because the amplifiers are spread over a large area. On the other hand, the beam gains generated by a large number of antennas are significantly increased. This means that less power is consumed from each power amplifier. In addition, as the capacity of the base stations increases, the total amount of energy consumed by the network decreases.

Use of small cell base stations, massive MIMO and adaptive beam forming technologies and use of RRH devices equipped with technologies such as envelope tracking and Doherty design GaN amplifier will enable lower power consumption.

3.13. An Example: LTE-Advanced Baseband Hardware Design

For today's advanced technology product design, the LTE-Advanced (Long Term Evolution) Baseband unit is an example. This unit has various peripheral cards such as baseband cards, backplane, redundant control card, fan card and power card. There are various high speed signals (Ethernet, Serial Rapid I/O, SGMII Ethernet) in the backplane of the unit with the highest 10 Gbps.

An exemplary Baseband Card is shown in Figure 9. This card contains 550 types of components, 62 of which are integrated circuit. The total number of components in the card is 4212. These electronic components are arranged on a 16-layer PCB with a width and length of 32 cm.

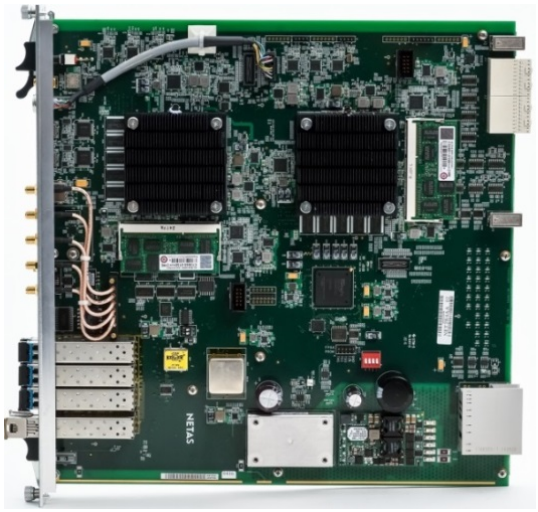


Figure 9 Baseband Card for 4G base station.

There are 10 different supply voltages between 5V and 0.75V, which supply the baseband card. The total current drawn from a 1V supply on one card can reach 40A.

There are various interfaces (CPRI-Common Public Radio Interface, Ethernet, reference clock) up to 4.9 Gbps on the baseband card. There are also various interfaces (hyperlink, serial rapid I/O, SGMII Ethernet) up to 6.25 Gbps among the master and slave processors of the card.

The SoC used in the baseband card has 8 DSP cores (1.2 GHz), 4 ARM cores (1.4 GHz) and 30 application specific coprocessors. Only the hardware documentation required for the hardware design constitutes a 5700 page document set.

There are 1517 pins of the SoC integrated circuit on the baseband card. In order to be able to perform the layout process, advanced production techniques such as blind and buried via, micro via, and bypass via are used.

Layout constraints for fast differential paths such as 10 GHz must be applied very rigidly, taking into consideration both the distances between the differential paths (at most 0.14 mm) and the distances between the source and the target of the signal (at most 0.24 mm). This situation is illustrated in the Figure 10 below.

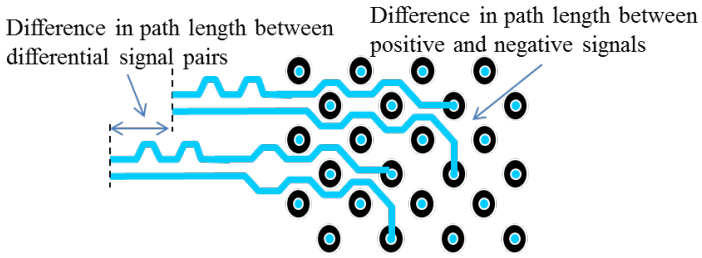


Figure 10 Path length definitions for differential signals.

Layout constraints for fast differential paths of the Baseband Card are illustrated in the Table below. Considering that 1 ps (pico second) is one in a trillionths of a second, and the thickness of the hair is 40 micrometers, it can be seen how difficult the following layout constraints can be achieved.

EMC/EMI precautions for high-speed interfaces must also be held in order for this card to function properly and the relevant processes must be implemented from the beginning of the hardware and mechanical design.

Table Layout constraints for fast differential paths.

Interface	Difference in path length between positive and negative signals	Difference in path length between differential signal pairs
1 Gbps Ethernet	< 5 ps – 700 µm	< 10 ps – 1.400 µm
4.915 Gbps CPRI	< 5 ps – 700 µm	< 10 ps – 1.400 µm
5 Gbps Serial Data	< 1 ps – 140 µm	< 15 ps – 2.100 µm
50 Gbps Hyperlink	< 1 ps – 140 µm	< 15 ps – 2.100 µm

4. Design for User Services: Software Design

Today, software has entered into all areas of life without exception. The diversity of emerging technologies and demands makes the software even more complex. The software now communicates more with peripherals and external systems. The development of network structures requires that software be more aware of each other. Manual data transfers over the past years are now done in real time. All of this has increased the compatibility issues of software.

It is said that the progress on hardware follows Moore's rule. But it is not possible to say that the software fits into this progress. In the early years when the software was growing and becoming more complicated, the designers faced more difficulties than the hardware because the processes were not fully matured [55]. For this reason, most software projects have been delayed for years. As a result, costly and unreliable, hard-to-maintain software is designed for these projects. During this period, software design has entered a crisis period. Hardware costs have fallen while software costs have increased.

Today, new techniques, design methods and tools are developed to control the natural complexity of large software systems. However, software design and test costs still constitute a large part of the design cost of ICT products today.

4.1. Software Development Processes

Companies working in software development are choosing different software development methods ranging from traditional approaches to agile approaches, depending on the needs of the projects they develop. The results of research on methodology selection in software projects show that traditional methods, including the Waterfall model, are still popular, although agile methodologies like Agile Unified Process and Scrum are more common than they were 10 years ago [55]. These surveys have shown that software development organizations also use a hybrid approach based on a large number of methodologies on projects.

Today, widely used software design methods are listed below. One or more of these methods may be preferred depending on the complexity of the software to be developed, the number of software engineers to work with, whether the customer requests are specific, the priority of the documentation, the design time, and whether the system requirements are evident clearly [55].

- Ad hoc
- Waterfall
- Evolutionary Process Models
    - Incremental
    - Spiral
    - Rational Integration Process
- Agile Methods
    - Feature-Driven Development
    - XP (Extreme Programming)

4.2. Current Trends for Software Development

Today, the complexity of software and the number of developers working on software projects are increasing. In addition, competition in this area is increasing. As the reliability requirements of the software increase, the duration of the project is also shortened. There are various developments in the field of software development, mainly open source code and framework usage, in order to enter the market quickly and cost effectively. This section describes the trends in software development today.

4.2.1. Open Source Software Development

Open source software is a software model that enables the source code to be reviewed, modified and distributed to third parties for free or for a fee. Today, almost all open source software is available for free. The fee for such codes is usually for additional services such as documentation and guides, technical support, maintenance, etc. provided with the source code of the software [56].

In the 1970s, open source software began to become widespread with the release of an open version of the Unix source code, an operating system with a large user base. This new type of software, which has a license that anyone can edit the source code as they like, has received great interest and has since become widespread.

Open source and traditional software are different in terms of total cost of ownership, flexibility, software quality, security, firm dependency and open standards. In some software products, such as Operating System, Web Browser, Web Server, Web Content Management, Mobile Operating Systems and Applications, the products developed as open source have become bigger market share than their alternatives. Today, the use of open source code has become increasingly widespread due to cost, and faster entry to market advantages.

4.2.2. Universal & Reusable Software Environments: Frameworks

The Framework contains many software libraries and allows developers to add their own libraries to them. It also guides how the software should be developed. In this way, the designers work with previously tested and reliable software. In addition, the designers can be guided on how to develop the software project to prevent the development mistakes.

Frameworks include libraries as well as design patterns, compilers and test tools. In this way, a development environment is also provided to designers to develop the software project. For example, big chip manufacturers in the embedded design world are no longer just producing chips. These companies also offer software development environments that include customized libraries, compilers suitable for the development environment, and even operating systems for the chips [57].

Chip manufacturers distribute their own product development environments free of charge, in order to promote the use of development environments and to dominate the market. Software project developers choose the framework they need to meet their requirements, leaving the details to the "framework".

Today, widely used libraries and frameworks are given below.

a) OpenCV (Open Source Computer Vision Library)

It is a cross-platform library originally developed by Intel for computer vision. Windows and Linux platforms are supported. Since this library includes advanced image processing algorithms, it is possible to develop image processing applications quickly using this library [58].

b) OpenCL (Open Computing Language)

It is a kind of framework designed to develop programs in heterogeneous systems including CPU, DSP, FPGA, GPU and similar hardware accelerators. OpenCL offers APIs (Application Programming Interfaces) to program and control these processing units. This framework also provides a standard interface for data and task based parallel programming [59].

c) .NET Framework

It is a software development framework originally developed by Microsoft for Windows installed machines. Using this framework is advantageous for designers who want to develop software on Windows. The use of this framework has now become standard for windows machines. Today, this framework is also attractive to iOS and Linux developers by providing cross-platform support. This platform continues to evolve with the principles: any designer, any application and any platform [60].

d) Coca Touch

It is a framework, which is developed by Apple to develop applications on iOS (IPhone, IPad), watcOS (AppleWatch) and tvOS (AppleTV) operating systems. Applications developed with this framework can only work on the MAC operating system. These applications are developed with the Objective C ++ language. This "framework" also provides hardware drivers and software support, and applications developed using this framework can be downloaded from the AppStore [61].

e) ActiveX

This framework is developed by Microsoft and is used to manage downloading of content over the internet. This framework is designed for downloading and updating the software modules over the internet. The first version was released in 1996. Today it is still supported on Windows 10 and by Internet Explorer 11 [62].

4.2.3. Virtualization

For virtualization, IT (Information Technology) and Telecom sectors have been conducting researches and realizations for many years. Virtualization can be defined as software-based representation of physical resources. The work in this area is focused on servers, data storage devices and communication networks. The goal is to allocate existing physical resources efficiently to services that use these resources, increase service continuity, and reduce installation and operating costs. Because of these advantages, cloud computing applications can be preferred in terms of installation and operating costs, especially with pay-as-you-go model for small and medium-sized businesses. Recently, new concepts such as "fog computing" have emerged in this area [63,64]. In this concept, resources that are used flexibly via the cloud are transported to devices at the edge of the network, for example, to WiFi hot-spot devices, sensors or IoT devices. In this model, approaches such as caching is used to optimize the use of bandwidth, reduce latency and network traffic. In this way, the service quality of the users in the access networks is being improved.

Network function virtualization is a special application of the concept of virtualization to communication networks. In this concept, network functions realized with special hardware are realized with software modules on standard server hardware. This is actually the application of virtualization techniques used in data processing centers to communication networks. Nowadays, this concept can be used in web servers and data processing centers. However, there are some problems to be solved to apply this concept to real-time network devices. The most important of these troubles are high performance requirements and data processing delay variations.

In addition, live migration of network functions between hardware platforms is underway from the point of view of performance of transport time and service interruption time.

The use of this architecture is becoming more widespread as NFV (Network Function Virtualization) technologies reduce investment and operating costs, especially power consumption costs.

4.2.4. Software Development for IoT (Internet of Things)

Despite the intensive software work for today's Internet of IoT platforms, a general software engineering approach to systematically developing IoT systems and applications does not exist [65]. So far, researches have been conducted on communication and interoperability issues in this field. Lately, researchers have been working to provide resources and user services with distributed services defined by the software. For example, within the vision of WoT (Web of Things), standard Web technologies are used to provide the development of coordinated IoT services [65]. There are several approaches to interface and programming support for the development of IoT systems and applications. However, work for a common model, methodology and discipline-based approach is still ongoing. There are also work on development environments and development tools in this field.

An article published on this field [66] proposed an IoT middleware that enables IoT systems and the functions of these systems to work together so that the related IoT systems work together. As the technologies used in the IoT become more mature and experience related to the subject grows, it is envisaged that more research on software engineering for IoT systems will be done in the coming years [65].

5. Methods for Product Validation & Test

The product validation is performed to verify if the system works correctly, that is if the product meets the specified requirements and standards, and to evaluate differences between the expected and observed results. The testing process begin by determining the requirements in product design and is performed in accordance with the following objectives:

- To ensure product quality
- To reduce re-development costs
- To increase customer satisfaction

The preparations of the test case and the proper setup are vital in order to follow test procedures, appropriately. System analysis and evaluation of the requirements are performed during this preparation step. Cooperation of different teams is vital to evaluate requirements in terms of testability. Any additional activities and required resources to meet the test objectives need to be planned. Possible risks are determined and backup solutions are planned. Test conditions and scenarios, as well as test exit criteria should be defined. A test plan including well-defined steps to follow should be prepared. After the preparation step, the application is taken.

In general, an important class of tests is referred to as box approach [67]. This approach concerns with whether the code or algorithm is tested by taking into account the internal components or only the inputs and outputs. While black-box and white-box approaches have been widely used, gray box approach has received considerable interest in recent years. The black box approach regards the function as a closed structure and is not interested in its design. It is only considered that the right outputs are produced for given inputs. The white box approach checks whether the internal variables and algorithm are appropriate as well as the function works correctly. The gray box approach is related to the design of the test. The gray box tests are applied like the same black box, but the internal data and algorithm structure of the function are considered when the test is designed.

Another point of view, software tests can be classified as follows [68].

- Unit Test
- Integration Test
- Functional Test
- Compliance Test
- Regression Test
- Performance Test

After the errors found in the verification tests are corrected, the test termination criteria are checked. The test is terminated if all criteria are found to be acceptable. After the test is finished, customer acceptance tests are started for the product. The same process is repeated for the faults, additions or changes requested by the customers. After this phase, the customer tests are repeated if necessary, and the test process is terminated.

Competition is increasing due to developments in the ICT sector. On the other hand, the quality of the product is getting more and more important. The importance given to the testing process has also increased for products to be presented to the user with fewer errors. Today, TMMI (Test Maturity Model Integration) model is defined to improve only the test processes, as CMMI (Capability Maturity Model Integration) model that measures software design processes [69].

In the traditional product design process, the first sample production is done at the end of the design and it is tested whether the requirements are met or not. If desired performance cannot be achieved, design change is made by returning to the design process and new sample production is made. The process is completed following this cycle. Constant sample production to test each design leads to both time loss and cost increase. In today's conditions, the number of initial samples is reduced by computer-aided design. In this way, both the time loss and the cost are reduced. This method is accepted as the modern product design process of today.

6. Mechanical Design & Material Engineering

In the mechanical design of the product, various criteria such as material selection, production methods and mold type and its mechanics should be evaluated in computer environment.

In the 1960s, metals were often used as engineering material. In the following years, demand for steel and cast iron was declined, and today generally, two main materials are used in the production of mechanical systems. These are metal and plastic materials. While metal materials are generally used in machining methods, plastic materials are used in injection molding technology [70,71].

Machining is the shaping process by removing material called sawdust from the inside of materials such as metal, plastic and wood. Along with this, unconventional manufacturing methods such as water jet and laser have started to be used.

Plastic molding is a manufacturing method, which involves pouring and shaping a melted plastic raw material into a mold with the help of temperature and cooling and removing it from the mold. Today, plastic materials can be shaped in a very short time.

One of the latest and greatest developments in manufacturing is the three-dimensional printer technology [72]. Although this method is currently used in rapid prototyping, production with this method in the future will not be surprising. Surface roughness cannot be achieved with the existing technology. For this reason, additional processing is required. Moreover, with this technology, very durable products cannot be produced due to the materials they can process.

In recent years, the production of high performance ceramics, polymers and composite industries has developed rapidly. Today, there are more than a hundred kinds of materials, and choosing materials from such a wide spectrum has become a separate engineering subject. Today, in material engineering and researches, various imaging techniques, thermal analysis methods and X-rays are widely used.

7. Manufacturing

Electronic components used in today's ICT devices are divided into two groups as SMD (Surface Mount Device) and through-hole. The through-hole components frequently used in previous years have been left almost entirely to SMD by the developing technology. The through-hole electronic components must be inserted and soldered in holes in electronic cards. The SMD elements are attached to the card surface and fired.

Surface mounting components can be manufactured much smaller than legged through-hole components [73]. Using small and lightweight components, it is possible to design cards that produce less heat and consume less power. The SMD components can be arranged on both surfaces of the card. Moreover, these components are more resistant to mechanical vibration and shock effects. The weakness of the surface mounting technology compared to the through-hole is the difficulty of manual intervention and repair.

BGA packages offer more pin connections than other SMT (Surface-Mount Technology) package types. For this reason, the BGA package is widely used for components with a high number of pins such as microprocessor and SoC. However, soldering to PCB is much more difficult than other types of packages and requires special stations.

In case of faulty installation, the relevant circuit element is dismounted by local heating method using special devices. At the next stage, the soldering balls are fixed and replaced by local heating under computer control. The BGA rework station is shown below. In this case, a baseband card is reworked for the SoC component.

In BGA packages, the pins are very close to each other and when they are soldered to the PCB, no pins can be seen. A short

circuit may occur between nearby pins, resulting in errors during production, or soldering.

BGA surface mount assembly is difficult to test, once it has been assembled. In this case, X-ray devices can be used to control the deformation and fracture of solder balls.

An X-Ray view of the SoC used in a Baseband card is given in Figure 11. In this view, the pins and the connections between them can be clearly seen. At the end of the review, it is confirmed that the pin connections are correct.

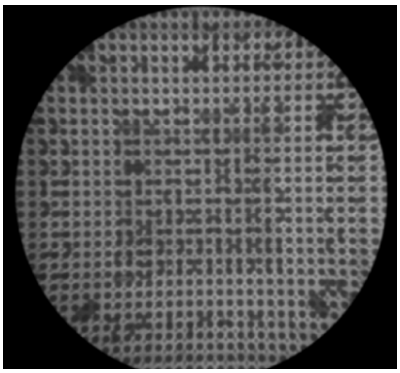


Figure 11 An X-Ray view of the SoC.

7.1. Environment Friendly Approaches for Manufacturing

As the electronic goods sector grows with industrialization and usage habits, the negative effects of electronic waste are increasing rapidly. In order to reduce these effects, environmentally sensitive design has become a mandatory legal requirement. In 2003, the European Union published two important directives for electronic producers. These are Restriction of Hazardous Substances (RoHS) and Waste Electrical and Electronic Equipment (WEEE) directives. These directives are concentrated in two main points. The first is to limit the use of toxic and harmful substances in newly manufactured electronic products and secondly to take into account recycling criteria in product design. In this context, the concept of environmentally friendly design, called eco-design, has emerged. In addition to the traditional design processes, eco-design includes processes from raw material acquisition, product recycling and disposal. For this reason, designers are expected to design by evaluating the recycling and dismount ability criteria of the products [74].

Under the RoHS directive, the use of lead, cadmium, mercury, chromium and brominated components is prohibited. The use of these components in electronic and electrical equipment is causing toxic particles to interfere with the environment. For example, the use of hydrocarbon fluoride to clean the PCB assembly is prohibited because it causes wear on the ozone layer.

8. A Practical Implementation: LTE-Advanced Baseband Unit

In this section, in order to illustrate some of the key features in a cutting edge ICT product design, a LTE-A Baseband Unit (BBU) is considered and the technical competencies, infrastructure and design processes required to develop this product are highlighted.

The first stage for the design of the BBU unit is system-level design, where functional features are derived from the expected characteristics, customer requirements, competitive product analysis and product roadmaps. This step consists of general block design and technical features such as interface characteristics, power supply and cooling. At this stage, it is decided to support some operator specific features in addition to 3GPP Release 10 & Release 11 standards.

The hardware design features of the BBU unit are given in detail in section 3.13. In order to design a high-tech product at this level, advanced technical expertise is required on the high density interconnect PCB production techniques using blind and buried via, micro via, and bypass via. The BBU hardware includes 82 CPU cores and is designed to provide the processing power required by the features specified in the system design phase. This corresponds to a capacity above 1 trillion operations per second. There are 18 CPRI (Common Public Radio Interface) interfaces for radio connections and 2 gigabit Ethernet interfaces for EPC (Evolved Packet Core) connections on this unit. Each radio interface is comply with the CPRI standard.

BBU unit software was developed on a System-on-Chip (SoC) specially developed for Macro Base Station products. Electronics Engineering, Telecommunications Engineering and Mathematical Engineering disciplines have been used together to develop highly optimized and real-time digital signal processing algorithms. BBU software which contains thousands of source files and millions of lines of source code, can be developed with roughly 100 software engineers and many years of design and coding work.

Testing infrastructure which requires tens of millions of dollars in investment cost is required for functional testing of the BBU unit. Using these infrastructures, thousands of functional test scenarios are implemented in the laboratory and in the field. Verification tests of a BBU unit can only be performed using high-tech and complex test equipment, examples of which are given above. For example, in order to use the multi user terminal simulator, it is necessary to examine over 1,000 pages of document and take the training for weeks. Some of the device types included in the test infrastructure required for functional testing of the BBU unit are Multi-user terminal simulators, Channel simulators, EPC (Evolved Packet Core) simulators, Air interface analysers, CPRI (Common Public Radio Interface) test devices, Remote radio units and antennas, Vector signal analyser and generator and Mobile packet analysers.

Finally, ICT products must comply with the relevant hardware, software, product integrity, reliability, environmental and EMC standards. To design LTE-Advanced BBU, tens of thousands of page standard, data sheet and papers have been studied and implemented.

9. Conclusion

Prior to the industrial revolution, goods and equipment to meet daily needs can be produced with simple techniques in small workshops. Later, large factories allowing mass production started to be established and needs for special expertise in various areas of manufacturing techniques have arisen. This trend continued increasingly in mechanical world until 1970s. From the mid-twentieth, first the emergence of transistor and later the invention of the integrated circuit caused a similar development in the electronic world. While in the beginning, mechanical expertise was still predominant in product design, the technical expertise on hardware and software design has become the most crucial parts, in the next two decades.

With the market launch of the first commercial microprocessor, electronic devices started to enter into many areas of our lives. This led to product diversity and consequently to the need for expertise in different areas. Until the 1980s, simple tools for ICT product design and testing were sufficient, but in the next 20 years, more sophisticated design tools and test equipment were required. Nevertheless, the tools and equipment needed to develop an ICT product remained affordable for a start-up company.

After the 1990s, the use of mobile devices in all areas of our lives led to the acceleration of development of technology in the ICT world. Especially, development of display and battery technologies has caused tremendous increase in the use of mobile devices. At the beginning of the new millennium, with the widespread use of internet and video-based services, more bandwidth was needed, which forced new solutions to be developed in cable and wireless communication technologies. However, the emergence of complex modulation techniques to address these demands caused additional hardware requirements, such as faster processors, faster memories, and faster peripheral components.

All of these effects have initiated the key design concepts: smaller, lighter and faster products. Therefore, the dimensions of the components have to be reduced despite the increasing performance and durability requirements. Today lightweight sensors which can be fitted even to swallow birds can be produced, and they can collect and store data for as long as a year. The first commercial microprocessor incorporated 2,300 transistors. Now, new generation processors have around 1.3 billion transistors. Nowadays, the processors produced in the size of one fourth of the 1 cm2 area for wearable devices, deliver 200 times more performance than the processors used in the first PCs. However, with the use of these new technologies in product design, the design process are becoming increasingly complex, engineering skills are becoming increasingly diverse, and the infrastructures needed are becoming increasingly expensive.

In this paper, all the design stages of ICT products are explained. Due to rapid development of hardware technologies, hardware design is getting more and more intriguing. As can be seen in the relevant sections, in order to design an ICT product, very deep technical expertise, expensive development and verification infrastructures are needed in the respective fields.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] C. Hamilton, “A Guide to Printed Circuit Board Design” Butterworth & Co. Ltd, England, 1984.

[2] E. D. Reilly, “Milestones in Computer Science and Information Technology” Greewood Press, USA, 2003.

[3] D. Greenstreet, “Enabling multistandard wireless base stations with TI’s KeyStone SoCs” Texas Instruments, Enabling multistandard wireless base stations with TI’s KeyStone SoCs, 2011.

[4] S. Narnakaje, Z. Lin, “Analysis of LTE Base Station Software Deployment on Multicore SoCs”, Engineers’ Guide to Multicore & Virtualization, s.24, 2012.

[5] A. P. Sage, Systems Engineering, John Wiley & Sons, Inc., USA, 1992.

[6] J. Brown, “The History of Printed Circuit Boards – Infographic”, http://www.pcb-solutions.com/pcb-market-monitor/the-history-of-pcb-infographic/

[7] K. Mitzner, “Complete PCB Design Using OrCad Capture and Layout” Elsevier, UK, 2007.

[8] H. Johnson, M. Graham, High-speed signal propagation : Advanced Black Magic. Upper Saddle River, New Jersey: Prentice Hall PTR. ISBN 0-13-084408-X, 2002.

[9] E. Bogatin, Signal and Power Integrity - Simplified, Second Edition. Upper Saddle River, New Jersey: Prentice Hall. ISBN 978-0-13-234979-6, 2009.

[10] https://en.wikipedia.org/wiki/List_of_device_bit_rates.

[11] Anritsu, “PAM4 Demands Accurate S-parameters”, 2016.

[12] H. Huff, Moore’s Law Beyond Planar Silicon CMOS, Springer Series in Materials Science, USA, 2009.

[13] https://www.extremetech.com/extreme/240901-tsmc-announces-plans-new-16-billion-foundry-push-envelope-5nm-3nm-process-nodes.

[14] https://www.extremetech.com/computing/221532-tsmc-will-begin-10nm-production-this-year-claims-5nm-by-2020.

[15] S. Dwivedi, N. R. Prakash, “A Study on Recent Advancements in VLSI Technology using FinFETs” in International Journal of Innovative Research in Science, Engineering and Technology, Vol. 4, Issue 12, 2015.

[16] S. H. Ok, Y. H. Lee, J. H. Shim, S. K. Lim, B. Moon “The Impact of 3D Stacking and Technology Scaling on the Power and Area of Stereo Matching Processors” in Licensee MDPI, Basel, Switzerland, 2017.

[17] A Brief History of the x86 Microprocessor, http://www.computerworld.com/article/2535019/computer-hardware/timeline--a-brief-history-of-the-x86-microprocessor.html.

[18] Graphics Processing Unit, http://www.nvidia.com/object/gpu.html.

[19] http://www.nvidia.com/object/multi-gpu-technology.html.

[20] R. Giladi, M. Kaufmann, Network Processors: Architecture, Programming, and Implementation, Elsevier, 2008.

[21] http://www.nvidia.com/object/tegra-4-processor.html.

[22] N.Blachford, "Lets Get Physical: Inside The PhysX Physics Processor", 2006.

[23] A. Lucian "Google's Big Chip Unveil For Machine Learning: Tensor Processing Unit With 10x Better Efficiency", Tom's Hardware, 2016.

[24] D. Nenni, P. McLellan, “Fables: The Transformation of the Semiconductor Industry”, 2013.

[25] http://www.cypress.com/event/psoc-6-purpose-built-iots.

[26] https://www.xilinx.com/products/silicon-devices/soc/zynq-ultrascale-mpsoc.html#productTable.

[27] S. Dwivedi, N. R. Prakash, “A Study on Recent Advancements in VLSI Technology using FinFETs” in International Journal of Innovative Research in Science, Engineering and Technology, Vol. 4, Issue 12, 2015.

[28] http://www.doityourself.com/stry/alkaline-battery-advantages-and-disadvantages.

[29] J. Bird, Electrical and Electronic Principles and Technology, Sixth Edition, Routledge, 2010.

[30] J. Lenz, S. Edelstein, “Magnetic Sensors and Their Applications” in IEEE Sensors Journal, vol. 6, no. 3, 10.1109/JSEN.2006.874493, June 2006.

[31] P. Ripka, M. Janosek, “Advances in Magnetic Field Sensors” in IEEE Sensors Journal, vol. 10, no. 6, June 2010.

[32] “Modern Magnetic Field Sensors – a Review”, Przeglad Elektrotechniczny, ISSN 0033-2097, 2013.

[33] M. Andrejasis, “MEMS Accelerometers”, University of Ljubljana Faculty for Mathematics and Physics Department of Physics, March 2008.

[34] O. J. Woodman, “An Introduction to Inertial Navigation” University of Cambridge Computer Laboratory, United Kingdom, August 2007.

[35] R. E. Hopkins, F. K. Mueller, W. Haeussermann “The Pendulous Integrating Gyroscope Accelerometer (PIGA) From the V-2 to Trident D5, the Strategic Instrument of choice”in Navigation, and Control Conference and Exhibit Montreal,Canada, 2001.

[36] R. Mehrotra ,“Thinner, Smaller, Better: Advances in Display Technology” in IEEE MultiMedia,vol. 8, 10.1109/MMUL.2001.939995, 2001.

[37] J. Gurski, L. M. Quach, “Display Technology Overview”, Lytıca White Paper, July 2005.

[38] F. Okano, M. Kanazawa, K. Hamasaki, M. Sugawara, M. Seino, A. Mochimaru, K. Doi, "Ultrahigh-Definition Television System with 4000 scanning lines ", in NAB BEC proceedings, p.437, 2004.

[39] K. Aoshima, K. Machida, D. Kato,H. Kinjou, K. Kuga, T. Mishina, "Advanced Display Systems Technologies",in Industry Applications Society Annual Meeting IEEE , 10.1109/IAS.2013.6682549, 2013.

[40] S. Methley, W. Webb, S. Walker, et al., 5G Candidate Band Study, Quotient Associates Limited, 2015.

[41] W. Roh, J. Seol, J. Park, et al., "Millimeter-Wave Beamforming as an Enabling Technology for 5G Cellular Communications: Theoretical Feasibility and Prototype Results" in IEEE Communications Magazine, 52(2),106-113, 2014.

[42] Z. Liguo, P. Jin, Y. Fang, et al., "A carrier leakage calibration and compensation technique for wideband wireless transceiver" in Journal of Semiconductors, 35(6), 065003, 2014.

[43] M. Racanelli, P. Kempf, "SiGe BiCMOS Technology for RF Circuit Applications" in IEEE Transactions on Electron Devices, 52(7), 1259-1270, 2005.

[44] H. Chuang, L. Yeh, P. Kuo, et al., "A 60-GHz Millimeter-Wave CMOS Integrated On-Chip Antenna and Bandpass filter" in IEEE Transactions on Electron Devices, 58(7), 1837-1845, 2011.

[45] M. Mardiguian, Controlling Radiated Emissions by Design, Springer, Third Edition, USA, 2001.

[46] X. C. Tong, Advanced Materials and Design for Electromagnetic Interference Shielding, CRC Press, USA, 2009.

[47] H. Johnson, "Why Johnny Can't Design a High-Speed Digital System", University of Oxford.

[48] https://yearbook.enerdata.net/electricity-domestic-consumption-data-by-region.html#electricity-domestic-consumption-data-by-region.html.

[49] S. Lambert, M. Pickavet, "Can the Internet Be Greener?", in Proceedings of the IEEE, vol. 105, no. 2, pp. 179-182, February 2017.

[50] Ericsson Mobility Report, https://www.ericsson.com/res/docs/2015/mobility-report/ericsson-mobility-report-nov-2015.pdf, November 2015.

[51] http://www.phonearena.com/news/Smartphone-battery-life-over-the-years-A-surprising-study_id82315.

[52] M.P. Mills, "The Cloud Begins With Coal", August 2013.

[53] G. Auer et al., "How Much Energy is Needed to Run a Wireless Network?" in IEEE Wireless Communications, vol. 18, no. 5, pp. 40-49, October 2011.

[54] J. Wu, Y. Zhang, M. Zukerman, E. K. Yung, "Energy-Efficient Base-Stations Sleep-Mode Techniques in Green Cellular Networks: A Survey", in IEEE Communications Surveys & Tutorials, vol. 17, no. 2, 2015.

[55] F.P. Brooks, "No Silver Bullet Essence and Accidents of Software Engineering", Computer Magazine, vol.20, no. 4, s.10-19, April 1987.

[56] What is Open Source, https://opensource.com/resources/what-open-source.

[57] D. Riehle, "Framework Design A Role Modeling Approach" Ph.D Thesis, Swiss Federal Instituteof Technology Zurich, 2000.

[58] OpenCV, http://opencv.org/.

[59] The open standard for parallel programming of heterogeneous systems, https://www.khronos.org/opencl/.

[60] Overview of the .NET Framework, https://msdn.microsoft.com/en-us/library/zw4w595w(v=vs.110).aspx.

[61] Cocoa Touch, https://developer.apple.com/library/content/documentation/General/Conceptual/DevPedia-CocoaCore/Cocoa.html.

[62] "Introduction to ActiveX Controls", MSDN, Microsoft. Retrieved June 1, 2016.

[63] O. Osanaiye, S. Chen, Z. Yan, R. Lu, K. Choo, M. Dlodlo, "From cloud to fog computing: A review and a conceptual live VM migration framework", in IEEE Access, 10.1109/ACCESS.2017.2692960, 2017.

[64] N. Iotti, M. Picone, S. Cirani Caligoo, G. Ferrari, "Improving Quality of Experience in Future Wireless Access Networks through Fog Computing" in IEEE Internet Computing, University of Parma, 10.1109/MIC.2017.38,Italy, 2017.

[65] F. Zambonelli, "Key Abstractions for IoT-Oriented Software Engineering", in IEEE Software, Volume: 34, Issue: 1, 10.1109/MS.2017.3, 2017.

[66] M.A. Razzaque et al., "Middleware for Internet of Things: A Survey" in IEEE Internet of Things J., vol. 3, no. 1, 2016.

[67] N. Li, Z. Li, X. Sun, "Classification of Software Defect Detected by Black-Box Testing: An Empirical Study" in IEEE Second World Congress on Software Engineering, 10.1109/WCSE.2010.28, December 2010.

[68] E. Suruchi, "Comparative analysis of testing techniques: A Review", in International Journal of Advanced Science and Research, vol. 1, Issue 5, Page No. 46-50, 2016.

[69] E. Venendaal, "Test Maturiy Model Integration (TMMi)", Version 3.1, TMMi Foundation, 2010.

[70] T. A. Osswald, L.S. Turng, P. Gramann, Injection Molding Handbook, Neuausgabe, 2007.

[71] M.P. Groover, Theory of Metal Machining: Fundamentals of Modern Manufacturing, John Wiley & Sons, Inc, 2002.

[72] E. Sachs, M. Cima, J. Cornie, "Three-Dimensional Printing: Rapid Tooling and Prototypes Directly from a CAD Model", Elsevier, Manufacturing Technology, Volume 39, Issue 1, Pages 201-204, 1990.

[73] R. Prasad, Surface Mount Technology: Principles and Practice, Springer, USA, 1997.

[74] P. Knight, O. J. Jenkins, "Adopting and applying eco-design techniques: a practitioners perspective" in Journal of Cleaner Production, s. 549-558, 2008.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1129-1136 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

A Novel Ultra High Speed and Configurable Discrete Wavelet Packet Transform Architecture

Mouhamad Chehaitly[1], Mohamed Tabaa[*2], Fabrice Monteiro[1], Abbas Dandache[1]

[1]*Laboratoire LGIPM, équipe ASE, Université de Lorraine, Île du Saulcy, 57045 METZ, France*

[2]*Laboratoire Pluridisciplinaire de Recherche et Innovation (LPRI), EMSI, Casablanca, Morocco*

ARTICLE INFO

Article history:
Received: 26 April, 2017
Accepted: 15 June, 2017
Online: 16 July, 2017

Keywords:
DWPTMallat-tree algorithm
Embedded Systems
FPGA
Parallel Architecture
VHDL-RTL modeling

ABSTRACT

This work is dedicated to present a new pipeline-parallel architecture of Discrete Wavelet Packet Transform (DWPT) for all wavelet family implemented in FPGA technology. The main target of our architecture is to provide an effective performance trade-off, where it significantly increases the throughput with a restricted amount of hardware. In this article, we propose two kinds of configurable architecture: first architecture with a very strict amount of hardware base of pipeline and sharing resource, and the second architecture provide an ultra-high speed by propose P-parallel DWPT and a parallel direct FIR filter under the strategy of pipeline-parallel and sharing resource. The pipeline and the clever sharing of the hardware resources are smartly connect based on low-pass and high-pass filters in the Mallat-tree algorithm. These architectures are fully configurable in synthesis according to parallel degree, the tree depth (number of tree levels), the order of the filters and the filter quantization coefficient. Consequently, the simulation results accelerated to an approximate value of P(Frequency). Furthermore, the tree depth and filters order has little impact (only due to place and route variations) on throughput. This architecture was synthesized using Altera Quartus prime lite edition targeting an Altera Cyclone IV - (FPGA) and it was developed in VHDL at RTL level modeling.*

1. Introduction

From the work of Coifman and Wickerhuaser in 1992, the wavelet packet transform has a lot of attention from both academic and industry establishments. Where in the literature, we found a lot of research dedicated of wavelet packet transform and there implemented in different domains: communication, image, video, data compression, wireless sensor network and other. But in the last years, we saw a rise in the amount of data that needs to trait and transmit. That requires a new throughput level and need a new platform or program logic device (PLD), which can provide a high speed, with low cost and low power consummation.

The powerful of Discrete Wavelet Packet Transform (DWPT) is due to the ability of represent both transient and stationary behaviors of data signal by a few of transform coefficients. The most implemented solution of wavelet transform is based on the FIR (Finite Impulse Response filter) filter banks concept [1]. To benefit as maximum possible of this characteristic, many works have addressed the implementation of DWPT in hardware device like processor or FPGA.

In this context, the first work try to implement the DWPT on processor was recorded in 1999 by [2]. However, DWPT (based of FIR filter banks) complexity makes it hard to fulfill high throughput and low area (Material resource) consumption. To provide these constraints, we find many works dedicated of the implementation of DWPT on FPGA like [3-16].

The architecture of Daubechies wavelet on scale 6 (DB6) based on algebraic integers is proposed in [3], the authors have developed architectures for one and two dimensions of DB6 with minimization of the number of used adders.

[*]Corresponding Author: Mohamed Tabaa, Laboratoire Pluridisciplinaire de Recherche et Innovation (LPRI), EMSI, Casablanca, Morocco
Email: med.tabaa@gmail.com

An integrated systolic architecture for computation of the 1-D of DWPT/IDWPT are presented in [4] where the architecture can limited the number of page requirements. In [5], DWPT/IDWPT are implemented using Embedded Instruction Codes (EIC) where the authors used two multipliers and four adders for the symmetric filters. [6] presented a word-serial pipeline architecture and parallel filter processing for DWPT/IDWPT. The same authors are followed their work in [7] where they are used high-pass and low-pass filters to speed up transformation. Then, the same authors are also developed the implementation of the wavelet packet transform on FPGA with three types of multipliers in [8]. In 2006, [9] presented a word-serial pipeline architecture with a new parallel FIR filter processor to perform the DWPT/IDWPT transform. And between the years 2013-2014, [10] presented a flexible architecture of DWP/IDWPT based directly on the registers and on an efficient multiplexing structure.

Another architecture of DWPT implement the tree search algorithm is presented in [11], then their implementation on FPGA is discussed in [12]. The same authors in [13] have developed parallelized architectures in two modes: serial-word and parallel- word of the direct and inverse DWPT transform.

The authors in [14] was proposed an architecture for implementing fast and configurable DWPT based on FIFO inputs, dual-port memory, multipliers and adders. [15] also presented a hardware design of DWPT based of frame-partitioned architecture. And in our previous work [16], we propose a fast and reconfigurable architecture but non parallel of DWPT.

This paper proposes a novel P-parallel (Parallel degree P) architecture of DWPT transform and also reconfigurable, based on the Mallat tree scheme [17], with a parallel direct FIR bank filter structure. It is very suitable for the implementation in global parallel DWPT architecture. The architecture aims to offer a suitable compromise between computational performance (throughput rate) and resource consumption (area cost). The core of our architecture is flexible and can provide an elegant sharing of computational resources (multipliers and adders) between the approximation and details related filter banks. The architecture is fully configurable in synthesis according to parallel degree, the tree depth (number of tree levels), the order of the filters and the filter coefficient quantization.

This paper describes in section II a brief theory of the wavelet and wavelet packet transform concepts. In section III, we propose our DWPT architecture compared with the results obtained from HDL coder. Our proposed P-parallel DWPT architecture in section IV. That summarizes the results obtained for this architecture on a FPGA implementation and compared with other previous work in literature. Finally in section V, we will set forth the conclusion part.

2. Review of wavelet packet transform

In the wavelet transform theory, a signal x(t) is decompose by projecting it into a family of functions $\Psi_{s,\tau}(t)$, all derived obtained from the same original single function called "Mother wavelet" by translation and dilation or contraction:

$$\Psi_{s,\tau}(t) = \frac{1}{\sqrt{s}}\Psi\left(\frac{t-\tau}{s}\right) \quad (1)$$

where $s \in R^*, \tau \in R$. For large s, the basis function becomes a stretched version of the prototype wavelet, that is a low frequency function, while for small s, the basis function becomes a contracted wavelet, that is a high frequency function.

Therefore, the discrete wavelets transform (DWT) are discretely scalable and translatable. This was achieved by modifying the wavelet representation in (1) to create Daubechies (1992) [18]:

$$\Psi_{j,k}(t) = \frac{1}{\sqrt{s_0^j}}\Psi\left(\frac{t-ks_0^j\tau_0}{s_0^j}\right) \quad (2)$$

So the Discrete Wavelet Transform (DWT) of signal $x(t)$ present in (3):

$$X(j,k) = 2^{-j/2}\int_{-\infty}^{+\infty} x(t)\Psi \quad (2^{-j}t-k)\,dt \quad (3)$$

DWT can be implemented by the convolution operation between signal x and mask $Mother\ wavelet$ " Ψ ". But also, DWT can be implemented by non-uniform filter banks. Where the input signal separates into low frequency component signal and high frequency component signal. The first is called smooth coefficient while the second part is called wavelet coefficients. The smooth coefficient is separated into two parts repeatedly. In the classical way, we indicate the two parts via low-pass digital filter H and a high pass-filter G. By using the scaling function and there corresponding mother wavelets, we obtain both digital filter H and G. We suppose H and G like a FIR filters non-recursive with L length, the transfer functions of H and G can be expressed as:

$$H(Z) = h_0 + h_1z^{-1} + h_2z^{-2} + \cdots + h_{L-1}z^{-(L-1)} \quad (4)$$

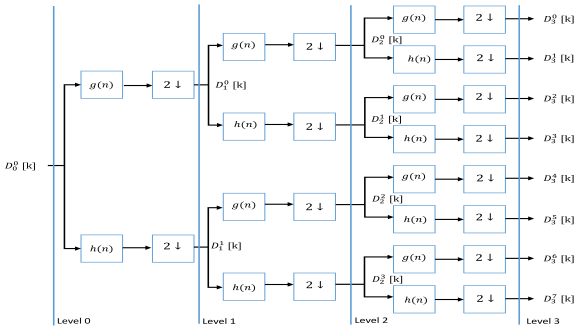
$$G(Z) = g_0 + g_1z^{-1} + g_2z^{-2} + \cdots + g_{L-1}z^{-(L-1)} \quad (5)$$

Mallat's tree algorithm or pyramid algorithm [19] can be used to find the multiresolution decomposition of DWT, the two scale relations (4) and (5) leads to scaling functions and wavelet functions similar to that in scalar wavelets. But the equations are two scale matrix equations and can be given as:

$$\phi(t) = \sum_n h(n)\phi(2t-n) \quad (7)$$

$$\Psi(t) = \sum_n g(n)\Psi(2t-n) \quad (8)$$

Where $\phi(t) = [\phi_1(t)\phi_2(t)\ldots\phi_r(t)]^T$ and $\Psi(t) = [\Psi_1(t)\Psi_2(t)\ldots\Psi_r(t)]^T$ forms the set of scaling functions and corresponding wavelets. The suffix r denotes the number of wavelets and is dubbed as multiplicity.



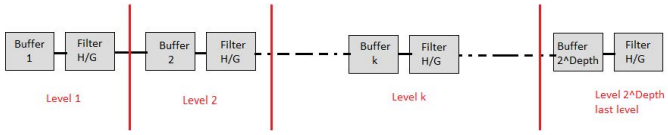
3. Our proposed architecture N°1: smart sharing and pipeline architecture

In the recent work, the most important challenge for developer is design a new DWPT architecture can ensure high speed data processing with low consummation resource. That leads to minimize the needed energy and size hardware (critical constraint in many application domain like WSN, Body area Network, and others).

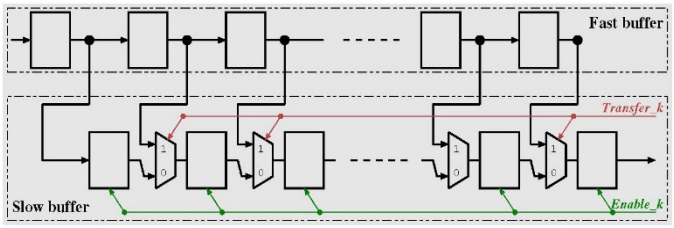
3.1. Fast and reconfigurable architecture unparalleled

To answer of their challenge, we are developed in the first time in [16], our first fast and configurable DWPT architecture based of Mallat tree approach (Figure 1) and transposed digital FIR filters in the filter banks.

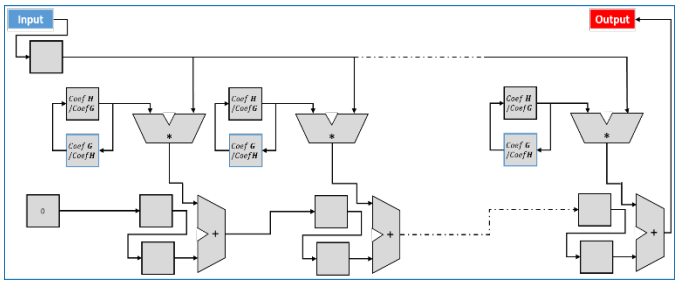
In Figure 2, we present the proposed unparalleled-pipeline architecture. That take advantage of this present in original Mallat tree where instead using a 2^k classic FIR filter in stage k (the amount of data to be processed by any filter of stage k is $2^{-k}\times$ the amount to be processed the filter or stage 1.), only one modified filter is implemented. The power of modified proposed filter can process all data in the same stage k and serve the functionality of low-pass filter and high pass in the same time.



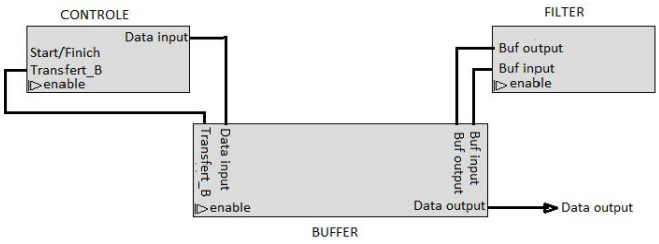
In each level, the simple architecture is based of two blocks: block "Buffer" and block d'un modified FIR filter "Filter H/G". The structure of the block buffer is shown in Figure 3. As her name buffer block, it is built up on two shift register: fast shift register represented by "Fast buffer" and slow shift register represented by "Slow buffer". An important characteristic of these blocs are the dynamic size i.e. the size of the block buffer depends on the stage k (where k parameter in the figure 2 and 3 is related to the number of the stage in which the buffer is implemented, example for k=1, we implement the first block buffer). The working mechanism of block buffer is the key word in our architecture where each buffer is built up as we mention bellow on two shift register of 2^{k-1} positions: a fast shit buffer that takes its data from the previous stage, and a slow buffer that feeds its own stage filter. The "fast buffer" achieve one shift on each clock while the slow buffer shift rate is one shift on each two clock cycles. As we show in figure3, the $enable_k$ signal controls the rate of slow buffer registers. And the $transfer_k$ signal is driven on each 2^k cycles, all the data in the fast buffer in parallel to the slow buffer.



In Figure 4, $block\ H/G$ are the modified blocks filters, they play the function of high pass filter and low pass filter and they are related to $H(Z)$ and $G(Z)$, in equation (4) and (5), respectively.



A control unit (figure 5) was also proposed in our work, to manager the BUFFER block and H/G FITLER block, we used a central CONTROLLE block provides the $enable_k$ and $transfer_k$ signals to the related K stage.



3.2. Results and synthesis

This proposed architecture was writing in VHDL-RTL modeling and we make a big effort to write a generic VHDL code (i.e. independent of target FPGA hardware). The synthesize of our architecture has been done by using Altera Quartus prime lite edition and it is targeted onto Altera FPGA belonging to Cyclone IV family. As we mentioned in [16], the architecture is fully configurable at synthesis in respect to the tree depth (number of tree levels), the order of the filters and the filter coefficient quantization.

Table 1. Implementation results of unparalleled DWPT

Design parameters	Clock frequency	Resource usage
(2, 2, 5)	207.68	(158, 74)
(3, 2, 5)	194.93	(252, 130)
(4, 2, 5)	200.28	(397, 233)
(2, 4, 5)	205.25	(296, 114)
(3, 4, 5)	181.23	(446, 180)
(4, 4, 5)	187.69	(647, 293)
(2, 16, 5)	178.99	(1103, 354)
(3, 16, 5)	179.73	(1598, 480)
(4, 16, 5)	170.36	(2138, 653)
(2, 2, 16)	104.14	(857, 228)
(3, 2,16)	101.31	(1332, 405)
(4, 2,16)	104.94	(1940, 717)
(2, 4,16)	102.54	(1676, 356)
(3, 4,16)	102.54	(1676, 356)
(4, 4,16)	92.06	(3473, 909)
(2, 16,16)	106.73	(1609, 1108)
(3, 16,16)	109.66	(2250, 1493)
(4, 16,16)	111.96	(3032, 2013)

Table 1, we obtained the result of needed area (consumption) and clock frequency for different values of the configuration parameters. In the table, design parameters are presented as a 3-tuple, depth (depth of Mallat tree), order (filter order), and quantization (number of bits of coefficient quantization). Clock frequency is given in MHz. And the resource usage is given as a 2-tuple, (l_e, l_r) where l_e stands for logic elements and l_r for logic registers.

3.3. Results obtained by HDL Workflow Advisor tools

In the last years, many tool are developed to facilate the designer work and to optomize the generate VHDL code of circuit. One of the important tools is HDL Workflow Advisor, this tool is developed by Mathwork society and it built in MATLAB/SIMULINk software. The aim in this section is to evaluate our unparalleled architecture by compare the results in table 1 with that obtained from optimizer HDL tool.

We developed a Simulink model of the some functional architecture of DWPT and then we are generated the optimized HDL code by using the HDL Workflow Advisor. The obtained VHDL code has been synthesized using the same software (Altera Quartus prime lite edition) and also target the same FPGA. The implementation results are present in table 2.

The synthesis in respect the parameter presented below in part B. but we add as result a furthest colon present the needed DSP block.

Table 2. Implementation results of DWPT from Simulink / Quartus.

Design parameters	Clock frequency	Resource usage	DSP block
(2, 2, 5)	286.29	(18,25)	0
(3, 2, 5)	207.25	(29,38)	0
(4, 2, 5)	127.29	(54,70)	0
(2, 4, 5)	134.66	(29,41)	0
(3, 4, 5)	96.72	(48,62)	0
(4, 4, 5)	64.24	(88,110)	0
(2, 16, 5)	31.2	(155,262)	12
(3, 16, 5)	21.66	(425,606)	28
(4, 16, 5)	14.03	(818,1380)	62
(2, 2, 16)	216.5	(44,69)	0
(3, 2,16)	165.4	(74,104)	0
(4, 2,16)	102.01	(131,180)	0
(2, 4,16)	103.77	(142,129)	0
(3, 4,16)	76.58	(258,210)	0
(4, 4,16)	50.36	(606,439)	0
(2, 16,16)	26.37	(839,823)	12
(3, 16,16)	20.65	(1985,1915)	28
(4, 16,16)	12.63	(4449,4361)	62

Base of implantation results presented in table I and II. It can easily make a comparison between our proposed architecture and other one from HDL tool. The synthesis of our architecture is done without any needed of DSP block when the optimized architecture (code) from Matalb/Simulink need from 12 to 62 DSP blocks depended of design parameters. When we increase the order of quantization from 5 to 16 that decrease approximately 50% the clock frequency but that decrease the Clock frequency to 95.5% with the optimized DWPT architecture by coded HDL tool from Matalb/ Simulink.

4. Our proposed architecture N°2: Parallel-Pipeline architecture with sharing resources

In this section, we propose a high throughput parallel architecture for DWPT, base of Mallat tree approach and using the direct P-parallel digital FIR filters in the filter banks. Figure 1 present a three-level decomposition tree for single parallel degree that provides eight frequency bands with P parallel degree. The filter bank consists of wavelet functions as the low pass filter $h(n)$ and its dual high-pass filter $g(n)$.

Considering the diagram in figure 6, the needed resource with classic Mallat tree is multiply by P. For example with degree of parallelism P=10 and depth=3, we need $P * (2^{depth+1} - 1) = 140$, we need 140 low/high pass filters where is a huge number. Furthermore, it can be observed that the data rate on any filter of given stage of the Mallat tree is just half the rate of any filter of the adjacent stage in input side, and twice the rate of any filter of the adjacent stage on the output side (factor 2 under-sampling from stage to stage). Indeed, instead of implementing $P * 2^k/2$ low pass filters and $P * 2^k/2$ low pass filters on bank filter stage$k$, only one and just modified one filter is implemented. As we show

in figure 7. It is important to mention that modified single filter (Block Filter H/G) will have to process all the data that would have been otherwise processed by the original $P * 2^k$ filters of stage$k$.

The structure for the modified filter is shown in figure 8. It follows the basic scheme. The main difference with the original filters is related to the handling of the filter coefficients and a smart shift to serve P sampling in each clock cycle. Respect the z^{-p} buffer order, it is the most sensitive and it is the core of the serial FIR filter to parallel operation. The diagram of P-input of n-tap transposed FIR filter as shown in Figure 4. Similar to coding theory, the Serial direct FIR filter is like a single-input single-output ($SISO$) system and parallel direct FIR filter is like a multiple-input multiple-output ($MIMO$) system. Each clock cycle, there are P parallel inputs signals, hence P outputs signals.

In our proposed, we don't have a limited of degree of parallel but it is necessary to choose degree of parallelism equal 2^x where $x \in \mathbb{N}^+$. The choice of $P = 2^x$is pivotal to organized data between different level.

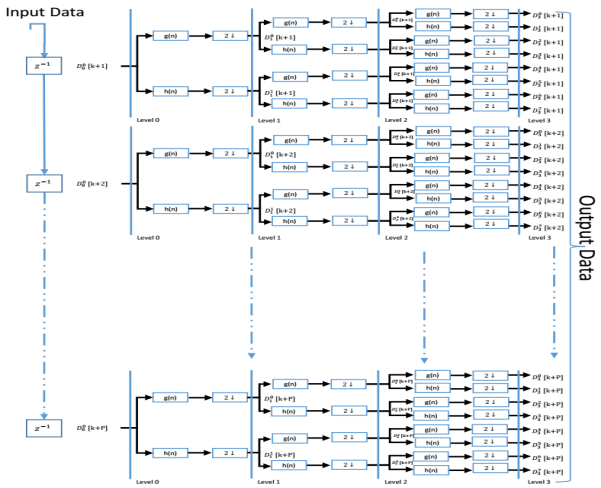


Figure 6 – P-parallel, Three-level DWPT decomposition tree

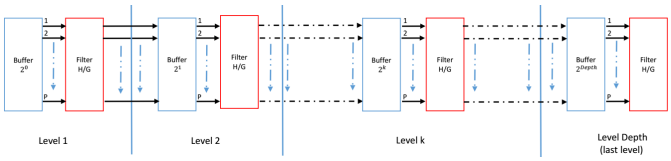


Figure 7. Block view of the proposed parallel DWPT architecture datapath

In Figure 8, $\frac{H}{G}$ $block$is the high pass filter and low pass filter and they are related to $H(Z)$ and $G(Z)$, in equation (4) and (5), respectively. Instead of using two filter with the same architecture, we use an alternatively process. Hence, the alternatively process (on consecutive cycles) force each filter to take a sample for the $H(Z)$ and the for$G(Z)$ and so on. This requires the filter to be feed by correctly scheduled data. This role is devoted to the buffers situated between the filters. The structure of a single buffer is shown in figure 9. The k parameter in the figure is related to the number of the stage in which the buffer is implemented. Then in each level k, the buffer is built up on two shift register of 2^{k-1} positions: a fast shit buffer that takes its data from the previous stage, and a slow buffer that feeds its own stage filter. And in global, the size buffer depends on parallel degree P and level k in which it is implemented. From its name fast buffer, it is fast than "slow buffer". The "fast buffer" achieve P-shift on each clock while the slow buffer shift rate is P-shift on each two clock cycles. To manage this latency, we use enable signal called "$enable_k$" xhich was driven the slower shift rate. For fast buffer, we used a counter signal equal 2^k control the $transfer_k$ signal which every 2^k cycles; all the data in the fast buffer is transferred in parallel to the slow buffer. This transfer operation its done P times on each stage.

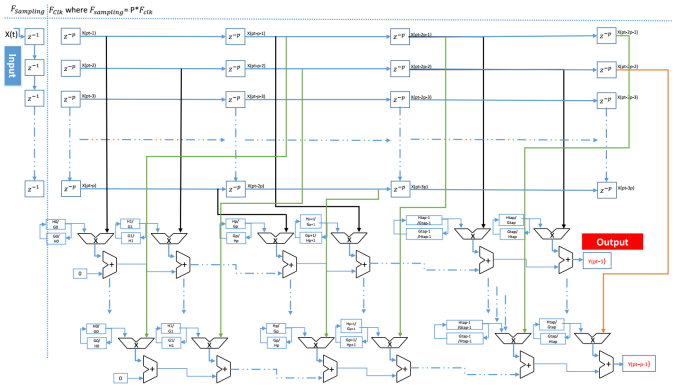


Figure 8. Architecture of a single modified FIR filter

Actually, we used those two signals "$enable_k$" *and* "$transfer_k$" allowing P-times on each stage to combine the down-sampling ($2\downarrow$) and to synchronize (sample selection/ordering) output data on stage (k-1) with input data on stage (k), because we respect the same samples in the original Mallat tree. In this new scheme, the output of all the related stage filters are stored together in the fast buffer (and hence in the slow buffer too) in an interleaved ordered disposition. Only half of the samples can be transferred from the fast to the slow buffer on each 2^k cycles as the buffers size is only half the number of samples being feed to the fast buffer during the same period of time. Furthermore, each sample flowing out from the fast buffer must be presented twice to the next stage in order to be processed by both the $H(z)$ and $G(z)$ sets of coefficients. This is granted by the low rate of the slow buffer. A control unit is used also like that used in figure 5 to manage different stage, and the interleaved on every stage called "$Control\_block$". "$Control\_block$" used to provides the "$enable_k$" and "$transfer_k$" signals to the related K stage.

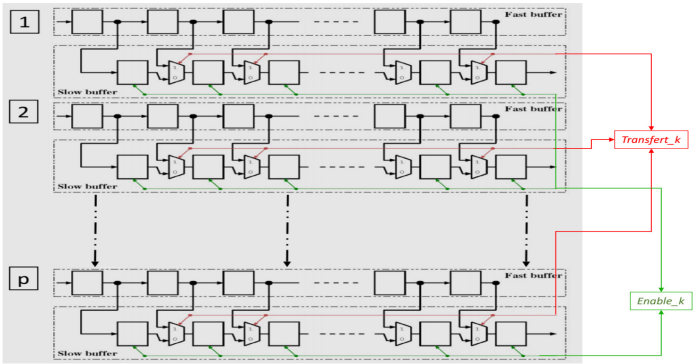


Figure 9. Structure of the buffers in stage k

4.1. Results and performances

The prove of the potential of our architecture isn't stop in the parallel features but also this architecture is fully configurable, so at synthesis the implementation structure diagram have implemented under four parameter: parallel degree (rate of parallelism), tree depth (number of tree levels), the order of the filters and the filter coefficient quantization. We presented the real hardware implementation of our architecture onto Altera 5CGXFC9E6F35C7 FPGA belonging to Cyclone V family with a speed grade of –7 and we have synthesized it by using Altera Quartus prime lite edition. The compilation results of clock frequency and area consumption for different values of the configuration parameters, i. e. Tree depth, Filter order, Coefficient quantization (in number of bits) will presented in table 1 and table 2 of two different parallel degree. In those table, design parameters are presented as a 3-tuple $(depth, order, quantization)$. *Clock frequency* is given in MHz is directly related to input sampling rate (half the clock frequency). Resource usage is given as a 2-tuple (l_e, l_r) where l_e stands for logic elements and l_r for logic registers.

Although the highest rate of our architecture isn't required any memory or DSP block. All results of implementation 4, 8 and 16 parallel DWPT is presented in Table 3, 4 and 5 respectively.

Table 3. Implementation results of 4-parallel DWPT

Design parameters	Clock frequency	Resource usage
(2, 2, 5)	415.36	(471,296)
(3, 2, 5)	400.86	(756,510)
(4, 2, 5)	389.56	(1204,899)
(2, 4, 5)	403.52	(879,456)
(3, 4, 5)	342.46	(1299,719)
(4, 4, 5)	375.24	(1941,1171)
(2, 16, 5)	357.99	(3299,1416)
(3, 16, 5)	369.83	(4794,1924)
(4, 16, 5)	3425.69	(6397,2614)
(2, 2, 16)	208.14	(2571,905)
(3, 2,16)	198.62	(4216,1599)
(4, 2,16)	209.89	(5850,2853)
(2, 4,16)	201.18	(5038,1324)
(3, 4,16)	194.60	(7521,2260)
(4, 4,16)	184.120	(10374,3636)
(2, 16,16)	180.41	(4902,4402)
(3, 16,16)	220.31	(6805,5729)
(4, 16,16)	213.63	(9107,7752)

Table 4. Presents the results of 8-parallel DWPT architecture

Design parameters	Clock frequency	Resource usage
(2, 2, 5)	583.03	(1109,504)
(3, 2, 5)	557.23	(1699,935)
(4, 2, 5)	562.25	(2754,1531)
(2, 4, 5)	586.20	(2120,897)
(3, 4, 5)	506.77	(3050,1197)
(4, 4, 5)	526.91	(4603,2023)
(2, 16, 5)	499.48	(7689,2447)
(3, 16, 5)	504.56	(12176,3166)
(4, 16, 5)	478.26	(14956,4571)
(2, 2, 16)	282.35	(6079,1696)
(3, 2,16)	294.41	(9279,2735)
(4, 2,16)	300.60	(13489,5011)
(2, 4,16)	289.86	(12032,2582)
(3, 4,16)	273.15	(17549,3965)
(4, 4,16)	238.45	(24311,6363)
(2, 16,16)	302.62	(11263,7856)
(3, 16,16)	325.85	(14750,11451)
(4, 16,16)	309.31	(21314,13091)

Table 5. Presents the results of 16-parallel DWPT architecture

Design parameters	Clock frequency	Resource usage
(2, 2, 5)	718.13	(3668, 652)
(3, 2, 5)	710.42	(6019, 960)
(4, 2, 5)	709.04	(8655, 1243)
(2, 4, 5)	536.55	(5991, 1689)
(3, 4, 5)	517.53	(8380, 2363)
(4, 4, 5)	516.33	(11181, 2601)
(2, 16, 5)	464.04	(30012, 4881)
(3, 16, 5)	454.15	(37172, 6575)
(4, 16, 5)	454.92	(38374, 7680)
(2, 2, 16)	455.49	(11116, 3395)
(3, 2,16)	447.30	(17679, 4330)
(4, 2,16)	446.25	(25389, 4830)
(2, 4,16)	336.75	(31336, 5137)
(3, 4,16)	331.99	(39361, 7764)
(4, 4,16)	341.59	(42687, 10342)
(2, 16,16)	302.46	(26408, 15572)
(3, 16,16)	303.44	(33859, 20959)
(4, 16,16)	298.79	(36348, 24456)

	Tze-Yun et al.[21]	Marino et al. [22]	Mohanty et al. [23]	Madishetty et al. [3]	Wang et al. [20]	Wu et al. [5]	Meihua et al. [15]	Proposed architecture N°1	Proposed architecture N°2
Wavelet	Daub-4	Quadri-filter	Daub-4	Daub-6	Lifting-based D4 wavelet	Quadri-filter	Quadri-filter	arbitrary	arbitrary
DWPT/IDWPT	Both	DWPT	DWPT	DWPT	Both	Both	DWPT	DWPT	DWPT
Logic cell	N/A	N/A	426	1040	N/A	30192 (logic gate)	1835 (logic unit)	158 (logic element)	3668 (logic element)
Technology	Xilinx XC2V4000	N/A	CMOS 90nm	Xilinx Virtex 6	CMOS 180 nm	0.35 µm	Alter EP20K200E	Alter Cyclone IV	Alter Cyclone V
Max. Freq. (MHz)	N/A	N/A	20	306.15	20	100	29	207.68	718.13
Memory	yes	N/A	yes	yes	yes	N/A	yes	No	NO
Quantization(bits)	N/A	N/A	N/A	N/A		32	N/A	5 and 16 (up to unlimited theatrically)	5 and 16 (up to unlimited theatrically)
DSP	N/A	N/A	N/A			N/A	N/A	No	No
Depth	3	N/A	N/A	4	3	3 (up to 6)	3	2,3 and 4 (up to unlimited theatrically)	2,3 and 4 (up to unlimited theatrically)

Table 6. Comparison of proposed architectures with existing DWPT architectures

4.2. Comparison: our architecture VS other work

In section III, we pare presented the potential of our first architecture of DWPT with the optimized HDL tools. In fact, that gave us strength to go far to propose a novel very fast, configurable, and P-parallel-pipeline a generic architecture of DWPT transform for all discrete wavelet family. To improve, the performance of these architectures, we must compare the obtained results with other recent works.

In table 6, we compared our proposed architectures with several DWPT transform designs achieved in literature. Well, a quantitative and comprehensive comparative of the most published of DWPT architectures are provides. This table present clearly the high frequency of our architecture with the other one and the full flexibility in all design parameters. Additionally, our architecture was implemented without used memory and present unlimited quantization or depth order, and they are suitable for all wavelet family (where the wavelet family are simplify by the filters orders).

5. Conclusion

We proposed a novel Discrete Wavelet Packet Transform (DWPT) architecture based on Mallat tree concept, by using a FIR filter banc. We have implemented unparalleled and P-parallel, very fast, and reconfigurable DWPT in VHDL-RTL model. We proposed also a P-Parallel architecture of modified direct FIR filter.

The final results indicate that it can efficiently increase the equivalent rate with reduced hardware resource consumption. And implement high-speed filtering, which cannot be achieved by the traditional Mallat tree or serial transformation. The architectures are fully configurable at synthesis in respect to P-parallel degree, depth (number of tree stages), filter order and filter coefficient quantization (generic parameters in the VHDL-RTL model). Filter coefficients are loaded dynamically during operation (after synthesis), providing high operational flexibility. The parallel model and also parallel direct FIR filter can be widely applied to various occasions where high-speed digital transformations are required. An FPGA Cyclone V the clock frequency is operational at 125MHz, after synthetized compiling we obtained 718.13MHz clock frequency of 2 level DWPT transform (order filter 2 and order of quantization 5) and intermediate level of parallelizing i.e. 16.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] G. Strang and T. Nguyen. « Wavelets and Filter Banks ». Wellesley-Cambridge Press, 1997.

[2] W. Xiaodong, L. Yongming and C. Hongyi. « Programmable wavelet packet transform processor ». In Electronics Letters, vol. 35, no. 6, pp. 449-450, 18 Mar 1999.

[3] S. K. Madishetty, A. Madanayake, R. J. Cintra and V. S. Dimitrov. « Precise VLSI Architecture for AI Based 1-D/ 2-D Daub-6 Wavelet Filter Banks with Low Adder-Count ». IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS-I: REGULAR PAPERS, VOL. 61, NO. 7, 1984 - 1993, JULY 2014.

[4] T. Acharya. « A Systolic Architecture for Discrete Wavelet Transforms ». IEEE, Digital Signal Processing Proceedings, 13th International Conference on Volume 2, 571 - 574 vol.2 , 1997.
[5] B-F. Wu and Y-Q. Hu. « An efficient VLSI implementation of the discrete wavelet transform using embedded instruction codes for symmetric filters ». IEEE Transactions on Circuits and Systems for Video Technology, vol. 13, no. 9, pp. 936–943, Sep. 2003.
[6] M. Farahani and M. Eshghi. « Architecture of a Wavelet Packet Transform Using Parallel Filters ». TENCON 2006 - IEEE Region 10 Conference, 1-4244-0548-3, 1 – 4, 2006.
[7] M. Farahani and M. Eshghi. « Implementing a new architecture of wavelet packet transform on FPGA ». Proceedings of the 8th WSEAS International Conference on Acoustics & Music: Theory & Applications, Vancouver, Canada. 2007.
[8] M.A. Farahani, S. Mirzaei and H.A. Farahani. « Implementation of a reconfigurable architecture of discrete wavelet packet transform with three types of multipliers on FPGA ». 24th IEEE Canadian Conference on.Electrical and Computer Engineering (CCECE), 2011.
[9] M. Farahani and M. Eshghi. « Architecture of a Wavelet Packet Transform Using Parallel Filters ». TENCON 2006 - IEEE Region 10 Conference, 1-4244-0548-3, 1-4, 2006.
[10] D. Garcia, M. Mansour and M. Ali. « A Flexible Hardware Architecture for Wavelet Packet Transform With Arbitrary Tree Structure ». IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—II: EXPRESS BRIEFS, VOL. 60, NO. 10, OCTOBER 2013.
[11] M. Trenas, J. López, M. Sanchez, F. Argüello and E. L. Zapata. « Architecture for Wavelet Pack et Transform with Best Tree Searching ». In Application-Specific Systems, Architectures, and Processors, 2000. Proceedings. IEEE International Conference on (pp. 289-298). IEEE, 2000.
[12] M. Trenas, J. López and E. Zapata. « FPGA Implementation of Wavelet Packet transform with Reconfigurable Tree Structure ». Euro micro Conference, 2000. Proceedings of the 26th Volume 1, 5-7, pp. 244 - 251 vol.1, Sept. 2000.
[13] M. Trenas, J. López, E. Zapata and F. Argüello. « A Configurable Architecture for the Wavelet Packet Transform ». Journal of VLSI signal processing systems for signal, image and video technology, Volume 32, Issue 3, pp 255-273, November 2002.
[14] A. Jamin and P. Mähönen. « FPGA implementation of the wavelet packet transform for high speed communications ». International Conference on Field Programmable Logic and Applications. Springer Berlin Heidelberg, 2002.
[15] X. Mei-hua, C. Zhang-jin, R. Feng, and C. Yu-lan. « Architecture research and VLSI implementation for discrete wavelet packet transform ». High Density Microsystem Design and Packaging and Component Failure Analysis, 2006. HDP'06. Conference on. IEEE, 2006.
[16] M. Chehaitly, M. Tabaa, F. Monteiro, and A. Dandache. « A Fast and Configurable Architecture for Discrete Wavelet Packet Transform ». Design of Circuits and Integrated Systems (DCIS), 2015 Conference on, P. 1 – 6. November 2015.
[17] S. Mallat. « A wavelet tour of signal processing". Academic Press, 1999.
[18] I. Daubechies et al. « Ten lectures on wavelets ». Philadelphia, SIAM, vol. 61, pp. 198–202, 1992.
[19] S. Mallat. "Multiresolution approximations and wavelet orthonormal bases of L2(R) ». Transactions of the American Mathematical Society, vol. 315, no. 1, pp. 69–87, 1989.
[20] C. Wang, J. Zhou,L. Liao, J. Lan, J. Luo, X. Liu and M. Je. « Near-threshold energy-and area-efficient reconfigurable DWPT/DWT processor for healthcare - monitoring Applications ». IEEE Transactions on Circuits and Systems II: Express Briefs, 62(1), 70-74, 2015.
[21] T.-Y. Sung, H.-C. Hsin, Y.-S. Shieh, and C.-W. Yu, « Low-power multiplierless 2-D DWT and IDWT architectures using 4-tap Daubechies filters ». In Proc. Seventh Int. Conf. PDCAT, pp. 185–190, 2006.
[22] F. Marino. « Two fast architectures for the direct 2-D discrete wavelet transform ». IEEE Trans. Signal Process, vol. 49, no. 6, pp. 1248–1259, 2001.
[23] B. K. Mohanty and P. K. Meher. « Memory-efficient high-speed convolution-based generic structure for multilevel 2-D DWT ». IEEE Trans. Circuits Syst. Video Technol., vol. 23, pp. 353–363, 2013.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1137-1146 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Cognitive Artificial Intelligence Method for Interpreting Transformer Condition Based on Maintenance Data

Karel Octavianus Bachri, Bambang Anggoro , Arwin Datumaya Wahyudi Sumari, Adang Suwandi Ahmad*

Cognitive Artificial-Intelligence Research Group, School of Electrical Engineering and Informatics, Bandung Institute of Technology, Bandung, 40132, Indonesia

ARTICLE INFO

Article history:
Received: 06 April, 2017
Accepted: 24 June, 2017
Online: 16 July, 2017

Keywords:
Cognitive Artificial-Intelligence
A3S
OMA3S
Information Fusion
Transformer Condition

ABSTRACT

A3S(Arwin-Adang-Aciek-Sembiring) is a method of information fusion at a single observation and OMA3S(Observation Multi-time A3S) is a method of information fusion for time-series data. This paper proposes OMA3S-based Cognitive Artificial-Intelligence method for interpreting Transformer Condition, which is calculated based on maintenance data from Indonesia National Electric Company (PLN). First, the proposed method is tested using the previously published data, and then followed by implementation on maintenance data. Maintenance data are fused to obtain part condition, and part conditions are fused to obtain transformer condition. Result shows proposed method is valid for DGA fault identification with the average accuracy of 91.1%. The proposed method not only can interpret the major fault, it can also identify the minor fault occurring along with the major fault, allowing early warning feature. Result also shows part conditions can be interpreted using information fusion on maintenance data, and the transformer condition can be interpreted using information fusion on part conditions. The future works on this research is to gather more data, to elaborate more factors to be fused, and to design a cognitive processor that can be used to implement this concept of intelligent instrumentation.

1. Introduction

In the earlier paper [1], transformer condition is used to estimate its end of life using instant data. In this paper, transformer condition is calculated using maintenance data from Indonesia Electric Company (PLN) and is interpreted using Cognitive-Artificial method. Each factor influencing the same component is assumed to have the same weight in the condition calculation. The condition is then interpreted to provide early warning system for the potential failure and to estimate the transformer end of life.

There are no single conventional method of transformer diagnosis can be used to define transformer condition accurately. Usually there are several methods combined to perform such a task. These methods are very expert-dependant and are not formulated. Therefore, an automated method for transformer condition monitoring is proposed. Using Cognitive-Artificial-Intelligence (CAI) method, the transformer condition interpretation can be accurately performed, and the expert-dependency can be reduced as well.

2. Transformer Condition Component

Transformer condition can be calculated using some factors. These factors are shown in Figure 1 [2].

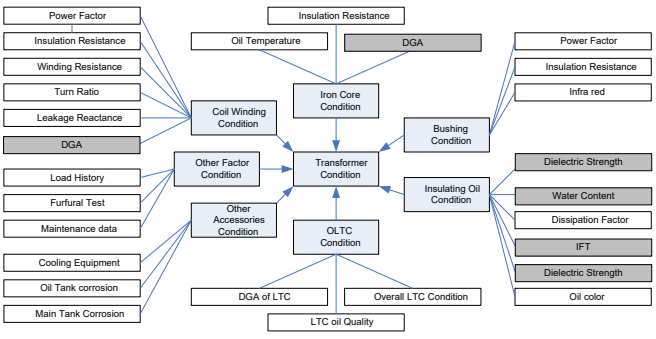


Figure 1 Transformer Condition Factors [2]

*Corresponding Author: Adang Suwandi Ahmad, Cognitive Artificial-Intelligence Research Group, Bandung Institute of Technology, Bandung, Indonesia
Email: adangsahmad@yahoo.com.

Data are collected from PLN and fused to obtain part conditions, part conditions are then fused to obtain transformer condition. Figure 2 explains how transformers degrade over time.

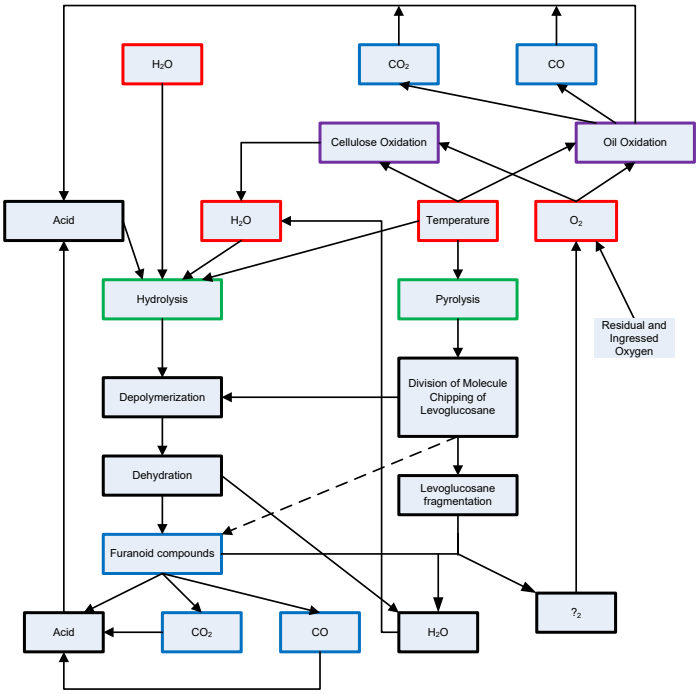


Figure 2 Transformer Degradation Diagram [3]

There are two main processes of the transformer degradation process, hydrolysis and pyrolysis. Hydrolysis is related to water, while pyrolysis is related to fire. The main cause of hydrolysis is water, acids and temperature causes hydrolysis as well. Pyrolysis is caused by temperature [3].

Hydrolysis causes depolymerization of transformer insulating system, and later produces furanoid compounds, which produces carbon dioxide and carbon monoxide, which is the main cause of acids [3]. Acid will then cause hydrolysis.

Pyrolysis causes levoglucosane fragmentation, which produces diatomic oxygen. Oxygen is the cause of oxidation in cellulose and oil, which leads to hydrolysis [3]. These two degradation processes and the compounds they produce makes the transformer degradation processes accelerate over time. The impact of diatomic oxygen will be discussed in Load Tap Changer.

3. The Mathematical Model of Arwin-Adang-Aciek-Sembiring (A3S) [4]

How knowledge grows in the system can be described using Figure 2 [4, 5]. There are two main parts of Knowledge-Growing System. The upper part of Figure 3 contains Information Fusion, while the lower part of the diagram contains knowledge fusion.

The system receives multi-source information from sensors and performs information fusion. When the information exceeds certain level of desirable Degree of Certainty, the information will be considered as knowledge.

The knowledge will be fused with the existing knowledge in the knowledge part of the system to obtain new knowledge and is stored. When the new knowledge exceeds certain level of DoC, it will become the ultimate knowledge.

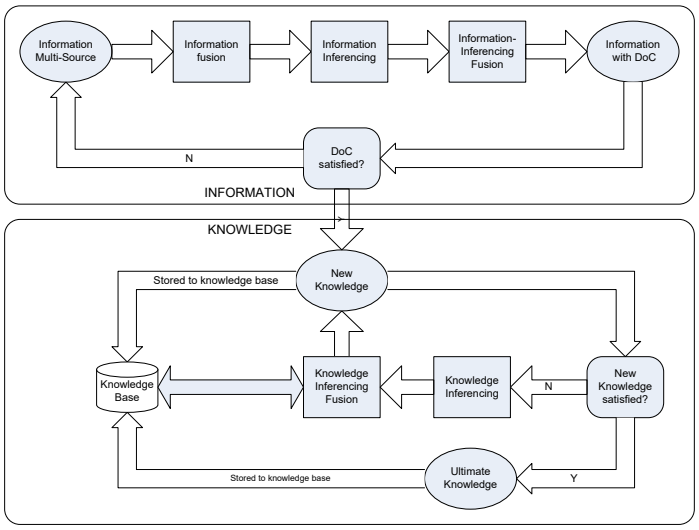


Figure 3 Knowledge-Growing System [5]

A3S (Arwin-Adang-Aciek-Sembiring) algorithm [5] is information fusion algorithm based on Bayesian Inference Method. When a problem occurs, the system collects information and fuses them to produce new knowledge. A3S starts with (1).

$$P\left(B_j \middle| A_i\right) = \sum_{i=1}^{n} \left(\frac{P\left(A_i \middle| B_j\right) P\left(B_j\right)}{\sum_{k=1}^{m} P\left(A_i \middle| B_k\right) P\left(B_k\right)} \right) \tag{1}$$

Where $P\left(B_j \middle| A_i\right)$ is the probability of B_j is true given the presence of the fusion or combination of all events A_i [4]. Maximum A Posteriori (MAP) is determined by (2)

$$P\left(B_j \underset{\max}{\boxplus} A_i\right)_{estimate} = \max_j \left(\frac{P\left(B_j \middle| A_i\right)}{n} \right) \tag{2}$$

It is then simplified to become (3)

$$P\left(\psi_1^j\right) = \frac{\sum_{i=1}^{\delta} P\left(\upsilon_i^j\right)}{\delta} \tag{3}$$

Where $P\left(\psi_i^j\right)$ will be the New Knowledge Probability Distribution (NKPD) at a certain observation time γ_1 [4]. The new knowledge will be obtained by applying (4).

$$P\left(\psi_1^j\right)_{estimate} = \max\left(\left[P\left(\psi_1^j\right)\right]\right) \tag{4}$$

The system will keep collecting information (NKPD) on each observation, $P(\psi_1^j)$, …, $P(\psi_\gamma^j)$, …, $P(\psi_\Gamma^j)$ [4]. The inferencing can be determined using (5).

$$P(\phi_\gamma^j)=\begin{cases}1, & P(\psi_\gamma^j)>\dfrac{P(\psi_\gamma^j)}{\lambda}\\ 0, & P(\psi_\gamma^j)\le\dfrac{P(\psi_\gamma^j)}{\lambda}\end{cases} \quad (5)$$

Where $P(\phi_\gamma^j)$ is inferencing of each information to the knowledge distribution.

Information-inferencing fusion will be calculated using OMA3S method, a dynamic version of A3S resulting NKPD over Time (NKPDT) [4].

$$P(\theta_j)=\frac{\sum_{\chi=1}^{\Gamma}P(\phi_\gamma^j)}{\Gamma} \quad (6)$$

$$P(\theta)_{estimate}=\max\left[P(\theta_j)\right] \quad (7)$$

4. Transformer condition calculation

System's block diagram is shown in Figure 4.

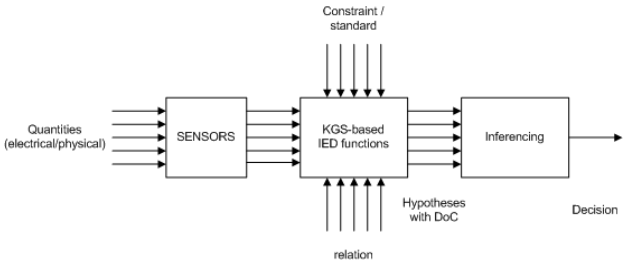


Figure 4 System's block diagram

Data is collected using sensors and is compared to standards and relation. The system fills the observation table using (1). After each observation, there will be new knowledge shown by New Knowledge Probability Distribution (NKPD). Each NKPD is fused with the previous NKPD to produce NKPD over time (NKPDT). Decisions are made based on NKPDT.

There are two kinds of faults in transformers. They are electrical faults and thermal faults [9]. Electrical faults are Partial Discharge, Low-Energy Discharge, and High Energy Discharge, while thermal faults are Thermal-Low and Thermal-High [9]. The proposed method is tested using previously published DGA dataset, which is classified based on the identified fault [9]. Table 1 shows Partial Discharge dataset, Table 2 shows Low-Energy Discharge, Table 3 shows High-Energy Discharge, Table 4 shows Thermal-Low, and Table 5 shows High-Energy Discharge.

Table 1 Partial Discharge dataset [9].

No.	H_2	CH_4	C_2H_6	C_2H_4	C_2H_2	CO
1	32930	2397	157	0	0	313
2	37800	1740	249	8	8	56
3	92600	10200	0	0	0	6400
4	8266	1061	22	0	0	107
5	9340	995	60	6	7	60
6	36036	4704	554	5	10	6
7	33046	619	58	2	0	51
8	40280	1069	1060	1	1	1
9	26788	18342	2111	27	0	704

Table 2 Low-Energy Discharge dataset [9].

No.	H_2	CH_4	C_2H_6	C_2H_4	C_2H_2	CO
1	78	20	11	13	28	0
2	305	100	33	161	541	440
3	35	6	3	26	482	200
4	543	120	41	411	1880	76
5	1230	163	27	233	692	130
6	645	86	13	110	317	74
7	60	10	4	4	4	780
8	95	10	0	11	39	122
9	6870	1028	79	900	5500	29

Table 3 High-Energy Discharge dataset [9].

No.	H_2	CH_4	C_2H_6	C_2H_4	C_2H_2	CO
1	440	89	19	304	757	299
2	210	43	12	102	187	167
3	2850	1115	138	1987	3675	2330
4	7020	1850	0	2960	4410	2140
5	545	130	16	153	239	660
6	7150	1440	97	1210	1760	608
7	620	325	38	181	244	1480
8	120	31	0	66	94	48
9	755	229	32	404	460	845

Table 4 Thermal-Low dataset [9].

No.	H_2	CH_4	C_2H_6	C_2H_4	C_2H_2	CO
1	1270	3450	520	1390	8	483
2	3420	7870	1500	6990	33	573
3	360	610	259	260	9	12000
4	1	27	49	4	1	53
5	3675	6392	2500	7691	5	101
6	48	610	29	10	0	1900
7	12	18	4	4	0	559
8	66	60	2	7	0	76
9	1450	940	211	322	61	2420

Table 5 Thermal-High dataset [9].

No.	H_2	CH_4	C_2H_6	C_2H_4	C_2H_2	CO
1	8800	64064	72128	95650	0	290
2	6709	10500	1400	17700	750	290
3	1100	1600	221	2010	26	0
4	290	966	299	1810	57	72
5	2500	10500	4790	13500	6	530
6	1860	4980	0	10700	1600	158
7	860	1670	30	2050	40	10
8	150	22	9	60	11	0
9	400	940	210	820	24	390

In order to analyze the DGA, there are several ratios required, they are [5]:

$$R_1 = [CH_4]/[H_2] \quad (6)$$

$$R_2 = [C_2H_2]/[C_2H_4] \quad (7)$$

$$R_5 = [C_2H_4]/[C_2H_6] \quad (8)$$

These ratios are put into groups based on Table 6.

Table 6 Gas Ratio grouping [9].

	R2	R1	R5
< 0.1	0	1	0
0.1 - 1.0	1	0	0
1.0 - 3.0	1	2	1
> 3	2	2	2

Estimated faults can be determined using the rules shown in Table 7 [9].

Table 7 Gas Ratio grouping [9].

No.	Characteristic Fault	R2	R1	R5
0	No fault	0	0	0
1	Partial Discharge	0 or 1	1	0
2	Low-Energy Discharge	1 or 2	0	1 or 2
3	High-Energy Discharge	1	0	2
4	Thermal-Low	0	0 or 2	0 or 1
5	Thermal-High	0	2	2

Datasets are made into Ratios and are put into groups as shown in Table 8 to Table 12.

Table 8 Ratios: Partial Discharge.

Ratio			Ratio Group		
R1	R2	R5	R1	R2	R5
2.72	0.01	2.67	1	2	0
2.30	0.00	4.66	1	1	0
1.69	0.03	1.00	0	2	2
27.00	0.25	0.08	0	2	0
1.74	0.00	3.08	0	1	0
12.71	0.00	0.34	0	1	0
1.50	0.00	1.00	1	0	0
0.91	0.00	3.50	1	1	0
0.65	0.19	1.53	0	0	0

Table 9 Ratios: Low-Energy Discharge.

Ratio			Ratio Group		
R1	R2	R5	R1	R2	R5
0.26	2.15	1.18	0	1	1
0.33	3.36	4.88	0	2	2
0.17	18.54	8.67	0	2	2
0.22	4.57	10.02	0	2	2
0.13	2.97	8.63	0	1	2
0.13	2.88	8.46	0	1	2
0.17	1.00	1.00	0	1	1
0.11	3.55	inf	0	2	2
0.15	6.11	11.39	0	2	2

Table 10 Ratios: High-Energy Discharge.

Ratio			Ratio Group		
R1	R2	R5	R1	R2	R5
0.20	2.49	16.00	0	1	2
0.20	1.83	8.50	0	1	2
0.39	1.85	14.40	0	1	2
0.26	1.49	inf	0	1	2
0.24	1.56	9.56	0	1	2
0.20	1.45	12.47	0	1	2
0.52	1.35	4.76	0	1	2
0.26	1.42	inf	0	1	2
0.30	1.14	12.63	0	1	2

Table 11 Ratios: Thermal-Low.

Ratio			Ratio Group		
R1	R2	R5	R1	R2	R5
2.72	0.01	2.67	2	0	1
2.30	0.00	4.66	2	0	2
1.69	0.03	1.00	2	0	1
27.00	0.25	0.08	2	1	0
1.74	0.00	3.08	2	0	2
12.71	0.00	0.34	2	0	0
1.50	0.00	1.00	2	0	1
0.91	0.00	3.50	0	0	2
0.65	0.19	1.53	0	1	1

Table 12 Ratios: Thermal-High.

Ratio			Ratio Group		
R1	R2	R5	R1	R2	R5
7.28	0.00	1.33	2	0	1
1.57	0.04	12.64	2	0	2
1.45	0.01	9.10	2	0	2
3.33	0.03	6.05	2	0	2
4.20	0.00	2.82	2	0	1
2.68	0.15	inf	2	1	2
1.94	0.02	68.33	2	0	2
0.15	0.18	6.67	0	1	2
2.35	0.03	3.90	2	0	2

Ratio Groups are arranged into observation table as shown in Table 13 to Table 17 where:

- H_PD: Hypothesis Partial Discharge.
- H_LE: Hypothesis Low-Energy Discharge.
- H_HE: Hypothesis High-Energy Discharge.
- H_TL: Hypothesis Thermal-Low.
- H_TH: Hypothesis Thermal-High.

Table 13 Observation: Partial Discharge.

Nth Obs.	sensors	Range Group	hypotheses				
			H_PD	H_LE	H_HE	H_TL	H_TH
1	R1	1	1	1	1	0	0
	R2	2	0	0	0	0	1
	R5	0	1	0	0	0	0
2	R1	1	1	1	1	0	0
	R2	1	1	0	0	0	0
	R5	0	1	0	0	0	0
3	R1	0	0	0	0	1	1
	R2	2	0	0	0	0	1
	R5	2	0	0	1	0	1
4	R1	0	0	0	0	1	1
	R2	2	0	0	0	0	1
	R5	0	1	0	0	0	0
5	R1	0	0	0	0	1	1
	R2	1	1	0	0	0	0
	R5	0	1	0	0	0	0
6	R1	0	0	0	0	1	1
	R2	1	1	0	0	0	0
	R5	0	1	0	0	0	0
7	R1	1	1	1	1	0	0
	R2	0	0	1	1	1	0
	R5	0	1	0	0	0	0
8	R1	1	1	1	1	0	0
	R2	1	1	0	0	0	0
	R5	0	1	0	0	0	0
9	R1	0	0	0	0	1	1
	R2	0	0	1	1	1	0
	R5	0	1	0	0	0	0

Table 14 Observation: Low-Energy Discharge.

Nth Obs.	sensors	Range Group	hypotheses				
			H_PD	H_LE	H_HE	H_TL	H_TH
1	R1	0	0	1	1	1	1
	R2	1	1	1	1	0	0
	R5	1	0	1	0	1	0
2	R1	0	0	1	1	1	1
	R2	2	0	1	0	0	1
	R5	2	0	1	1	0	1
3	R1	0	0	1	1	1	1
	R2	2	0	1	0	0	1
	R5	2	0	1	1	0	1
4	R1	0	0	1	1	1	1
	R2	2	0	1	0	0	1
	R5	2	0	1	1	0	1
5	R1	0	0	1	1	1	1
	R2	1	1	1	1	0	0
	R5	2	0	1	1	0	1
6	R1	0	0	1	1	1	1
	R2	1	1	1	1	0	0
	R5	2	0	1	1	0	1
7	R1	0	0	1	1	1	1
	R2	1	1	1	1	0	0
	R5	1	0	1	0	1	0
8	R1	0	0	1	1	1	1
	R2	2	0	1	0	0	1
	R5	2	0	1	1	0	1
9	R1	0	0	1	1	1	1
	R2	2	0	1	0	0	1
	R5	2	0	1	1	0	1

Table 15 Observation: High-Energy Discharge.

N-th Obs.	sensors	Range Group	hypotheses				
			H_PD	H_LE	H_HE	H_TL	H_TH
1	R1	0	0	1	1	1	1
	R2	1	1	1	1	0	0
	R5	2	0	1	1	0	1
2	R1	0	0	1	1	1	1
	R2	1	1	1	1	0	0
	R5	2	0	1	1	0	1
3	R1	0	0	1	1	1	1
	R2	1	1	1	1	0	0
	R5	2	0	1	1	0	1
4	R1	0	0	1	1	1	1
	R2	1	1	1	1	0	0
	R5	2	0	1	1	0	1
5	R1	0	0	1	1	1	1
	R2	1	1	1	1	0	0
	R5	2	0	1	1	0	1
6	R1	0	0	1	1	1	1
	R2	1	1	1	1	0	0
	R5	2	0	1	1	0	1
7	R1	0	0	1	1	1	1
	R2	1	1	1	1	0	0
	R5	2	0	1	1	0	1
8	R1	0	0	1	1	1	1
	R2	1	1	1	1	0	0
	s	2	0	1	1	0	1
9	R1	0	0	1	1	1	1
	R2	1	1	1	1	0	0
	R5	2	0	1	1	0	1

Table 16 Observation: Thermal-Low.

Nth Obs.	sensors	Range Group	hypotheses				
			H_PD	H_LE	H_HE	H_TL	H_TH
1	R1	2	0	0	0	1	1
	R2	0	0	0	0	1	0
	R5	1	1	1	0	1	0
2	R1	2	0	0	0	1	0
	R2	0	0	0	0	1	0
	R5	2	0	1	1	0	1
3	R1	2	0	0	0	1	0
	R2	0	0	0	0	1	0
	R5	1	1	1	0	1	0
4	R1	2	0	0	0	1	0
	R2	1	1	1	1	0	0
	R5	0	0	0	0	1	0
5	R1	2	0	0	0	1	0
	R2	0	0	0	0	1	0
	R5	2	0	1	1	0	1
6	R1	2	0	0	0	1	0
	R2	0	0	0	0	1	0
	R5	0	0	0	0	1	0
7	R1	2	0	0	0	1	0
	R2	0	0	0	0	1	0
	R5	1	1	1	0	1	0
8	R1	0	0	1	1	1	1
	R2	0	0	0	0	1	0
	R5	2	0	1	1	0	1
9	R1	0	0	1	1	1	1
	R2	1	1	1	1	0	0
	R5	1	1	1	0	1	0

Table 17 Observation: Thermal-High.

Nth Obs.	sensors	quantity	hypotheses				
			H_PD	H_LE	H_HE	H_TL	H_TH
1	R1	2	0	0	0	1	1
	R2	0	0	0	0	1	1
	R5	1	1	1	0	1	1
2	R1	2	0	0	0	1	1
	R2	0	0	0	0	1	1
	R5	2	0	1	1	0	1
3	R1	2	0	0	0	1	1
	R2	0	0	0	0	1	1
	R5	2	0	1	1	0	1
4	R1	2	0	0	0	1	1
	R2	0	0	0	0	1	1
	R5	2	0	1	1	0	1
5	R1	2	0	0	0	1	1
	R2	0	0	0	0	1	1
	R5	1	1	1	0	1	1
6	R1	2	0	0	0	1	1
	R2	1	1	1	1	0	0
	R5	2	0	1	1	0	1
7	R1	2	0	0	0	1	1
	R2	0	0	0	0	1	1
	R5	2	0	1	1	0	1
8	R1	0	0	1	1	1	0
	R2	1	1	1	1	0	0
	R5	2	0	1	1	0	1
9	R1	2	0	0	0	1	1
	R2	0	0	0	0	1	1
	R5	2	0	1	1	0	1

New Knowledge Probability Distribution (NKPD) is shown in Table 18 to Table 22.

Table 18 NKPD: Partial Discharge.

H_PD	H_LE	H_HE	H_TL	H_TH
0.44	0.11	0.11	0.00	0.33
0.78	0.11	0.11	0.00	0.00
0.00	0.00	0.17	0.17	0.67
0.33	0.00	0.00	0.17	0.50
0.67	0.00	0.00	0.17	0.17
0.67	0.00	0.00	0.17	0.17
0.44	0.22	0.22	0.11	0.00
0.78	0.11	0.11	0.00	0.00
0.33	0.11	0.11	0.28	0.17

Table 18 shows hypothesis H_PD has the highest value of Degree of Certainty (DoC) on seven out of nine samples, while the other two do not provide an accurate interpretation.

Table 19 NKPD: Low-Energy Discharge.

H_PD	H_LE	H_HE	H_TL	H_TH
0.11	0.36	0.19	0.25	0.08
0.00	0.36	0.19	0.08	0.36
0.00	0.36	0.19	0.08	0.36
0.00	0.36	0.19	0.08	0.36
0.11	0.31	0.31	0.08	0.19
0.11	0.31	0.31	0.08	0.19
0.11	0.36	0.19	0.25	0.08
0.00	0.36	0.19	0.08	0.36
0.00	0.36	0.19	0.08	0.36

Table 19 shows hypothesis H_LE has the highest value of Degree of Certainty (DoC) on nine out of nine samples with most of them showing other hypothesis with significant DoC.

Table 20 NKPD: High-Energy Discharge.

H_PD	H_LE	H_HE	H_TL	H_TH
0.11	0.31	0.31	0.08	0.19
0.11	0.31	0.31	0.08	0.19
0.11	0.31	0.31	0.08	0.19
0.11	0.31	0.31	0.08	0.19
0.11	0.31	0.31	0.08	0.19
0.11	0.31	0.31	0.08	0.19
0.11	0.31	0.31	0.08	0.19
0.11	0.31	0.31	0.08	0.19
0.11	0.31	0.31	0.08	0.19

Table 20 shows hypothesis H_HE has the highest value of Degree of Certainty (DoC) on nine out of nine samples with all of them showing significant H_LE and other hypothesis with less significant DoC.

Table 21 NKPD: Thermal-Low.

H_PD	H_LE	H_HE	H_TL	H_TH
0.11	0.11	0.00	0.61	0.17
0.00	0.11	0.11	0.67	0.11
0.11	0.11	0.00	0.78	0.00
0.11	0.11	0.11	0.67	0.00
0.00	0.11	0.11	0.67	0.11
0.00	0.00	0.00	1.00	0.00
0.11	0.11	0.00	0.78	0.00
0.00	0.19	0.19	0.42	0.19
0.22	0.31	0.19	0.19	0.08

Table 21 shows hypothesis H_TL has the highest value of Degree of Certainty (DoC) on eight out of nine samples with some of them showing other hypotheses with less significant DoC.

Table 22 NKPD: Thermal-High.

H_PD	H_LE	H_HE	H_TL	H_TH
0.08	0.08	0.00	0.42	0.42
0.00	0.11	0.11	0.33	0.44
0.00	0.11	0.11	0.33	0.44
0.00	0.11	0.11	0.33	0.44
0.08	0.08	0.00	0.42	0.42
0.11	0.22	0.22	0.17	0.28
0.00	0.11	0.11	0.33	0.44
0.11	0.33	0.33	0.11	0.11
0.00	0.11	0.11	0.33	0.44

Table 22 shows hypothesis H_TH has the highest value of Degree of Certainty (DoC) on eight out of nine samples with all of them showing other hypotheses with significant DoC.

The overall accuracy of A3S algorithm is 91.1% as shown in Table 23.

Table 23 NKPD: Thermal-High.

Fault Identification	Accuracy (%)
Partial Discharge	77.8
Low-Energy Discharge	100
High-Energy Discharge	100
Thermal Low	88.9
Thermal-High	88.9
Average	**91.1**

The algorithm is then used to calculate transformer condition and make interpretation of the calculated condition based on DGA data. The condition of a transformer depends on several factors. In this research, the condition is calculated is based on DGA data. Table 24 shows Dissolved Gas Analysis (DGA) data collected from PLN and Table 25 shows other quantities included in transformer condition factors.

Table 24 DGA data

	gas concentration (ppm)					
t(days)	H_2	CH_4	C_2H_6	C_2H_4	C_2H_2	CO
0	20.00	94.54	62.87	0.00	0.00	27.87
486	20.00	0.00	80.97	15.54	0.00	6.58
551	58.01	120.32	169.23	3.26	0.00	205.42
586	439.47	137.21	121.72	63.06	57.12	0.00
873	380.04	137.39	122.89	71.24	75.69	135.97
884	315.90	156.73	123.19	56.56	79.04	55.53
961	276.17	172.71	114.88	74.86	70.50	171.51
985	48.24	94.83	83.03	47.84	29.93	0.00
990	165.21	156.14	163.52	72.09	40.07	120.90
1011	42.91	152.12	157.53	58.60	36.10	226.49
1374	39.84	36.84	199.27	60.37	14.07	35.55
1692	20.00	172.92	208.30	51.90	6.63	305.14
1882	20.00	162.52	179.81	34.24	0.00	0.00
2034	321.12	146.04	210.90	35.13	0.00	516.79
2203	20.00	70.33	236.11	10.74	0.00	0.00
2315	20.00	26.75	87.71	10.93	0.00	0.00
2316	20.00	48.67	75.74	10.58	0.00	100.40

Table 25 Dielectric, water content, and acid number data

Dielectric Breakdown (kV/2.5 mm)	Water Content (ppm)	IFT (dyne/cm)	Acid Number (mg KOH/g)
-	5.96	41.00	0.08
-	5.96	30.00	0.10
-	5.96	29.00	0.14
-	5.96	31.00	0.17
-	5.96	30.00	0.24
-	5.96	30.00	0.23
-	5.96	31.00	0.14
-	5.96	31.00	0.15
-	5.96	32.00	0.15
-	5.96	32.00	0.22
68.40	5.96	32.00	0.20
66.00	5.96	34.20	0.14
50.20	5.96	33.80	0.13
52.60	4.13	33.40	0.13
50.10	4.13	32.20	0.11
55.20	4.13	32.90	0.11
43.10	4.13	32.60	0.11

Data are compared to standards [7, 8] and are given scores. The results are shown in Table 26 and Table 27.

Table 26 Condition grouping and scoring based on DGA

	Condition Scoring Based on DGA					
t (days)	H_2	CH_4	C_2H_6	C_2H_4	C_2H_2	CO
0	100	100	100	100	100	100
486	100	100	67	100	100	100
551	100	67	0	100	100	100
586	67	67	33	67	0	100
873	67	67	33	67	0	100
884	67	67	33	67	0	100
961	67	67	33	67	0	100
985	100	100	67	100	33	100
990	67	67	0	67	0	100
1011	100	67	0	67	0	100
1374	100	100	0	67	33	100
1692	100	67	0	67	67	100
1882	100	67	0	100	100	100
2034	67	67	0	100	100	67
2203	100	100	0	100	100	100
2315	100	100	67	100	100	100
2316	100	100	67	100	100	100

Table 27 Condition scoring based on other factors

	Condition Scoring Based on Other Factors			
t (days)	Dielectric Strength	Water Content	IFT	Acid Number
0	-	100	100	50
486	-	100	0	50
551	-	100	0	50
586	-	100	33	0
873	-	100	0	0
884	-	100	0	0
961	-	100	33	50
985	-	100	33	0
990	-	100	33	0
1011	-	100	33	0
1374	-	100	33	0
1692	100	100	33	50
1882	100	100	33	50
2034	75	100	33	50
2203	75	100	33	50
2315	75	100	33	50
2316	75	100	33	50

The total condition of the transformer is shown in Table 28.

Table 28 Transformer total condition

t (days)	total condition
0	94
486	80
551	69
586	52
873	48
884	48
961	57
985	70
990	48
1011	52
1374	59
1692	68
1882	75
2034	66
2203	77
2315	84
2316	84

The total condition can be drawn in form of Figure 5.

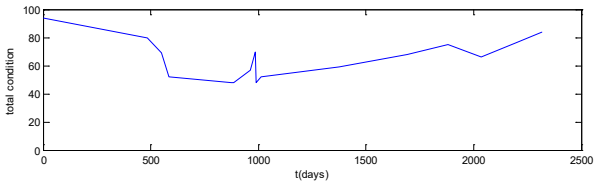


Figure 5 Condition curve

At first, transformer condition has high value, as time passes, it decreases due to the degradation process. The degradation process produces gases, acid, water, which accelerate the degradation process.

There is a sudden increase in condition and then followed by a sudden decrease, this is probably caused by the maintenance process and the setting process after maintenance. After maintenance, the condition began to increase with small gradient.

The gradient of condition is shown in Figure 6.

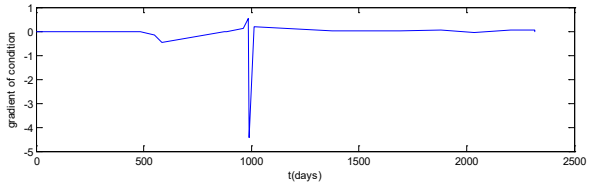


Figure 6 Gradient of condition

The gradient of condition depends on condition. In general, it is relatively stable. There is a sudden decrease at day 990 and then followed by sudden increase.

The observation table is shown in Table 29.

Table 29 Observation table

t (days)	cond	gradient	SC	SG	NKPD	
					HW	HL
0	94	0.0000	1	0	1	1
486	80	-0.0305	1	0	1	1
551	69	-0.1709	1	1	0	1
586	52	-0.4762	1	1	1	0
873	48	-0.0129	0	0	1	1
884	48	0.0000	0	0	1	1
961	57	0.1203	1	1	1	0
985	70	0.5401	1	1	0	1
990	48	-4.4444	0	1	1	0
1011	52	0.1764	1	1	1	0
1374	59	0.0204	1	0	1	1
1692	68	0.0277	1	0	1	1
1882	76	0.0390	1	0	1	1
2034	66	-0.0624	1	1	1	1
2203	77	0.0657	1	1	1	1
2315	85	0.0661	1	1	1	1
2316	85	0.0000	1	0	1	1

Where:

- Cond: condition of the transformer
- Gradient: condition change over time
- SC: cond, after compared to standard
- SG: gradient, after compared to standard

NKPD is knowledge of the system at each time of observation. To obtain knowledge growth, NKPD is fused with the previous NKPD, in this case, NKPD from the beginning of observation. This process produced NKPD over time (NKPDT)

Knowledge growth is represented by NKPDT can be calculated using (6) and is shown in Table 30. Knowledge growth can also be represented using Figure 7.

Table 30 NKPDT

t (days)	NKPD	
	HW	HL
0	0.5000	0.5000
486	0.5000	0.5000
551	0.3333	0.6667
586	0.2500	0.7500
873	0.3000	0.7000
884	0.3333	0.6667
961	0.2857	0.7143
985	0.2500	0.7500
990	0.3333	0.6667
1011	0.3000	0.7000
1374	0.3182	0.6818
1692	0.3333	0.6667
1882	0.3462	0.6538
2034	0.3214	0.6786
2203	0.3000	0.7000
2315	0.2813	0.7188
2316	0.2941	0.7059

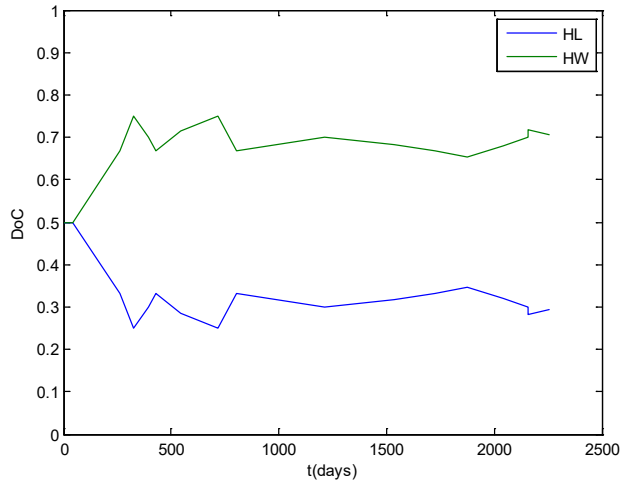


Figure 7 Hypotheses/Knowledge growth

Where HL shows Hypothesis-Life and HW shows Hypothesis-Warning. Both hypotheses show the same value at the first observation. As time passes, HW increases, while HL decreases. It indicates that the transformer condition is in the limit and there are some changes in the gradient. The positive gradient shows maintenance, while negative condition shows degradation.

Spikes in Figure 7 shows there is an occurrence of a phenomenon indicating a hypothesis. In this case, there is a change in gradient of condition, making it in a warning condition. As gradient is dependent to condition, the hypothesis HW is dependent to HL as well.

5. Concluding Remarks

A3S has successfully interpreted DGA data to identified fault based on the classified dataset. It has successfully identified not only the main fault, which has the most significant DoC. It has successfully identified the fault(s) occurred along with the main fault, which has less significant of DoC. This feature acts as the early warning system.

OMA3S has successfully interpreted transformer condition by fusing the parameter Condition (SC) and Gradient (SG) to produce Hypothesis-Life (HL) and Hypothesis-Warning (HW). HL and HW both are fused with the previous values to obtain knowledge growth.

In the next research, more parameters will be included and elaborated to increase accuracy. The number of hypothesis will be added as well to reduce the direct impact of the change of one hypothesis to the other when using only two hypotheses.

This algorithm will be implemented in form of a processor called cognitive processor. Using a special purposes processor will have advantages, such as energy efficiency and minimize disturbance caused by electromagnetic transmission.

Conflict of Interest

Hereby, The authors would like to declares that this article “Cognitive Artificial Intelligence Method for Interpreting Transformer Condition Based on Maintenance Data“ is an original work of our research group and has no conflict of interests.

Acknowledgment

The first author would like to appreciate Yokeu Wibisana for the maintenance data and Harry Gumilang of PLN for the discussion and information provided.

References

[1] A. S. Ahmad and K. O. Bachri, "Cognitive artificial intelligence method for measuring transformer performance," *2016 Future Technologies Conference (FTC)*, San Francisco, CA, 2016, pp. 67-73.

[2] L. Po-Chun and G. Jyh-Cherng,”Research on Transforming Condition-Based Maintenance System using the Method of Fuzzy Comprehensive Evaluation”, *World Academy of Science, Engineering and Technology, vol. 5, 2011*

[3] H. Gumilang, "Hydrolysis process in PLN P3BJB transformers as an effect of oil insulation oxidation," *2012 IEEE International Conference on Condition Monitoring and Diagnosis*, Bali, 2012, pp. 1147-1150.

[4] A. D. W. Sumari, A. S. Ahmad, A. I. Wuryandari, J. Sembiring, “Brain Inspired Knowledge Growing System: Towards a True Cognitive Agent”, IJCSAI Vol. 2 Issue 1, 2012

[5] K. O. Bachri, B. Anggoro, A. S. Ahmad, “Transformer Performance Calculation Using Information Fusion based on DGA Interpretation”, 2016 IEEE International Symposium on Electronics and Smart Devices (ISESD), Lembang, 2016.

[6] Z. Qi, W. Zhongdong, and P. A. CROSSLEY, “Power Transformer End-of-life Modeling: Linking Statistical Approach with Physical Ageing Process”, CIRED Workshop, June 2010.

[7] IEEE Guide for the Interpretation of Gases Generated in Oil-Immersed Transformers," in *IEEE Std C57.104-2008 (Revision of IEEE Std C57.104-1991)* , vol., no., pp.1-36, Feb. 2 2009

[8] IEEE Guide for Acceptance and Maintenance of Insulating Oil in Equipment," in *IEEE Std C57.106-2006 (Revision of IEEE Std C57.106-2002)* , vol., no., pp.1-36, Dec. 6 2006.

[9] M. Duvall, A. dePablo,”Interpretation of Gas-In-Oil Analysis Using New IEC Publication 60599 and IEC TC 10 Databases” IEEE Insulation Magazine, Vol. 17, No. 2, March/April 2001, Feature Article.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1147-1151 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Selective Detrending using Baseline Drift Detection Index for Task-dependant fNIRS Signal

Jinung An*, Gihyoun Lee, Seung Hyun Lee, Sang Hyeon Jin

Convergence Research Center for Wellness, Daegu Gyeongbuk Institute of Science and Technology, 42988, Korea

A R T I C L E I N F O

Article history:
Received: 02 May, 2017
Accepted: 28 June, 2017
Online: 16 July, 2017

Keywords:
fNIRS
Selective detrending
Baseline drift
Wavelet analysis

A B S T R A C T

A functional near-infrared spectroscopy (fNIRS) can be employed to investigate brain activation by measuring the absorption of near-infrared light through the intact skull. The general linear model (GLM) as a standard model for fMRI analysis has been applied to functional near-infrared spectroscopic (fNIRS) imaging analysis as well. The GLM has drawback of failure in fNIRS signals, when they have drift globally. Wavelet based detrending technique is very popular to correct the baseline drift (BD) in fNIRS. However, this method globally distorted the total multi-channel signals even if just one channel's signal was locally drifted. This paper suggests the selective detrending method using BD detection index to indicate BD as an objective index. The experiments show the performance of the proposed method as graphic results and objective evaluation index with current detrending algorithms.

1. Introduction

A functional near-infrared spectroscopy (fNIRS) can detect regional cerebral blood flow (rCBF) variations through the absorption changes of the near-infrared light at wavelengths between 650–950 nm [1, 2]. fNIRS has important advantages of low cost, portability, and the ability to extend research to young children and within more ecological environments [3]. fNIRS shows similar results to blood-oxygenation-level-dependent (BOLD) outputs of fMRI for same cortical regions and fNIRS also has quite similar cortical activation with fMRI from same BOLD signal [4]. fNIRS analysis has been to adapt well-established methods optimized for fMRI such as statistical parameter mapping (SPM) based on the general linear model (GLM) [5]. However, there is noises, the GLM often fails to find cortical activations. During NIRS measurements, the noises are accrued various reasons, such as subject movement, blood pressure variation, and instrumental instability. When subjects move, the fNIRS optodes may shift relative to the head and alter the coupling between the optodes and scalp, and this results in an artifact that abruptly changes, potentially inducing spikes with amplitude much larger than the true signal [6]. These issues can generate system noise, baseline drift (BD), and baseline jumping in fNIRS signals. Especially, the BD is a very important component which can generate errors of brain activation detection. In order to eliminate the BD high-pass filtering is usually used. However, the frequency response of the hemodynamic response can also be affected during high-pass filtering [8]. To overcome this problem, there were many studies that apply a wavelet-based detrending algorithm. They have good performance of detrending and adjusting baseline. However, they have a problem that cannot distinguish between an increase of fNIRS signal by blood flow and BD at single task stimuli.

This paper suggests a new selective detrending method using BD detection index. The proposed method included the BD detection method using an objective index which use an increase-to-decrease ratio of the fNIRS signal trend. According to the BD detection index, the baseline drifted channels are detected and detrended.

2. Theory and Method

2.1. General Linear Model for fNIRS

In general, most analysis of functional hemodynamic changes in neuroimaging such as fNIRS and fMRI is based on an assumption of the linear addition of hemodynamic changes. The GLM has been used as a standard analyzing method for fNIRS signal analysis, and statistical parametric mapping (SPM) [9] is widely used programs based on the GLM. The GLM describes an

*Corresponding Author: Jinung An, Convergence Research Center for Wellness, Daegu Gyeongbuk Institute of Science and Technology, 42988, Korea
Email: robot@dgist.ac.kr

expression of the noisy oxy-hemoglobin signal in terms of a linear combination of L explanatory variables and each error terms:

$$y_{HbO}(r,t) = x_1(t)\beta_1 + \cdots + x_L(t)\beta_L + \varepsilon_{HbO}(r,t), \quad (1)$$

where β_i denotes an unknown strength of response, and $x_i(t)$ is an explanatory variable from a model of hemodynamic responses [10]. Let y and ε denote the vector of the time series of the hemodynamic signal and noise at the location r:

$$\mathbf{y} = [y_{HbO}(r,t_1) y_{HbO}(r,t_2) \cdots y_{HbO}(r,t_N)]^T, \quad (2)$$

$$\boldsymbol{\varepsilon} = [\varepsilon_{HbO}(r,t_1) \varepsilon_{HbO}(r,t_2) \cdots \varepsilon_{HbO}(r,t_N)]^T. \quad (3)$$

The corresponding GLM model is represented by a matrix form as follows:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (4)$$

where $\mathbf{y}$ is an N-dimensional column vector, $\boldsymbol{\varepsilon}$ is error vector, and $\boldsymbol{\beta}$ is L-dimensional column vector that represented from unknown strengths of the response.

2.2. Modifed GLM with Baseline Drift

When there exist BD of a signal, the GLM often fails in the fNIRS measurements [11]. In a novel detrending algorithm for fNIRS, the BD is included as part of the GLM:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} + \theta, \quad (5)$$

where θ is additional global BD. According to the modified Beer-Lambert law (MBLL), the change of the noisy measured optical density (OD) data is described by the following matrix formulation for fNIRS:

$$\begin{bmatrix} \Delta OD(r,t;\lambda_1) \\ \Delta OD(r,t;\lambda 2) \end{bmatrix} = d(r)l(r) \begin{bmatrix} a_1(\lambda_1) & a_2(\lambda_1) \\ a_1(\lambda_2) & a_2(\lambda_2) \end{bmatrix} \begin{bmatrix} \Delta c_{HbO}(r,t) \\ \Delta c_{HbR}(r,t) \end{bmatrix} + \begin{bmatrix} w(r,t;\lambda_1) \\ w(r,t;\lambda 2) \end{bmatrix} + \begin{bmatrix} \widetilde{\theta}(r,t;\lambda_1) \\ \widetilde{\theta}(r,t;\lambda 2) \end{bmatrix}, \quad (6)$$

Where $d(r)$ is differential path length factor (DPF) and $l(r)$ is distance between the sources to detector at position r. $a_i(\lambda_i)$ is the extinction coefficient of the i'th chromophore at wavelength (λ). $\widetilde{\theta}(r,t;\lambda_i)$ denotes the global trend for wavelength (λ_i) at the location r. Let $\mathbf{y}_{HbO}$, $\boldsymbol{\varepsilon}_{HbO}$, and $\boldsymbol{\theta}_{HbO}$ denote vectors of the oxy-hemodynamic response signal, noise signal, and global trend signal. The modified GLM is represented for fNIRS as follows:

$$\mathbf{y}_{HbO} = \mathbf{X}_{HbO}\boldsymbol{\beta}_{HbO} + \boldsymbol{\varepsilon}_{HbO} + \boldsymbol{\theta}_{HbO}. \quad (7)$$

2.3. Proposed Selective Detrending Method

The measured oxy-hemodynamic noisy signal (ŷ) can be represented by time-channel form as follows:

$$\hat{\mathbf{y}}(t,c) = \begin{bmatrix} y(t_1,c_1) & y(t_2,c_1) & \cdots & y(t_L,c_1) \\ y(t_1,c_2) & y(t_2,c_2) & \cdots & y(t_L,c_2) \\ \vdots & \vdots & \ddots & \vdots \\ y(t_1,c_N) & y(t_2,c_N) & \cdots & y(t_L,c_N) \end{bmatrix}, \quad (8)$$

where L is length of acquisition time, c is acquisition channel, and N is number of channel from NIRS signal acquisition system.

On the other hands, wavelet transform is widely used in various signal processing because it is very efficient and powerful for frequency analysis of a signal. In this paper, the measured oxy-hemodynamic noisy signal (ŷ) is decomposed using modified wavelet transform based on Daubechies6 wavelet basis. The modified wavelet transform consist of 11 level tree and 12 sub-bands of the wavelet coefficient. The decomposed oxy-hemodynamic noisy signal (wjŷ(t,c)) is the wavelet coefficient of the j'th level at time (t) and channel (c), and it can be expressed in 3-demensional matrix as follows:

$$w^j\hat{\mathbf{y}}(t,c) = \begin{bmatrix} w^1\hat{\mathbf{y}}(t,c) \\ w^2\hat{\mathbf{y}}(t,c) \\ \vdots \\ w^{11}\hat{\mathbf{y}}(t,c) \end{bmatrix}. \quad (9)$$

The wavelet coefficient of the 6th level (w6ŷ) is well represented on the signal trend of hemodynamic response, and the 11th level (w11ŷ) is represented on the signal slope of the global baseline. The BD generates a difference between length of increase sections and decrease sections. Using the difference of the wavelet coefficients, we calculate a new index for the BD detection. Let γ+ is increase and γ- decrease length of the wavelet coefficient of the 6th level (w6ŷ):

$$\gamma^+ = \int diff[w^6\hat{\mathbf{y}}]^+, \quad \gamma^- = \int diff[w^6\hat{\mathbf{y}}]^-, \quad (10)$$

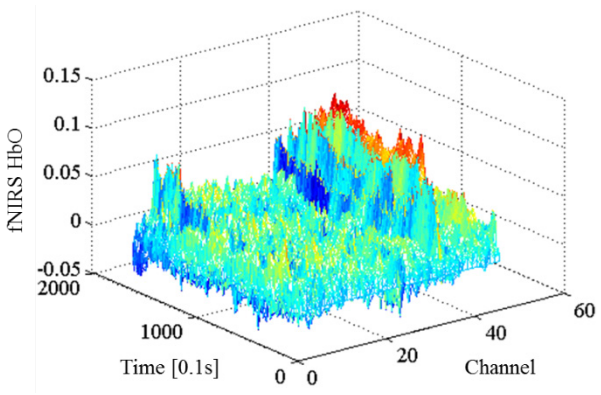
where diff[w6ŷ]+ is a positive differential value of w6ŷ, and [w6ŷ]- is negative differential value of w6ŷ. The BD index (BI) defined as follows:

$$BI(c) = \frac{\gamma^+}{\gamma^-} \cdot \sum_{i=1}^{L} [w^{11}(i,c) - w^{11}(i-1,c)], \quad (11)$$

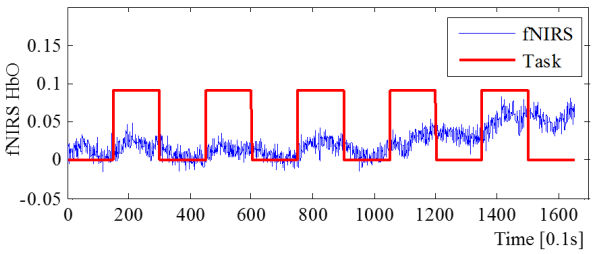
here, BI(c) objectively indicates a degree of the BD for each measured channels. And then, the detrending process is selectively performed depending on a BD threshold, and it is determined by abrupt increase changing point on sorted channel. If an index (BI) of some channels have high values, the detrending process of the fNIRS signals of the channels are selectively performed using wavelet-MDL (minimum description length) [10].

3. Experimental Result

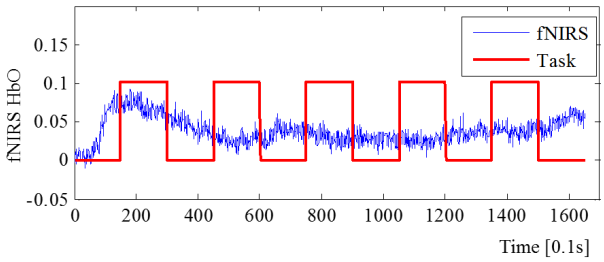
The fNIRS data were obtained by fNIRS brain imaging system (FOIRE-3000, Shimadzu, Japan) at 10 Hz sampling frequency. All of data were approved by the Institutional Review Board (IRB) at Daegu Gyeongbuk Institute of Science & Technology (DGIST). To perform experiment of the proposed BD index, a subject who is right-handed without known neurological, physical illness participated. Fig. 1 shows the task-dependant test data 1 which is obtained from general motions of upper-limb rehabilitation in a block design (5 repetitions with 15s task and 15s rest). BD can be found in channel 6, 33, and 43 in Fig. 1 (a). Oxy-hemoglobin (HbO) signals for the selected channels are shown in Fig. 1 (b)-(d).



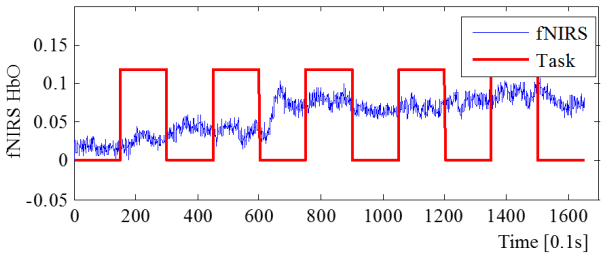
(a)



(b)

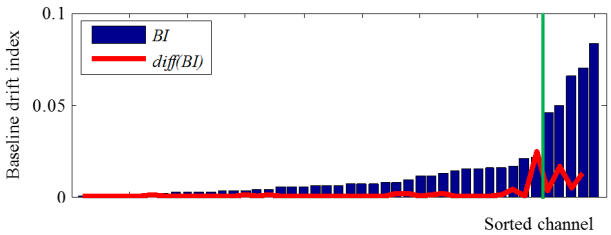


(c)

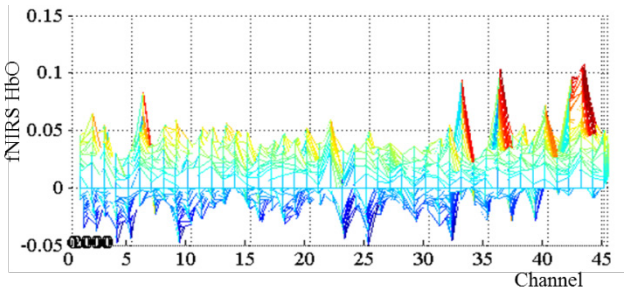


(d)

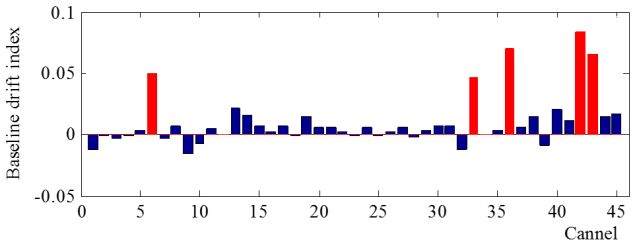
Figure 1. The signals of test data 1, (a) fNIRS signal on time-all channels, (b) the signal of channel 6, (c) the signal of channel 33, and (d) the signal of channel 43,



(a)



(b)

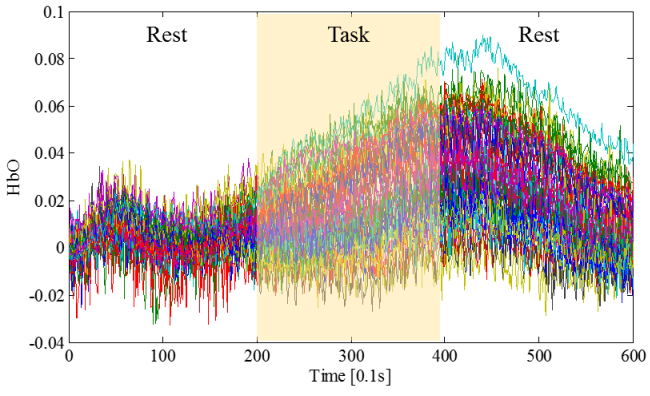


(c)

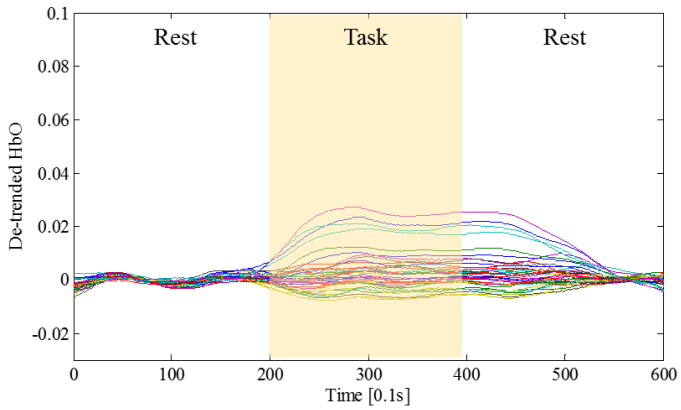
Figure 2. BD detection result from test data 1, (a) *BI* on sorted channel, (b) fNIRS signal on each channels, and (c) the result of BD detection.

To evaluate the proposed BI, the result of BD detection is presented in Fig. 2. The change of BI on sorted channel and a threshold of BD are shown in Fig. 2 (a). As can be seen Fig. 2 (b) and (c), the BI of channel 6, 33, 36, 42, and 43 have more high score than the threshold, and baseline drifted channels for detrending can be detected.

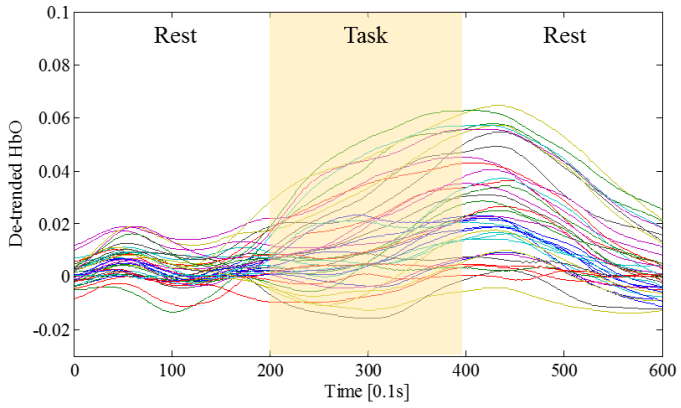
The obtained fNIRS signal is analysed by HRF (hemodynamic response function) smoothind and wavelet-MDL. HRF smoothing and wavelet-MDL have good performance for denoising and detrending, and the detrending of wavelet-MDL can adjust BD. However, wavelet-MDL can often distort a NIRS signal at a specific experimental environment of only single task stimuli. The graphical results of fNIRS detrending are shown in Fig 3 to compare with wavelet-MDL. In order to evaluation performance of the detrending, task-dependant test data 2 that is obtained from general walking motions in a block design (1 repetitions with 20 s task and 20 s rest)



(a)



(b)



(c)

Figure 3. Detrending of test data 2, (a) the measured original fNIRS signals, (b) the result of detrended fNIRS signal using HRF smoothing and wavelet-MDL, and (c) the result of detrended using proposed method.

Fig. 3 (a) is the measured original fNIRS signals of 40 channels which measured on a whole brain. (b) shows the result of detrended fNIRS signal using HRF smoothing and wavelet-MDL on all channels, and can be confirmed that it has too much signal distortion to compare with Fig. 3 (a). However, Fig. 4 (b) shows the result of detrended fNIRS signal using proposed selective detrending method, it keeps original characteristic and hemodynamic response of fNIRS signals on Fig. 3 (a).

In order to evaluate performances of the algorithms, the contrast-to-noise ratio (CNR) is used to the fNIRS signals. If large difference of each part and small variation of one part, the CNR has high value. The CNR indicate that the higher CNR value of fNIRS signal, the better performance of the algorithm. Table 1 shows detailed CNR of real measured whole brain fNIRS signals of 40 channels that used on Fig. 3.

Table 1 The performance of fNIRS detrending obtained using CNR.

		Original	Wavelet-MDL	The proposed method
CNR (dB)	Measured noise	0.231	0.128	0.327
Improvement rate		-	55 %	142 %

As can be see Table 1, the proposed method has high CNR than both of the original fNIRS signal and the wavelet-MDL. Moreover, wavelet-MDL has low CNR than the original signal because of the signal distortion. It shows that the proposed method has a better detrending performance than conventional algorithm on the task-oriented fNIRS signals.

4. Conclusions

This paper dealt with detrending of globally drifted fNIRS signal. The baseline drift detection method was suggested to selectively extract the drifted fNIRS channels by quantifying objective indicator of baseline drift. It was determined by quantitatively manipulating wavelet coefficients based on a stability of the baseline. The proposed detrending highly performed the identification of selective drifted channels and enhancement of fNIRS singal. This method will be carried out in advance of global detrending of fNIRS signals for enhancing quality of task-dependant fNIRS signal post-processing including neuro-cortical mapping, brain signal decoding, etc. The autonomous optimal baseline index determination and following noise elimination should be further discussed.

Ethical Approval

This study was reviewed and approved by the institutional review board at the Ethics Committee of Daegu Gyeongbuk Institute of Science & Technology (DGIST_151211_HR-020-06).

Acknowledgment

This work has been funded by the R&D Program 10045164 of the Ministry of Trade, Industry and Energy, KOREA, and this work was supported by Institute for Information & Communications Technology Promotion (IITP) grant funded by the Korea government (No. 2017-0-00451).

References

This paper is an extension of work originally presented in Brain-Computer Interface (BCI) 2017 5th International Winter Conference [1].

[1] G. Lee, S. H. Lee, S. H. Jin, and J. An, "Baseline drift detection index using wavelet transform analysis for fNIRS signal," in *Brain-Computer Interface (BCI), 2017 5th International Winter Conference on*, 2017, pp. 73-76.

[2] H. Santosa, M.J. Hong, S.P. Kim, and K.S. Hong, "Noise reduction in functional near-infrared spectroscopy signals by independent component analysis," Review of Scientific Instruments, vol. 84, no. 7, 2013.

[3] T. Fekete, D. Rubin, J.M. Carlson, and L.R. Mujica-Parodi, "The NIRS analysis package: noise reduction and statistical inference," PloS one, vol. 6, no. 9, 2011.

[4] S. Jin, S.H. Lee, G.H. Jang, Y.J. Lee, J. An, and Y.S. Lee, "Cortical activation by fNIRS and fMRI during grasping in patient with traumatic brain injury: a case study," In Brain-Computer Interface (BCI), 2014 International Winter Workshop on IEEE, pp. 1-4, 2014.

[5] K.J. Friston, P. Fletcher, O. Joseph, A. Holmes, M.D. Rugg, and R. Tunner, "Event-related fMRI : characterizing differential responces," NeuroImage, vol. 7, pp. 30–40, 1998.

[6] J.C. Ye, S. Tak, K.E. Jang, J. Jung, and J. Jang, "NIRS-SPM: statistical parametric mapping for near-infrared spectroscopy," Neuroimage, vol. 44, no. 2, pp. 428-447, 2009.

[7] B. Abibullaev, and J. An, "Classification of frontal cortex haemodynamic responses during cognitive tasks using wavelet transforms and machine learning algorithms," Medical engineering & physics, vol. 34, no. 10, pp. 1394-1410, 2012.

[8] K. J. Friston, J. Ashburner, S. Kiebel, T. Nichols, and W. Penny, Eds., Statistical Parametric Mapping: The Analysis of Functional Brain Images, Academic Press, San Diego, CA, 2006.

[9] K J. Friston, A P. Holmes, K.J. Worsley, J.P. Poline, C.D. Frith, and R.S.J. Frackowiak, "Statistical parametric maps in functional imaging: a general linear approach," Hum. Brain Mapp, vol. 2, no. 4, 189-210, 1995.

[10] K.E. Jang, S. Tak, J. Jung, J. Jang, Y. Jeong, and J.C. Ye, "Wavelet minimum description length detrending for near-infrared spectroscopy," Journal of biomedical optics, vol. 14, no. 3, 2013.

[11] M. Cope and D.T. Delpy, "System for long-term measurement of cerebral blood and tissue oxygenation on newborn infants by near infra-red transillumination," Med. Biol. Eng. Comput, Vol. 26, pp. 289–294, 1988.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1152-1158 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Autonomous Micro-Air-Vehicle Control Based on Visual Sensing for Odor Source Localization

Kenzo Kurotsuchi*, Mitsuharu Tai, Hiromasa Takahashi

Center for Exploratory Research, Hitachi, Ltd., 1-280, Higashi-Koigakubo, Kokubunji, Tokyo 185-8601, Japan

ARTICLE INFO

Article history:
Received: 11 May, 2017
Accepted: 28 June, 2017
Online: 16 July, 2017

Keywords:
odor source localization
autonomous
micro air vehicle
drone
gas sensor
vision

ABSTRACT

In this paper, we propose a novel control method for autonomous-odor-source localization using visual and odor sensing by micro air vehicles (MAVs). Our method is based on biomimetics, which enable highly autonomous localization. Our method does not need any instruction signals, including even global positioning system (GPS) signals. An experimenter simply blows a whistle, and the MAV will then start to hover, to seek an odor source, and to keep hovering near the source. The GPS-signal-free control based on visual sense enables indoor/underground use. Moreover, the MAV is light-weight (85 grams) and does not cause harm to others even if it accidentally falls. Experiments conducted in the real world were successful in enabling odor source localization using the MAV with a bio-inspired searching method. The distance error of the localization was 63 cm, more accurate than the target distance of 120 cm for individual identification. Our odor source localization is the first step to a proof of concept for a danger warning system. These localization experiments were the first step to a proof of concept for a danger warning system to enable a safer and more secure society.

1. Introduction

As the performance of sensors is improved and as the internet of things (IoT) is expanded, new social systems utilizing sensors are being designed to improve safety and security. Although odor sensing is one of the most important kinds of information for living things, the sensitivity of gas sensors in the past could not come close to that of a living thing, such as dogs, which can respond to a parts-per-trillion (ppt) odor [1]. Gas sensors with parts-per-quadrillion (ppq) sensitivity have been reported [2]. As gas sensors are improved, new safety systems and disease-detection systems are being designed using gas sensors [2, 3].

This paper is an extension of work originally presented in 2016 23rd International Conference on Mechatronics and Machine Vision in Practice (M2VIP 2016) [4]. We challenged to discover issues and consider countermeasures for constructing a danger warning system to enable a safer and more secure system utilizing autonomous odor source localization by experiments. The results of the experiments were mainly presented in the previous paper.

*Corresponding Author: Kenzo Kurotsuchi, Center for Exploratory Research, Hitachi, Ltd.; 1-280, Higashi-Koigakubo, Kokubunji, Tokyo 185-8601, Japan; Tel: +81-42-323-1111; Fax: +81-42-327-7699; E-mail: kenzo.kurotsuchi.qs@hitachi.com

Micro air vehicles (MAVs) are promising carriers of gas sensor devices in safety systems because these vehicles can move quickly enough to chase dubious characters and fly over pedestrians without any interference. In safety systems, plural MAVs will work in a coordinated manner and communicate with other systems such as security cameras and audience guide systems. If a MAV detects danger, the MAVs will give a warning of the danger using an alarm bell and/or notification to other security systems to call a security guard. While the MAVs need a lot of information, central control computer and network performance is limited. To enable robust control in complex situations, an autonomous-MAV-control system is required. Therefore, we studied autonomous-MAV control.

Odor source localization has been studied widely for detecting explosives, for narcotics control, and for detecting gas leaks to ensure social safety and security [5–12]. Odor source localization using MAVs is a challenging task due to the non-uniform dispersion of odor, low sensitivity of commercially-available-gas sensors, and constraints of MAVs. Odor from its source is moved by the wind and is divided into many odor filaments. Localizing the source by following the concentration gradient of odor is difficult. In addition, MAV control is difficult due to unstable flight performance, low performance cameras for self-localization,

small payload, and limited flight time. Therefore, a robust control method for odor source localization is needed. Many algorithms for odor source localization have been proposed. The first algorithm uses geometric control such as spiral patterns [6]. The second one is based on Bayesian inference theory [8]. The third mimics biological models such as cells [14], bacteria [9], insects [15], moths [10, 11], and ants [12]. Insects and animals use their olfactory senses with ingenuity for collecting food, avoiding enemies, and finding mates [10]. In this study, a bio-inspired algorithm, specifically a bacterium model, was utilized to achieve stable odor source localization in difficult circumstances. We thought that simple and effective control of bacterium model is valid as the first step in biomimetic control for odor source localization.

In addition, mobile robots—including MAVs—must have a self-localization method so that they can determine their position with respect to known locations in the environment in order to navigate effectively and to reach the goal region. For self-localization, two methods have been proposed. One is vision-based control, and the other is global positioning system (GPS) based control [5]. The safety system has applicability to large facilities, such as sports stadiums, stations—including underground train stations—and airports. The aforementioned places may be out of reach of the GPS-based control. Therefore, we selected a method that is GPS signal free and that has vision-based self-localization. This approach enhances the independence of the MAV.

2. Micro-Air-Vehicle Platform

2.1. Micro Air Vehicles

The Phenox 2 (Phenox Lab, referred to as MAV in this article) MAVs were selected because they are light-weight, they have high computing performance for vision-based self-localization, and they can be extended by attaching odor sensors and enabling control program flexibility [13, 16]. Figure 1 shows the Phenox 2, which flies using four electric motors and has a diameter of 176 mm. The specifications of the MAVs are summarized in Table I. The light-weight MAVs were designed so as not to cause harm to others even if they accidentally fall. One of them weighs 85 grams, including the odor sensor module, LiPo battery, and propeller guard. In some countries, there are regulations for drone whose weight is above standard (e.g. 200 grams in Japan). The MAV is lighter than the standard and easy to use. The controller chip is Zynq 7000 (Xilinx, XC7Z010), which consists of FPGA and dual-core CPUs. The FPGA conducts fast image processing for controlling self-positioning. The CPUs are Cortex-A9 667 MHz dual-core CPUs based on the ARM architecture. One CPU (CPU1) is used for real-time operation such as low-level-flight control. The other (CPU0) runs an odor-source-localization program written in C language on the Ubuntu/Linux OS. 256 MB DDR3 memory and a 4 GB microSD card are mounted on the MAV for working memory and storage.

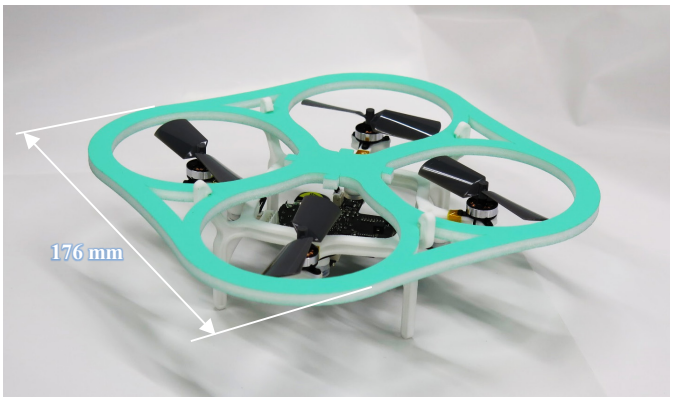


Figure 1. Photograph of the MAVs.

Table I. Specifications of the MAV

MAV name	Phenox 2
Developer by	Phenox Lab
Microprocessor	Xilinx Zynq XC7Z010 (Dual core Cortex-A9 processer + FPGA)
Camera	CMOS Image Sensor (320×240)

The flight control relies on cameras, an onboard inertial measurement unit (IMU, 9-DOF), range sensor, microphone, and gas sensor. The front and bottom cameras are CMOS cameras with 320×240 pixels, respectively. In this experiments, we use only bottom camera. The IMU consists of a three-axis accelerometer and a three-axis rotation rate sensor. In addition, a sonar sensor is used to control the altitude of the MAV. The MAV has a mounted microphone that receives a 3 kHz whistle signal blown by the experimenter. Communication is established using a Wi-Fi to send flight logs to the ground station. The MAV supports I^2C communication to extend the hardware capabilities. We used I^2C communication to connect to an odor sensor module board.

Figure 2 shows a block diagram of the MAV. The bottom camera mounted on the MAV collects image data. The front camera is unused for controlling the MAV. The FPGA conducts image processing of feature point detection based on the FAST algorithm [13, 20] and feature point description based on the BRIEF algorithm [21]. CPU1 calculates the *current position* based on the feature points for self-localization, and it notifies CPU0 of the *current positio*n. CPU0 determines the *target position* based on

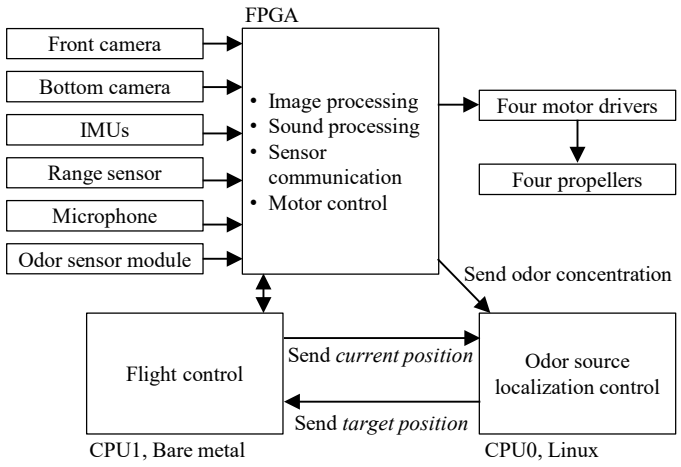


Figure 2. Block diagram of MAV control system based on vision and odor sensing.

the *current position* and current odor concentration, and it notifies CPU1 of the *target position*. CPU1 calculates the parameters of the four motor drivers based on the *current position*, the *target position*, and information on the IMUs and range sensor. The FPGA controls the motor drivers to enable propellers to move the MAV.

2.2. Odor sensor module board

The odor sensor module board was designed for light-weight, low-power consumption. A photograph of the odor sensor module is shown in Figure 3. The gas sensor (Figaro Engineering, TGS8100) converts the concentration of odor to the current magnitude. We selected a MEMS-type gas sensor, which makes the sensor small and light-weight and enables it to have low-power consumption. A constant voltage is applied to the gas sensor. If hydrogen or alcohol is present, the resistance of the gas sensor is reduced, and the current flows more. For example, the current magnitude of 10 ppm ethanol is five times higher than that of clean air. The sensitivity of the commercially-available gas sensor is much lower than the nose of living things. For example, dogs can respond to a parts-per-trillion (ppt) odor [2]. Light-weight odor sensor modules with low-power consumption were fabricated

The analog front-end (Texas Instruments, LMP91000) converts the current magnitude to voltage magnitude. The analog-digital converter (Texas Instruments, ADS1100) converts the voltage magnitude to a digital signal and transmits to MAV's CPU through the I/O. The I/O consists of four pins. Two pins are used for I^2C communication. The other two pins are used for a voltage supply at 3.3 volts. The linear regulator (Texas Instruments, LP5907MFX-1.8) converts the supply voltage (3.3 volts) to voltage for the heater of a gas sensor (1.8 volt).

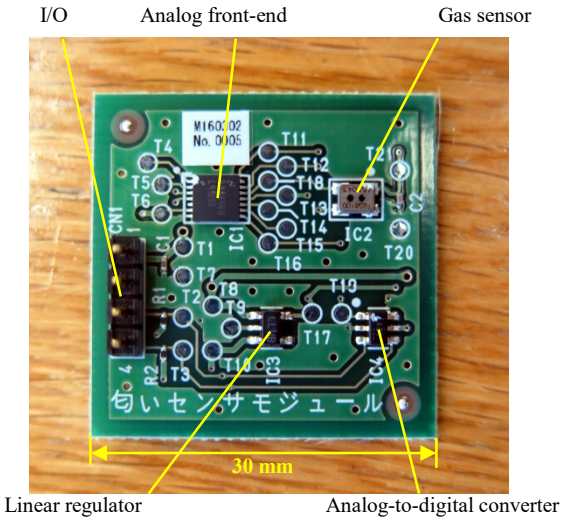


Figure 3. Photograph of odor sensor module.

If the odor concentration increases, the magnitude of the current flow in the gas sensor and the output voltage of the analog front-end increase. Threshold output voltage V_T for deciding how to finish the odor source localization is determined as follows:

$$V_T = V_I + 0.7 \text{ mV} \quad (1)$$

Initial voltage V_I was measured at the start of the experiment; namely, the V_I means the voltage of the clean air.

2.3. MAV arena

Reproducible experimental conditions are required to compare different odor source localization methods. Figure 4 shows a real-

Figure 4. Real MAV arena used for the experiments.

world experiment that was carried out in a 7 × 12 m training room without special lighting. A fabric (Asahi Kasei, Bemcot) sprayed with ethanol for disinfection (ethanol 76.9–81.4% v/v) was set on the top of a light stand as the odor source. Plural markers 30 cm in diameter were put into place to help the image feature extraction and to stabilize flight.

The MAV was placed on the floor at a position 1.8 m away from the odor source. Two external cameras were set outside of the arena. One camera with a fisheye lens was placed on the floor to take a movie. After the experiment, the trajectory of the flight was calculated. Another camera with a normal lens was set on a tripod to take a movie of the experiment. No signals were sent to the MAV for controlling it.

3. Bio-Inspired-Search Algorithm

3.1. Algorithm

We utilized a bio-inspired algorithm, specifically a bacterium model, to achieve stable odor source localization in difficult circumstances due to the limited power of the MAV's onboard computer, the low image quality of the camera, and the low sensitivity of the gas sensor.

The cells undergo chemotaxis to get close to acceptable circumstances [19]. Some bacteria, such as E. coli, have several flagella, which can rotate in two ways. One is counter-clockwise (CCW) rotation, which aligns the flagella into a single rotating bundle, causing the bacterium to swim in a straight line (Straight swim). The other is clockwise (CW) rotation that breaks the flagella bundle apart, causing the bacterium to tumble in place (Tumbling). The chemotaxis is controlled by switching the rotation depending on the concentration gradient.

Figure 5 shows a pseudocode of the firmware for controlling the MAV. At the start of the experiment, the MAV was placed on the floor. The MAV started its takeoff after recognizing the whistle signal blown by the experimenter (pseudocode line 1). The cruising altitude was controlled to 1.2 m. After the whistle, the MAV flew autonomously, and the experimenter did not control the MAV. In addition, the MAV did not receive any external control signals, including a GPS signal. Next, the MAV hovered for eight seconds (line 2 and 3), and measured the initial odor concentration based on the output value of the analog-digital converter. The initial odor concentration determined the threshold odor concentration for finishing the odor source localization.

Also, the MAV randomly decided the *traveling direction* (line 5) from 360° to start odor source localization. If the odor concentration was less than the threshold, the processes of lines 7 to 23 were repeated. In these processes, the *target positions* were updated based on the *traveling direction*. The value of *Speed* was constant. Next, the MAV measured the odor concentration (line 9). If the odor concentration decreased (line 10), the MAV turned back with a probability of 20% (lines 11–15) (tumbling). However, if the odor concentration did not decrease, the MAV went forward (lines 16-19) (straight swim). Please note that the *traveling direction* made a small change with a probability of 5% (line 17). If the MAV became likely to fly out of the arena based on self-position localization using onboard camera and onboard CPUs, the MAV changed the *traveling direction* (line 20, 21) randomly from 180°. The size of the arena was limited to 400×700 pixels in the MAV camera, which corresponds to about 3×6 m in real space. The MAV was controlled to travel to the *target position* (line 22).

```
1  MAV waits for the whistle
2  MAV hovers
3  MAV waits 8 seconds
4  MAV measures the initial odor concentration
5  traveling direction ← random
6  while (Concentration of odor is lower than threshold)
7      target_x_position += Speed * Sin(traveling direction)
8      target_y_position += Speed * Cos(traveling direction)
9      MAV gets odor concentration
10     if (Concentration of odor decreased)
11         // Tumbling
12                 if (random < 20%)
13             // Go backward
14             traveling direction += 135º–225º
15     else
16         // Straight swim
17         if (random < 5%)
18             // Change pathway
19             traveling direction += -45º–45º
20     if (MAV_position == outside)
21         traveling direction ← random
22     lead MAV to (target_x_position, target_y_position)
23     MAV waits 80 msec
24 MAV continues hovering
```

Figure 5. Pseudocode of MAV firmware to detect odor source.

These processes were repeated every 80 ms (line 23). If the odor concentration exceeded the threshold, the MAV continued hovering in the position (line 24), and notified the experimenter that the MAV found the odor source. The *flight time* was defined as the search time duration from finishing the initial hovering in eight seconds to starting the hovering near the odor source.

3.2. Comparison in simulation

The bio-inspired algorithm and random-walk algorithms for MAV's control were compared in a simulation. The time required for odor source localization varied significantly and randomly. Therefore, many trials were required for studying the algorithms. Computer simulations, enabling results to be obtained in a short time, were useful.

In the simulation, the flight area of the MAV was 5×5 m, and the speed was 0.4 m/s. The distance between the MAV and the odor source was set to 1.8 m. We assumed the MAV could detect an odor concentration within 150 cm of the odor source and that the device succeeded in the odor source localization if it approached within 63 cm of the odor source. These values of the distances were determined experimentally.

The bio-inspired algorithm is shown in Figure 5. The random-walk algorithm is based on the bio-inspired algorithm excluding the "Go backward" process (tumble in place). A total of 19 trials were conducted for each algorithm.

The simulation results are shown in Figure 6. The results show the search time based on the bio-inspired algorithm was shorter. The median time required for odor source localization was 17 seconds in the bio-inspired algorithm and 36 seconds in the random-walk algorithm, respectively. The search time based on the bio-inspired algorithm was as much as 47% shorter than that based on the random-walk algorithm. The efficient bio-inspired algorithm was utilized for a real-robot search using the MAV.

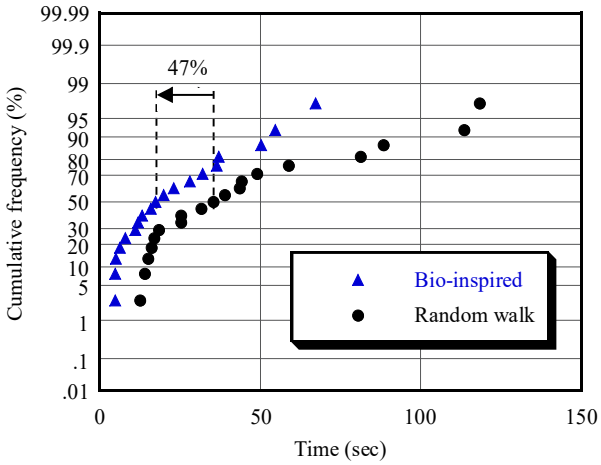


Figure 6. Simulation results of odor-source-localization time based on the control of bio-inspired algorithm and random-walk algorithm.

4. Results

The results of the experiments are shown in Table II. In the four experiments, the MAV succeeded in finding the odor source. In the 5th experiments, the MAV flied away from arena. The median distance of the MAV from the odor source was 63 cm. The median *flight time* required for searching the odor source was 35 seconds.

The results of the first trial are described in detail using Figures 7 to 9. The trajectory of the MAV is shown in Figure 7. The *actual position* was calculated using the recorded movie taken by the external camera with fisheye lens. The MAV took off from the placed position with various inclination angles due to deviation from the center of gravity in the MAV. The start position shows the position in which search began after taking off and hovering for 8 seconds. The MAV moved from position 1→2→3→4→5 and hovered, which shows that the MAV found the odor source. Figure 8 shows the time variation in the trajectory of the MAV. From 8.2 to 15.5 seconds, the MAV flew over the horizon of the external camera. Figures 8-(c) and (d) show the *target position* set by the MAV autonomously and the *current position* recognized by the MAV. At 5.7 seconds, the MAV finished hovering and started to search for the odor source.

Table II. Experiment results for odor source localization

# of trials	Distance from odor source to MAV's goal (cm)	Flight time (sec)
1	73	34
2	42	36
3	79	29
4	53	77
5	Fly away from arena	
Median	63	35

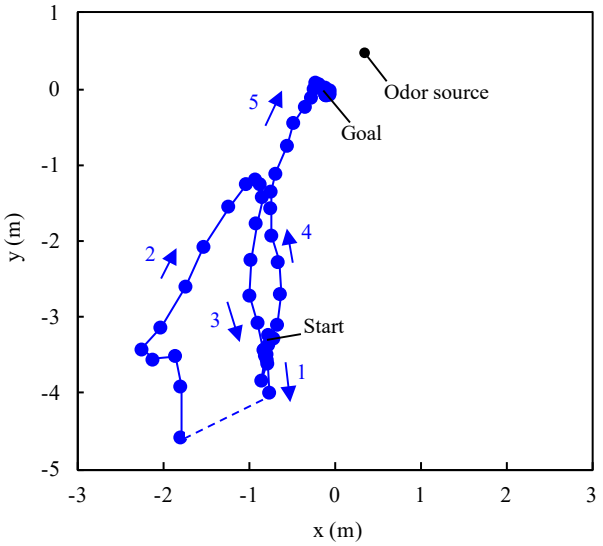


Figure 7. MAV flight trajectory. The *actual position* of the MAV was determined using the external camera.

Figure 8 shows the *actual position*, *target position*, and *current position*. The *actual position* in Figure 8-(a) and (b) was calculated from the footage taken by the external camera. The *target position* and *current position* in Figure 8-(c) and (d) were logged by the MAV. Figure 8 shows that the *actual position*, *target positio*n, and *current position* were attuned to each other.

Figure 9 shows the variation in time of the output voltage of the analog front-end, which indicates the odor concentration. The initial voltage V_I at the start of the odor source localization, namely the voltage of the clean air, was 919.3 mV. Therefore, the threshold voltage V_T for finishing the search was set to 920.0 mV. Figure 9 shows the time of the MAV turn. The MAV operated the change pathway three times and the go backward three times. At 40.2 seconds, the output voltage exceeded the threshold voltage V_T and notified the MAV to finish searching and started hovering.

The deviation distance of 63 cm means this localization method can achieve personal identification because the distance was lower than the personal space of 120 cm [17]. Most people feel discomfort, annoyance, or some level of anxiety when someone or something else encroaches on their personal space. In addition, the deviation distance allows for people to escape from danger found by a MAV. Our results are the first step towards improving social safety and security systems. The reasons for the deviation between the odor source and the position specified by the MAV can be stated as follows: (1) Odor diffusion by the wind; (2) A searching algorithm that terminates the search if the odor concentration exceeds a threshold and does not find the position with highest odor concentration; and (3) Utilization of a proportional and derivative (PD) control method that excludes integration (I) control.

In this experiment, the deviation between the *target position* and *current position*, which depends on the *travelling direction*, was observed. For example, Figure 8-(d) shows that if *target position* y was larger than the *current position* y, the deviation increased. The reason is that the estimated center of gravity in the MAV is located behind the ideal position, i.e., the device is weight biased. Therefore, forward moving, which decreases the *current position*, is slower than backward moving.

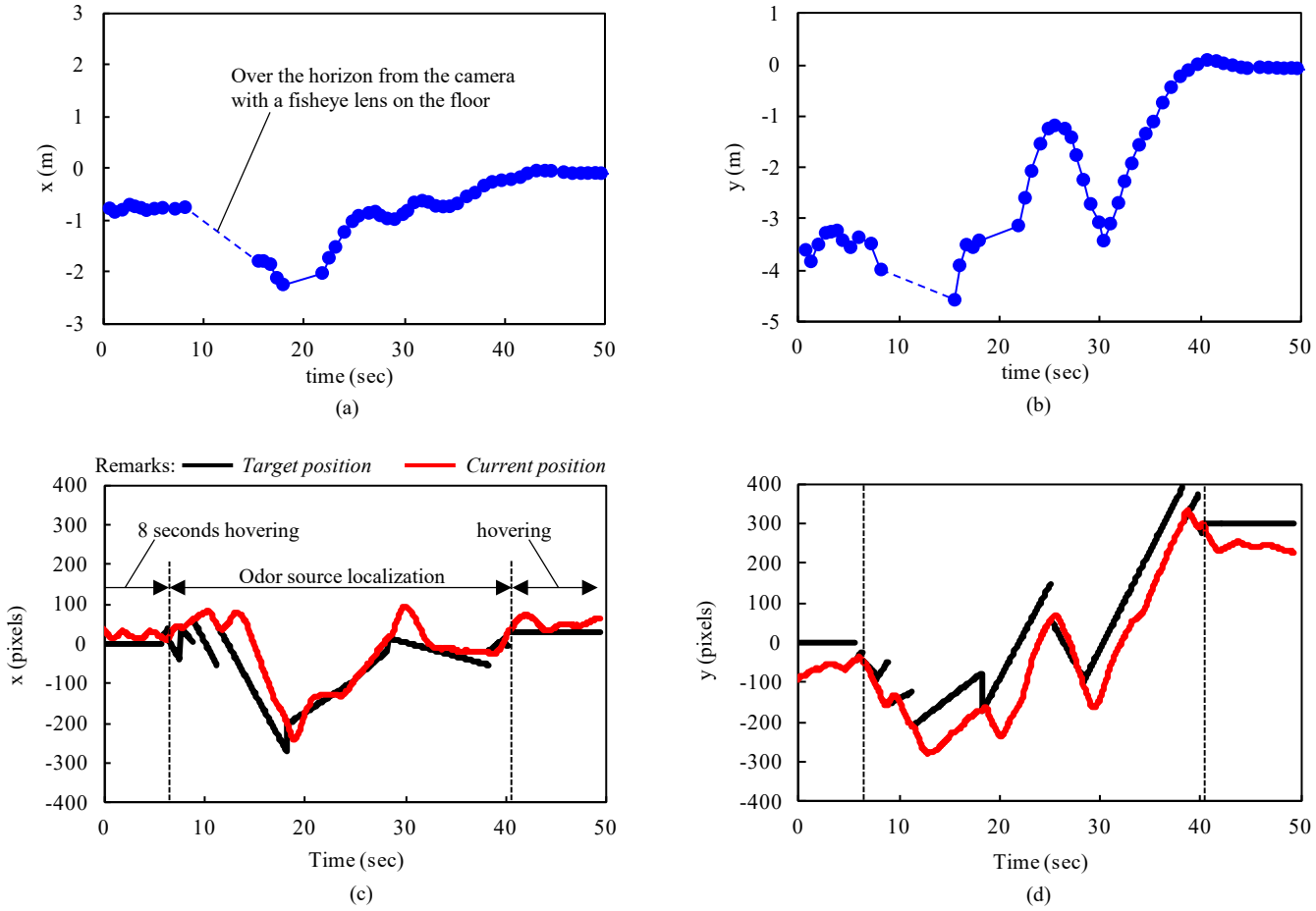


Figure 8. x/y MAV flight paths. The *actual position* of the (a) and (b) was determined using the external camera. The *current position* of (c) and (d) was calculated and logged by onboard CPU using the onboard camera. The *target position* of (c) and (d) was determined and logged by the onboard CPU.

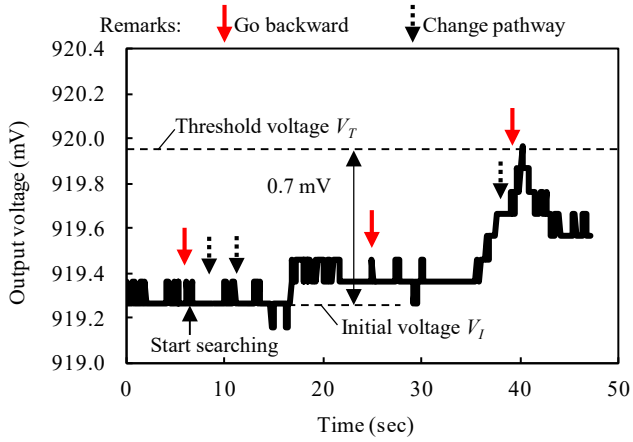


Figure 9. Time – output voltage plot. The odor concentration was measured by the gas sensor and converted to the output voltage using the analog front-end.

The search time in the real-MAV-experiment results was longer than that in the simulation results. The reason for that was presumed to be the simulation assuming that the velocity of the MAV was constant, whereas the real-MAV-experiment took sufficient time to turn around due to the PD control.

5. Discussion

The issues obtained by this experiment for constructing a danger warning system to enable a safer and more secure society with MAVs are summarizes below:

- Vision sensing
- MAV's balancing
- Odor sensing
- Swarm control

In the vision sensing, we used the markers. In the actual safety system, the markers often blocked by pedestrians. Therefore, more robust self-positioning method will be required. Dynamic detection method of crowd was reported [22]. In addition, the 5th experiment was failed by flying away from arena. This reason was the MAV lost the self-positioning estimation. The arena was illuminated by a strong afternoon sun. The bottom camera did not take photo of marker. The sun is thought to have disturbed photography of the pattern of the marker by the bottom camera of the MAV. Therefore, cameras with a larger dynamic range are required. A CMOS imaging array with wide dynamic range pixels was proposed [23]. It is necessary to examine the control method of the wide dynamic range camera for the MAVs.

In the MAV's balancing, the MAV was attached self-produced odor sensor module and adjusted the balance by battery position. However, it was difficult to achieve a balance it as if the speed was different depending on the direction of movement. Therefore, improvement in exercise performance of MAVs is required.

In the odor sensing, the size of the arena was 7 × 12 m. In the actual system, the size will be much larger (e.g. 100,000 square meters in sports stadium). Therefore, ultra-sensitive odor sensors will be required. Gas sensors with parts-per-quadrillion (ppq) sensitivity have been reported [3].

In the swarm control, in large covered areas without any omission and robust exploration, swarm flight will be required.

In the future, we will improve the location accuracy of the odor source localization using person recognition and increase the amount of information collected to enable MAVs to search for suspects using their built-in cameras [18].

6. Conclusion

Localization of odor source using a MAV with vision and gas sensing was achieved. The MAV used vision and odor localization with ingenuity and succeeded in localizing odor sources like living things can. Our approach does not even require GPS-based control. The location accuracy of odor source localization based on a bio-inspired localization method was 63 cm, enabling individual identification. The MAV was controlled with highly autonomously. An experimenter just blows a whistle, and the MAV starts to hover, to seek odor sources, and to keep hovering during flight near the odor source. The MAV is light-weight and does not cause harm to others even if it accidentally falls. These localization experiments were the first step to a proof of concept for a danger warning system. These results will be applied to the system to enable a safer and more secure society.

References

[1] D. B. Walker, J. C. Walker, P. J. Cavnar, J. L. Taylor, D. H. Pickel, S. B. Hall, and J. C. Suarez, "Naturalistic quantification of canine olfactory sensitivity," Appl Anim Behav Sci, vol. 97, pp. 241–254 (2006)

[2] A. Lichtenstein, E. Havivi, R. Shacham, E. Hahamy, R. Leibovich, A. Pevzner, V. Krivitsky, G. Davivi, I. Presman, R. Elnathan, Y. Engel, E. Flaxer, and F. Patolsky, "Supersensitive fingerprinting of explosives by chemically modified nanosensors arrays," Nature Communications, vol. 5, Article number 4195 (2014)

[3] R. Vishinkin and H. Haick, "Artificially Intelligent Nanoarrays for Disease Detection via Volatolomics," IEEE International Electron Devices Meeting (IEDM 2015) 33.4 (2015)

[4] K. Kurotsuchi, M. Tai, and H. Takahashi, "Vision-based Autonomous Micro-Air-Vehicle Control for Odor Source Localization," International conference on mechatronics and machine vision in practice (M2VIP 2016)

[5] P. P. Neumann, V. H. Bennetts, A. J. Lilienthal, M. Bartholmai, and J. H. Schiller, "Gas source localization with a micro-drone using bio-inspired and particle filter-based algorithms," Advanced Robotics, vol. 27, No. 9, pp. 725–738 (2013)

[6] AT Hayes, A Martinoli, and RM Goodman, "Distributed Odor Source Localization," IEEE Sensors Journal, vol. 2, pp. 260–271 (2002)

[7] H. Jiu, J. Li, S. Pang, and B. Han, "Odor plume source localization with a Pioneer 3 Mobile Robot in an indoor airflow environment," IEEE Region 3 South East Conference (SoutheastCon 2014) (2014)

[8] G. Ferri, M. V. Jakuba, A. Mondini, V. Mattoli, B. Mazzolai, D. R.Yoerger, and P. Dario, "Mapping multiple gas/odor sources in an uncontrolled indoor environment using a Bayesian occupancy grid mapping based method," Robotics and Autonomous Systems, vol. 59.11, pp. 988–1000 (2011)

[9] A. Dhariwal, G. S. Sukhatme, and A. A. G. Requicha "Bacterium-inspired robots for environmental monitoring," Proceedings of IEEE International Conference on Robotics and Automation (ICRA '04), vol. 2, pp. 1436–1443 (2004)

[10] H. Ishida, H. Hayashi, M. Takakusaki, T. Nakamoto, T. Moriizumi, and R. Kanzaki, "Odour-source localization system mimicking behavior of silkworm moth," Sensors and Actuators A, vol. 51, pp. 225–230 (1996)

[11] A. Liberzon, R. Gurka, A. Harari, and G. Zilman, "Robust moth olfactory search algorithm based on dispersion of pheromone puffs," arXiv:1310,1411 (2013)

[12] Y. Zou and D. Luo, "A modified ant colony algorithm used for multi-robot odor source localization," Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence 5227, pp. 502–509 (2008)

[13] R. Konomura and K. Hori, "Phenox: Zynq 7000 based quadcopter robot," in 2014 International Conference on ReConFigurable Computing and FPGAs (ReConFig14) (2014)

[14] Tetsuya J. Kobayashi, "Implementation of dynamic Bayesian decision making by intracellular kinetics," Physical Review Letters 104, 228104 (2010)

[15] J. Murlis, J. S. Elkinton, and R. T. Carde, "Odor plumes and how insects use them," Annual review of entomology, vol. 37, pp. 505–532 (1992)

[16] PHENOX LAB: Phenox, http://phenoxlab.com/?page_id=601

[17] Edward Hall, "The hidden dimension," Anchor Books (1966)

[18] K. Al-Mutib, M. Emaduddin, M. Alsulaiman, H. Ramdane, and E. Mattar, "Motion Periodicity based Pedestrian Detection and Particle Filter based Pedestrian Tracking using Stereo Vision Camera," in Mechatronics and Machine Vision in Practice (M2VIP), 19th International Conference, pp. 32–37 (2012)

[19] R. B. Bourret and A. M. Stock, "Molecular information processing: lessons from bacterial chemotaxis," Journal of Biological Chemistry 277.12, pp. 9625–9628 (2002)

[20] G. Klein and D. Murray, "Parallel Tracking and Mapping for Small AR Workspaces," International Symposium on Mixed and Augmented Reality (ISMAR'07, Nara) (2007)

[21] E. Rosten, R. Porter, and T. Drummond, "Faster and better: A machine learning approach to corner detection," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 32:105119 (2010)

[22] H. Chebi, D. Acheli, and M. Kesraoui, "Dynamic detection of abnormalities in video analysis of crowd behavior with DBSCAN and neural networks," Advances in Science, Technology and Engineering Systems Journal (ASTESJ), vol. 1, No. 5, pp. 56-63 (2016)

[23] S. Decker, R. D. McGrath, K. Brehmer, and C. G. Sodini, "A 256/spl times/256 CMOS imaging array with wide dynamic range pixels and column-parallel digital output", IEEE Solid-State Circuits (JSSC), vol. 33, Issue 12 (1998)

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1159-1164 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

EMI Shielding Performance For Varies Frequency by Metal Plating on Mold Compound

Min Fee Tai[*1, 2], Swee Leong Kok[2], Kenichiroh Mukai[1], Stephan Hotz[3], Patrick Brooks[3], Diego Reinoso Cocina[3]

[1]*Atotech (M) Sdn Bhd, Malaysia*

[2]*Faculty of Electronic and Computer Engineering, Universiti Teknikal Malaysia Melaka*

[3]*Atotech Deutschland GmbH Erasmusstraße 20, 10553 Berlin, Germany*

ARTICLE INFO

Article history:
Received: 05 March, 2017
Accepted: 03 May, 2017
Online: 16 July, 2017

Keywords:
EMI shielding
Mold compound
Conformal plating

ABSTRACT

Conformal metalization on mold compound offers new possibility for IC package design to improve features such as rigidization of the flexible core, heat sink capability, 3D-circuit patterning and the electromagnetic interference (EMI) shielding. With the unique processes, the fabrication technology had enabled to achieve the high reliable performance and had passed the electrical test. Following research after the reliability concern, this paper further study the shielding effectiveness of varying coating thickness with respect to laboratory simulated EMI condition, using radio frequency from 10MHz to 5.8 GHz. Different metal namely pure nickel, nickel-phosphorous and pure plated copper are studied for their effectiveness of EMI sheilding. Our first result showed over 35-40dB of shielding effectiveness is achievable on high frequency 868-5800MHz. Nevertheless on low frequency of 10MHz, the shielding effectiveness achievement is below than 25dB. To overcome the shielding need for lower frequency, we further expanded our test by choosing ferromagentic material Nicke/Ironl-alloy in combination with thick copper plating. With this new metal combination, EMI shielding effectiveness for lower frequency is improved to 40dB.

1. Introduction

This paper is an extension of work originally presented in conference name electromagnetic interference (EMI) Shielding Performance by Metal Plating on Mold Compound [1]. EMI shields work as isolation of components, minimizes crosstalk, and reduce EMI susceptibility without impacting system speed. Electromagnetic shielding is the practice of reducing the electromagnetic field in a space by blocking the field with barriers made of conductive or magnetic materials. To achieve good EMI shielding on integrated circuit (IC) assembly, manufacturer is switching from metal capping to other methods such as conductive painting, sputtering or chemical vapor deposition (CVD) on top of mold compound with the intention to miniaturize electronic device. Printing is an easy and fast method, however to control the thickness of the shielding layer within 50-150um is challenging; on the other hand on thin film solution , sputtering or CVD can be a better option.

Plating was not a preferred process in the past due to its difficulty in achieving good adhesion between the plated metal over the surface of a mold compound. Adhesion was greatly improved recently with the introduction of adhesion promoter developed by Atotech, which have achieved a peel strength up to 4-5N/cm complied with industrial standard of moisture sensitivivity level 1 (MSL1) and thermal cycling 1000 (TC1000) [2]. Plating approach offers lower cost solution, ease of thickness control on thin layer (1-10um), and more important, provides uniform coating coverage on all the surfaces of a mold compound with a thickness variation of <20% regardless of the shape, even on structure of high aspect ratio (Figure 1).

The inevitable trends for the circuits inside the modules are fast edge rates and high clock frequencies, which may cause significant electromagnetic interference (EMI) and radio frequency interference (RFI) problems. Recently mobile phone manufacturers have started in applying EMI shielding technology to a variety of digital chips, radio frequency (RF), connectivity (wireless LAN, Bluetooth) chips and others that also include AP (application processor) and modems. Although the use of this

* Corresponding Author: Min Fee Tai, Atotech (M) Sdn Bhd, Malaysia
Email: mftai@yahoo.com

technology on printed circuit boards (PCB) and connectors is not novel, however it is the first time whereby the application of plating technology is extended to most of the major chips [3]. The decision to use this technology stems from interference in wireless communication. The use of that fabrication technology for others components such as those with sensitive clock signals of digital chips have also increasing recently. EMI shielding by metal plating is an important measure to improve device performance whilst decreasing electromagnetic interference to major chips and to relieve consumer fear of electromagnetic waves from smartphones. As EMI is an increasingly important source of unreliable functionality of all electronic technologies, EMI is rapidly becoming a major functional safety issue. Malfunctions caused by EMI in safety-related electronics can lead to serious safety risks for people or the environment in general as well as financial risks for manufacturers and service providers [4].

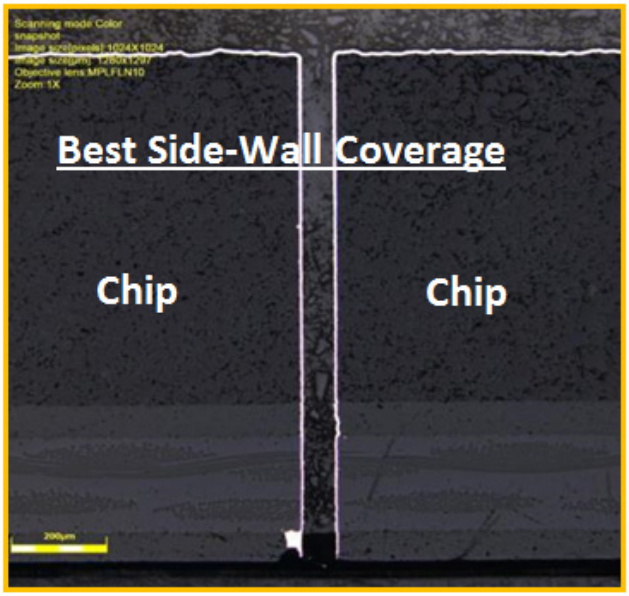


Figure 1: Excellent Throwing Power of Coating by Plating Technology.

Figure 2 shows the process flow of plating on mold compound which is proposed by Atotech's standard process. Theoretically, shielding effectiveness of a material is contributed from three main loss mechanisms: reflection, absorption, and multi-reflection. When an incident wave travels to the shielding material, it is partially reflected with a portion of the wave penetrating into the material. The penetrated wave can propagate inside the material and get attenuated due to the absorption of the material. Once the remaining transmitted wave arrives at the other surface of the shielding material, the second reflection occurs and the wave gets bounced back, resulting in multiple reflections between the two surfaces of the shielding material.

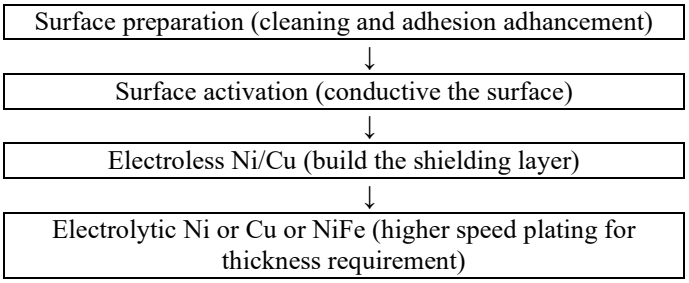


Figure 2: Process Flow of Plating on Mold Compound Shielding Effectiveness

Due to the fact that the wave is attenuated to $1/e$ at one skin depth, the multi-reflection loss can be ignored for the cases where the shielding thickness is much larger than the skin depth. In summary, the shielding effectiveness in dB, SE_{dB} is expressed in equation 1,

$$SE_{dB} = R_{dB} + A_{dB} + M_{dB}, \qquad (1)$$

where R_{dB} is the reflection loss, A_{dB} is the absorption loss, and M_{dB} is the multi-reflection loss. Equation is further explained by Figure 3 [5].

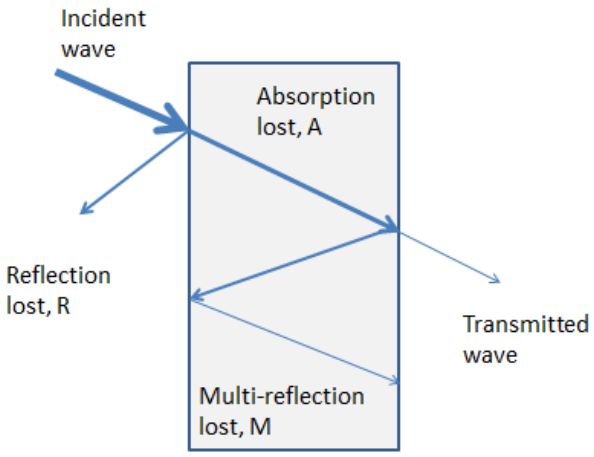


Figure 3: Electromagnetic Shielding Schematic Diagram

The reflection loss is

$$R = 20log_{10}\left|\frac{(Z_0+Z_1)^2}{4Z_0Z_1}\right| \qquad (2)$$

where Z_0 represents the wave impedance of air and Z_1 represents the wave impedance of the metal sheet. As

$$Z_1 \approx \sqrt{\frac{\mu_1\omega}{\sigma_1}}(1+j) \ll Z_0 = 120\pi\omega \qquad (3)$$

$\mu_1$and σ_1 are the relative permeability and conductivity of the metal sheet, respectively, formula (2) can be estimated as

$$R = 20log_{10}\left|\frac{Z_0}{4Z_1}\right| = 20log_{10}\left|\frac{30\pi}{1+j}\right|\sqrt{\frac{\sigma_1}{\mu_1\omega}} \qquad (4)$$

The reflection loss is directly proportional to the conductivity of the metal sheet but inversely proportional to the relative permeability of the metal sheet.

The absorption loss is

$$A = 20log_{10}e^{\frac{L}{\delta}} \qquad (5)$$

where L is the thickness of the metal sheet and

$$\delta = \sqrt{\frac{2}{\mu_1\omega\sigma_1}} \qquad (6)$$

is the skin depth. Formula (5) can be estimated as

$$A = 20log_{10}e^{\sqrt{\frac{\mu_1\omega\sigma_1}{2}}L} \qquad (7)$$

The absorption loss is directly proportional to the thickness, conductivity and relative permeability of the metal sheet.

The multi-reflection loss is

$$B = 20log_{10}\left|1 - \frac{(Z_0 - Z_1)^2}{(Z_0 + Z_1)^2} e^{-2(1+j)\frac{L}{\delta}}\right| \quad (8)$$

As Z>>Z1, formula (8) can be estimated as

$$B = 20log_{10}\left|1 - e^{-2(1+j)\frac{L}{\delta}}\right| \quad (9)$$

If the thickness of the metal sheet is much larger than skin depth, incident wave will be attenuated significantly so that the multi-reflection loss will be negligible [6].

The EM shielding capability of a shield is escribed as shielding effectiveness, *SE*, which is defined as the ratio of the electric fields and magnetic field at a point without and with the shield as equation 10a and 10b,

$$SE = 20log_{10}\frac{E_1}{E_0} \quad (10a)$$

$$SE = 20log_{10}\frac{H_1}{H_0} \quad (10b)$$

where E_1, H_1 is the electrical and magnetic field at an arbitrary point without the shield, and E_2, H_2 is the electrical and magnetic field at the same point with the shield [7].

The shielding effectiveness value depends on the shielding material, its thickness, size of the shielded volume, grounding structures, and the frequency of interest.

2. Preparation of Test

In order to investigate the performance of metal plating on EMI sheilding, test matrixes are prepared with electrolytic plated nickel (pure nickel), electroless plated nickel (nickel-phophorous, P at 7-10%), and pure plated copper as the shielding layer on top of mold compound. For measurement of the electromagnectic field, a copper circuit is designed with substrate material Isola IS410, whereby the spiral flow is made on top of dielectric according to the specifications of ASTM D3123 (Figure 4).

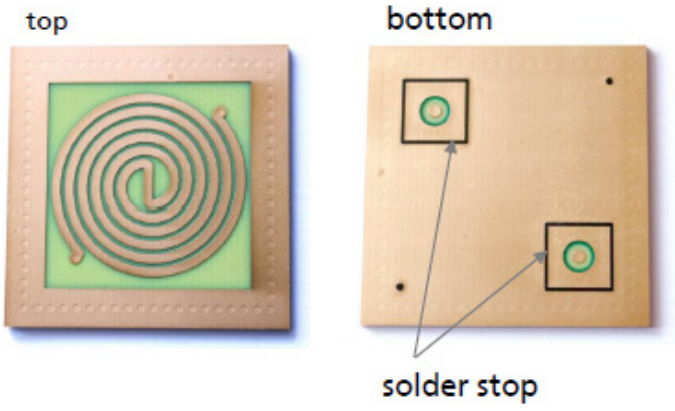
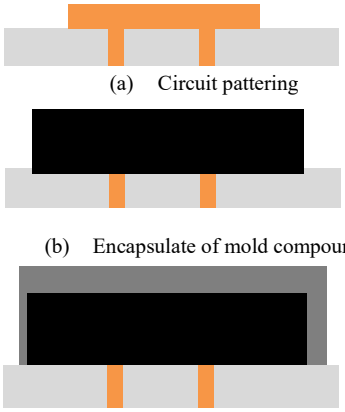


Figure 4: Spiral Circuit formed for near field EMI shielding test

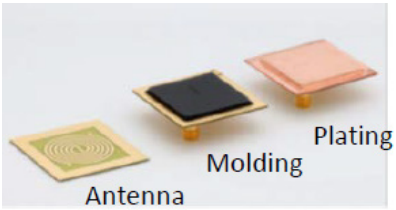
Circuit is then encapsulated by mold compound Sumitomo EME_G770H Type F, with the leg connected to power input is located in opposite side of the dielectric. After that, the sample will go through the process as illustrated in Figure 5.



(a) Circuit pattering

(b) Encapsulate of mold compound

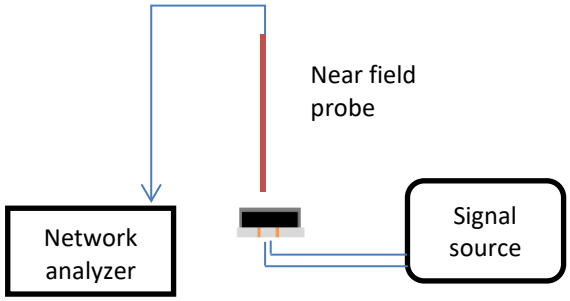
(c) Plating on mold compound



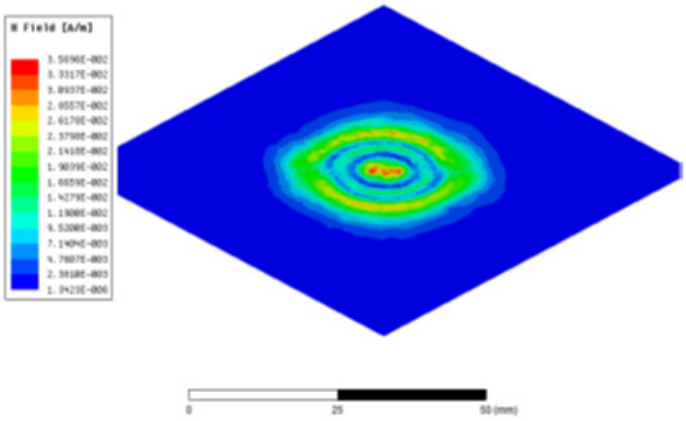
(d) Overall actual sample

Figure 5: Preparation of test coupon.

From signal source, frequencies ranging from low to high frequency are loaded into the test coupon, magnetic field generated by spiral flow circuit is measured using the near field probe, and recorded in network analyzer. The setup of the experiment is shown in Figure 6.



(a) Setup of EMI shielding measurement



(b) Generated Field Distribution

Figure 6: Shielding Effectiveness Measurement Setup

3. Experimental Results and Analysis

A series of samples with increasing thickness of plated Cu from lower to higher side, as indicated on Table 1 are fabricated and tested.

Table 1: Metal Thickness DOE

Material	Thickness, um
Electroless Cu	1
Electrolytic Cu	5, 10, 15

All data recorded by the plated samples with different thickness and materials is compared versus a control sample (sample without any plating on mold compound). The ratio of 2 readings is used for calculation of shielding effectiveness. Formula 10b is applied on these calculations.

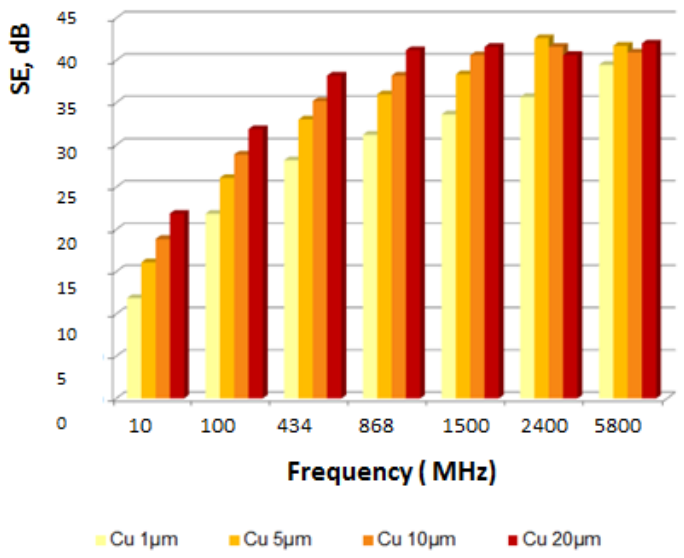


Figure 7: Shielding Effectiveness on Copper with various frequency test

On data analysis of metal shielding by copper (Figure 7), 4 samples are prepared, which are 1 um Cu sample plated by electroless Cu, 5um, 10um, 20um, samples are plated by electrolytic Cu. Higher electroless Cu thickness (> 1um) s not prepared considering its high running cost and high leadtime during mass production. Signal input with frequency 10MHz, 100 MHz, 434 MHz, 868 MHz, 1.5 GHz, 2,4 GHz, and 5.8 GHz are tested. The experimental results show that shielding performance is improved by increasing the metal thickness, especially in lower frequency (< 1 GHz). At higher frequency, 5.8 GHz, thickness of Cu shows little impact on EMI shielding effectiveness. Hence further increasing is not helping to improve shielding effectiveness (*SE*) at high frequency range.

Similar test matrix is applied with nickel, shows in Table 2. Samples with electrolytic Ni 1 um, 3 um, and 5 um; as well as electroless NiP 5um have been evaluated (5um NiP is reasonable in cost and can be plated within 25 min). Same signal input with frequency 10MHz, 100 MHz, 434 MHz, 868 MHz, 1.5 GHz, 2,4 GHz, and 5.8 GHz are applied to allowed the direct comparison. NiP is selected as matrix of study due to it is electroless process which required non conductive point. Besides, NiP with physical property of high hardness and good corrosion resistance make it suitable to be used as final surface.

Table 2: Metal Thickness DOE

Material	Thickness, um
Electroless NiP	5
Electrolytic Ni	1, 5, 10

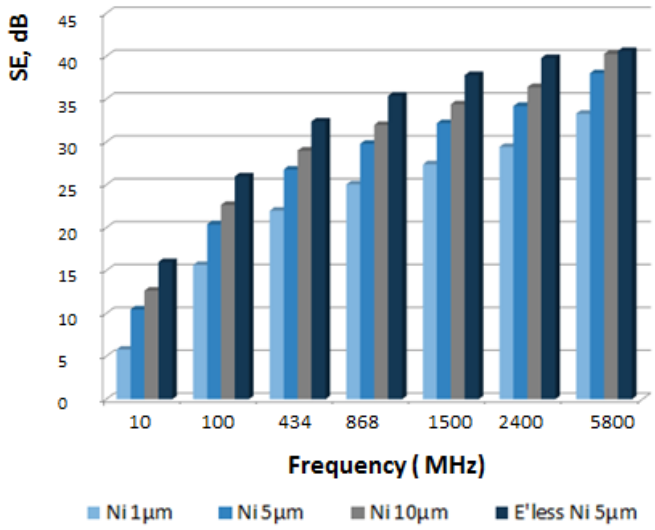


Figure 8: Shielding Effectiveness on Nickel with various frequency test

In gerenal, result from Figure 8 shows a more consistent trend for shielding effectiveness improvement over the thickness increase. Nickel alloy with phosporuos content is showing relatively better perfomance on *SE* versus pure nickel.

Comparing 1um and 5um of pure copper vs pure nickel at high frequency 5800 MHz, we see copper has higher SE (40dB) compare to nickel. As from equation 7, the absorption loss is directly proportional to the thickness, conductivity and relative permeability of the metal sheet. Electromagnetic radiation from the high frequency penetrates only at the near surface region of conductor layer. From equation 6, the skin depth, δ, decreases with increased frequency and increase conductivity and permeability. For copper, when relative conductivity, $\mu_r = 1$, and $\sigma = 5.8 \times 10^7\ \Omega^{-1}m^{-1}$, $\delta = 2.43 \times 10^{-9}$ at 5800 MHz; and for nickel, when relative conductivity, $\mu_r = 100$, and $\sigma = 1.15 \times 10^7\ \Omega^{-1}m^{-1}$, $\delta = 0.547 \times 10^{-9}$ at 5800 MH.

Table 3: Electrical conductivity relative to copper (σ_r) and relative magnectic permeability (μ_r) of selected material. [8]

Material	σ_r	μ_r	$\sigma_r\mu_r$	σ_r/μ_r
Silver	1.05	1	1.05	1.05
Copper	1	1	1	1
Gold	0.7	1	0.7	0.7
Aluminum	0.61	1	0.61	0.61
Brass	0.26	1	0.26	0.26
Bronze	0.18	1	0.18	0.18
Tin	0.15	1	0.15	0.15
Lead	0.08	1	0.08	0.08
Nickel	0.2	100	20	2×10^{-3}
Stainless steel (430)	0.02	500	10	4×10^{-5}
Mumetal (at 1 kHz)	0.03	20,000	600	1.5×10^{-6}
Superpermalloy (at 1 kHz)	0.03	100,000	3,000	3×10^{-7}

Lower frequency shows a more challenging task to shield. In copper material, a maximum 40 dB can be obtain at frequency 5.8GHz at 5um thickness, but for 10mHz, the best shielding performance is only 22 dB at 20um thickness. By taking 5 um as benchmark to compare the material performance, Ni or NiP and Cu show close SE result at 5.8 GHz; but in lower frequency mode (10MHz) Cu and NiP are performing better than pure Ni; which Cu and NiP SE is 15 dB, versus pure Ni SE at 10 dB.

Our target to fulfill industrial requirement is to achieve minimum 30dB SE for varies high and low requency applications. Higher frequency interference is normally shielded by a thin conductive metallization layer. Simple electrically conductive layers however are transparent to low frequency magnetic fields, as shown in Figure 7. This low frequency magnetic interference can be emitted from sources such as switches, motors, power supplies, and transformers and is typically a challenging EMI shielding problem. If low frequency magnetic shielding is required ,it is normally achieved through the use of special ferromagnetic metal alloys with high magnetic permeability (see Figure 9). In order to further improve the SE on lower frequency signal device (such as radio), additional Ni/Fe alloy (Ni75/Fe25 in weigh percentage) is plated on top of Cu. Ni/Fe alloy with high ferromagnetic property can improve electromagentic shielding in low frequency mode.

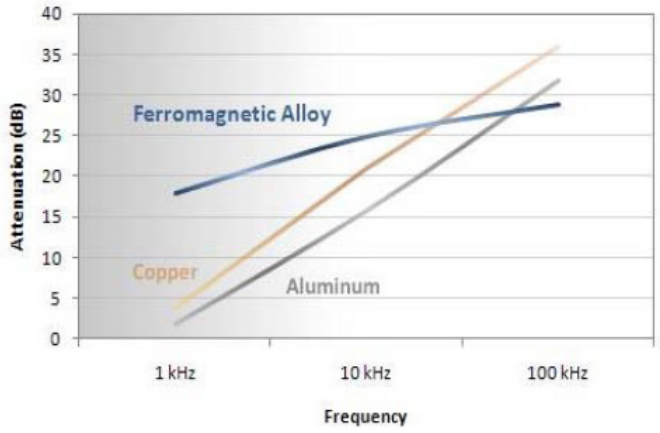


Figure 9 - Magnetic shielding ability at different frequencies [8]

Materials with high magnetic permeability protect sensitive devices from the electrical noise caused by magnetic fields by redirecting the magnetic field through the shielding material and away from the protected device. In the same way, the devices that emits low frequency magnetic fields can also be isolated through the use of materials with high magnetic permeability. [9]

Table 4: Additional Metal Thickness DOE

Matrix	Metal combination, um
1	1um electroless Cu seed + 20um Electrolytic Cu + 4um Ni/Fe
2	1um electroless Cu seed + 10um Electrolytic Cu + 4um Ni/Fe
3	1um electroless Ni seed + 10um Electrolytic Cu + 4um Ni/Fe
4	1um electroless Ni seed + 20um Electrolytic Cu + 4um Ni/Fe
5	1um electroless Ni seed + 20um Ni/Fe

New results showed that plated Cu in combination with ferromagnetic Ni/Fe gives significant improvement on shielding over low frequency signal (Figure 10). Combination of 1um electroless Cu seeding, plus 20um electrolytic Cu and with final finishing layer 4um of Ni/Fe could achieve 40dB of SE over the 10MHz wave. From the seeding method selection between electroless Cu and Electroless Ni, we notice electroless Cu gives slight better SE performance compare to electroless Ni, with around 3-4dB extra of SE. Plating 20of Cu give extra 6-7 dB SE compare to 10um Cu. Plating 20um Ni/Fe alone will give only 20dB of SE.

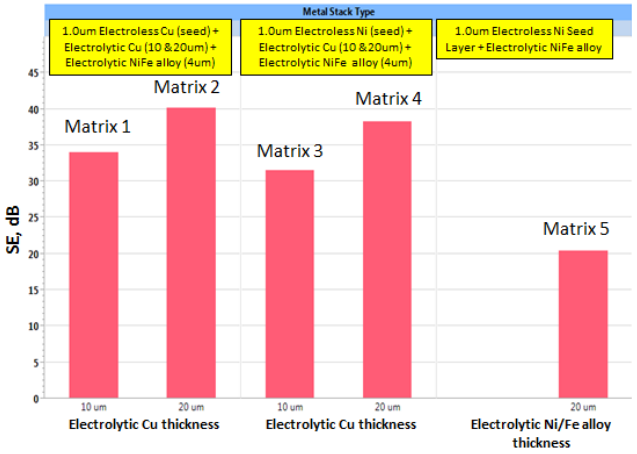


Figure 10: Shielding Effectiveness of Cu Plating combined with Ni/Fe alloy plating on Lower Frequency 10MHz

4. Conclusions

Plating on mold compound provides alternative option on EMI shielding with the proven reliability performance. The solution show features such as thin film coating capability, high throwing power in multiple dimension and high cost advantage compare to sputtering process. Selection of material and metal coating thickness are factors of EMI shielding effectiveness. Experiments show at high frequency of 5.8 GHz, shielding effectiveness as high of 35-40dB can be obtained by plating of

elctrolytic Ni, electrolytic Cu or electroless NiP. Meanswhile, for lower frequency signal down to 10MHz, 35- 40dB SE can be achieved by electrolytic Cu plating and capped by 4um Ni/Fe alloy. With the correct combination on metals and thicknesses, conformal metal plating performs a good solution meeting EMI shielding's requirement over high and low frequency.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The authors would like to acknowledge the support of this work by Atotech USA, Atotech Berlin and Universiti Teknikal Malaysia Melaka (UTeM) in sharing the information, facilities and materials

References

[1] EMI Shielding Performance by Metal Plating On Mold Compound, 2016 IEEE 37th International Electronics Manufacturing Technology (IEMT) & 18th Electronics Materials and Packaging (EMAP) Conference

[2] ICSE 2014, [Adhesion Enhancement for Electroless Plating on Mold Compound for EMI Shielding with Industrial Test Compliance]

[3] Etnews, Korea IT news, Feb16, 2016, http://english.etnews.com/20160216200001

[4] Yi He, Jun Li, Gengxin Tian, Fengman Liu, and Liqiang Cao, IEEE Transactions On Electromagnetic Compatibility, Vol. 58, No. 2, April 2016 [Study on a Conformal Shielding Structure With Conductive Adhesive Coated on Molding Compound in 3-D Packages]

[5] Liao Kuo-Hsien, Alex, Chi-Hong, Chan, Shen Chia Hsien, Lin, I-Chia; Huang Hsin Wen, CPMT Symposium Japan, 2012 2nd IEEE [Novel EMI Shielding Methodology on Highly Integration SiP Module]

[6] Fankun Jiang, Ming Li, Liming Gao, 2014 IEEE, Research on conformal EMI shielding Cu/Ni layers on package, School of Materials Science and Engineering Shanghai Jiao Tong University.

[7] Nozad Karim, Jingkun Mao, Jun Fan 2010 Asia-Pacific International Symposium on Electromagnetic Compatibility, April 12 - 16, 2010, Beijing, China [Improving Electromagnetic Compatibility Performance of Packages and SiP Modules Using a Conformal Shielding Solution]

[8] DDL Chung, ASM International, 5 Jan2000 [Material for Electromagnetic Interference Shielding]

[9] Rich Emrich, September 21, 2010, Integran Technologies Inc., Toronto, ON [Low Frequency Magnetic Shielding – An Integrated Solution]

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 2, No. 3, 1165-1171 (2017)
www.astesj.com
Special Issue on Recent Advances in Engineering Systems

ASTES Journal
ISSN: 2415-6698

Comparative Study of Adaptive Consensus Control of Euler-Lagrange Systems on Directed Network Graph

Yoshihiko Miyasato*

The Institute of Statistical Mathematics, Tachikawa, Tokyo, 190-8562, Japan

ARTICLEINFO

Article history:
Received: 28 April, 2017
Accepted: 28 June, 2017
Online: 20 July, 2017

Keywords :
Adaptive Control
Multi-Agent System
H_∞ Control

ABSTRACT

A comparative study between adaptive consensus control of multi-agent systems composed of fully actuated mobile robots which are described as a class of Euler-Lagrange systems on directed network graphs based on the notion of inverse optimal H_∞ control criterion (Controller I), and the similar control strategy without H_∞ control criterion (Controller II), is given in this paper. Controller I is deduced as a solution of certain H_∞ control problem, where estimation errors of tuning parameters are considered as external disturbances to the process. Asymptotic properties, stability and robustness to unknown parameters are discussed for Controller I and Controller II.

1 Introduction

Motivated by crowd behaviors of animals, birds and fish, cooperative control problems of multi-agent systems are active research fields, and plenty of control strategies have been proposed in those areas, such as formation control, task assignment, traffic control, and scheduling *et al.* (for example, [1]-[11]). Among those, distributed consensus tracking of multi-agent systems with restricted communication networks, has been a basic and important issue, and various research results have been developed for various processes and under various conditions such as [1], [12]-[21]. In those works, adaptive control methodologies were also investigated in order to deal with uncertainties of agents, and stability of control systems was assured via Lyapunov function analysis. Furthermore, robustness properties of the control schemes were also examined. However, so much attention does not have been paid on control performance such as optimal property or transient performance in those research works, and especially, the case of multi-agent systems composed of processes with unknown and different system parameters on directed information network graphs, does not have been investigated in detail in the previous works.

The purpose of the paper is to provide a comparative study of two types of adaptive consensus control schemes (Controller I and Controller II) of multi-agent systems composed of fully actuated mobile robots which are described as a class of Euler-Lagrange systems [22] on directed network graphs. Controller I is constructed based on the notion of inverse optimal H_∞ control criterion [23], [24], and Controller II is synthesized without H_∞ control criterion. Controller I is derived as a solutios of certain H_∞ control problem, where estimation errors of tuning parameters are considered as external disturbances to the process. Asymptotic properties, stability and robustness to unknown parameters are discussed for those two control strategies (Controller I and Controller II).

The proposal of Controller I is an extension of our previous work [25], [26], in which the first-order or second-order linear or nonlinear regression models on directed network graphs were considered. This manuscript is also an extended version of the conference paper [1] (proposal of Controller I), and especially the present paper adds a comparative study of the two types of adaptive consensus control strategies (proposal and comparison of Controller I and Controller II), and highlights robustness properties of Controller I.

2 Multi-Agent Systems and Network Graphs

Mathematical preliminaries on information network graph of multi-agent systems are summarized [17], [20], [21]. As a model of interaction among agents, a weighted directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is considered,

*Corresponding Author : Yoshihiko Miyasato, Email: miyasato@ism.ac.jp

https://dx.doi.org/10.25046/aj0203147

where $\mathcal{V} = \{1, \cdots, N\}$ is a node set which corresponds to a set of agents, and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is an edge set. An edge $(i, j) \in \mathcal{E}$ means that agent j can obtain information from i, but not necessarily vice versa. In the edge (i, j), i is called as a parent node and j is called as a child node, and the in-degree of the node i is the number of its parents, and the out-degree of i is the number of its children. An agent which has no parent (or with the in-degree 0), is denoted as a root. A directed path is a sequence of edges in the form (i_1, i_2), (i_2, i_3), $\cdots$ $(\in \mathcal{E})$, where $i_j \in \mathcal{V}$. The directed graph is called strongly connected, if there is always a directed path between every pair of distinct nodes. A directed tree is a directed graph where every node has exactly one parent except for a unique root, and the root has directed paths to all other node. An directed spanning tree $\mathcal{G}_S = (\mathcal{V}_S, \mathcal{E}_S)$ of the directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is a subgraph of $\mathcal{G}$ such that $\mathcal{G}_S$ is a directed tree and $\mathcal{V}_S = \mathcal{V}$.

Associated with the edge set $\mathcal{E}$, a weighted adjacency matrix $A = [a_{ij}] \in \mathbf{R}^{N \times N}$ is introduced, and the entry a_{ij} of it is defined by

$$a_{ij} = \begin{cases} > 0 & : \ (j, i) \in \mathcal{E}, \\ 0 & : \ \text{otherwise.} \end{cases}$$

For the adjacency matrix $A = [a_{ij}]$, the Laplacian matrix $L = [l_{ij}] \in \mathbf{R}^{N \times N}$ is defined such as

$$l_{ii} = \sum_{\substack{j=1 \\ j \neq i}}^{N} a_{ij},$$
$$l_{ij} = -a_{ij}, \ (i \neq j).$$

Laplacian matrix has at least one zero eigenvalue and all nonzero eigenvalues have positive real parts. Especially, the Laplacian matrix has a simple 0 eigenvalue with the associated eigenvector $\mathbf{1} = [1 \cdots 1]^T$, and all other eigenvalues have positive real parts, if and only if the corresponding directed graph has a directed spanning tree.

In this manuscript, a consensus control problem of leader-follower type is considered, where y_0 is a leader which each agent $i \in \mathcal{V}$ should follow (i is called a follower). Concerned with the leader, a_{i0} is defined such as

$$a_{i0} = \begin{cases} > 0 & : \ \text{leader's information is available} \\ & \quad \text{to follower } i, \\ 0 & : \ \text{otherwise,} \end{cases} \tag{1}$$

and from a_{i0} and L, the matrix $M \in \mathbf{R}^{N \times N}$ is defined by

$$M = L + \mathrm{diag}\,(a_{10} \cdots a_{N0}). \tag{2}$$

It is shown that $-M$ is a Hurwitz matrix, if and only if 1. at least one a_{i0} $(1 \leq i \leq N)$ is positive, and 2. the graph $\mathcal{G}$ has a directed spanning tree with the root $i = 0$.

Hereafter, we assume that

1. The graph has a directed spanning tree with the root $i = 0$.
2. At least one a_{i0} $(1 \leq i \leq N)$ is positive, that is, the information of the leader y_0 ($\dot{y}_0$, $\ddot{y}_0$), is available to at least one follower i.
3. For the leader, y_0, $\dot{y}_0$, $\ddot{y}_0$ are uniformly bounded.

In the following, two adjacency matrices $A = [a_{ij}]$, $C = [c_{ij}] \in \mathbf{R}^{N \times N}$ are introduced for a directed graph $\mathcal{G}$, and the corresponding matrices are denoted as L_a, L_c (Laplacian matrices), and M_a, M_c, respectively.

3 Problem Statement

A multi-agent system composed of N fully actuated mobile robots which are described as a class of Euler-Lagrange systems [9], [10], are considered such that

$$M_i(y_i)\ddot{y}_i + C_i(y_i, \dot{y}_i)\dot{y}_i = \tau_i, \quad (i = 1, \cdots, N), \tag{3}$$

where $y_i \in \mathbf{R}^n$ is an output (a generalized coordinate), $\tau_i \in \mathbf{R}^n$ is a control input (a force vector), $M_i(y_i) \in \mathbf{R}^{n \times n}$ is an inertia matrix, and $C_i(y_i, \dot{y}_i) \in \mathbf{R}^{n \times n}$ is a matrix of Coriolis and centripetal forces. Each component has the next properties as a Euler-Lagrange system.

Properties of Euler-Lagrange Systems [22]

1. $M_i(y_i)$ is a bounded, positive definite, and symmetric matrix.
2. $\dot{M}_i(y_i) - 2C_i(y_i, \dot{y}_i)$ is a skew symmetric matrix.
3. The left-hand side of (3) can be written into

$$M_i(y_i)a_i + C(y_i, \dot{y}_i)b_i = Y_i(y, \dot{y}_i, a_i, b_i)\theta_i, \tag{4}$$

where $Y_i(y_i, \dot{y}_i, a_i, b_i)$ is a known function of y_i, $\dot{y}_i$, a_i, b_i (a regressor matrix), and θ_i is an unknown system parameter vector.

The control objective is to achieve consensus tracking of the leader-follower type for the unknown multi-agent system (unknown θ_i) such as

$$y_i \to y_j, \ (i, j = 1, \cdots, N), \tag{5}$$
$$y_i \to y_0, \ (i = 1, \cdots, N), \tag{6}$$

under the limited communication structure $\mathcal{G}$ among agents.

Remark 1 More generalized Euler-Lagrange systems including damping terms and gravitational forces, can be also considered in the present framework. However, for simplicity of notations, the description (3) is to be adopted hereafter, since the generalized Euler-Lagrange systems are written in the similar form to (4).

4 Control Law and Error Equation

As the first step of the design of Controller I and Controller II, an estimator of $\dot{y}_0$ (the leader's information) is developed via available data from the follower i. A similar estimation procedure was presented in [16].

$$\dot{\hat{z}}_i(t) = -\beta \sum_{\substack{j=1 \\ j\neq i}}^{N} c_{ij}\{\hat{z}_i(t) - \hat{z}_j(t)\} - \beta c_{i0}\{\hat{z}_i(t) - \dot{y}_0(t)\} + n_{i0}\ddot{y}_0(t), \tag{7}$$

where $\hat{z}_i$ is a current estimate of $\dot{y}_0$, and is synthesized from the data available to the follower i. $c_{ij}(1 \le i \le N, 0 \le j \le N)$ is defined as the entry of the adjacency matrix C and (1) deduced from the directed graph $\mathcal{G}$, and $\beta > 0$ is a design parameter. Associated with c_{i0}, n_{i0} is defined such as

$$n_{i0} = \begin{cases} 1 & : \quad c_{i0} > 0, \\ 0 & : \quad \text{otherwise.} \end{cases} \tag{8}$$

By employing the estimate $\hat{z}_i$, the control scheme is constructed as follows:

$$\dot{y}_{ri}(t) = \hat{z}_i(t) - \alpha \sum_{\substack{j=0 \\ j\neq i}}^{N} a_{ij}\{y_i(t) - y_j(t)\}, \tag{9}$$

$$s_i(t) = \dot{y}_i(t) - \dot{y}_{ri}(t), \tag{10}$$

$$\tau_i(t) = Y_i(t)\hat{\theta}_i(t) + v_i(t), \tag{11}$$

$$Y_i(t) \equiv Y_i(y, \dot{y}_i, \ddot{y}_{ri}, \dot{y}_{ri}), \tag{12}$$

where $a_{ij}(1 \le i \le N, 0 \le j \le N)$ is defined similarly from the entry of the adjacency matrix A and (1) deduced from $\mathcal{G}$, and $\alpha > 0$ is a design parameter. $\hat{\theta}_i$ is denoted as a current estimate of unknown θ_i, and v_i is a stabilizing signal which is to be determined later based on the notion of inverse optimal H_∞ control criterion (Controller I), or without H_∞ control criterion (Controller II). An estimation error between the leader $\dot{y}_0$ and the estimate $\hat{z}_i$ is defined by

$$\tilde{z}_i(t) \equiv \hat{z}_i(t) - \dot{y}_0(t), \tag{13}$$

and the following relations are deduced for s_i and $\tilde{z}_i$.

$$\dot{\tilde{z}}_i(t) = -\beta \sum_{\substack{j=1 \\ j\neq i}}^{N} c_{ij}\{\tilde{z}_i(t) - \tilde{z}_j(t)\} - \beta c_{i0}\tilde{z}_i(t) + (n_{i0} - 1)\ddot{y}_0(t), \tag{14}$$

$$M_i(y_i)\dot{s}_i(t) + C_i(y_i, \dot{y}_i)s_i(t) = v_i(t) + Y_i(t)\{\hat{\theta}_i(t) - \theta_i\}. \tag{15}$$

Then, the overall representations of the multi-agent system are given as follows:

$$\dot{\tilde{z}}(t) = -\beta(M_c \otimes I)\tilde{z}(t) + \{(N_0 - \mathbf{1}) \otimes I\}\ddot{y}_0(t), \tag{16}$$

$$M\dot{s}(t) + Cs(t) = Y(t)\{\hat{\theta} - \theta(t)\} + v(t), \tag{17}$$

where

$$\tilde{z} = [\tilde{z}_1^{\mathsf{T}}, \cdots, \tilde{z}_N^{\mathsf{T}}]^{\mathsf{T}}, \tag{18}$$

$$s = [s_1^{\mathsf{T}}, \cdots, s_N^{\mathsf{T}}]^{\mathsf{T}}, \tag{19}$$

$$M = \text{block diag}(M_1, \cdots, M_N), \ (M_i \equiv M_i(y_i)), \tag{20}$$

$$C = \text{block diag}(C_1, \cdots, C_N), \ (C_i \equiv C_i(y_i, \dot{y}_i)), \tag{21}$$

$$Y = \text{block diag}(Y_1, \cdots, Y_N), \tag{22}$$

$$\theta = [\theta_1^{\mathsf{T}}, \cdots, \theta_N^{\mathsf{T}}]^{\mathsf{T}}, \tag{23}$$

$$N_0 = [n_{10}, \cdots, n_{N0}]^{\mathsf{T}}, \tag{24}$$

$$\mathbf{1} = [1, \cdots, 1]^{\mathsf{T}}, \tag{25}$$

$$v = [v_1^{\mathsf{T}}, \cdots, v_N^{\mathsf{T}}]^{\mathsf{T}}, \tag{26}$$

and $\otimes$ denotes Kronecker product.

5 Adaptive H_∞ Consensus Control for Euler-Lagrange Systems (Controller I)

In this section a design scheme of adaptive consensus control based on inverse optimal H_∞ criterion (Controller I) is provided. Stability analysis of the overall control system and deduction of Controller I are composed of four steps. As the first step, for stability analysis of s and the related terms, a positive function W_0 is defined such as

$$W_0(t) = \frac{1}{2}s(t)^{\mathsf{T}}Ms(t). \tag{27}$$

Then, the time derivative of W_0 along its trajectory is derived as follows:

$$\dot{W}_0(t) = s(t)^{\mathsf{T}}[Y(t)\{\hat{\theta} - \theta(t)\} + v(t)]. \tag{28}$$

By considering the evaluation of $\dot{W}_0$ (28), the next virtual system is introduced.

$$\dot{s} = f + g_1 d + g_2 v, \tag{29}$$

$$f = 0, \tag{30}$$

$$g_1 = Y, \ g_2 = I, \tag{31}$$

$$d = (\hat{\theta} - \theta). \tag{32}$$

The virtual system is to be stabilized via a control input v based on H_∞ criterion, where d is considered as an external disturbance to the process. For that purpose, the following Hamilton-Jacobi-Isaacs (HJI) equation and its solution V_0 are introduced.

$$\mathcal{L}_f V_0 + \frac{1}{4}\left\{\frac{\|\mathcal{L}_{g_1}V_0\|^2}{\gamma^2} - (\mathcal{L}_{g_2}V_0)R^{-1}(\mathcal{L}_{g_2}V_0)^{\mathsf{T}}\right\} + q = 0, \tag{33}$$

$$V_0 = \frac{1}{2}s^{\mathsf{T}}s, \tag{34}$$

where q and R are a positive function and a positive definite matrix respectively, and those are deduced from HJI equation based on the notion of inverse optimality for the given solution V_0 and a positive constant γ. The substitution of the solution V_0 (34) into HJI equation (33) yields

$$\frac{1}{4}s^{\mathsf{T}}\left\{\frac{YY^{\mathsf{T}}}{\gamma^2} - R^{-1}\right\}s + q = 0. \tag{35}$$

From (35), R and q are obtained such as

$$R = \left(\frac{YY^{\mathsf{T}}}{\gamma^2} + K\right)^{-1}, \tag{36}$$

$$q = \frac{1}{4}s^{\mathsf{T}}Ks, \tag{37}$$

where K is a diagonal positive definite matrix (a design parameter), From R, v is derived as a solution of the corresponding H_∞ control problem as follows (Controller I):

$$\begin{aligned} v &= -\frac{1}{2}R^{-1}(\mathcal{L}_{g_2}V_0)^{\mathsf{T}} = -\frac{1}{2}R^{-1}s \\ &= -\frac{1}{2}\left(\frac{YY^{\mathsf{T}}}{\gamma^2} + K\right)s. \end{aligned} \tag{38}$$

Then, via HJI equation, the time derivative of W_0 (28) is evaluated as follows:

$$\begin{aligned} \dot{W}_0 &= -q - v^{\mathsf{T}}Rv \\ &\quad + \left(v + \frac{1}{2}R^{-1}s\right)^{\mathsf{T}} R\left(v + \frac{1}{2}R^{-1}s\right) \\ &\quad + \gamma^2\|d\|^2 - \gamma^2\left\|d - \frac{Y^{\mathsf{T}}s}{2\gamma^2}\right\|^2, \end{aligned} \tag{39}$$

and it follows that s is bounded for bounded $\hat{\theta}$ and for the stabilizing signal v (38).

Next, for stablity analysis of the estimation error $\tilde{z}$, a positive function V_1 is introduced such as

$$V_1 = \tilde{z}^{\mathsf{T}}(P_c \otimes I)\tilde{z}, \tag{40}$$

$$P_cM_c + M_c^{\mathsf{T}}P_c = I, \ \ (P_c = P_c^{\mathsf{T}} > 0). \tag{41}$$

There exists a positive definite and symmetric matrix P_c satisfying (41), since $-M_c$ is Hurwitz. Then, the time derivative of V_1 along its trajectory is evaluated as follows:

$$\begin{aligned} \dot{V}_1 &= -\beta\|\tilde{z}\|^2 - 2\tilde{z}^{\mathsf{T}}(P_c \otimes I)\{(N_0 - \mathbf{1}) \otimes I\}\ddot{y}_0 \\ &\le -\frac{\beta}{2}\|\tilde{z}\|^2 + \frac{2}{\beta}\|P_c \otimes I\|^2\|\{(N_0 - \mathbf{1}) \otimes I\}\ddot{y}_0\|^2, \end{aligned} \tag{42}$$

and it is shown that $\tilde{z}$ is bounded for bounded $\ddot{y}_0$.

Thirdly, for stability analysis of the control error $y_i - y_0$ and the related terms, $\tilde{y}_i$, $\tilde{y}$ are defined by

$$\tilde{y}_i = y_i - y_0, \tag{43}$$

$$\tilde{y} = [\tilde{y}_1^{\mathsf{T}}, \cdots, \tilde{y}_N^{\mathsf{T}}]^{\mathsf{T}}. \tag{44}$$

Then, the following relation holds

$$\dot{\tilde{y}} = s + \tilde{z} - \alpha(M_a \otimes I)\tilde{y}, \tag{45}$$

and $-M_a$ is shown to be Hurwitz because of the assumption of the network graph $\mathcal{G}$. From that property, a positive function V_2 is defined by

$$V_2 = \tilde{y}^{\mathsf{T}}(P_a \otimes I)\tilde{y}, \tag{46}$$

$$P_aM_a + M_a^{\mathsf{T}}P_a = I, \ \ (P_a = P_a^{\mathsf{T}} > 0). \tag{47}$$

Similarly to the previous case (M_c), there exists a positive definite and symmetric matrix P_a satisfying (47), since $-M_a$ is Hurwitz. Then, the time derivative of V_2 along its trajectory is evaluated as follows:

$$\begin{aligned} \dot{V}_2 &= -\alpha\|\tilde{y}\|^2 + 2\tilde{y}^{\mathsf{T}}(P_a \otimes I)(s + \tilde{z}) \\ &\le -\frac{\alpha}{2}\|\tilde{y}\|^2 + \frac{4}{\alpha}\|P_a \otimes I\|^2(\|s\|^2 + \|\tilde{z}\|^2). \end{aligned} \tag{48}$$

From the three stages of stability analysis (the evaluations of $\dot{W}_0$, $\dot{V}_1$, $\dot{V}_2$), the next theorem is obtained.

Theorem 1 *The nonlinear control system composed of the control laws (7), (9), (10), (11), (12), (38) (Controller I) is uniformly bounded for an arbitrary bounded design parameter $\hat{\theta}_i$, and bounded y_0, $\dot{y}_0$, $\ddot{y}_0$, and v is an optimal control input which minimizes the following cost functional J.*

$$\begin{aligned} J(t) \equiv \sup_{d_i, d_2, d_3 \in \mathcal{L}_2} \Bigg[&\int_0^t \{q + v^{\mathsf{T}}Rv\}d\tau + W_0(t) \\ &-\gamma^2 \int_0^t \|d\|^2 d\tau \Bigg]. \end{aligned} \tag{49}$$

Also the next inequality holds.

$$\begin{aligned} &\int_0^t \{q + v^{\mathsf{T}}Rv\}d\tau + W_0(t) \\ &\qquad \le \gamma^2 \int_0^t \|d\|^2 d\tau + W_0(0). \end{aligned} \tag{50}$$

Theorem 1 denotes the properties of the proposed nonlinear control system (7), (9), (10), (11), (12), (38) (Controller I), where the tunings of $\hat{\theta}$ is not included (or not necessarily required).

Next, the tuning law of $\hat{\theta}$ is determined as follows:

$$\dot{\hat{\theta}}(t) = \mathrm{Pr}\left\{-\Gamma Y(t)^{\mathsf{T}}s(t)\right\}, \tag{51}$$

where $\mathrm{Pr}(\cdot)$ is a projection operation in which the tuning parameter $\hat{\theta}$ is constrained to a bounded region deduced from upper-bounds of $\|\theta\|$ [27]. As the fourth step of stability analysis of the overall control system, a positive function W_1 is defined by

$$W_1(t) = \frac{1}{2}s(t)^{\mathsf{T}}s(t) + \frac{1}{2}\left\{\hat{\theta}(t) - \theta\right\}^{\mathsf{T}}\Gamma^{-1}\left\{\hat{\theta}(t) - \theta\right\}, \tag{52}$$

and the time derivative of W_1 along its trajectory is evaluated such as

$$\dot{W}_1(t) \le -\frac{1}{2}s(t)^{\mathsf{T}}R^{-1}s \le 0. \tag{53}$$

From the four stages of stability analysis (the evaluations of $\dot{W}_0$, $\dot{W}_1$, $\dot{V}_1$, $\dot{V}_2$), the next theorem is obtained.

Theorem 2 *The adaptive control system composed of the control laws (7), (9), (10), (11), (12), (38), and the tuning law of $\hat{\theta}$ (51) (Controller I), is uniformly bounded for bounded y_0, $\dot{y}_0$, $\ddot{y}_0$, and it follows that*

$$\lim_{t\to\infty} s(t) = 0. \tag{54}$$

Especially, if $\ddot{y}_0(t) = 0$ or the information of the leader $\ddot{y}_0$ is available for all followers ($\{(N_0 - \mathbf{1}) \otimes I\}\dot{y}_0 = 0$), then it follows that

$$\lim_{t\to\infty} \tilde{y}(t) = \lim_{t\to\infty} \dot{\tilde{y}}(t) = 0, \tag{55}$$

and the asymptotic consensus tracking is achieved. Otherwise, when $\ddot{y}_0(t) \neq 0$ and the information of $\ddot{y}_0$ is not available for all followers ($\{(N_0 - \mathbf{1}) \otimes I\}\ddot{y}_0 \neq 0$), then the next relation holds.

$$\|\tilde{y}\| \sim \text{const.} \cdot \frac{1}{\alpha\beta}\|\{(N_0 - \mathbf{1}) \otimes I\}\ddot{y}_0\|, \tag{56}$$

$$\|\dot{\tilde{y}}\| \sim \text{const.} \cdot \frac{1}{\beta}\|\{(N_0 - \mathbf{1}) \otimes I\}\ddot{y}_0\|. \tag{57}$$

Theorem 2 denotes the properties of the proposed adaptive control system (7), (9), (10), (11), (12), (38), (51) (Controller I), and states that the asymptotic consensus tracking is achieved under the specified condition ($\{(N_0 - \mathbf{1}) \otimes I\}\dot{y}_0 = 0$), and also shows that the approximate consensus tracking with the ratios of $1/(\alpha\beta)$, $1/\beta\,(>0)$ is still assured, even if that condition is not satisfied ($\{(N_0 - \mathbf{1}) \otimes I\}\ddot{y}_0 \neq 0$).

Remark 2 It should be noted the proposed control scheme and the adaptation scheme (Controller I) are all implemented in a distributed fashion, where availabilities of signals for each agent i are highly restricted and prescribed by the directed graph $\mathcal{G}$.

Remark 3 The objective of the design of Controller I, is to obtain an adaptive control structure whose stability is not seriously affected by the adaptation scheme, or whose performance is not significantly degraded by the estimation errors of the tuning parameters. For that purpose, in order to attenuate the effects of the estimation errors, the control scheme is to be deduced as a solution of certain H_∞ control problem, where L_2 gains from the estimation errors of tuning parameters to the generalized output is prescribed by a positive constant γ (design parameter).

Remark 4 Of course, J in **Theorem 1** is a fictitious cost functional, since d is not an actual disturbance but an estimation error of the tuning parameter, and since it is not generally included in $L^2[0,\infty)$. Nevertheless, v, which is derived as a solution for that fictitious H_∞ control problem (Controller I), attain the inequality in **Theorem 1**, stabilize the total system, and it means that the L^2 gain from the disturbance d to the generalized output $\sqrt{q + v^T R v}$ is prescribed by the positive constant γ.

6 Adaptive Consensus Control for Euler-Lagrange Systems without H_∞ Criterion (Controller II)

The adaptive consensus control strategy without H_∞ control criterion (Controller II) is easily deduced by setting $\gamma \to \infty$ in the synthesis of v (38) such that

$$v = -\frac{1}{2}Ks. \tag{58}$$

In this case, the tuning scheme of $\hat{\theta}$ (51) is required to assure uniform boundedness of the total control system. The time derivative of W_1 (52) with the tuning law (51) gives

$$\dot{W}_1(t) \leq -\frac{1}{2}s(t)^{\mathsf{T}}Ks \leq 0. \tag{59}$$

Then, by applying the same stability analysis via the evaluations of $\dot{W}_1$, $\dot{V}_1$, $\dot{V}_2$, the following theorem is obtained.

Theorem 3 *The adaptive control system composed of the control laws (7), (9), (10), (11), (12), (58), and the tuning law of $\hat{\theta}$ (51) (Controller II), is uniformly bounded for bounded y_0, $\dot{y}_0$, $\ddot{y}_0$, and it follows that*

$$\lim_{t\to\infty} s(t) = 0. \tag{60}$$

Especially, if $\ddot{y}_0(t) = 0$ or the information of the leader $\ddot{y}_0$ is available for all followers ($\{(N_0 - \mathbf{1}) \otimes I\}\dot{y}_0 = 0$), then it follows that

$$\lim_{t\to\infty} \tilde{y}(t) = \lim_{t\to\infty} \dot{\tilde{y}}(t) = 0, \tag{61}$$

and the asymptotic consensus tracking is achieved. Otherwise, when $\ddot{y}_0(t) \neq 0$ and the information of $\ddot{y}_0$ is not available for all followers ($\{(N_0 - \mathbf{1}) \otimes I\}\ddot{y}_0 \neq 0$), then the next relation holds.

$$\|\tilde{y}\| \sim \text{const.} \cdot \frac{1}{\alpha\beta}\|\{(N_0 - \mathbf{1}) \otimes I\}\ddot{y}_0\|, \tag{62}$$

$$\|\dot{\tilde{y}}\| \sim \text{const.} \cdot \frac{1}{\beta}\|\{(N_0 - \mathbf{1}) \otimes I\}\ddot{y}_0\|. \tag{63}$$

Remark 5 Asymptotic properties of the adaptive consensus control without H_∞ criterion ((60), (61), (62), (63) (Controller II)) are the same as the adaptive H_∞ consensus control version ((54), (55), (56), (57) (Controller I)). However, it should be noted that uniform boundedness of the overall control system via Controller II is not assured for arbitrary bounded design parameters $\hat{\theta}$. Instead, the tuning of $\hat{\theta}$ (51) is necessary to attain the stability of consensus tracking by Controller II. Furthermore, Controller II in Theorem 3 does not contain an optimal property for the cost functional (49), and the inequality (50) does not hold.

7 Numerical Example

In order to show the effectiveness of the two types of adaptive consensus control schemes, numerical experiments for Euler-Lagrange systems are performed.

A multi-agent system composed of simple Euler-Lagrange systems is considered as follows:

$$m_i\ddot{y}_i(t) = \tau_i(t), \quad (i = 1, 2, 3),$$
$$(y_1(0) = 1, \ y_2(0) = 0, \ y_3(0) = -1),$$

where $y_i \in \mathbf{R}$, $\tau_i \in \mathbf{R}$, and $m_i \in \mathbf{R}$ is an unknown system parameter. Associated with the information network structure (Fig.1), the adjacency matrix $A = [a_{ij}]\,(= C)$ and $a_{i0}\,(= c_{i0})$ are chosen such that

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix},$$
$$a_{10} = 1, \ a_{20} = a_{30} = 0.$$

The control objective is to achieve consensus tracking

$$y_i \to y_j,\ \dot{y}_i \to \dot{y}_j,$$
$$y_i \to y_0,\ \dot{y}_i \to \dot{y}_0,$$
$$(i, j = 1, 2, 3),$$

where the virtual leader y_0 is determined such as

$$\ddot{y}_0 + 2\dot{y}_0(t) + y_0(t) = \sin t.$$

The design parameters are chosen as follows:

$$\Gamma = 10I,\ K = 5I,\ \alpha = \beta = 10,$$
$$\gamma_i = 1.$$

As system parameters m_i $(i = 1 \sim 3)$, we consider both time-invariant and time-varying cases such that

$$m_1 = 1,\ m_2 = 2,\ m_3 = 3,\ \text{(time – invariant case)},$$
$$m_1 = f_m(t),\ m_2 = 2f_m(t),\ m_3 = 3f_m(t),$$
$$\text{(time – varying case)},$$

where

$$f_m(t) = \begin{cases} 2 & 0 \le t < 2.5,\ 5 \le t < 7.5,\ \cdots, \\ 1 & 2.5 < t \le 5,\ 7.5 < t \le 10,\ \cdots. \end{cases}$$

Numerical simulations were carried out by utilizing MATLAB and Simulink of the MathWorks, Inc., and the solver is Dormand-Prince method with the adaptive step-size integration algorithm (ode45).

The simulation results of Controller I (Theorem 2) are shown in Fig.2 (time-invariant case) and Fig.4 (time-varying case). For comparison, the cases of Controller II (the adaptive control systems which do not contain H_∞ control scheme), are shown for both cases; Fig.3 (time-invariant case) and Fig.5 (time-varying case). In those figures (Fig.2 $\sim$ Fig.5), the plots were response curves of y_1, y_2, y_3 and y_0 (vertical axises) versus time t (horizontal- axises).

From those results, it is seen that Controller I (H_∞ adaptive control strategy) achieves better tracking convergence property together with robustness to abrupt changes of the system parameters compared with the non-H_∞ control scheme (Controller II), and those are owing to disturbance attenuation properties of Controller I, since the performance of Controller I is not significantly degraded by the estimation errors of the tuning parameters in the settings of both time-invariant and time-varying cases (see also Remark 3).

8 Concluding Remarks

The comparative study of two types of adaptive consensus control (Controller I and Controller II) of multi-agent systems composed of fully actuated mobile robots which are described as a class of Euler-Lagrange systems on directed network has been given in this paper. Controller I is deduced based on the notion of inverse optimal H_∞ control criterion, and is derived as a solution of certain H_∞ control problem, where estimation errors of tuning parameters are considered as external disturbances to the process. On the contrary, Controller II is synthesized without H_∞ control criterion. Asymptotic properties, stability and robustness to unknown parameters are discussed for those two control strategies (Controller I and Controller II). Although asymptotic properties of those two controllers are similar to each others, Controller I achieves better tracking convergence property together with robustness to abrupt changes of the system parameters compared with Controller II, and those are owing to disturbance attenuation properties of Controller I and the problem setting of H_∞ criterion.

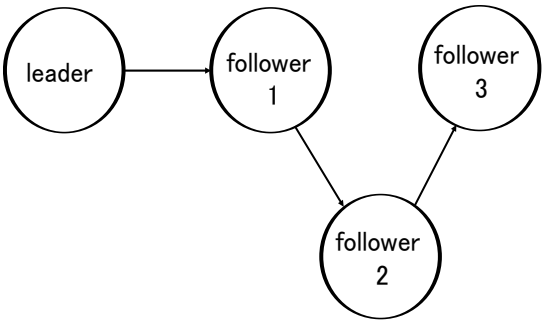


Fig. 1 Information Network Graph

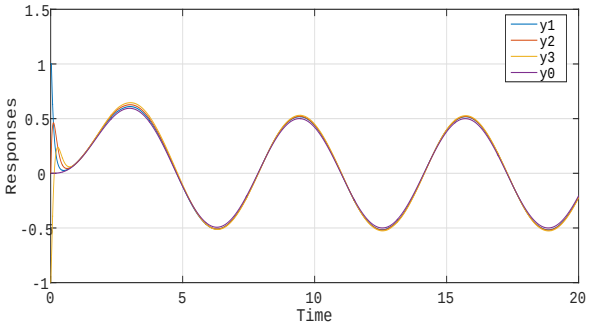


Fig. 2 Simulation Result for Time-Invariant Case with H_∞ Control Scheme (Controller I)

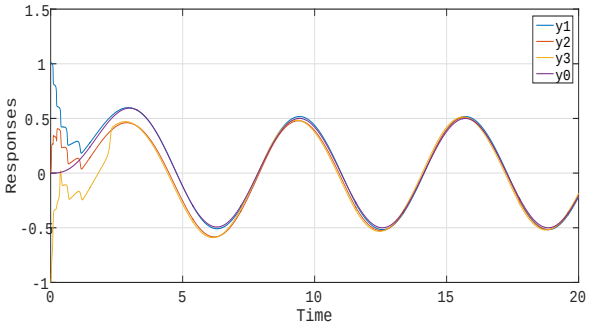


Fig. 3 Simulation Result for Time-Invariant Case without H_∞ Control Scheme (Controller II)

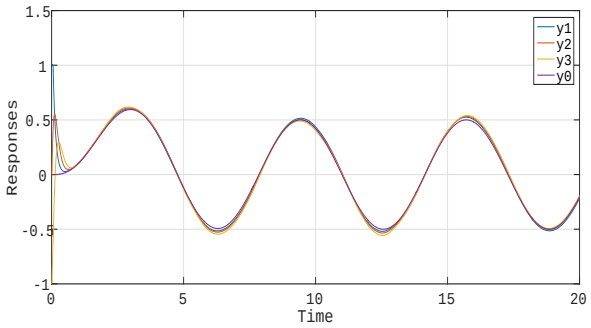


Fig.4 Simulation Result for Time-Varying Case with H_∞ Control Scheme (Controller I)

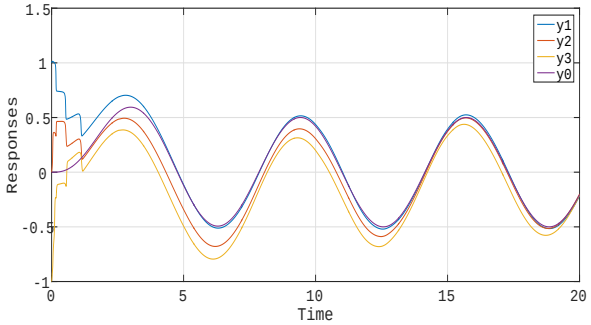


Fig. 5 Simulation Result for Time-Varying Case without H_∞ Control Scheme (Controller II)

References

[1] Y. Miyasato, "Adaptive H_∞ Consensus Control of Euler-Lagrange Systems on Directed Network Graph", *Proceedings of IEEE Symposium Series on Computational Intelligence (IEEE SSCI 2016)*, 2016. https://doi.org/10.1109/SSCI.2016.7849871

[2] M.A. Tan and K. Lewis, "Virtual Structures for High-Precision Cooperative Mobile Robot Control", *Proceedings of the 1996 IEEE/RSJ INternational Conference Intelligent Robots and Systems*, Vol.1, pp.132-139, 1996.

[3] T. Balch and R.C. Arkin, "Behavior-Based Formation Control for Multirobot Teams", *IEEE Transactions on Robotics and Automation*, Vol.14, pp.926-939, 1998.

[4] G. Inalhan, D.M. Stipanovic, and C. Tomlin, "Decentralized Optimization with Application to Multiple Aircraft Coordinate," *Proceedings of the 41st IEEE Conference on Decision and Control*, pp.1147-1155, 2002.

[5] J. Yao, R. Ordóñez and V. Gazi, "Swarm Tracking Using Artificial Potentials and Sliding Mode Control", *Proceedings of the 45th IEEE Conference on Decision and Control*, pp.4670-4675, 2006.

[6] J. Shao, G. Xie and L. Wang, "Leader Following Formation Control of Multiple Mobile Vehicles", *IET Control Theory Appl.*, Vol.1, pp.545-552, 2007.

[7] R.M. Murray, "Recent Research in Cooperative Control of Multivehicle Systems," Journal of Dynamic Systems, *Measurement, and Control*, Vol.129, No.5, pp.571-583, 2007.

[8] K.L. Moore and D. Lucarelli, "Decentralized Adaptive Scheduling Using Consensus Variables, " *International Journal of Robust and Nonlinear Control*,Vol.17, pp.921-940, 2007.

[9] A.R. Pereira and L. Hsu, "Adaptive Formation Control Using Artificial Potentials for Euler-Lagrange Agents", *Proceedings of the 17th IFAC World Congress*, pp.10788-10793, 2008.

[10] C.C. Cheah, S.P. Hou and J.J.E. Slotine, "Region-Based Shape Control for a Swarm of Robots", *Automatica*, Vol.45, pp.2406-2411, 2009.

[11] Y. Miyasato, "Adaptive H_∞ Formation Control for Euler-Lagrange Systems, *Proceedings of the 49th IEEE Conference on Decision and Control*, pp.2614-2619, 2010.

[12] D. Kingston, W. Ren, and R. Beard, "Consensus algorithms are input-to-state stable," *Proceedings of 2005 American Control Conference*, pages 1686-1690, 2005.

[13] R. Olfati-Saber, J. A. Fax, and R. M. Murray. Consensus and Cooperation in Networked Multi-Agent Systems. *Proceedings of the IEEE*, Vol.95, pp. 215-233, 2007.

[14] W. Ren, K.L. Moore, and Y. Chen, "High-Order and Model Reference Consensus Algorithms in Cooperative Control of Multi-Vehicle Systems," *Transaction of American Society of Mechanical Engineers*, Vol.129, pp.678-688, 2007.

[15] Y. Cao and W. Ren, "Distributed Coordinate Tracking via a Variable Structure Approach - Part I : Consensus Tracking," *Proceedings of 2010 American Control Conference*, pp.4744-4749, 2011.

[16] J. Mei and W. Ren, "Distributed Coordinate Tracking with a Dynamic Leader for Multiple Euler-Lagrange Systems," *IEEE Transactions on Robotics and Automation*, Vol.56, pp.1415-1421, 2011.

[17] W. Ren and Y. Cao, *Distributed Coordination of Multi-Agent Networks*, Springer, 2011.

[18] G. Wen, Z. Duan, W. Yu and G. Chen, "Consensus in Multi-Agent Systems with Communication Constraints", International Journal of Robust and Nonlinear Control, Vol.22, pp.170-182, 2012.

[19] W. Yu, W. Ren, W.X. Zheng, G. Chen and J. Lü, "Distributed Control Gains Design for Consensus in Multi-Agent Systems with Second-Order Nonlinear Dynamics", *Automatica*, Vol.49, pp.2107-2115, 2013.

[20] J. Mei, W. Ren and J. Chen, "Consensus of Second-Order Heterogeneous Multi-Agent Systems under a Directed Graph," *Proceedings of 2014 American Control Conference*, pp.802-807, 2014.

[21] J. Mei, W. Ren, J. Chen B.D.O. Anderson, "Consensus of Linear Multi-Agent Systems with Fully Distributed Control Gains under a General Directed Graph", *Proceedings of the 53rd IEEE Conference on Decision and Control*, pp.2993-2998, 2014.

[22] M.W. Spong and M. Vidyasagar, *Robot Dynamics and Control*, John Wiley & Sons, 1989.

[23] M. Krstić and H. Deng, *Stabilization of Nonlinear Uncertain Systems*, Springer, 1998.

[24] Y. Miyasato, "Adaptive Nonlinear H_∞ Control for Processes with Bounded Variations of Parameters - General Relative Degree Case -", *Proceedings of the 39th IEEE Conference on Decision and Control*, pp.1453-1458, 2000.

[25] Y. Miyasato, "Adaptive H_∞ Consensus Control of Multi-Agent Systems on Directed Graph", *Proceedings of the 39th IEEE Conference on Decision and Control*, pp.1453-1458, 2015.

[26] Y. Miyasato, "Adaptive H_∞ Consensus Control of Multi-Agent Systems on Directed Graph by Utilizing Neural Network Approximators", *Proceedings of IFAC Workshops, ALCOSP and PSYCO 2016*, 2016.

[27] P.A. Ioannou and J. Sun, *Robust Adaptive Control*, PTR Prentice-Hall, NJ; 1996.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 2, No. 3, 1172-1180 (2017)
www.astesj.com
Special Issue on Recent Advances in Engineering Systems

ASTES Journal
ISSN: 2415-6698

Future Contract Selection by Term Structure Analysis

Vasco Grossmann*, Manfred Schimmler

Kiel University, Technical Computer Science, Department of Computer Science, Faculty of Engineering, 24098, Germany

ARTICLEINFO

Article history:
Received: 30 May, 2017
Accepted: 09 July, 2017
Online: 20 July, 2017

Keywords:
Computational finance
Data mining
Trading cost analysis
Future rolling

ABSTRACT

In futures markets, a single asset is generally represented by several contracts with different maturities. The selection of specific contracts is an inevitable task that also creates new opportunities, especially in terms of speculative trading. Evaluating immediate and upcoming trading costs for all considered contracts might lead to a significantly improved performance. Among that, even possible market inefficiencies might be taken into consideration. This research introduces and evaluates a new algorithm for the contract selection. The results are benchmarked and compared with established methods using a Monte Carlo simulation on different commodity and index futures.

1 Introduction

Contractual agreements over prospective commodity flows already were established in the ancient. The commitment of transanctions to a fixed maturity reliably enabled long-term planning and reduced risks. The standardization of these so-called forwards to exchange-traded futures therefore only has been a question of time. Accordingly, futures have quickly spread after their first release at the Chicago Board of Trade in the 1860s. Modern futures markets offer high volumes and attractive trading conditions for numerous assets. This situation generates a highly interesting trading environment for speculative investors.

This paper is an extension of work originally presented in SSCI 2016 [1] in which different contract selection strategies have been introduced and tested in commodity markets. This research deepens the findings and improves the presented algorithm by refining the selection process. For this purpose, the new algorithm separates the ask and bid price structure to furtherly improve a case-dependent trading cost analysis. The results are evaluated by Monte Carlo simulations on sets of arbitrary trading instructions on three future classes in both, commodity and index futures. The analysis is settled on a broad fundament of historical data that enables a wider range of benchmarks than the previous paper. The results allow the conclusion of market inefficiencies in the analyzed markets.

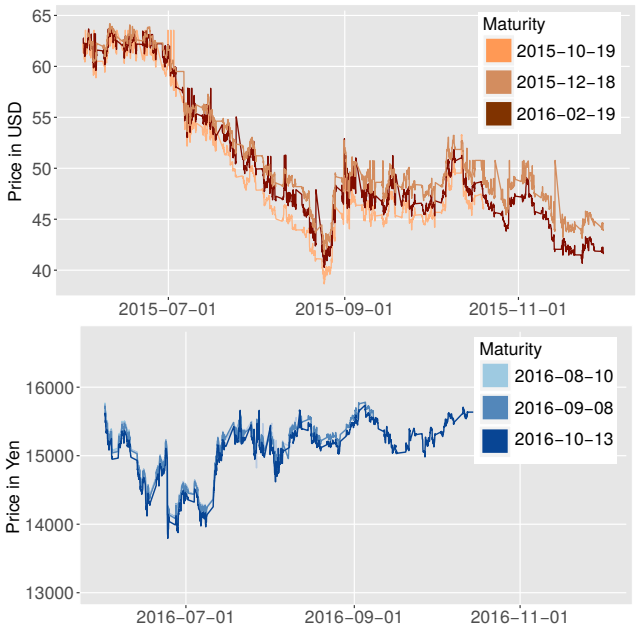


Figure 1: Prices of three WTI future contracts (top chart) and three Nikkei future (bottom chart) contracts with different maturities

Among other properties, every future class is defined by a fix maturity interval in which corresponding contracts expire (typically one or three months). The life span of a contract however is an issue of supply and demand and generally considerably longer. Consequently, there are several coexistent contracts with different price structures and liquidities – all for

*Vasco Grossmann, Kiel University, Germany, vgr@informatik.uni-kiel.de

https://dx.doi.org/10.25046/aj0203148

the same asset, but with different maturities. Due to individual risks and demand changes, different underlyings may exhibit significant disparities in their term structure properties. While consumer goods futures like natural gas show strong seasonal fluctuations, differences of index futures tend to more constant deviations [2]. Figure 1 exemplarily shows quotes for different WTI future contracts during the period from June to December 2016.

The application of conventional trading strategies in futures markets mandatorily generates the challenge of defining a strategy for the contract selection. As advantageous properties of a contract systematically result in other properties being unfavorable, the selection is always a nontrivial compromise with the objective of the best possible balance.

Of course, the prices of future contracts are significantly influenced by the current price of the underlying asset. However, an adequate model requires several additional factors to be taken into consideration. The definition of a formal relationship between future and spot price has been researched precisely and several models have been proposed [3–5]. The cost-of-carry model explains the price $F_{t,\tau}$ of a future contract with maturity τ at time t as a function of the spot price S given by

$$F_{t,\tau} = S_t \cdot e^{(r+s-c)\cdot(\tau-t)}. \quad (1)$$

The interest rate is divided into three parts: the risk-free interest rate r and storage cost s. Arising inaccuracies are explained by the so-called convenience yield c that covers all movements that are created by changing market expectations. Thereby, market situations like contango and backwardation can be constituted.

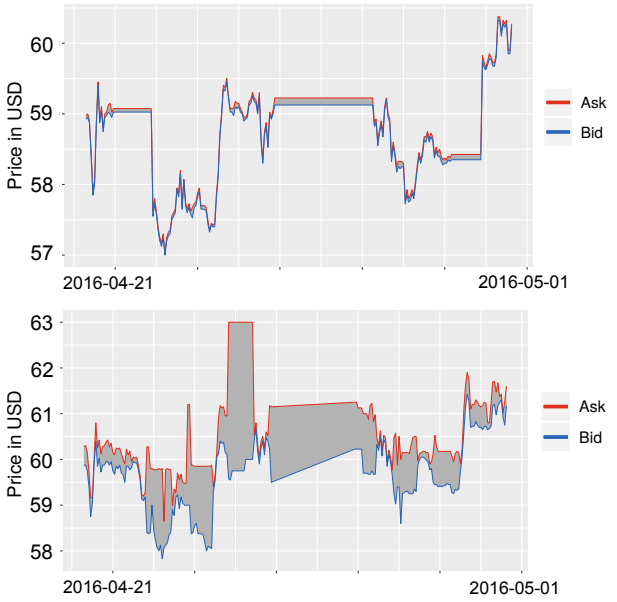


Figure 2: Ask and bid quotes from two WTI Mini contracts with maturities July (above) and September (below) 2016 show huge differences in market liquidity and estimated trading costs

Figure 4 shows ask and bid quotes for two WTI Mini future contracts with maturities July and September 2016 during April 2016. Obviously, the spread highly depends on the time left to the maturity. The lower and much more persistent spread of the near contract yields favorable conditions. The notably smaller liquidility of contract markets with a more distant maturity generates an averagely higher spread. As the trading volume tremendously rises with a decreasing time span to the maturity, spreads tend to be the lowest for the nearest future contracts. While commissions follow a plain pattern, spreads between ask and bid prices strongly depend on market capitalization.

However, the acquisition of contracts with short-term maturities for long-running positions require an extension of the expiry by re-opening positions for the same underlying asset. This procedure is called future rolling and results in additional trades and accordingly trading costs.

So, a suitable future contract selection strategy for speculative investments should balance the contracts in a way that the benefit from low trading costs is as high as possible. Besides, the chance of upcoming future rollings should be as low as possible to avoid additional costs.

One established selection strategy is the so called front month rolling strategy [6]. The method restricts itself to the trade of the nearest future contract whose maturity is more than one month in the future. After exceeding that point in time, contracts are rolled to the subsequent maturity. Of course, the month interval might be scaled depending on average trading intervals. Next to the benefit of most likely addressing liquid markets, it is simple to use and backtesting requires only the two nearest future contracts to be considered.

However, the constant period may be to rigid to fit the needs of the underlying system. Especially highly varying holding times or partial reorganizations of a long-running portfolio may cause the necessity of future rolls. The front month strategy disregards these factors and also ignores several information that might improve the selection process. Even if more distant contracts suffer from higher trading costs in general, their evaluation might depict situations in which transactions of distant contracts create new opportunities. Besides, current portfolio positions have a significant impact during the liquidation process and should be evaluated – their consideration can result in a better contract selection. Also statistical parameters of the trading strategy might be taken into consideration. Correspondingly, this research introduces a procedure that optimizes the selection of futures by investigating the influence of a larger set of parameters. The described problem will be formalized in the following section.

2 Optimal contract selection

The application of a trading strategy based on a conventional non-expiring asset in futures markets is non-trivial due to the additional task of selecting maturities. This research addresses this contract selection problem and introduces algorithms that assign trading decisions for non-expiring assets to expiring contracts. It is searched for the optimal selection of specific contracts for a given future class $\mathbb{F}$ in a defined environment. The objective of the presented method is to allocate contracts in a way that maximizes the overall wealth. As the selection process should work on arbitrary trading strategies on a future class, the latter are sufficiently characterized by deviated trading decisions.

Let a future class $\mathbb{F}$ be a set of related future contracts $F_{\tau_1}, F_{\tau_2}, \ldots, F_{\tau_n}$ that only differ in their maturities $\tau_1, \tau_2, \ldots, \tau_n \in \mathbb{T}$.

Let $(\mathcal{D}_t)_{t\in\mathbb{T}}$ with $\mathcal{D}_t \in \mathbb{Z}$ be the sequence of trading decisions. $\mathcal{D}_t$ contains the number of positions to be bought or sold (depending on sign) at time t – the value 0 represents no transactions.

Let Ω be the set of all possible market scenarios and the finite sequence $(\mathcal{F}_t)_{t\in\mathbb{T}}$ be a filtration on the space Ω, so that the element $\mathcal{F}_t$ represents the known and relevant information at time t.

Let $q : \mathcal{F} \times \mathcal{D} \to \mathbb{F}$ be a selection strategy that returns a specific future contract F_τ by evaluating the known information set $\mathcal{F}_t$ for a given trading decision $\mathcal{D}_t$. Let Q be the set of all selection strategies.

Let $w : \mathcal{D} \times Q \to \mathbb{R}$ be the wealth function that calculates the return after the application of all trading decisions in $\mathcal{D}$ with a selection strategy in the set Q. Thereby, the function gives a measure that enables the comparison of different selection strategies.

The problem is then to find a contract selection strategy $\hat{q} \in Q$ at time t that maximizes the expected wealth after the execution of all trading decisions:

$$w(\mathcal{D}, \hat{q}) = \max_{q\in Q}(w(\mathcal{D}, q)) \tag{2}$$

The optimal selection strategy $\hat{q}$ depends on several factors. The variety of contracts creates an opportunity of lesser trading costs – providing that trading costs (and possible future roll costs) are at least rateable. Thus, the analysis of observed spreads is a mandatory step for a suitable prediction that is discussed in section 3.2. Statistical information of the underlying trading strategy is then included to estimate the temporal uncertainty of trading decisions. These distributions are introduced to identify the time window in which a prospective order may be created by a trading strategy so that emerging costs can be predicted as precise as possible 3.4. In a final step, the overall minimum trading costs are evaluated by the analysis of all discrete trading paths over all possible combinations (section 3.5).

3 Minimizing trading costs

The identification of a selection strategy that maximizes the overall wealth not only depends on minimal trading costs but as well on the question whether the underlying market is efficient in regard to the market-efficiency hypothesis [7]. While cointegrated processes may be used to furtherly improve trading results in inefficient markets, the following section focuses on the minimization of trading costs.

As there is a strong connection between distance to the maturity and the spread of a contract, trading costs can be predicted quote precisely. Figure 3 shows the average development for six WTI contracts during the last 180 trading days (maturities from July to December 2015) and clearly reveals this correlated relation. Not only the value but its volatility decreases for near maturities which is the typical evolution for the price structure [8]. The convergence to a spread of 1‰ of the base price and enables almost exact estimates for the spread during the last months. The spread oscillation is a result of liquidity discrepancies between regular and irregular trading hours. As these tremendous differences may generate a huge impact on the overall performance of the trading strategy, the contract selection is a crucial topic .

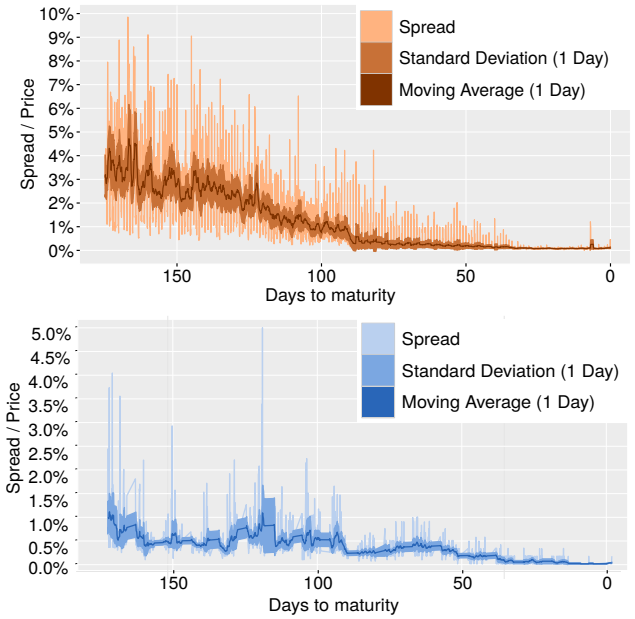


Figure 3: Spread over the last six months to maturity for WTI Mini contracts (top chart) and Nikkei contracts (bottom chart)

Figure 3 clearly displays the advantage of short-term contracts. However, as one contract per month exists for WTI, it also reveals that at least the last three contracts may have similar spreads at times. Considering long-term trading strategies, the acquisition of near contract positions create a high probability of necessary future rolls. This effect appears even stronger for different Nikkei contracts. Although there are frequently high spreads for distant futures, the high fluctuation may yield opportunities as well.

3.1 Spread composition analysis

High spreads are a symptom of market illiquidity. However, they do not automatically represent a similarly disadvantageous situations for both, buyers and sellers. For example, it is absolutely possible that a high spread of a future contract is an expression of a low supply for either long or short positions in comparison to other contracts. Still, the other side may offer attractive prices anyhow. That is the reason why an alignment of spread structures of all contracts is a reasonable step to identify trading opportunities. Therefore, a robust estimation for evolutions of the log prices is applied that generate the least absolute residuals.

Let $A_t^{F_\tau}$ and $B_t^{F_\tau}$ be the ask and bid log prices of F_τ at time t. The evaluation of the estimated log price X_{t+1} is given by

$$\begin{aligned} \mathrm{d}X_{t+1} =& \min_{\epsilon\in\mathbb{R}} \sum_{F\in\mathbb{F}} \max\left(X_t - A_{t+1} + \epsilon, 0\right) + \\ & \min\left(B_t - X_{t+1} + \epsilon, 0\right). \end{aligned} \tag{3}$$

Figure 4 shows the application of the robust quote estimation for the two presented WTI contracts. By interpreting all current contracts, price evolutions are computed that are placed between ask and bid prices. This creates comparable quotes for all future contracts. The differences of ask, bid and robust estimation enables the separate rating of supply and demand. For instance, the significant increase of the spread at April 24 for the December contract is primarily generated by an unfavorable ask price that cannot be traced back to a general movement. The bid price does not show any illiquidities in this case.

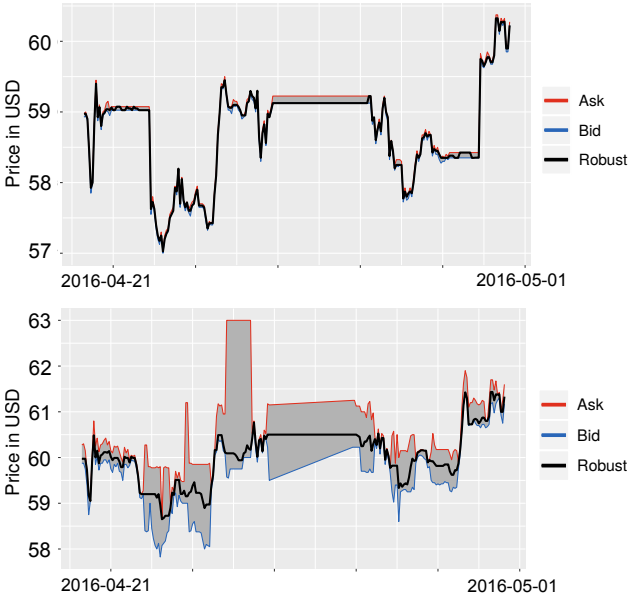


Figure 4: Robust quote estimation

The resulting time series is mandatorily less volatile than both, ask and bid quotes. It only contains the minimum movement that both time series make and is robust concerning every one-sided break. This also enables to separate non-exploitable mean reverse tendencies from real mispricings. This property will be used to evaluate under- and overvaluations in chapter 4.

3.2 Relation between spread and maturity

To formally include the remaining time span of a contract, let $\mathbb{D} \subset \mathbb{Z}$ be the set of durations. Let $d_\tau : \mathbb{T} \to \mathbb{D}$ with $d_\tau(t) = \tau - t$ be the remaining time span between a time t and the maturity date τ of a future contract F_τ.

Let $\mathrm{d}\tau \in \mathbb{D}$ be the time interval between maturities of two consecutive future contracts. It is assumed to be constant for a future class (e.g. 1 month for WTI).

Let $c\colon \mathbb{D} \to \mathbb{R}$ with $c(d), d \geq 0$ then be the average observed trading costs that incur d time steps before the maturity.

Further, let

$$c(d) = \underbrace{c(0) + c(\mathrm{d}\tau)}_{\text{Future roll cost}} + \underbrace{c(d + \mathrm{d}\tau)}_{\text{Liquidation cost}} \qquad \forall\, d < 0. \tag{4}$$

be the trading costs for an exceeded maturity. Future rolls consist of two trades: closing the expiring position at the last valid point before the corresponding maturity ($t = c(0)$) and re-opening the same position in a the consecutive contract ($c(\mathrm{d}\tau)$). The actual liquidation itself must be included as well and is delayed by $\mathrm{d}\tau$. Therefore, the overall trading cost is a sum of three values as shown in equation 4. Figure 6 illustrates these additional future roll costs for exceeded maturities.

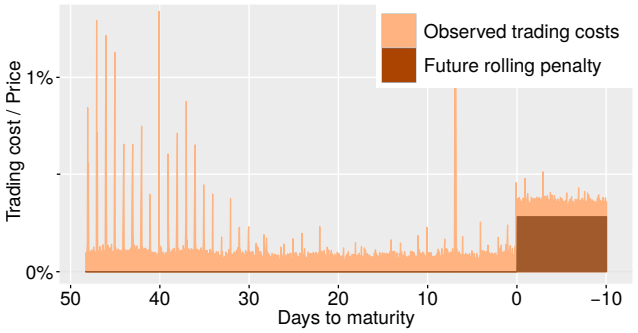


Figure 5: The average trading costs $c(d)$ over the last 50 days for WTI Mini contracts are displayed. An exceeded maturity results in additional trades by future rolling. Therefore, the expected trading costs increase considerably for $t < 0$.

It can be summarized that the a low future roll probability opposes low trading costs – the least costs will generally arise in a balance between these objectives. The probability of future rolls not only depends on maturities but is primarily based on the relationship between maturity and trading frequency. Therefore, statistical information about the trading strategy must be evaluated.

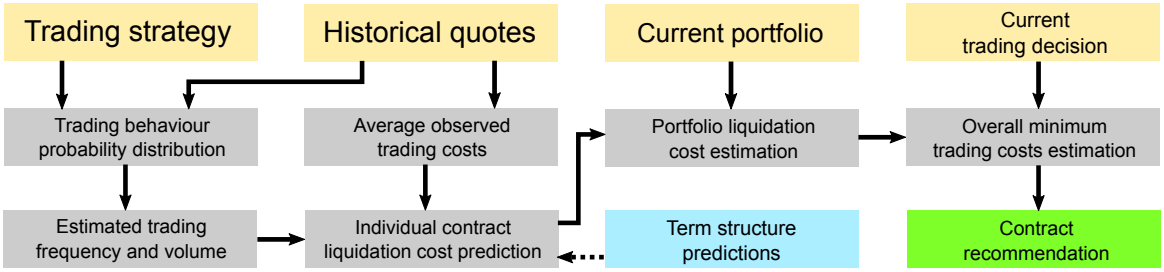


Figure 6: Overview over the evaluation pipeline that leads to a contract recommendation for a given trading decision.

3.3 Statistical trading strategy evaluation

Depending on the trading strategy, prospective trading times may be perfectly predictable. Even if there are no temporal connections, statistical methods allow the estimation based on past observations. Thus, it is assumed that a probability density function of the number of trades per time interval can be computed. In this case, it is assumed to be normally distributed with an average interval between two trades μ and a standard deviation σ. These values enable the calculation of the number of trades up to a prospective point in time. The time of the n-th upcoming trade is then represented by a random variable $t_n \sim \mathcal{N}(t + n\mu, n\sigma)$ for the last observed trading time t. The expected number of trades in an interval from s to t is $\frac{t-s}{\mu}$ with a standard deviation of $\frac{t-s}{\mu} \cdot \sigma$.

3.4 Portfolio liquidation costs

Next to immediate trading costs, the attractivity of contracts is strongly connected to the question how expensive the liquidation of the acquired position will be. This section focuses on the prediction of these costs. A minimization procedure is introduced with regard to the contract selection problem. It is based on the evaluation of the following information:

- future contract positions in portfolio
- estimation of spread fraction (section 3.1)
- estimation of trading cost function c (section 3.2)
- statistical information frequency and volume of the underlying trading strategy (section 3.3)

To connect the trading cost function c with the statistical information, we define $v : \mathbb{T}\times\mathbb{T}\times\mathbb{T} \to \mathbb{R}$ to be expected trading costs. The value $v(\tau, t, s)$ displays the trading cost estimate for a future contract F_τ at time t at calcuation time s with $s < t$. These costs are independent from s for this first researched method and directly arise from the average observed trading costs:

$$v(\tau, t, s) = c(\tau - t). \tag{5}$$

The future trading costs can be predicted by convoluting the probability density function f (in this case a normal-distributed approximation) and a given estimate v. The temporal uncertainty rises with increasing variance of the distribution for larger distances. This effect reduces the accuracy for more distant predictions that reveals itself by observably smoother movements representing the convolution kernel. The function $C_\tau : \mathbb{T} \to \mathbb{R}$ illustrates this relationship and enables the estimation of transaction costs of prospective trades at an approximate time t with the information of time s by

$$\mathbb{E}(C_\tau(t) \mid \mathcal{F}_s) = (f * v)(t) \tag{6}$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\left(\frac{t-s}{\mu}\cdot\sigma\right)^2}} \cdot \exp\left(-\frac{x^2}{2\left(\frac{t-s}{\mu}\cdot\sigma\right)^2}\right) \cdot v(\tau, t+x, s)\ \mathrm{d}x. \tag{7}$$

In summary, statistical information about the trading strategy, observed trading costs and the estimated probability of future roll costs are weighted to calculate the most likely trading costs for a prospective trading decision.

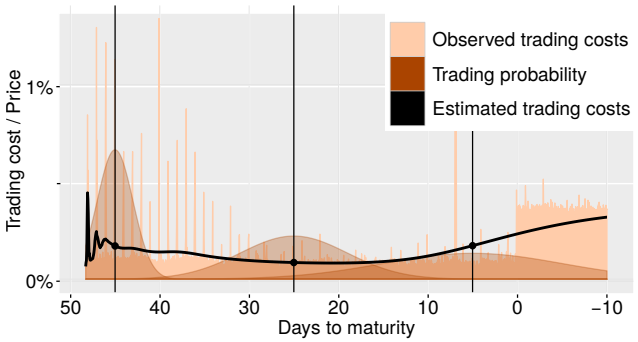


Figure 7: Estimated trading costs $\mathbb{E}(C_\tau(t) \mid \mathcal{F}_s)$ for future F_τ 55 days before its maturity and $\frac{\sigma}{\mu} = \frac{1}{5}$. Normal distributions for three exemplary points (45, 25 and 5 days left) are schematically figured to denote the uncertainty of the effective trading time.

The diverse trading cost effects are exemplarily demonstrated in Figure 7. It shows predictions for uprising trades for a future contract 48 days before its maturity. While the anticipated market liquidity prognosticates averagely lower costs up to 20 days before the maturity, the increased probability of future

rolls yields higher estimates afterwards. While near forecasts apparently embed short-term fluctuations, estimated trading costs for distant trading events are strongly smoothed.

Indeed, these predictions are only suitable to calculate trading costs for discrete future contracts. However, individual estimates can be applied to compare different outcomes in complex scenarios of many consecutive trading decisions. This enables the evaluation of all potential executions and yields the opportunity of a trading cost minimization by prioritizing these estimates. An efficient procedure to calculate the optimal order in such a scenario is introduced in the following.

3.5 Minimizing of portfolio liquidation costs

In the following, the minimum liquidation costs by partial successive trades of a given portfolio iares measured. They are computed by an iterative calculation of optimal liquidation orders for sub-portfolios according to the dynamic programming principle. Starting with the complete portfolio, possible trading decisions are converted to orders that are used to successively remove positions. By comparing each path, an optimal trading path can be evaluated for every sub-portfolio and accordingly, for the empty portfolio. The optimal first liquidation step for the given portfolio is then revealed by the first step of the path with overall minimum trading costs.

Let the set of necessary transactions for the full liquidation of a given portfolio with N futures at time $t \in \mathbb{T}$ be given by $L_t = \{L_t^1, L_t^2, \ldots, L_t^N\} \in \mathbb{N}_0^N$.

Let $\mathcal{C}_s(L_t) \in \mathbb{R}$ at time $s \in \mathbb{T}, s \leq t$ be the estimate of the trading costs for all transactions in L_t.

These transactions will happen at sequent points in future. As described, their order and their costs are predicted recursively. The remaining trading costs are for an empty portfolio have obviously the value 0:

$$\mathcal{C}_s(L_t) = 0 \text{ if } L_t^j = 0 \;\forall\; j \in [1, N]. \tag{8}$$

For every sub-portfolio, there may be up to $N + 1$ actions for each upcoming trading decision:

- if the direction of the order (long or short) is contrary to an existing position of the N given futures, a position of L_t can be reduced (actions will be indexed by $j = [1, N]$)
- choose a different transaction with minimal trading costs that does not contribute to the portfolio liqudation (indexed by $j = 0$)

The latter case might be reasonable if a future provides significantly better trading conditions, but it always increases the probability of future rollings as the number of necessary liquidation steps remains from t to the next trading decision at time $t + \mu$. Therefore, $L_{t+\mu} = L_t$ and accordingly

$$\mathcal{C}_s\left(L_{t+\mu}\right) = \mathcal{C}\left(L_t\right) \text{ if } j = 0. \tag{9}$$

holds. In the other case, contracts of the future with maturity τ_j are liquidated ($j = [1, N]$). The generated trading costs $\mathbb{E}\left(C_{\tau_j}(t) \mid \mathcal{F}_s\right)$ contribute to the overall estimation. They are added to result of the costs for the partial liquidation of the sub-portfolio.

Let the function $\mathcal{L}\colon \mathbb{Z}^N \times F \to \mathbb{Z}^N$ represent contract liquidation. For a given portfolio in $\mathbb{Z}^N$, it returns a portfolio with a reduced position in a specific future of F. The overall trading costs are defined by

$$\begin{aligned}\mathcal{C}_s\left(L_{t+\mu}\right) = \mathcal{C}_s\left(\mathcal{L}\left(L_t, \tau_j\right)\right) \\ + \mathbb{E}\left(C_{\tau_j}(t) \mid \mathcal{F}_s\right) \text{ if } j \in [1, N].\end{aligned} \tag{10}$$

The estimated costs $\mathbb{E}\left(C_{\tau_j}(t) \mid \mathcal{F}_s\right)$ for the liquidation of the specific contract are added to liquidation costs of the residual sub-portfolio $\mathcal{L}\left(L_t, \tau_j\right)$. These cases can be summarized to one evaluation method for the minimum liquidation costs:

$$\mathcal{C}_s\left(L_{t+\mu}\right) = \min_{j \in [0,N]} \begin{cases} 0 & \text{if } L_t^j = 0 \\ & \forall\; j \in [1, N] \\ \mathcal{C}_s(L_t) & \text{if } j = 0 \\ \infty & \text{if } L_t^i = 0 \\ \mathcal{C}_s\left(\mathcal{L}\left(L_t, \tau_j\right)\right) + & \\ \mathbb{E}\left(C_\tau(t) \mid \mathcal{F}_s\right) & \text{else.} \end{cases} \tag{11}$$

Equation 11 contains four cases, three of them are already explained. Only the third case which assigns infinite trading costs to the acquisition of positions for already liquidated contracts has not yet been mentioned. This (obviously wrong) estimate avoids the examination of further transactions after a successful liquidation. Thus, the resulting portfolio is always empty and therefore, the recursive evaluation will always stop its calculation with case one (zero transaction costs for an empty portfolio).

Every sub-portfolio can be reached by a finite number of different order combinations. The dynamic programming principle grants all possible paths to be evaluated. The resulting trading path has the minimum transaction costs according to all intermediate predictions. All in all, $\mathcal{O}(t_{\max} \cdot \prod_{i=1}^{|L|} L_t^i)$ elements must be calculated. The best path can then easily be calculated by backtracking.

The selection of the best liquidation points is unambiguous for a single future when all future costs for its contract are predicted. The best liquidation is directly specified by the order subset with the minimum cumulated trading costs. The complexity arises when several liquidations seem to be optimal at the same time. The best compromise may significantly differ from a greedy selection strategy. As every future generates a new dimension in our evaluation table, such collisions become more probable when more futures

are taken into consideration. The task of finding the optimal path for a single future is trivial, the complexity lies in effecting a compromise between conflicting future liquidations. The resulting path for one future is directly specified by the subset of trades with the minimal cumulated trading costs. However, possible collisions in which several transactions seem optimal at the time yield the problem of selecting specific ones. Therefore, the actual decision is based on the evaluation of all further trading steps as well. It is then given by the first trade of the best trading path.

The process of avoiding liquidations due to better trading costs of other acquisitions extends the time until the portfolio is cleared. Therefore, no finite point in time can be identified and the theoretical number of recursions has no limit. However, as future roll costs cumulate for long-running positions, liquidations of contracts near to their maturity are automatically preferred. The recursive date extension will not exceed $t_{\max} = \max(s + k \cdot \mu, \tau_k)$ with k successive futures for this reason.

4 Mean reverse inefficiencies

The following researches focus on the relationship between a spot price S_t the concerning future prices $F_{t,\tau}$. We expect the future prices to converge to the spot price S_t with a simplified cost-of-carry approach. The differences are explained by specific general interest rates $r_{t,\tau}$, so that

$$F_{t,\tau} = S_t \cdot e^{r_{t,\tau} \cdot (\tau - t)}. \tag{12}$$

Several finance products are only traded as futures, so that spot prices must be approximated by the evaluation of future prices. The following approach averages interest rates between all future combinations. The spot price can be extrapolated to the current time:

$$r_{t,\tau} = avg_{i \neq j} \frac{\ln F_{t,\tau_i} - \ln F_{t,\tau_j}}{\tau_i - \tau_j} \tag{13}$$

The interest rates between spot prices and future prices may strongly fluctuate in short periods. The set of different contract prices is referred to as term structure and several researches concerning its internal systematics have been proposed [9, 10]. As even the smallest under- or overvaluations may be used to enhance the performance of a trading strategy, an ineffiency analysis is a crucial step during the future selection process. For this purpose, it will be analysis whether cointegration between the related price series exist. Different from pure arbitrage strategies, there is no lower limit for the intensity of these inefficiencies. They do not need to be high enough to counterbalance additionally emerging trading costs as all trading decisions are already settled. However, the success of this method does not only require the existence of mean reverse effects but also their persistence to ensure a measurable predictability.

We assume the deviations from the average interest rate $\mathrm{d}r_{t,\tau}$ to be potentially inefficient. In case of over- or undervaluations, mean reverse reactions should be taken into consideration. Therefore, we try to characterize the term evolutions of $\mathrm{d}r_{t,\tau}$ by Ornstein-Uhlenbeck processes with

$$\mathrm{d}r_{t,\tau} = \theta_\tau \left(\mu_\tau - \mathrm{d}r_{t-\mu,\tau} \right) dt + \sigma_\tau dW_t \tag{14}$$

The potential mispricings $\mathrm{d}r_{t,\tau}$ converge to the level μ_τ with the mean reversion speed θ_τ with a standard Wiener process $(W_t)_{t \in \mathbb{T}}$. σ_τ represents the influence of random noise. The parameters of this Ornstein-Uhlenbeck process are fitted with a maximum likelihood estimation [11]. The estimated approximation $\beta_{t,\tau}$ can be used to introduce a future counterbalance of mispricings to the average interest rate in the calculation of $v(\tau, t, x)$ (see Equation 6):

$$\beta_{t,\tau} = F_{t,\tau} \cdot \mathrm{e}^{\mathrm{d}r_{t+\mu,\tau} - \mathrm{d}r_{t,\tau}} \tag{15}$$

$$v(\tau, t, x) = \begin{cases} c((\tau - t) - x) + \beta_{t,\tau} & \text{for a buy} \\ c((\tau - t) - x) - \beta_{t,\tau} & \text{for a sell} \end{cases} \tag{16}$$

By the modification of $v(\tau, t, x)$, the acquisition of a presumably overvalued future contract is penalized in the same way in which a sell of such a future is favored.

5 Results

The objective to reduce trading costs in futures markets by selecting specific futures is targeted by two introduced strategies. They are compared with the established front month rolling strategy for a detailed analysis. This evaluation is based on a Monte Carlo simulation with 10000 tests for each strategy that are based on sets of 100 random trading decisions each. These simulations use real time data from the time period from July 2015 to January 2016.

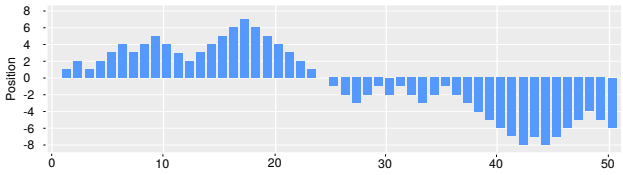


Figure 8: Random portfolio series created by the first 50 trades of an exemplary simulation

Figure 8 shows the resulting portfolio after the first 50 trades of a simulation with a constant volume of one position. Accordingly, the shown portfolio evolution requires 50 contract selections at preassigned times. The result of the front month rolling strategy is shown schematically in Figure 9.

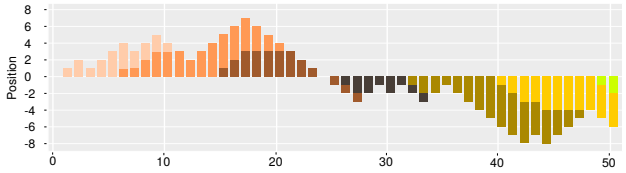


Figure 9: Portfolio evolution of front month rolling strategy

Every color transition represents the change to the subsequent future. Noticeable is the periodic process in which no future is specifically preferred. As every contract is applied for one month, the number of trades is directly dependent from the trading frequency during this period. There are always not more than two different futures in the portfolio. This is induced and without exception, because overlaps directly cause future rollings in this case.

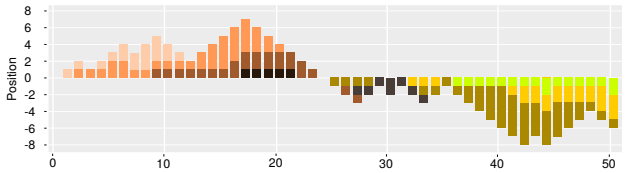


Figure 10: Portfolio evolution of the minimum trading cost contract selection

Figure 10 shows the conversion of trading decisions with the introduced minimum spread approach. The preferred contract changes frequently and yield a more diverse composition with occasional formations with more than two contracts. The expected overall transaction costs are temporary estimated to be the lowest for contracts whose maturity is more than one month in the future, so that transactions partially happen noticeably earlier.

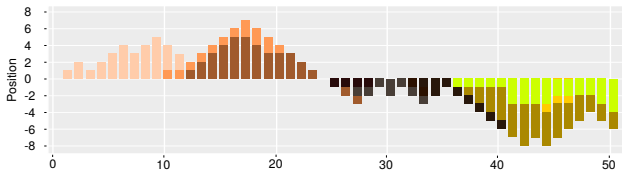


Figure 11: Portfolio evolution of the minimum trading cost contract selection with the Ornstein-Uhlenbeck modification

Mean reverse predictions are added to the selection mechanism in Figure 11. The Ornstein-Uhlenbeck model favors some maturities and leads to situation where some futures are almost skipped. This behavior seems likely as some futures might be seen as under- or overvalued. Even small trading costs might be no compensation in these cases.

The following research examines five different strategies:

1. Front Month rolling strategy
2. Min Spread: strategy that minimizes trading costs as described in chapter 3.5
3. Ornstein-Uhlenbeck: the minimum trading cost estimation (chapter 3.5) with an additional mean reverse expectation term calculated with an Ornstein-Uhlenbeck model
4. Min Spread (SAB): Method 2, but all trading cost evaluations are based on processed spread data according to the ask / bid separation introduced in chapter 3.1. The term *SAB* stands for a separate ask / bid quote analysis during the selection process.
5. Ornstein-Uhlenbeck (SAB): Method 3 with the same ask / bid separation

Table 1: Performance of random sets of trading decisions on three commodities measured in index points.

Return	**Front Month**	**Min Spread**	**Ornstein-Uhlenbeck**
WTI	2.6695	2.5692	3.1958
Natural Gas	-0.1542	0.0245	-0.0625
Silver	0.6951	1.0625	1.4265
Costs	**Front Month**	**Min Spread**	**Ornstein-Uhlenbeck**
WTI	1.4896	1.4126	1.5533
Natural Gas	0.5045	0.3782	0.5913
Silver	1.2405	1.0863	1.5235
Return	**Min Spread (SAB)**		**Ornstein-Uhlenbeck (SAB)**
WTI	2.9684		4.1251
Natural Gas	0.3118		0.8352
Silver	1.2014		1.6495
Costs	**Min Spread (SAB)**		**Ornstein-Uhlenbeck (SAB)**
WTI	1.3840		1.7802
Natural Gas	0.5403		0.4821
Silver	1.2114		1.2247

Table 2: Performance of random sets of trading decisions on three national index futures.

Return	**Front Month**	**Min Spread**	**Ornstein-Uhlenbeck**
S&P 500	-31.8806	-28.4124	-19.1266
Nikkei 225	-760.1611	-680.2393	-580.1591
DAX	-74.1688	-61.5443	-40.5191
Costs	**Front Month**	**Min Spread**	**Ornstein-Uhlenbeck**
S&P 500	35.1198	30.9552	35.9812
Nikkei 225	712.1519	663.0504	634.9534
DAX	80.1419	70.4560	74.2401
Return	**Min Spread (SAB)**		**Ornstein-Uhlenbeck (SAB)**
S&P 500	-5.9915		4.1958
Nikkei 225	-531.0053		-579.5910
DAX	-10.5242		20.5994
Costs	**Min Spread (SAB)**		**Ornstein-Uhlenbeck (SAB)**
S&P 500	32.7817		35.8914
Nikkei 225	649.9088		691.1145
DAX	70.1414		75.0079

The Tables 1 and 2 show the results of Monte Carlo simulations for the different contract selection strategies. Averaged returns and trading costs are depicted for three commodities and three national indices. The cells show cumulated values that are averaged over all simulations with 100 trades each during June 2015 to January 2016.

A return of -760 points with a cost value of 712 points (Nikkei, Front Month) i.e. implies an average

loss of approximately 7 index points per trade that is mainly created by a spread near 7 index points. The loss is expected as the simulations are based on 100 random trades. The corresponding minimum spread approach reduces the cumulated trading costs to 663 points and generated an average loss of 680 points for the same 100 random trading events. The implication of mean reverse tendencies generally increases trading costs, but nevertheless accomplishes higher returns in five cases. These apparently contradicting results encourage the consideration that the term structure temporarily shows mean reverse inefficiencies.

The separation of spreads to supply and demand achieves further trading cost reductions in some cases. These are interestingly more noticeable in the minimum spread approach, the Ornstein-Uhlenbeck method hardly benefits in this category. The return raises clearly in exchange in some cases. The front month rolling strategy averages at a return of -74 points for the DAX index, but generates a win of over 20 points for the improved Ornstein-Uhlenbeck approach. Holding the same positions in different contracts results in a difference of almost 100 index points during the reviewed half year. Especially the newly introduced trading cost separation enhances the result in this case, raising it by approximately 60 points.

6 Conclusion

The selection of specific contracts is a mandatory step in futures markets that may have a huge impact on trading costs and return. Due to a complex spread structure, there are conflictive reasons to acquire contracts with near or distant maturities – the optimal selection depends on many factors. This research opposes five different selection strategies that show considerable disparities in their performance.

The study reveals that a statistical analysis of the used trading strategy and its effects on the portfolio evolution can significantly improve the performance. Therefore, not only detailed analyses of trading times and frequencies but also estimations of upcoming trading costs are contributive. Of course, this advantage requires trading events to be at least approximately predictable. The separate research on ask and bid prices for buy and sell events furtherly improve the trading performance. This holds for commodities as well as for national index futures. The introduction of mean reverse motions after possible over- or underreactions lead to higher trading costs as well as to higher return in the Monte Carlo simulation. On the one hand side, the objective of cost minimization is degraded by the introduction of further influences. On the other hand side, the consideration of over- or undervaluations increases the average return and suggests their existence in the reviewed interval.

References

[1] V. Grossmann and M. Schimmler, "Portfolio-based contract selection in commodity futures markets," in *Computational Intelligence (SSCI), 2016 IEEE Symposium Series on*. IEEE, 2016, pp. 1–7.

[2] H. Geman and C. Kharoubi, "Wti crude oil futures in portfolio diversification: The time-to-maturity effect," *Journal of Banking & Finance*, vol. 32, no. 12, pp. 2553–2559, 2008.

[3] B. Cornell and K. R. French, "The pricing of stock index futures," *Journal of Futures Markets*, vol. 3, no. 1, pp. 1–14, 1983. [Online]. Available: http://EconPapers.repec.org/RePEc:wly:jfutmk:v:3:y:1983:i:1:p:1-14

[4] H. Hsu and J. Wang, "Price expectation and the pricing of stock index futures," *Review Of Quantitative Finance and Accounting*, 2004.

[5] E. S. Schwartz, "The stochastic behavior of commodity prices: Implications for valuation and hedging," *Journal of Finance*, vol. 52, no. 3, pp. 923–73, 1997. [Online]. Available: http://EconPapers.repec.org/RePEc:bla:jfinan:v:52:y:1997:i:3:p:923-73

[6] K. Tang and W. Xiong, "Index investment and the financialization of commodities," *Financial Analysts Journal*, vol. 68, no. 5, pp. 54–74, 2012.

[7] E. F. Fama, "Efficient Capital Markets: A Review of Theory and Empirical Work," *Journal of Finance*, vol. 25, no. 2, pp. 383–417, May 1970. [Online]. Available: https://ideas.repec.org/a/bla/jfinan/v25y1970i2p383-417.html

[8] G. H. Wang, J. Yau *et al.*, "Trading volume, bid-ask spread, and price volatility in futures markets," *Journal of Futures markets*, vol. 20, no. 10, pp. 943–970, 2000.

[9] J. C. Cox, J. E. Ingersoll Jr, and S. A. Ross, "A theory of the term structure of interest rates," *Econometrica: Journal of the Econometric Society*, pp. 385–407, 1985.

[10] L. P. Hansen and R. J. Hodrick, "Forward exchange rates as optimal predictors of future spot rates: An econometric analysis," *journal of Political Economy*, vol. 88, no. 5, pp. 829–853, 1980.

[11] J. C. G. Franco, "Maximum likelihood estimation of mean reverting processes," *Real Options Practice*, 2003.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1181-1187 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Optimal Rotor Design of Line Start Permanent Magnet Synchronous Motor by Genetic Algorithm

Bui Minh Dinh*

Affiliation of electrical engineer, Department of Electronic electrical equipment, Ha Noi University of Science and Technology, Vietnam

ARTICLE INFO

Article history:
Received: 06 March, 2017
Accepted: 01 June, 2017
Online: 20 July, 2017

Keyword:
Line Start Permanent Magnet Synchronous Motor (LSPMSM)
Induction Motor(IM)
Genetic Algorithm (GA)

ABSTRACT

Line start permanent magnet synchronous motor (LSPMSM) is one of the highest efficiency motors due to no rotor copper loss at synchronous speed and self-starting. LSPMSM has torque characteristics of both induction motor IM and Permanent Magnet Synchronous Motor-PMSM. Using Genetic Algorithm (GA) for balancing magnetic cost and for copper loss minimization, the magnetic sizes and geometry parameter of stator and rotor are found and manufactured for industrial evaluation. This article is also taking account practical manufacturing factors to minimize mass production cost. In order to maximize efficiency, an optimal design method of cage-bars and magnet shape has to be considered. The geometry parameters of stator and rotor can be obtained by an analytical model method and validated by FEM simulation. This paper presents the optimal rotor design of a three-phase line-start permanent magnet motor (LSPM) considering the starting torque and efficiency. To consider nonlinear characteristics, the design process is comprised of the FEM and analytical method. During this study, permanent-magnets and cage bars were designed using the magnetic equivalent circuit method and the barriers that control all magnetic flux were designed using the FEM, and the tradeoff of starting torque and efficiency is controlled by weight function in Taguchi method simulation. Finally, some practical results have been obtained and analyzed based on a LSPMSM test bench.

1. Introduction

Line start permanent magnet synchronous motor (LSPMSM) are well known as high efficiency motor due to outstanding advantages compared to other types of induction motor such as high efficiency [1-3], robust structure and high power density. LSPMSM has rotor cage and permanent magnet to maximum starting torque and efficiency because rotor bar loss will be minimized at synchronous operation. The LSPMSM can apply for many areas of ventilation and fan drives. Most of industrial drive motor can be replaced by LSPMSM high efficiency, it saves from 3 to 5% energy of total electric motor consumption. However, it has some drawbacks of starting torque and efficiency performance based permanent magnet sizes and rotor structure. The stator consists of stacked steel laminations with windings placed in the slots whereas the rotor is made of embedded permanent magnet that can vary from two to eight pole pairs with alternatenorth and south poles. In LSPMSM design, rotor magnetic structure has a significant effect of starting torque performances [4-6]. The torque of rotor cage is to speed up motor reach synchronous operation. The permanent magnet will pull rotor to synchronous speed with high efficiency. To maximize efficiency and starting torque, a practical optimal rotor structure have been implemented by Genetic Algorithm (GA) with different factors of cost, sizes and losses. The practical optimization is material and manufacturing cost of LSPMSM from 2.2kw to 11kW for mass production.

2. Basic parameter of LSPMSM

2.1. Electromagnetic design

In order to determine operation points of permanent magnet circuit, some basic parameters of magnetic circuit have calculated in an analytical model. This point depends on remnant flux density and silicon steel material. In this paper, the magnet of LSPMSM has been carried out by analytical model in fig 1 and NdFeB35 is magnet material used due to its good thermal stability and remnant flux density (~1.3T) allowing its use in applications exposed to high temperature about 180^0C. The flux density is estimated about 0.88 T.

*Corresponding Author: Bui Minh Dinh, Ha Noi University of Science and Technology, Email: dinh.buiminh@hust.edu.vn

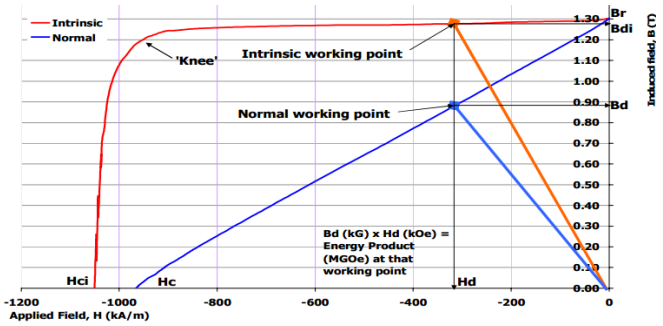


Figure 1 Magnetic properties of NdFeB35

Based on the analytical method, some geometry parameters of stator and rotor can be calculated as follow chart in fig 2. An analytical model was undergone many calculation steps to define basic parameters. Based on torque volume density TVR from 20 to 30 kNm/m3 [5], if we assume rotor diameter equal to rotor length, the rotor diameter D and length L sizes of LSPMSM is determined as follow:

$$D = L = \sqrt[3]{\frac{T}{\frac{p}{4}, TVR}} = \sqrt[3]{\frac{47.75}{\frac{3.14}{4}, 25}} \quad (1)$$

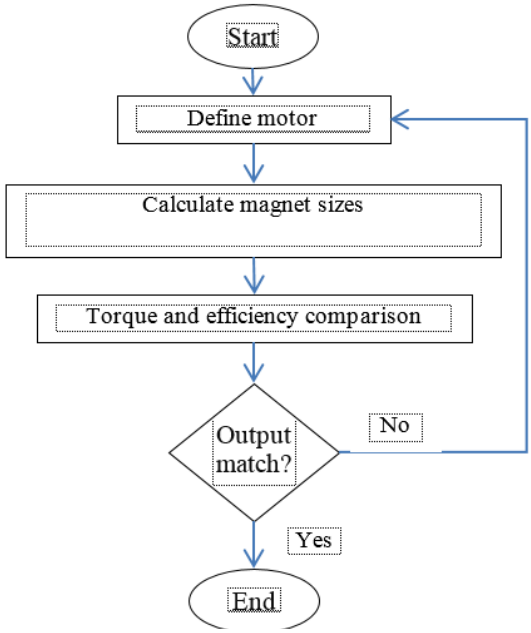


Figure 2 Calculation process

In general, the design process of LSPMSM is similar to that of induction motor. The main parameters (such as outer diameter, rotor diameter, motor length, stator slot, air gap length) are defined by taken into account some practical factors with desired input requirements [3]. The main part of the process is to design the rotor configuration which is embedded permanent magnet. The PM configuration needs to create sufficiently magnetic voltage for magnetic circuit. In fact, there are some possible configurations sorted by the shape and position of PM inside rotor as listed below:

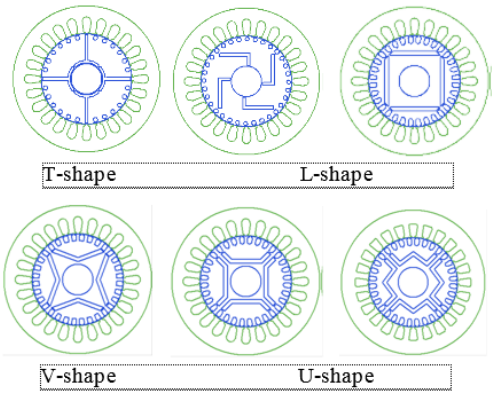


Figure 3 LSPM PM configurations

According to Ilhan Tairimer's research in 2009 [4], the PM configuration has a great influence on motor efficiency. He also pointed out that the V-shape PM gives the highest efficiency. However, the decision on choosing PM configuration depends on the possibility of manufacturing it. In other word, the technique (includes the cost of procedure) and material properties are crucial criteria that impact the configuration selection. By considering them, the I-shape PM is applied in this design.

Table I. Motor parameters

Parameter	Value	Unit
Air gap length	0.5	mm
Stator inner diameter	155	mm
Stator outer diameter	240	mm
Rotor outer diameter	154	mm
Rotor shaft diameter	40	mm
Rotor/stator length	135	mm
PM thickness	8	mm
PM width	60	mm
Number of PM piece	4	-
Stator slot	36	-
Slot space factor	0.48	-
Air gap flux density	0.95	T
speed	1500	rpm

Copper and iron losses are determined efficiency and copper losses in stator winding are biggest, other losses are also calculated and shown in Table II.

Table II. Loss calculation

Loss	Value	Unit
Copper loss	160.5	W
Stator teeth loss	8.1	W
Stator yoke loss	7.6	W
Bearing loss	12.0	W
Windage loss	14.4	W
Additional loss	6.4	W
Total loss	219	W
Efficiency	94.5	%

Some layout of LSPMSM with I-shape are shown in fig 4 due to manufacturing technologies and high flux density in air gap of this rotor configuration.

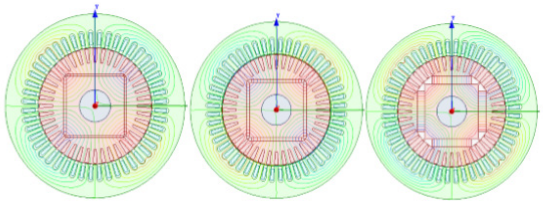


Figure 4 Layout of LSPMSM rpm with different magnet sizes

3. Optimal design of LSPMSM by genetic algorithm

To apply the GA approach, an objective function has to be defined to evaluate how good each motor design is. This objective function may include all the geometrical dimensions of the motor and a large subset of constraints (geometrical constraints) have to be formulated to ensure the physical feasibility of the motor. LSPMSM efficiency is determined as in (2):

$$\eta = \frac{P_{out}}{P_{out}+\sum \Delta P} \quad (2)$$

With: P_{out} is output power which is calculated by:

$$P_{out} = \frac{n_{DC}.M_T}{9.55} \quad (3)$$

n_{DC}: motor speed in steady state (rpm) (determined by FEA)

M_T: load torque in steady state (N.m) (determined by FEA)

$\sum \Delta P$ is total loss in motor.

$$\sum \Delta P = \Delta P_{stator} + \Delta P_{Fe} + \Delta P_{mec} + \Delta P_{add} \quad (4)$$

ΔP_{Fe} core loss in stator and rotor laminations (by FEA)

ΔP_{stator} mechanical loss

ΔP_{stator} additional loss

In order to evaluate efficient objective in 10 scale, the objective function is determined by the following:

$$G_1 = 1 + 9 * (1 - \eta) \quad (5)$$

When $\eta = \eta_{min} = 0$ function value will be 10

$\eta = \eta_{max} = 1$ function value will equal to 1

Second objective function is permanent magnet volume function.

Due to the motor is considered in 2-D planar environment, the permanent magnet (PM) volume calculation is transformed into PM cross-sectional area and PM stacking length is considered constant.

To evaluate this objective in 10 scale, objective function is determined by the following:

$$G_2 = 1 + \frac{9}{S_{max}-S_{min}}.(S_{nc} - S_{min}) \quad (6)$$

Finally, the cost function

$$Cost = \varepsilon_1 \cdot W_1 + \varepsilon_2 \cdot W_2 \quad (7)$$

Where:

$\varepsilon_1 = (G_1 - 1)$ error of efficiency target

$\varepsilon_2 = (G_2 - 1)$ error of magnet volume target

W_1, W_2 weight factors.

Because the permanent magnet cost is more important than other components, the factors are selected W_1=0.8 and W_2=0.2.

4. Optimal results comparison

4.1. Optimized Program Results

Cost factors can be obtained from magnet sizes, based on those results, optimal design parameters will be determined in Table III.

Table III. Cost factor results

No	**Rib (mm)**	**Lm (mm)**	**Wm (mm)**	**Cost**
270	1.35	2.595	54.13	0.57922
323	1.828	2.615	55.03	0.58458
190	1.542	2.5	58.25	0.58826
273	1.756	2.657	54.27	0.58846
157	1.782	2.735	51.98	0.59382
320	1.887	2.57	58.45	0.60239
317	1.497	2.685	49.52	0.60727
327	1.313	2.622	57.71	0.60985
382	1.889	2.913	48.7	0.6134
301	1.281	2.926	49.09	0.61626
245	1.498	2.95	49.11	0.61809
184	1.002	2.92	48.25	0.61933
213	1.666	2.975	48.49	0.62213
123	1.483	3.062	46.76	0.63403
375	1.015	3.06	47.41	0.63436
282	1.923	2.818	55.09	0.63493
251	1.89	2.632	60.08	0.6444
221	1.087	2.785	47.14	0.64621
163	1.822	2.934	53.06	0.64817
244	1.679	3.188	44.48	0.65168

The optimized value distribution results are obtained by the following figure 5.

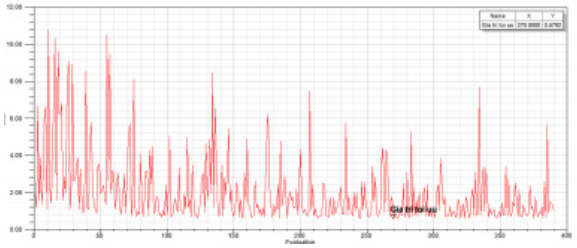


Figure 5 Distribution of optimized value of LSPMSM 7,5 kW

The optimized value is 270 which has the cost valued equals to 0.5792.

4.2. Revaluate optimized results

With the optimized results:

R_{ib} = 1.35 (mm), L_m=2.595 (mm); W_m = 54.13(mm)

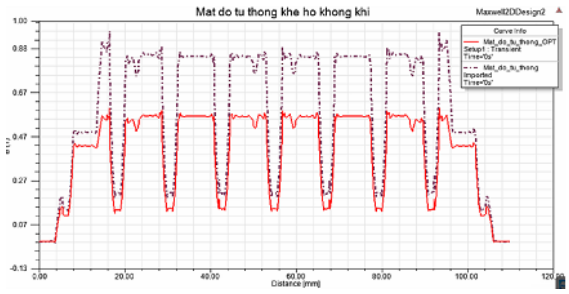
The volume has been decreased 2.5 times compared to original design, notice that PM depth is still unchanged.

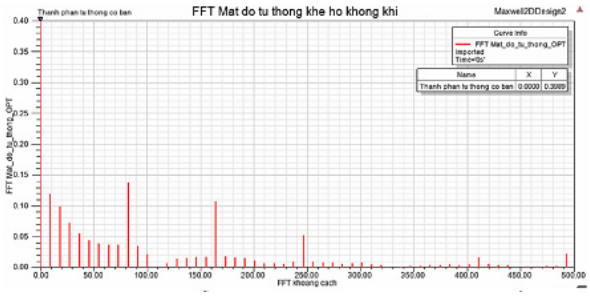
$$V_{nc}^{opt} < V_{nc}$$

In order to be convenient to compare, another results will be shown between before and after running optimization program.

After running simulation by electromagnetic circuit model with optimized parameters, the results are:

a) Flux density and e.m.f





Air gap flux density after optimization is smaller than before due to the decrement of PM width. This proves motor doesn't require large PM energy as before optimized results.

$$W_m^{opt} < W_m$$

$$B_g^{opt} < B_g$$

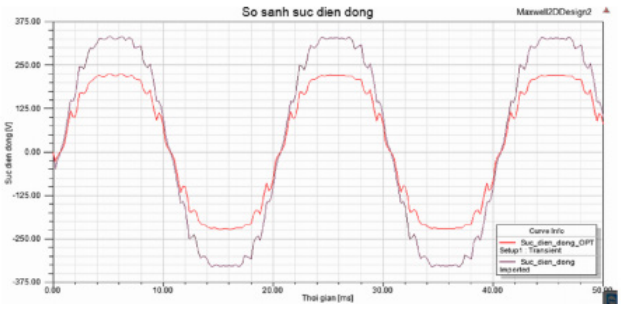


Figure 8 Comparison between e.m.f before and after optimization

Optimized e.m.f is smaller than before optimization corresponding to smaller air gap flux density. Electromagnetic force will cause negative effect of braking torque and dynamic characteristics will be better, faster starting time with the same load torque. In the other hand, peak wave shape will be smoother.

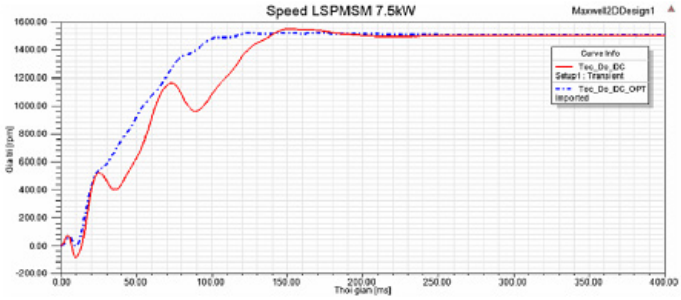


Figure 9 Comparison between speeds starting characteristic before and after optimization

Speed starting characteristic of LSPMSM 7.5 kW after optimization is much better than the previous value with faster stable speed and less ripple.

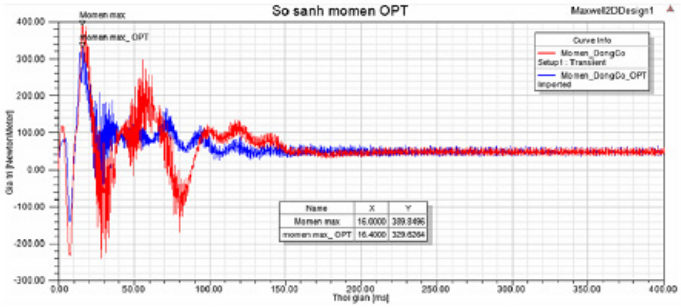


Figure 10 Comparison between speed starting characteristic before and after optimization

Torque starting characteristic of LSPMSM 7.5 kW after optimization is also better than before optimization value, with smaller torque ripple amplitude. It can also be considered from speed starting characteristic.

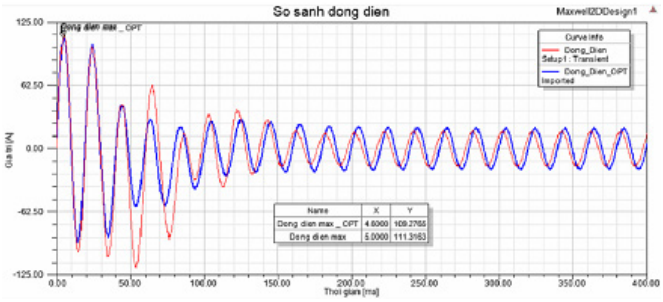


Figure 11 Comparison between starting current before and after optimization

From the figure, it can be seen that the starting current after optimization is smaller than the before value but higher steady-state value due to higher power factor.

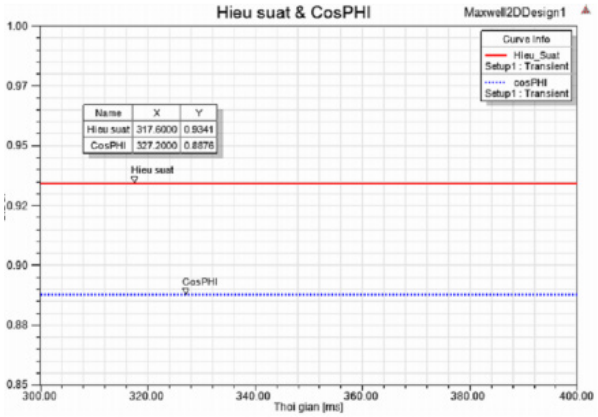


Figure 11 Efficiency and power factor after optimization

According to figure, it can be seen that, after optimization parameters provide higher efficiency. Smaller power factor is caused by the remarkable reduction in PM volume after optimization, however, the new value satisfies requirement that power factor has to be higher than 0.85.

$$\cos \varphi_{rate}^{opt} < \cos \varphi_{rate} \quad (9)$$

5. Calculate temperature rise using thermal model

There are currently some thermal models can be applied to calculate operating temperature of motors. In this research, we chose Lump parameter model [5] to calculate temperature rise of 8 common motor nodes due to its uncomplicated method and rough precision.

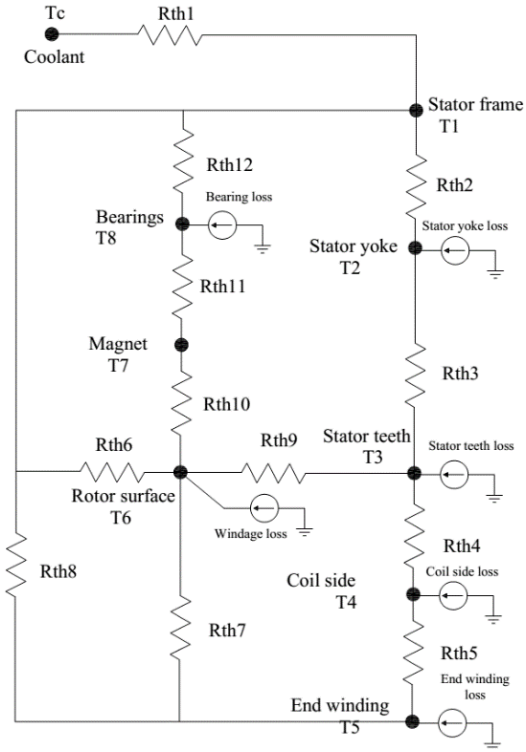


Figure 12 IPM thermal model network

For simplicity, it proposes some following assumptions [3]:

- The heat flows from the slots to the teeth but not directly to the yoke.

- The temperature of copper is assumed constant in every cross-section of the winding, but it varies in the axial direction.

- In the tooth, the temperature is assumed constant in the axial direction.

- Resistive losses are distributed uniformly in the winding.

In order to evaluate the temperature rise, we firstly calculate individually thermal resistance of motor parts, which is affected by its material and parameters. Then by taking them into the following equation with particular losses, we obtain the temperature of each desired node:

Table IV Temperature calculation results

Tc	T1	T2	T3	T4	T5	T6	T7	T8
50	50.0	50.7	51.1	56.0	56.4	68.8	71.0	71.2

If the coolant temperature is assumed to be 50 0C, temperature of the hottest points (usually the permanent magnet and bearings) are under 71 0C. That means with air cooling, the temperature rise is 21 0C at maximum. This result allows the PM to operate well and gets rid of demagnetization.

However, this design is still not really reliable. To validate the design, we'll input all these parameters into RMxprt to see if the output is matched or not.

Using all the calculated parameters and initial input (rated speed, rated voltage, rated output power) as input for RMxprt, we obtained the operating statistic of the motor. Because the design of permanent magnet can have major influence on the motor. An insufficient size of PM could force the magnet to operate in higher flux density which is not recommended or even not adapt the rated output power if the value exceeded remnant flux. Thus, what we must care most about are output power, RMS line current, efficiency and magnetic data.

6. FEM Model Results

FEM has been applied to investigate magnetic performance of LSPMSM design. The flux density of stator and rotor has been validated by FEA model for one pole in fig 5. Flux density curve vs rotor angle is shown in fig 6. Average values are 0.5 tesla in good agreement with calculation.

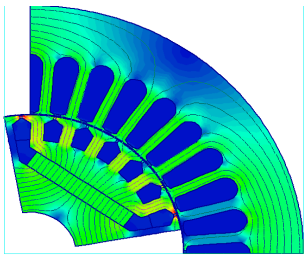


Figure 13 Flux density distribution

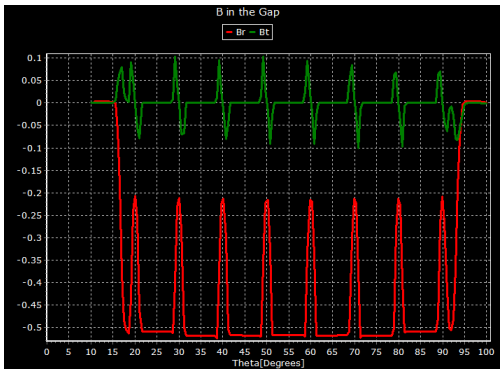


Figure 14 Flux density in air gap results

Magnetic circuit is obviously not saturated and magnet flux density is also adequately high. That allows the motor to operate in overload mode which has not been defined yet.

Torque and efficiency of LSPMSM design play an important role on implementation of this motor. The torque curves of rotor cage, magnet and motor are shown in fig 7.

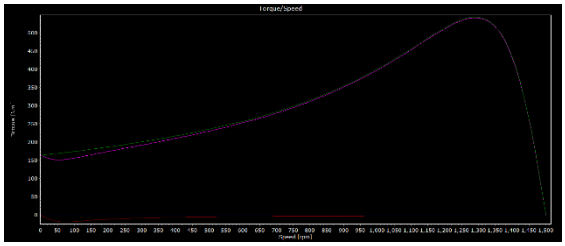


Figure 15 Torque of cage, magnet and motor results

Efficiency of 91% with 10% overload in fig 8. The power shaft is 8 kW started direct from grid. From those results can be concluded that LSPMSM is higher efficiency with IE 2, 3 than induction motor. Total cost of LSPMSM is more expensive than IM but lower operation cost of electric energy can help LSPMSM payback in short time.

Temperatures of rotor and stator tooth, yoke and windings have been calculated under natural convention. The results of analytical and model are good agreements.

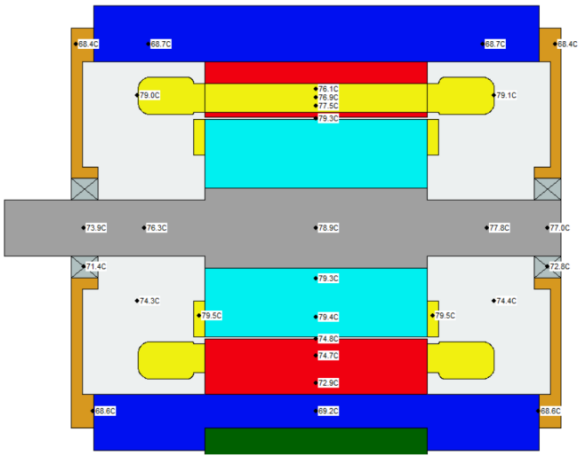


Figure 18 Temperatures of stator and rotor

7. Hardware test bench setup

Rotor magnetic slots have manufacture by wire cutting after die-casting rotor bars and shaft assembly as fig 9.

Figure 16 Rotor magnet slots

The whole hardware of LSPM motor was built together as fig 10.

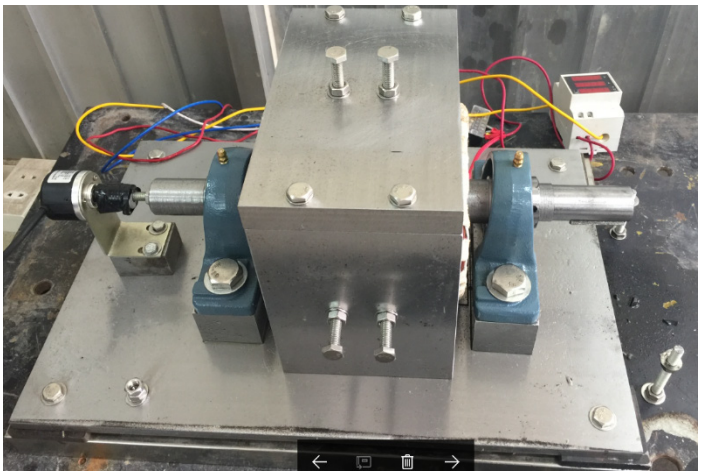


Figure 17 Hardware of LSPM motor

In order to control air gap of rotor and stator, adjustable screw have used to change stator position in three sides. The rotor shaft was connected to encoder to measure speed of motor though NI card and LabVIEW software. Especially, it can measure the dynamic speed performances to validate period of time for speed synchronizing. The LSPM motor was setup to evaluate synchronizing speed under different load and voltage by autotransformer of 10 kW. The transient time for constant speed is less than 0.8 second.

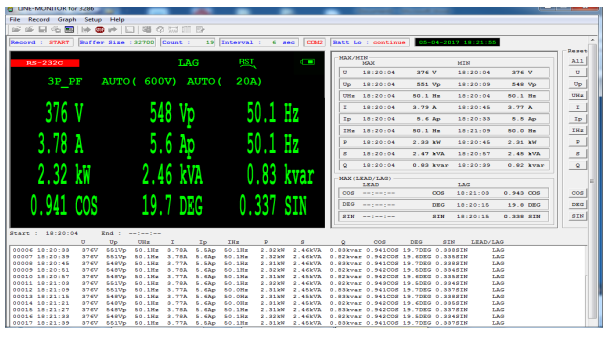


Figure 19 Experiment results of LSPMSM test

8. Conclusion

The paper has presented a comprehensive design of a LSPMSM for industrial applications. The design was calculated by analytical method, validated by Ansysmaxwell software. Particularly, thermal calculation was carried out to assure that the permanent magnet cannot be demagnetized under steady operation. Optimal design parameters have been calculated by GA program. Those results have evaluated by analytical and model methods. The hardware test bench has been built to evaluate synchronizing speed under different loads. The GA program results are LSPMSM design and manufacture cost with lowest price for mass production similar induction motor prices and transfer to industrial companies.

Acknowledgment

This research was supported the National Research Foundation of Vietnam funded by the Ministry of Science and Technology under Program KC.05/16-20.

References

[1] Kwangsoo Kim; Seung Joo Kim; Won Ho Kim; Jong Bin Im; Suyeon Cho; Ju Lee, *"The optimal design of the rotor bar for LSPMSM considering the starting torque and magnetic saturation",*Electromagnetic Field Computation (CEFC), 2010 14th Biennial IEEE Conference on Year: 2010.

[2] Sorgdrager, A. J.; Wang, R-J.; Grobler, A. J. *"Transient performance investigation and Taguchi optimization of a line-start PMSM",* Electric Machines & Drives Conference (IEMDC), 2015 IEEE International, On page(s): 590 – 595.

[3] A.H. Isfahan, S.V. Zadeh (2011) Effects of Magnetizing Inductance on Start-Up and Synchronization of Line-Start Permanent-Magnet Synchronous Motors. IEEE Transactions on magnetics, vol. 47, no. 4.

[4] A.H. Isfahani, S.V. Zadeh, M.A. Rahman (2011) *Evaluation of Synchronization Cappability in Line Start Permanent Magnet Synchronous Motor.* IEEE International Electric Machines & Drives Conference.

[5] J.R.Hendershot, T.J.E. Miller. "Design of brushless Permanentmagnet motors". Magna Physics publishing and Clarendon press-Oxford1994.

[6] P. Ji, W. Song,andY.Yang, "Overview on application ofpermanent magnet brushless DC motor," Electrical MachineryTechnology,vol.40, pp.32-36,Feb.2003.

[7] John Wiley & Sons Ltd, "Design of rotating electrical machines", 2008, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom.

[8] Ilhan Tarimer (2009), "Investigation of the Effects of Rotor Pole Geometry and Permanent Magnet to Line Start Permanent Magnet Synchronous Motor's Efficiency", Muğla Turkey.

[9] Joachim Lindstrom, "Thermal Model of a Permanent Magnet Motor for a Hybrid Electric Vehicle", April 1999, Department of Electric Power Engineering, Chalmers University of Technology, Goteborg, Sweden.

[10] Soo-whang Baek, Byung-taek Kim, and Byung-il Kwon,"Practical Optimum Design Based on Magnetic Balance and Copper Loss Minimization for a Single-Phase Line Start PM Motor" IEEE TRANSACTIONS ON MAGNETICS, VOL. 47, NO. 10, OCTOBER 2011.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1188-1195 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Fault-Tolerant in Embedded Systems (MPSoC): Performance Estimation and Dynamic Migration Task

Kamel Smiri[*1,3], Habib Smei[1,2], Nourhen Fourati[1,3], Abderrazak Jemai[1,4]

[1]*Université de Tunis El Manar, Faculté des Sciences de Tunis, Laboratoire LIP2, 2092, Tunis, Tunisie*

[2]*Direction Générale des Etudes Technologiques, Institut Supérieur des Etudes Technologiques de Rades, Rades, Tunisie*

[3]*Université de Manouba, Institut Supérieur des Arts Multimédias Manouba, Campus Universitaire Manouba, 2010, Tunisie*

[4]*Université de Carthage, INSAT, B.P. 676, 1080 Tunis, Cedex, Tunisie*

ARTICLE INFO

Article history:
Received: 09 April, 2017
Accepted: 24 June, 2017
Online: 20 July, 2017

Keywords:
Hybrid Model
Fault tolerance
Dynamic migration Task
Multiprocessor Systems-on-Chip (MPSoC)

ABSTRACT

Multiprocessor Systems-on-Chip (MPSoC) allow the implementation of heterogeneous architectures with a high integration capacity. In recent years, computational requirements MPSoC are increasing exponentially. This complexity, coupled with constantly evolving specifications, has forced designers to consider intrinsically flexible implementations. Deploying applications typical of multimedia domains is difficult, not only due to the heterogeneous parallelism in the platforms, but also due to the performance constraints that typify these systems. An application can be modeled as a set of cooperative tasks. A task can be implemented in software or in hardware depending on its complexity and the involved cost. Our proposal is a fault tolerance approach which combines the results of a performance model and a technical's fault tolerance. We interest of the dynamic migration task to resolve the Fault-Tolerant for Multiprocessors Embedded System. We exploited an example of multimedia application (MJPEG decoder) to find optimal Fault tolerance systems. Our aim in this paper is to exploit the classic technique of fault tolerance. The solution chosen is the transformation of software processing into hardware processing. And also, exploitation of hybrid models (simulation/analytics). The goal is to have a Fault Tolerant Embedded System.

1. Introduction

This paper is an extension of work originally presented in 11th International Design & Test Symposium, IDT 2016 [1].

Multiprocessors System-on-Chip (MPSoC) is a system that contains several heterogeneous processors interconnected around advanced communication architecture as a Network on Chip (NoC). Nowadays, Multimedia applications use MPSoC architectures in order to achieve high performance. These architectures may include different types of computation units (DSP, Microcontroller...) and use different diagrams of communication.

We distinguish in the literature [2,4] three design approaches for MPSoC: (1) platform based design, (2) components based design and (3) synthesis system. In this paper we are going to focus on the first approach that targets the implementation of an application on configurable prototype architecture.

We follow along our work the MPSoC hardware/software design flow proposed by [5]. The software and the hardware are then gradually refined and simulated on four levels of abstraction. These levels are System Level, Virtual Architecture, Transaction Accurate and Cycle Accurate. System Level (SL) is the highest level of design flow. It provides an environment for parallel execution. The application is described as a set of processes that exchange data only through the FIFO blocking point to point. Virtual Architecture (VA) level is more detailed than the standard SL. The major difference is that it includes information on the architecture. For abstracting the underlying hardware architecture, the code of tasks using APIs known as HdS (Hardware dependent Software). Primitives of these APIs (ex. Send/Receive) access to explicit communication components. Transaction Accurate (TA) level is the second intermediate level. In this level, the tasks are linked to an Operating System (OS) and a library of communication to implement the communication and management services tasks. This software layer uses primitives of Hardware Abstraction Layers (HAL). The data transfer uses explicit addresses. The hardware platform is more detailed, including a communications network, the devices accessed by HAL-API, and an execution model of abstract processor. Cycle

[*]Corresponding Author: Kamel Smiri, Université de Tunis El Manar, Faculté des Sciences de Tunis, Laboratoire LIP2, 2092, Tunis, Tunisie
Email: smiri_kamel@yahoo.fr

Accurate (CA) level is the lowest level. At this level, the HAP-API is implemented. This implementation is done by adding a HAL software layer and a processor for each sub-system (for example, we create an implementation and a model for ARM, MIPS, and SPARC...).

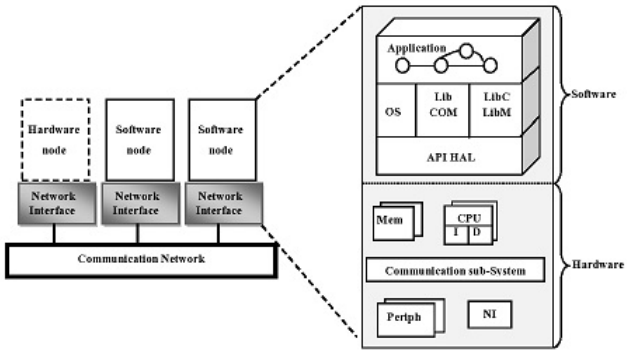


Figure 1 Typical architecture of MPSoC systems

Figure 1 present a typical architecture of MPSoC which is composed by software and hardware nodes connected by a communication network. Hardware nodes are component which does not have the ability of programming. A communication network connects all the nodes together. Software nodes provide the execution environment for applications tasks.

The software node is structured into several layers:

- Application Layer, this layer represents a set of tasks of the application which are carried out in parallel, in order to profit to the maximum of the parallelism offered by architecture multiprocessors. This layer communicates with the software node by calling on primitives of the operating system (OS).

- The Operating System OS, the operating system makes it possible to provide an interface between hardware equipment and the application.

- Libraries, the objective to use the library of communication is to call on the channels of communication and to use the various features of the library C standard and the library of mathematics.

- Hardware Abstraction Layer, this layer makes it possible the operating system to interact with the material peripherals on a level abstract rather than on a detailed material level.

The motivation of the paper is to explore the design space (Design for Space Exploration DSE) by early in the design cycle. We present in this paper a methodology for the migration of a software task in hardware component of a multimedia application at Transaction Accurate level and at Cycle Accurate level. The chosen application to illustrate our approach is the MJPEG decoder that is characterized by its high treatment density and important data exchange. This paper is organized as follows. Section 2 presents the related work. Section 3 gives an overview of Fault-Tolerant in Embedded Systems (MPSoC) approach. Section 4 presents the experimentation of MJPEG Decoder, with the objective to show the efficiency of our approach.

We propose in this paper an extension of the paper conference (IDT 2016) [1]. Mainly, we have added the following parts: a) An advanced synthesis on related works, b) A detailed description of our dynamic migration approach (based on Fault-Tolerant in Embedded Systems (MPSoC)), and c) A series of experiments to illustrate the effectiveness of our approach. We have added some experimentation that shows the effectiveness of our approach. The MJPEG decoder (which is a refresh application) is chosen to get me the dynamic migration efficiency. The IDCT task is chosen to show the gain during dynamic migration of the software to the hardware.

2. Related works

Due to the complexity of multiprocessor systems, the probability of failure is all the more important that it requires special consideration. Indeed, during a failure, a part of the application state disappears and the application may pass in an inconsistent state that prevents it from continuing normal execution. To perform effectively and properly, it must withstand stresses related to the execution environment. It must be able to tolerate the failure of one or more components. Fault tolerance is the property that keeps the smooth functioning and continuity of the system during a failure. So the intensity of the fault is not proportional to the decrease of the system load operation since found contents a very small mistake can cause a total blockage in a system designed.

There are many causes for failures in distributed systems. In [2], the author focused on solving crash failures in GRID computing, their proposal aims to ease the process of recovery when system failures are detected at runtime avoiding the necessity for application restarts. Their proposal works through a set of services that performs transparent task migration over the computing nodes, hiding the complexity related with error handling when a hybrid programming model based on Open MPI and OpenCL is employed.

In [3], the author proposes scheduling algorithm of the adaptive fault tolerant tasks which is a combination of two fault tolerant algorithm TMR (Triple Modular Redundancy) and DMR (Dual Modular Redundancy) and it also benefits from EDF and LLF task scheduling algorithms for decreasing miss rate of tasks. The main goal of these algorithms is to find a proper tasks assignment on the cores and tolerate processing component failures which could either be homogeneous or heterogeneous, so this algorithm is more perform than DMR and TMR methods in average 35% which can guarantee the proper execution of running tasks and reduce the total time of task execution.

In [4], the authors proposes an offline task remapping technique to minimize the communication energy and task migration overhead of an application mapped on a heterogeneous multiprocessor system for all processor fault-scenarios. This technique involves two steps: the first is communication Energy driven Design Space Exploration (CDSE) to select an initial mapping and the second is communication energy and Migration overhead aware Task Mapping (CMTM) for different fault-scenarios. The results show that the proposed CDSE reduces space exploration time of Conception 99% and the reduced energy CMTM 35% average communication and migration over by an average of 20% for all scenarios of single and double faults compared to tolerance techniques existing faults.

In [6], the authors proposes a design flow to explore mapping strategies and shape of the NoC network to improve synchronization performance and the power consumption, then for each step of the flow they used a heterogeneous platform (PC and

FPGA) to fulfill them. The exploration flow is provided to help designers NoC to choose a technique appropriate mapping tasks on a NoC appropriate form for a particular application. The experiments demonstrate that the most appropriate task mapping strategy and the most suitable NoC shape strongly depend on the algorithm used. Depending on the timing latency results obtained and the FPGA resources used, the designer can select the appropriate task mapping strategy on the suitable shape in a short exploration time and with precise timing evaluation.

In [5], the author proposed a methodology to master the migration process at transaction level and they demonstrate the various levels of the design process to deals with the impact of task migration as an alternative to meet performance constraints in the design flow. And, they use a SDF3 tools to provide performance estimation at transaction level. Using the SDF3 tool, the authors model a multimedia application using SDF graphs. Secondly, they target an MPSoC platform. The authors take a performance constraint to achieve 25 frames per second.

Table 1: synthesis works of fault tolerant

	[3]	[4]	[5]	[6]	[7]	[8]
GC	++	- -	- -	- -	- -	--
MPs	- -	++	++	++	++	++
AM	- -	- -	- -	- -	++	--
SM	- -	- -	- -	- -	++	--
OS	- -	++	+	+	++	+
SMt	++	- -	++	++	++	--
DMt	++	- -	++	++	++	++
IPb	- -	- -	- -	++	++	--
FT	++	++	++	++	- -	+

GC: Grid computing SMt: Static Migration task

AM: Analytic Model DMt: Dynamique Migration task

SM: Simulation Model IPb: IP blocks

OS: Operating Systems, FT: Fault-Tolerant, MPs:MPSoC systems

In [7], the team of Tahir maqsood are conducted a detailed quantitative evaluation of the selected dynamic task mapping algorithms for NoC based MPSoCs for a wide range of mesh sizes with varying task and communication loads. Comparisons are conducted with varying network load, number of tasks, and network size for constantly running applications. Moreover, they propose an extension to communication-aware packing based nearest neighbor (CPNN) algorithm that attempts to reduce communication overhead among the interdependent tasks. The results indicate that proposed mapping algorithm reduces communication cost, average hop count, and end-to-end latency as compared to CPNN especially for large mesh NoCs. And the proposed scheme achieves up to 6% energy savings for smaller mesh NoCs. Finally, results of formal modeling indicate that proposed model is workable and operates according to specifications.

Synchronous Data Flow SDF is a particular case of the data flow and Petri network. Formalism SDF is often used for the entire predictable systems, in which the number of samples of data consumed and produced by each node with each execution, is specified a priori. Graphs SDF constitute the central element of modeling by SDF3 [8]. Moreover, SDF3 tool integrates several commands, for each one has a precise function. Indeed, the command sdf3flow is most interesting of tool SDF3 since it makes it possible to deploy an application graph with a constraint on architecture graph [9].

System on Chip Library (SoCLib) is an open platform for the virtual prototyping of MPSoC systems. SoCLib is a tool based on the simulation which it allows the development of platforms of virtual prototyping and facilitates the exploration of architectures by means of software tools to conceive embedded applications and of library Open Source is made up of models of high level simulation written in SystemC for the material components, which allow fast simulations [10].

Following our synthesis established in Table 1, we propose to use the contributions proposed in the paper (4) and paper (5), which we will be useful to propose a new fault tolerance approach based on dynamic migration for Embedded Systems.

3. Fault Tolerance Approach for Embedded Systems

Fault tolerance approach for embedded systems multiprocessor (Multiprocessor System on Chip MPSoC), combines the results of a hybrid model (proposed by LIP2 laboratory, in [6]) and a technical fault tolerance that is based on the dynamic migration. This approach provides an effective solution to the current problems of modern embedded systems.

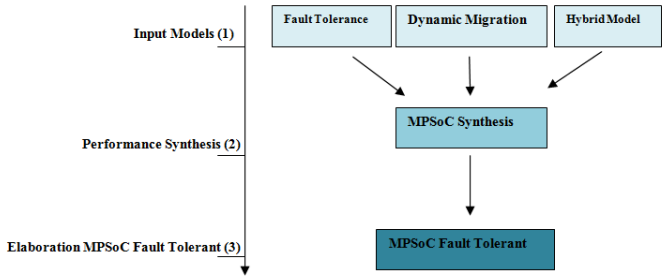


Figure 2 Fault tolerance approach for Embedded Systems

Our Fault tolerance approach for Embedded Systems is structured in three steps: 1) implementation of input models (Hubrid model, Dynamic migration technique, Fault tolerance technique), 2) performance synthesis and 3) Elaboration MPSoC Fault Tolerant.

3.1. First step: Input Models

The first step shows the essentials components of our approach citing hybrid model and fault tolerance techniques based on dynamic migration. The hybrid model (MH) resulting hybridization of two performance estimation methods: Analytical Method (MA) and Method based on simulation (MS). This model allows the exchange of information between the two methods MA and MS through an intermediate bonding layer whose goal is to treat various types of systems and reduce the time estimation and evaluation of performance and a low error rate. So, the aims of the MH model, is to combine the strengths of both MA and MS methods.

Note that the analytical method is characterized by the speed of optimal solution. It is based on the exploitation of SDF graph,

this formalism to model the software applications, hardware and software requirements for deploying tasks on hardware resources. The analytical model used in this approach is a synchronous data stream which is a particular case of the data stream (Data Flow) and Petri network, to generate the mapping of the software application on the hardware platform, not only but also the usage rate of hardware resource which can be regarded as performance indicators.

The simulation-based method is characterized by the precision of the results. So in this model the designer must perform three tasks: Define the structure of the software application, Set the hardware architecture and Check the deployment of the software application on the hardware platform. Concerning technical fault tolerance that is based on dynamic migration, we propose to use migration software / hardware in a multiprocessor architecture, as a solution for fault tolerance.

3.2. Second step: performance synthesis

For the second step is the synthesis step we will make a complete architectural exploration of which will choose a multiprocessor hardware architecture by specifying these settings, A software application and we will follow the simulation and implementation of this application on the architecture while retaining control over the brutal mistakes that makes system failures.

3.3. Tree step: Elaboration MPSoC Fault Tolerant

The last step of our approach that will get a MPSoC fault tolerant. This step involves a component fault monitor CCP noted that the state's objectives of monitoring the system running. This hardware component in two missions which should take care of: detecting hardware failures and handling. CCP's interest is to eliminate any dependency on the rest of the system, consequently control failures will be independent of the processing part or in cases of failures remains functional and it keeps the communication is with the rest of system or with the outside (network). At the end of our proposed approach to fault tolerance in embedded systems multiprocessor allows to restructure the system into two subsystems: the first subsystem dedicated to the execution of the application and the second a dedicated subsystem fault tolerance.

Migration is the solution for fault tolerant problem in MPSoC. The dynamic migration is structured on four steps: (1) modeling software by KPN; (2) mapping and execution; (3) performance evaluation migration.

- *Modeling software by KPN*

First of all, we have modeled a software application on a target multiprocessor architecture. The application is a set of tasks interconnected by communication channels (FIFO) each architecture used contain one or more hardware accelerators where migration can be activated for critical tasks. And, we also fix the performance constraints that must be satisfied.

- *Mapping and execution*

In the second step, we make mapping of the software part on the hardware platform that contains the hardware component. We rely on profiling results of the application to detect the task requiring migration. This task will be transformed into hardware coprocessor to accelerate the processing of the complete system.

- *Performance evaluation of Migration*

To migrate a task from the software to a hardware accelerator, we propose three steps that need to be gone through: (1) Hardware description, it is to design a hardware component and integer throughout the MPSoC system; (2) The instantiation of the task as a hardware task; (3) The deployment of the task as hardware task.

When you decide to execute hardware a task, we start the backup from its current state including the control and the task selected in the shared memory information. Then the task migrated will be deleted or suspended on the source tile.

4. Experimentation: MJPEG Decoder

The figure 3 present Our architecture is formed by two processors ARM, a coprocessor for the TG task, a coprocessor for the RAMDAC task and a coprocessor for the migrate task.

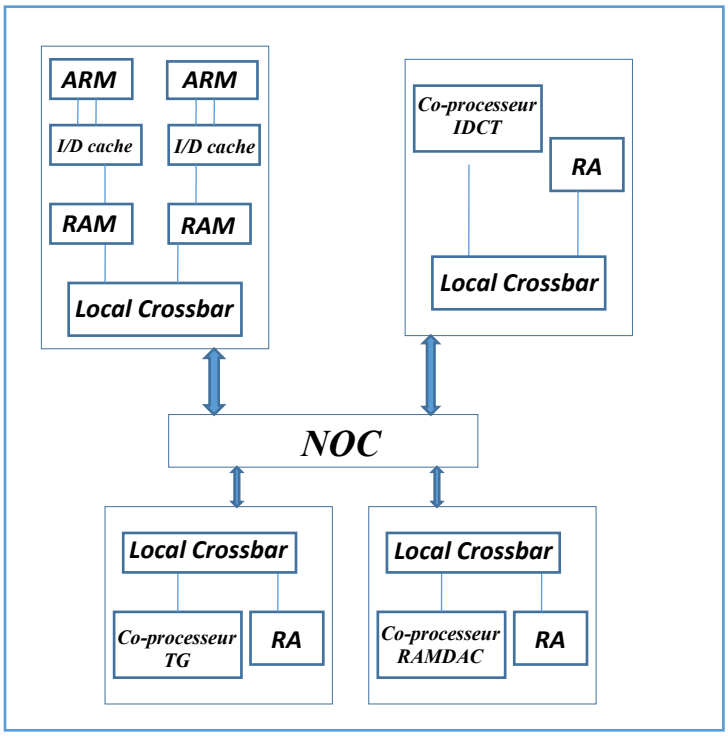


Figure 3 Architecture Hardware Platform for Embedded systems

One has validated the Fault Tolerance Approach for Embedded Systems using a Motion JPEG decoder. One wants to find an implementation of the MOTION JPEG decoder realising 25 frames per second (fps) as a functional constraint and using 50MHz processors as a non-functional constraint. The MOTION JPEG decoder reads a stream of JPEG images from an input peripheral: a traffic generator named TG and writes pixels on an output peripheral: a digital-to-analogue converter named RAMDAC. The first functional bloc DEMUX dispatches the input stream to the other blocs. The decoding chain (Fig. 5.a) begins with the decompression bloc (VLD), then the zigzag rearrangement bloc (ZZ), followed by the inverse quantification bloc (IQ). Finally, the stream passes to the inverse discrete cosine transformation (IDCT).

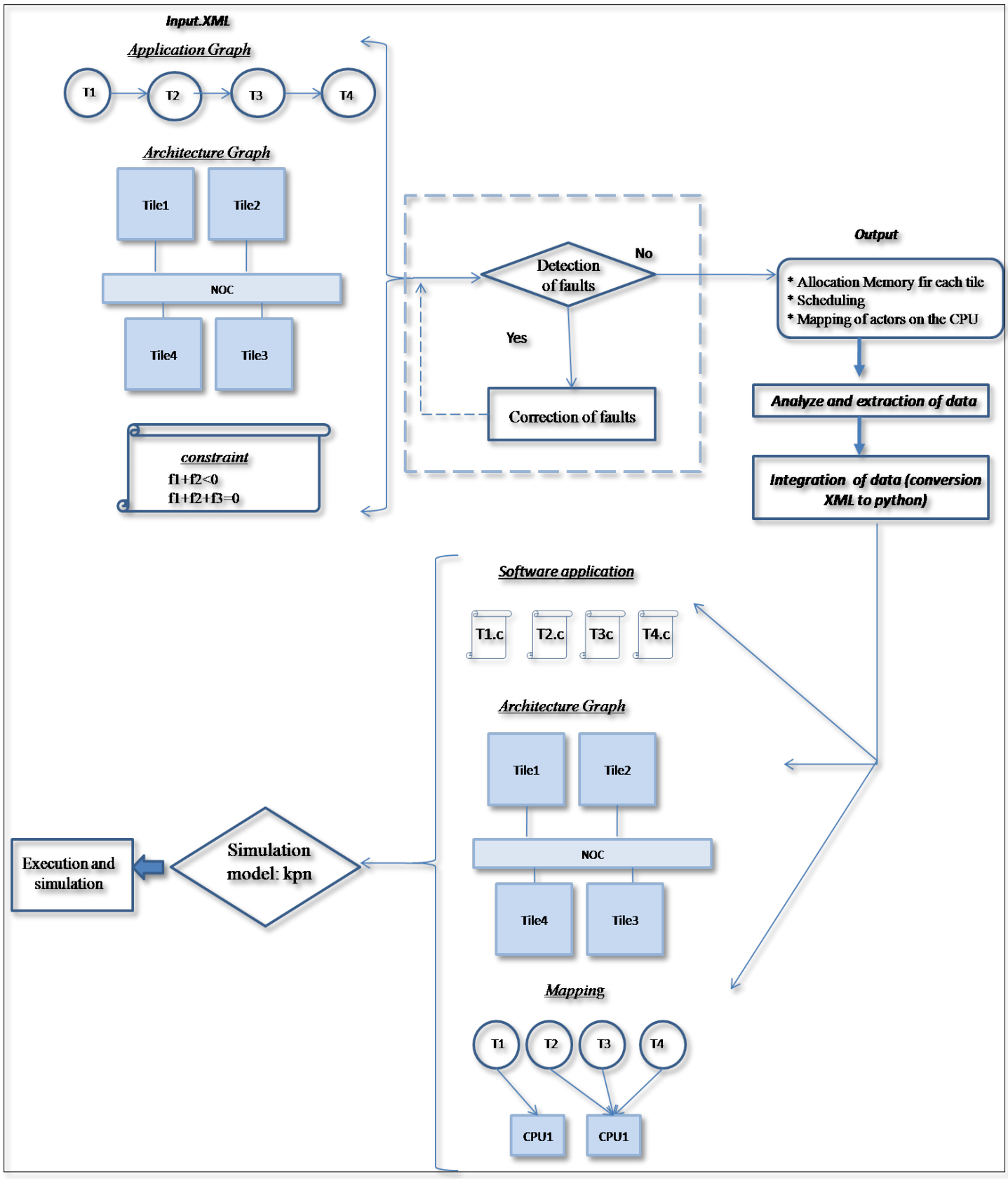
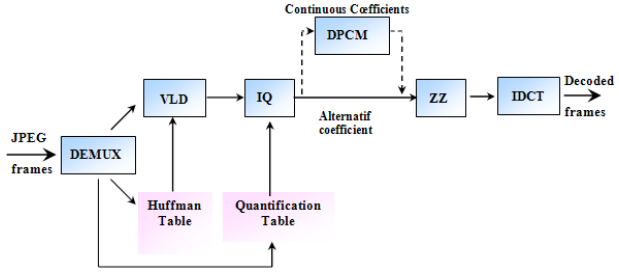
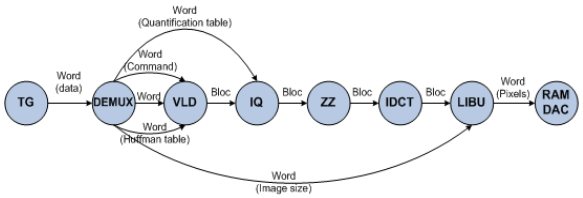


Figure 4 Fault Tolerance Approach for Embedded Systems

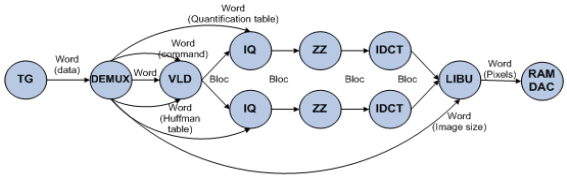
One has validated the Fault Tolerance Approach for Embedded Systems using a Motion JPEG decoder. One wants to find an implementation of the MOTION JPEG decoder realising 25 frames per second (fps) as a functional constraint and using 50MHz processors as a non-functional constraint. The MOTION JPEG decoder reads a stream of JPEG images from an input peripheral: a traffic generator named TG and writes pixels on an output peripheral: a digital-to-analogue converter named RAMDAC. The first functional bloc DEMUX dispatches the input stream to the other blocs. The decoding chain (Fig. 5.a) begins with the decompression bloc (VLD), then the zigzag rearrangement bloc (ZZ), followed by the inverse quantification bloc (IQ). Finally, the stream passes to the inverse discrete cosine transformation (IDCT).



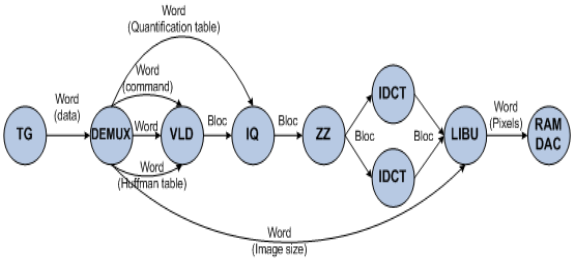
(a)



(b)



(c)



(d)

Figure 5: (a) Motion JPEG decoding principle, (b) model of Motion JPEG decoder, (c) 1st parallelization solution, (d) 2nd parallelization solution

Based itself on migration diagram presented in Fig. 4, we transform the software task IDCT to hardware component. The validation of hardware IDCT is carried out in two stages. The first is the validation of behavioral IDCT; by comparing the results of the implementation software IDCT in C with the equipment in SystemC. The second step is the validation of the communication interface that interconnects IDCT block with the rest of MPSoC system. The integration is made by adding a software/hardware channel between VLD-IDCT and a hardware/software channel between IDCT-LIBU. We limit ourselves to adapt to existing protocols KPN channels to have a coherent refinement. We have structured hardware IDCT component on three layers: (1) an adapter for easy connection to the rest of MPSoC system, (2) the interface module that manages the exchange of data between SystemC module and the rest of system, and (3) the true behavior of hardware component. The interface module is composed of two SystemC modules *Write* and *Read*, and two FIFO with finite sizes (64 KBytes). *Write* module written in the first FIFO data from the VLD. *Read* module permits to read the data processed by IDCT and sends it to the LIBU. Primitives wait ($t_{\_Write}$), wait ($t_{\_Write}$) and wait ($t_{\_IDCT}$) allow modeling of execution time modules respectively *Write*, *Read*, *IDCT*.

For aim of determining the timing of hardware IDCT and functions Write and Read on target technology (Xilinx, Altera ...), we explored GAUT tool, which is among architectural synthesis tools under constraints. For a cadence = 100ns and period = 10ns, the execution time of the IDCT is $t_{\_IDCT}$ = 100ns for the treatment of Macro-Block (8 × 8 pixels). We estimate that the execution time of each module Write and Read are equal and is $t_{\_Write} = t_{\_Read}$ = 10ns.

2) Performance migration at CA level

The architecture that we synthesize at Cycle Accurate (CA) level is relative to the constraints laid down at the beginning of the problem. The first constraint is the functional constraint on the decoding speed of 25 frames per second (25 fps). The second constraint is the non-functional constraint on the platform that uses MIPS processors running at 50 MHz. The Operating System *Mutek* is a multiprocessor multithread kernel supporting POSIX APIs. This *Mutek* kernel proposes FIFO based communication layer allowing KPN communications. It includes three different schedulers. In the Symmetric Multiprocessor (SMP) kernel, there is only one scheduler allowing the KPN processes to migrate between processors. The Asymmetric centralized Scheduler (Non_SMP_CS) is only one scheduler affecting statically processes to processors. The Asymmetric distributed scheduler (Non_SMP_D) instantiates one scheduler to every processor and the processes are statically affected to processors. The design space exploration synthesis of kernel scheduling policies and software/hardware mapping are presented in Table 1. This table summarizes the performance of implementation of the MJPEG decoder in number of frames per second (fps).

Table 2: Performance implementation of MJPEG decoder (fps)

proc	SMP-CS				Non SMP-CS		SMP - CS		
	Seq	KPN	ZZ + IQ + IDCT	Double IDCT	Double IDCT	ZZ + IQ + IDCT	IDCT hard	IDCT+V LD hard	IDCT+ VLD+I Q hard
1	12,2	6,72	---	6,78	---	---	7,7	17	24,12
2	12,61	12,01	12,16	12,86	---	12,68	11,49	24,07	32,45
3	12,63	14,1	17,65	18,63	20,49	---	11,55	36,73	32,56
4	12,6	13,86	22,35	22,09	25,53	---	---	---	---

We decide to transform IDCT task in hardware component, which allows us to achieve a speed decoding of 24.07 fps with 2 MIPS processors and 36.73 fps with 3 MIPS processors on a simulation platform. The optimal architecture will use 2 MIPS processors configured with 4 KB memory cache, RAM memory, Pi-Bus and an interruption controller. The 2 MIPS processors are dedicated to execute DEMUX, VLD, IQ, ZZ and LIBU processes with an SMP scheduling policy. In addition, an input peripheral is needed for reading video stream TG and an output peripheral for display RAMDAC. For achieving the 25 fps functional constraint, hardware implementation of the IDCT processes is necessary. If the performance is meadows of 25 fps (for example 24 fps), one accepts the solution under condition to optimize the codes of application until to get the performance of 25 fps.

3) Performance migration at TA level

The modeling of the IDCT in hardware, adding time performance metrics ($t_{\_Write}$, $t_{\_Read}$ and $t_{\_IDCT}$) and the integration of the component with the rest of system, allowed us to perform simulations for three different input images (see Fig. 6.). We note that the best performance is obtained for a configuration of two CPU and hardware IDCT. This configuration agrees well with the results obtained at CA level.

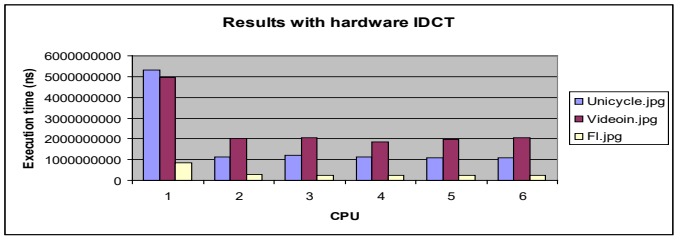
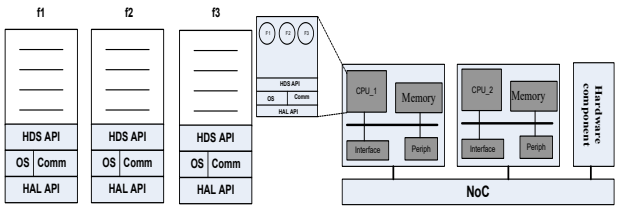


Figure 6 Comparison of simulation Results with IDCT hardware for three different images

The first stage of our approach is analytical modeling. The latter provide the estimation performance in earlier stage of MPSoC design. In our methodology, analytical modeling is based on the exploitation of graph SDF. This formalism makes it possible to model an application graph, an architecture graph and, it allows to model the constraints (we interested in the throughput constraint as an indicator of performance) under which the deployment will be run (mapping the software tasks on the processors) via SDF3 tool. In order to calculate the throughput constraint, we multiply the number of frames per second by a cycle time.

Throughput = 1/(frequency)(number of frames)*

The first stage of the diagram is the modeling of the application on a target platform at Transaction Accurate level. The Fig. 7.a presents the structure in layers of the software part and the Fig. 7.b shows an example of hardware architecture of a target platform. The application is modeled via the Khan Process Network (KPN) at TA level. The software stack is composed of a set of tasks, HdS-API, Operating System (OS) and HAL-API. Communications at this abstraction level is done through specific addresses by the primitive types *Read()* and *Write()*.



(a) (b)

Figure 7 Transaction Accurate level: (a) software layers, (b) model of the target platform

We are going to try in this section to validate our hybrid model of estimation of the performances for the design of the embedded systems. We begin with the first part which is the analytical modeling and the data, which will be used by this model, can be obtained via tools of profiling either from the traces of a tool of simulation (we choose this method).

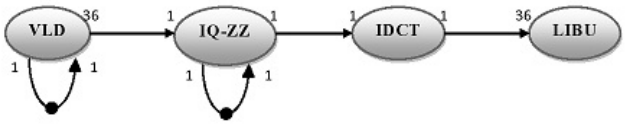


Figure 8 Graph of the MJPEG decoder application

The figure 8 represents the SDF model of the decoder MJPEG which is used during the estimations of performance. The actor VLD produces 36 tokens which correspond to 36 macros block of the image "plane.jpg" with a size of 48 pixels by 48. These are then transmitted to the actor IQ-ZZ, who reorganizes pixels according to the coding inverse zigzag then transmits it to the actor IQ-ZZ, who realizes the inverse quantification. The macro block is then treated by the actor IDCT who makes the inverse discreet transformed in cosine. Finally, the actor LIBU handles the pixels of the image to adapt them to the ring peripheral of exit, it is for it, he requires 36 macro blocks to reconstitute the whole image. Auto-arcs on knots VLD and IQ-ZZ represents respectively the tables of Huffman and quantification auto-passed on executions after execution. Once the actors are modeled, it is necessary to determine the requirements in terms for memory and the execution times of the actors of the graph. These properties are also described in an XML file. This code makes it possible to carry out actor IQZZ on the processor mips0_0 with 174763 cycles like an execution time for this task.

We add the attribute name for the element memory to facilitate the assignment or the allowance of a segment of memory for each actor in the SoCLib simulation tool. The latest information needed for the application description in its XML file is the throughput constraint which must be respectful when the decoder MJPEG mapped on the target platform. This following code presents how can define the throughput constraint in description file.

```
<timeConstraints>
    <throughput>0.0000005</throughput>
</ timeConstraints>
```

In this pseudo code, the objective of constraint is to reach a rate of 25 frames per second. So, the Application must be performed 25 iterations per second and the processor being clocked at 50 MHz (the time unit = 0.2 nanoseconds).

5. Conclusion

In this paper, one presented the technique of Fault-Tolerant in Embedded Systems (MPSoC). We interest of the dynamic migration task to resolve the Fault-Tolerant for Multiprocessors Embedded System. Multiprocessor Systems-on-Chip (MPSoC) allow the implementation of heterogeneous architectures with a high integration capacity. In recent years, computational requirements MPSoC are increasing exponentially. This complexity, coupled with constantly evolving specifications, has forced designers to consider intrinsically flexible implementations. Deploying applications typical of multimedia domains is difficult, not only due to the heterogeneous parallelism in the platforms, but also due to the performance constraints that typify these systems. An application can be modeled as a set of cooperative tasks. A task can be implemented in software or in hardware depending on its complexity and the involved cost. In this paper, our proposal is a fault tolerance approach which combines the results of a performance model and a technical's fault tolerance. Mainly our approach deals with three techniques: 1) Performance estimation in embedded systems, 2) Exploitation of hybrid models (which are based on technical simulation and analytical model), 3) Fault tolerance (based on duplication of software/hardware processing).

This approach has been tested for a Motion JPEG case study, to find out an optimal hardware and software implementation. However, the actual design of MPSoC Fault-Tolerant lacks for a high level performance evaluation model and a technical's fault tolerance. We propose a Fault-Tolerant approach based on the use of three techniques: 1) performance estimation, 2) exploitation of hybrid models (simulation / analytics) and 3) fault tolerances. The goal is to have a Fault Tolerant Embedded System. The limit of our approach is that it will only be applied in the MPSoC design approaches platform based design.

References

[1] K. Smiri, S. Bekri, H. Smei,"Fault-Tolerant in Embedded Systems (MPSoC): Performance estimation and dynamic migration tasks", in 11th International Design & Test Symposium, IDT 2016, Hammamet, Tunisia, December 18-20, 2016. IEEE 2016, ISBN 978-1-5090-4900-4 (IDT 2016), pp1-6.

[2] D. Sender Rocha dos A. Santos, L. M. Jorge, Adaptive Intelligent Systems applied to two-wheeled robot and the effect of different terrains on performance, Advances in Science, Technology and .Engineering Systems Journal Volume: 2 Issue: 1 Pages: 1-5 Published: 2017

[3] S. Shiravi, M. E. Salehi, Fault tolerant task scheduling algorithm for multicore systems. In Electrical Engineering (ICEE), 2014 22nd Iranian Conference on (pp. 885-890). IEEE.

[4] A. Das, A. Kumar, B. Veeravalli, Communication and migration energy aware task mapping for reliable multiprocessor systems. Future Generation Computer Systems, 30, pp 216-228, 2014.

[5] A. Jemai, K. Smiri, H. Smei, Task Migration in Embedded Systems: Design and Performance. Embedded Computing Systems: Applications, Optimization, and Advanced Design: Applications, Optimization, and Advanced Design, 2013.

[6] K. Pang, V. Fresse, S. Yao, O. A. De Lima, Task mapping and mesh topology exploration for an FPGA-based network on chip. Microprocessors and Microsystems, 39(3), pp 189-199, 2015.

[7] T. Maqsood, S. Ali, S. U. Malik, S. A Madani, Dynamic task mapping for Network-on-Chip based systems. Journal of Systems Architecture,61(7), pp 293-306, 2015.

[8] K. Smiri, A. Jemai, "NoC-MPSoC Performance Estimation with Synchronous Data Flow (SDF) Graphs", Autonomous and Intelligent Systems (AIS'2011), June 22-24, Burnaby, BC, Canada, pp. 406-415, 2011.

[9] Ghamarian A.H., Stuijk S., Basten T., Geilen MGW. And Theelen B.D., "Latency Minimization for Synchronous Data Flow Graphs", Published in Digital System Design Architectures, Methods and Tools, 2007.

[10] Website SoCLib http://www.soclib.fr, Juin 2017.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1196-1201 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Incremental Control Techniques for Layout Modification of Integrated Circuits

Patrik Vacula[*1, 2], Vlastimil Kotě[1, 2], Adam Kubačák[1], Milan Lžičař[1], Radek Zelený[1], Miroslav Husák[2], Jiří Jakovenko[2]

[1]*STMicroelectronics, IBC Building, Pobřežní 620/3, Prague 8,186 00, Czech Republic*

[2]*Department of Microelectronics, Faculty of Electrical Engineering, Czech Technical University in Prague, Technická 2, Prague 6, 166 27, Czech Republic*

ARTICLE INFO

Article history:
Received: 28 April, 2017
Accepted: 11 June, 2017
Online: 20 July, 2017

Keywords:
Layout
Mouse and Keyboard Control
Gesture Control

ABSTRACT

Analog layout is created in integrated circuits (IC) computer aided design (CAD) environment and during realization requires a lot of modifications of database objects. Because database objects modification is time consuming, then any improvement and simplification of modification flow in IC CAD environment can increase layout productivity. A proposed new modification concept can speed-up layout work and is very intuitive and easily put into practice. In advance, it allows to apply modern gesture tracking devices for control. By simplifying control of layout object modifications, the productivity of analog layout creation has been improved in the range of 23% to 66%.

1. Introduction

Virtuoso Layout Suite is very universal and user friendly environment for physical implementation of integrated circuits (IC) [2]. This IC computer aided design (CAD) environment supports the Cadence SKILL programming language, which allows implementation of new features [3]. In process of an analog layout creation, engineers spend a lot of time by modifying layout objects properties, inserting accurate values into forms, and confirmation of different windows. Common concept for modification of selected layout objects is through the Edit Properties window as it is depicted in Figure 1.

A width of selected wire can be changed by using a bindkey to call the Edit Properties window, then type a new value for the wire width and press the OK button to apply. This default concept is working well but consist of four steps. If layout object modifications are done very often, then an improvement of the default control concept should speed up the analog layout creation. Considerable effort has been devoted to the creation of a user-friendly and effective interface [4] where principles of IC CAD environment have been described. Since then, evolution of graphic user interface has led to more complex control of CAD systems. This paper introduces a new effective way of IC CAD environment control reducing pop-up windows and forms

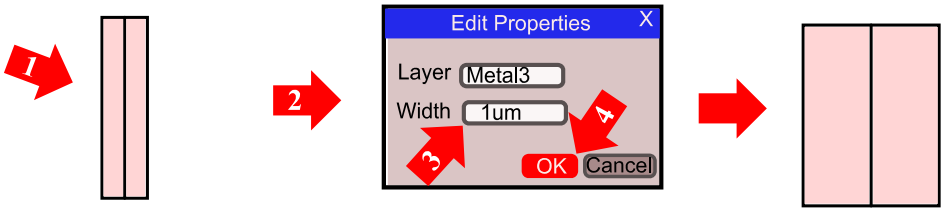


Figure 1 The example of standard four steps modification flow for wire width change in the Virtuoso Layout Suite

*Corresponding Author: Patrik Vacula, STMicroelectronics, IBC Building, Pobřežní 620/3, Prague 8,186 00, Czech Republic
Email: patrik.vacula@st.com

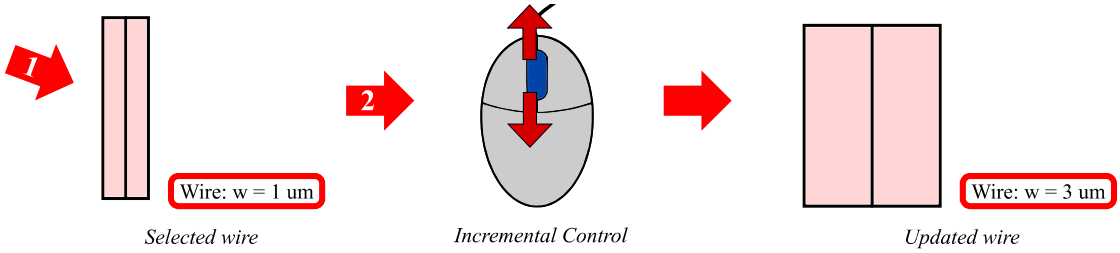


Figure 2 Two steps for the wire width modification [1]

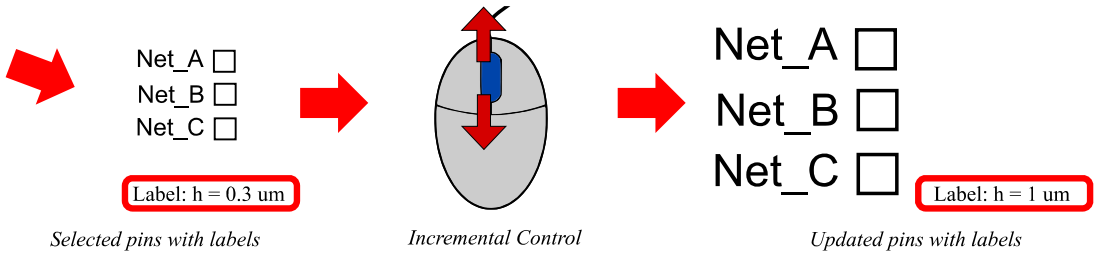


Figure 3 Height modification of pins with labels

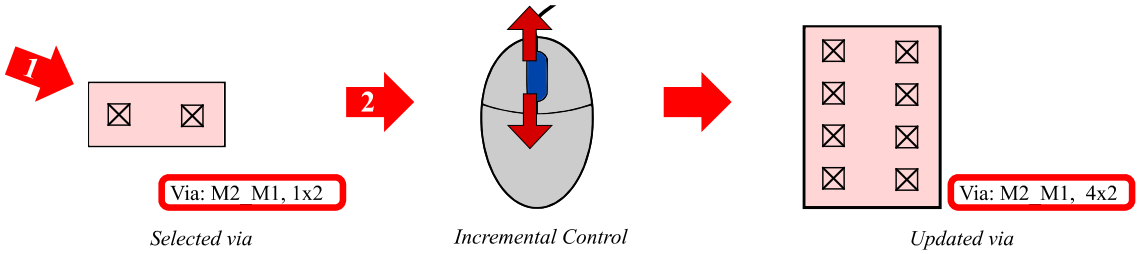


Figure 4 Via modification of rows, columns or both

2. New Incremental Control Concept

A new control concept for changing of layout object properties is defined. The main principles are as follows:

- Replacement of numeric value typing by incremental approach.
- Suppression of the pop-up Edit Properties windows for the most frequent operations.
- Display of changed values in a status bar.
- Application of the same control bindkeys on different layout objects.
- Application of different bindkey modifiers on different layout object properties.
- Possibility of using classic and modern control methods for incremental operations.

This new control concept is called the Incremental Control, where an increment value is applied directly instead of typing a numeric value [1]. After application on a selected wire, the width modification flow is simplified to scrolling of a mouse wheel, which is shown in Figure 2. The exact width value is displayed in the status bar, which allows a user to have all the modifications under control [1].

The same concept of the Incremental Control is also applied on modifications of other layout objects such as pins and their labels (Figure 3). To enlarge or shrink selected pins and their labels, scrolling of the mouse wheel is just needed. In the status bar, the user is informed about pin dimensions or the actual labels height.

The Incremental Control on additional layout objects like vias can be applied (Figure 4). Because the mostly changed properties of via object are rows and columns, control of two independent parameters is needed. To do so, a different bindkey modifier is used to differentiate between rows and columns increment. For example bindkey "*Ctrl + mouse wheel*" is used to change via columns while bindkey "*Shift + mouse wheel*" is used to change via rows.

In the status bar, the user is informed about via array dimensions to immediately know the exact count of rows and columns. Information in the status bar is for the Incremental Control very important because it contains accurate actual values.

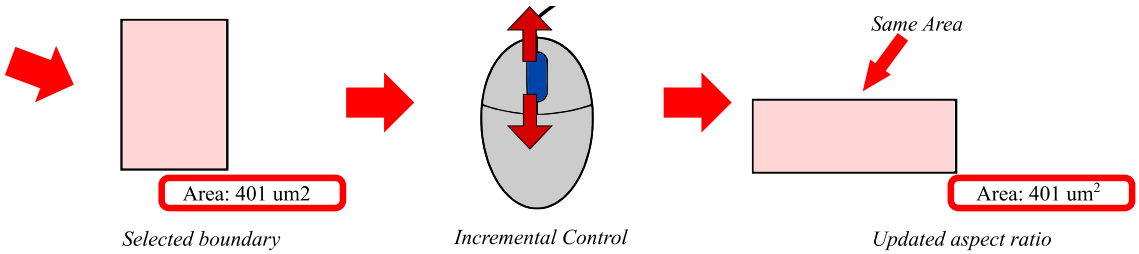


Figure 5 Aspect ratio modification of rectangular objects

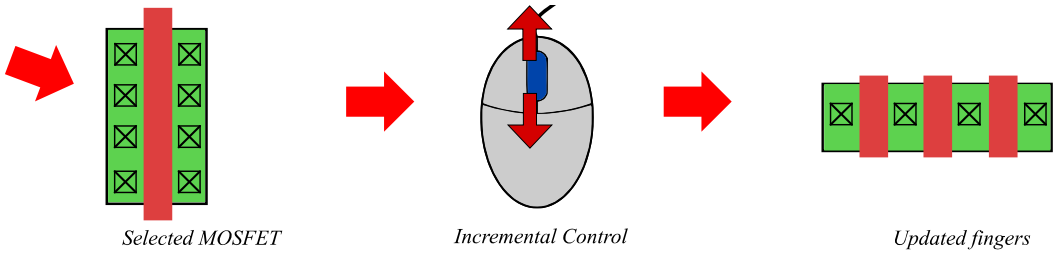


Figure 6 NGF modification of MOSFET

Another practical case for use of the Incremental Control is ability to reshape rectangular layout objects while keeping fixed area. This helps especially during floorplan tasks when the user is able to very quickly modify the aspect ratio of boundaries or rectangles circumscribed to sub-cells to fit expected area. The Incremental Control approach with mouse scrolling is very effective way how to do so (Figure 5).

The Incremental Control allows layout engineers to get simple access to a reshape utility which keeps a rectangular object area and thus has positive influence on productivity of floorplaning as layout engineers can quickly browse among all variants of a rectangle shape to find an optimum. Each modification of the rectangle aspect ratio with the Incremental Control is from user point of view ergonomic and intuitive. In the status bar, the user is informed about an area of the selected rectangle. Such information is very important for floorplaning and is available after each transformation with the Incremental Control.

Now main analog layout objects are discussed. The most important ones are metal-oxide-semiconductor field effect transistors (MOSFETs). Even if there are a lot of different MOSFETs in a process design kit (PDK), there is one common parameter a number of gate fingers (NGF) which is very often modified to allow a MOSFET transformation. The Incremental Control applied on MOSFET allows layout engineers to quickly browse among all possible variants which speeds up the time needed to find an optimal MOSFET shape with the relevant "number of gate fingers" (Figure 6). The Incremental Control applied on MOSFET does not change length and width of a transistor channel but only adjusts the "number of gate fingers" parameter.

Other often used analog layout objects which can be modified by the Incremental Control are capacitors. Goal is to change the aspect ratio of a capacitor while a capacitance or an area is fixed. Generally a capacitor object usually consists of multiple layers, each with different design rules. Due to rounding after each modification, the capacitance of the capacitor is usually slightly changed. This capacitance rounding error has cumulative behavior and for the uncompensated Incremental Control flow can produce a significant shift of the value. Without a correction mechanism the Incremental Control would not be able to precisely keep the existing capacitance value. (Figure 7).

The Incremental Control concept can be used on any other layout object and its properties. If a requested value of a layout object property is very far from current value an adaptive increment or a predefined list of allowed values is used to speed up transformations and improve user experience.

There is an additional constrain influencing user experience. Robustness against a human error. If a way of modifications is very simple, then immunity is lower, for example in a case that all layout object are selected and the mouse wheel is scrolled. Therefore to prevent unwanted modifications, the bindkey modifier such as "*Ctrl*" is combined with the mouse wheel, in the terms of "two hand control".

3. Status Bar for Effective Incremental Control

The status bar is used to retain simplicity and precision of modifications performed by the Incremental Control [1]. In the status bar the user is informed about a current value of a modified property. Information value of the status bar is significantly improved by displaying electrical properties of selected layout objects. For example for a wire or a path the status bar displays not only actual width but even more useful electrical quantities such as resistance, maximum electromigration current (EMC) for direct current (DC) or for root mean square (RMS) and also connectivity. Similarly for vias the status bar displays not only array dimensions but also useful electrical properties like electrical resistance or maximum DC or RMS EMC and connectivity. The status bar is shown in Figure 8. Refreshing all these values is done after each modification and improves productivity.

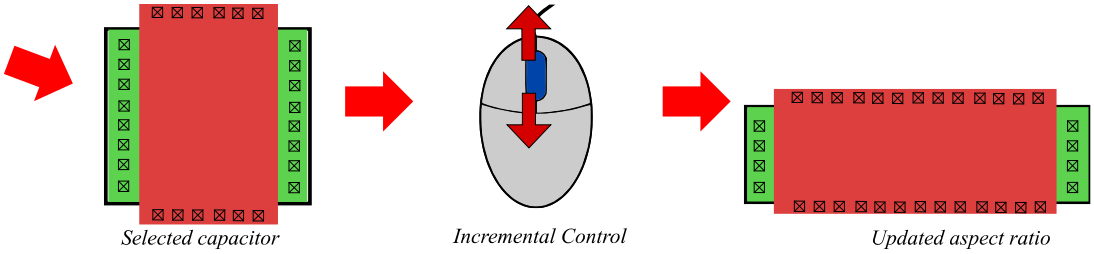


Figure 7 Capacitor transformation

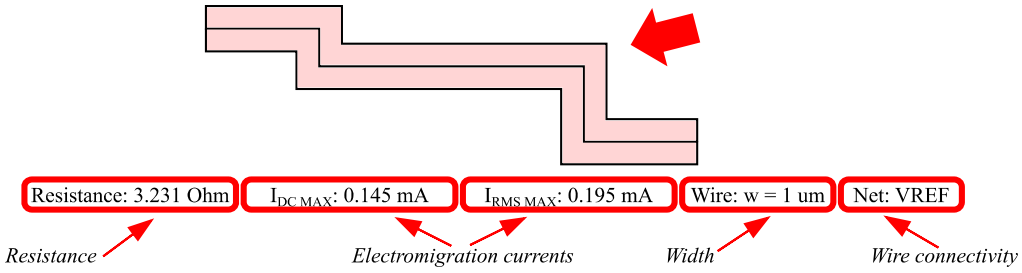


Figure 8 The selected wire electrical and physical properties displayed in the status bar [1]

Figure 9 Gesture control mapped to the incremental interface

4. Gesture Control with Incremental Control

Discussed new control technique is compatible with the IC CAD environment and with current control devices such as keyboard, mouse, or track-ball. In addition to it is also compatible with new gesture tracking systems similar as in [5] and [6]. The Incremental Control is usually mapped to mouse wheel but has been experimentally mapped to gesture recognition camera Creative Senz3D using [7].

By using camera gestures “thumb up” and “thumb down” mapped to Incremental Control bindkeys, a wire width is updated as shown in Figure 9. Based on experience, when gestures were swapped or not recognized and had to be repeated, precision improvement of tested device must be put in place to be suitable for real analog layout application [1].

5. SKILL Implementation of Incremental Control

A flow chart describing the Incremental Control implementation in the programming language Cadence SKILL is shown in Figure 10. SKILL provides direct database access and allows environment customization as described in [3].

The program starts with a user action when the user selects a request type (RQT) by pressing a modifier and performs an increment or a decrement by the mouse wheel. Then the program detects the request type and information about the increment. After the request type and the increment detection, the check is performed, if objects are selected. When not, the program ends. When objects are selected, the program detects object IDs from the CAD (Computer aided design) database (CAD-DB) and reads the Incremental Control Database (IC-DB) which contains a list of changeable properties of each object type and determines the way of all possible modifications. For case of continuous values to be entered, the IC-DB also contains a list of predefined values which the program reads incrementally. If selected objects are members of the IC-DB, then the program loads actual values of the changeable properties from layout objects and loads technique of layout object modification from the IC-DB.

In some cases during layout object modification, an actual value of a changeable property can be slightly different than desired value in order to align layout objects to design grid. After repeated modifications, the actual value can shift from the desired value and a noticeable error can occur. Hence, an initial value of the changeable property is introduced. It allows that the actual value to be always calculated from the original correct value. Hereby the maximal possible accuracy of incremental changes done by the program is ensured. If the initial value is required and not yet set, the program stores the initial value as a parameter of the modified layout object and together with this also creates a record of the last updated value, which serves as a decisive criterion, whether the layout object has been changed in different way by this program.

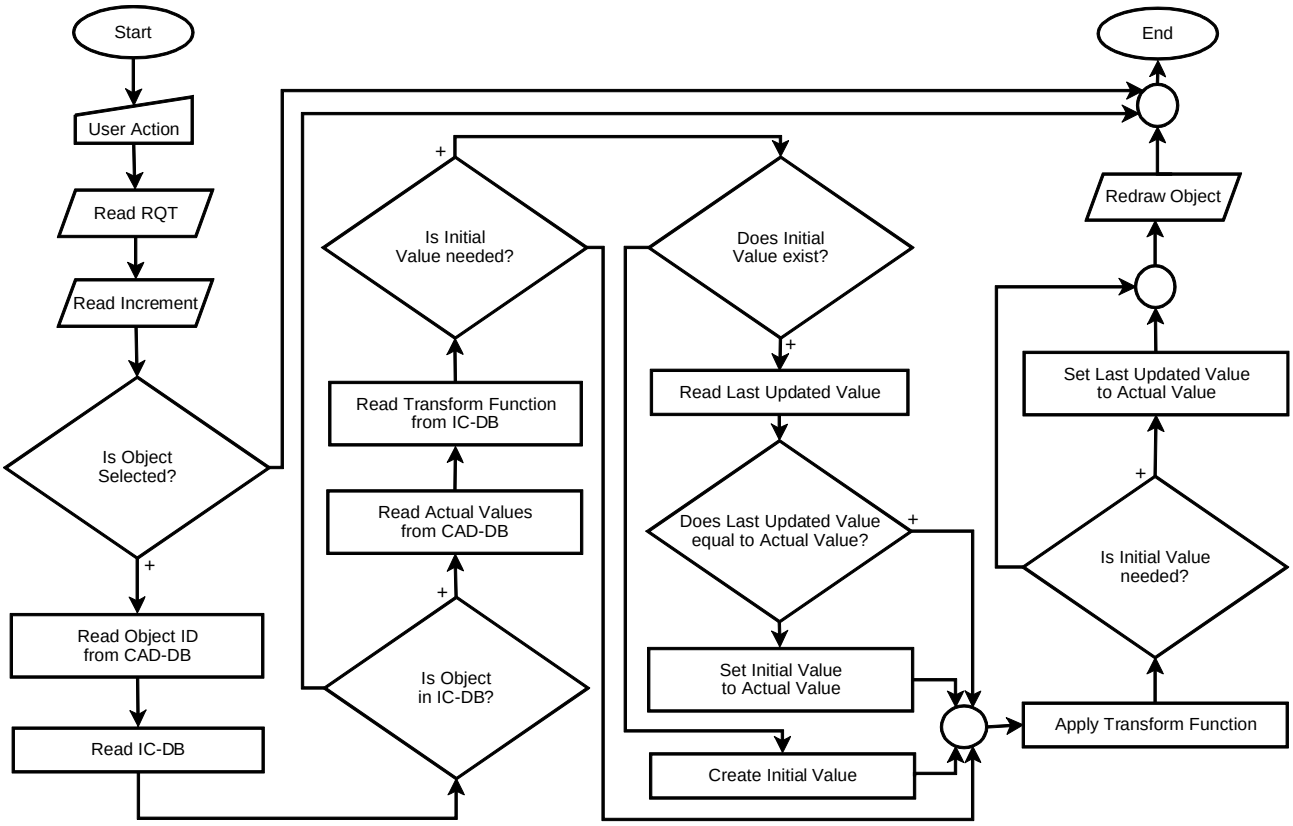


Figure 10 The flow chart of the Incremental Control

Table 1 Layout productivity measurement results [1]

Layout object	t_{SC} (s)	t_{IC} (s)	$(t_{IC} - t_{SC}) / t_{SC}$ (%)
via	5.6	2.0	-64.7
wire	4.6	3.5	-23.4
pin	7.0	4.5	-36.1
label	5.8	3.5	-39.9
pin + label	15.1	5.0	-66.7
rectangle	5.2	3.2	-38.5
transistor	5.6	3.7	-33.6
capacitor	5.4	3.5	-35.2

When the initial value is set, then is read and compared with the actual value. If these values equal, the layout object has not been modified by an external intervention and the initial value is untouched. If these values are different, the external intervention has been performed and the initial value is set to the actual value [1]. It means that the program respects the external intervention.

After these steps, update of a current state is executed based on obtained parameters and the function which has been read from the IC-DB. When the layout object initial value is mandatory, the last updated value is set to the actual value. So, the last updated value is the parameter of the layout object which allows that its value to be used in another run of this program. In the last step, the layout object is redrawn and the actual values are printed into the status bar. By cyclic applications of this program, the layout object is controlled incrementally. This program is able to operate with the OpenAccess and also with the CDBA databases.

6. Layout Productivity Gain

As already mentioned, the Incremental Control allows variety of different layout object modifications by intuitive common control concept. This concept reduces the number of actions needed to reach optimal result. To quantify layout productivity gain value, the average time of standard control modification t_{SC} and of Incremental Control modification t_{IC} to reach target value of layout objects are introduced. The measurement was performed under the same conditions and by a very experienced analog layout engineer with 14 years of experience. At least ten objects of each type were modified. Initial values of properties, which were changed, were set randomly for all objects. Then, the property values were modified to reach the target value by the both ways and the introduced durations t_{SC} and t_{IC} were measured.

Relative difference of t_{IC} and t_{SC} was used for calculation of average layout productivity gain for each layout object type. Based on measurement which is noted in Tab. 1, we can say that layout productivity time saving for the Incremental Control is in the range of 23 % to 66 %. From Tab. 1., it is evident that higher efficiency is reached when different objects are modified simultaneously.

7. Conclusion

The Incremental Control technique was developed to modify various database object properties. This control concept is very intuitive, hence it can be easily put into practice. The Incremental Control can be applied on different properties of database objects such as wires, pins, labels, vias, rectangles, MOSFETs, capacitors or other, which is the main advantage. Incrementally modified properties of objects can be controlled using keyboard, mouse or gesture tracking devices.

Main simplification of IC CAD modification flow is achieved due to removing of typing in object property window and replacing it by value increment. Additional time saving is achieved by suppression of pop-up Edit Properties windows for the most frequent operations. For additional flow improvement Incremental Control was mapped to mouse wheel. Then, productivity of analog layout has been improved in the range of 23 % to 66 %.

In advance, an alternative control was used experimentally when the Creative Senz3D gesture control camera was mapped to Incremental Control bindkeys. For example, width of wire was changed by “thumb up” and “thumb down” gestures. Based on our testing, the gesture tracking device still needs an additional improvement of precision to be suitable for analog layout creation.

Acknowledgment

This work is part of the CTU SGS grant no. SGS17/188/OHK3/3T/13.

References

[1] P. Vacula, V. Kotě, A. Kubačák, M. Lžičař, R. Zelený, M. Husák, J. Jakovenko, “Incremental Control Techniques for Layout Modification” Advanced Semiconductor Devices & Microsystems (ASDAM), Smolenice, Slovakia, 2016.

[2] “Virtuoso® Layout Suite XL User Guide”, Cadence Design Systems, Product Version 6.1

[3] “Cadence® User Interface SKILL Functions Reference”, Cadence Design Systems, Product Version 6.1.6, August, 2015

[4] N. Munro, “ECSTASY – A Control System CAD Environment,” Proceedings of the International Conference on Control, 1988. CONTROL 88, Oxford, United Kingdom, April 13–15, pp. 76–80, 1988

[5] W. H. Chen, Y. H. Lin and S. J. Yang, “A generic framework for the design of visual-based gesture control interface,” Proceedings of the 5th IEEE Conference on Industrial Electronics and Applications (ICIEA), Taichung, Taiwan, June 15–17, pp. 1522–1525, 2010.

[6] D. Ionescu, B. Ionescu, C. Gadea and S. Islam, “An intelligent gesture interface for controlling TV sets and set-top boxes,,” Proceedings of the 6th IEEE International Symposium on Applied Computational Intelligence and Informatics (SACI), Timisoara, Romania, May 19–21, pp. 159–164, 2011.;

[7] “Intel® Perceptual Computing SDK – Human Interface Guidelines”, Intel Corporation, Revision 3.0, February 25, 2013, [cited 10. 3. 2016], available from: https://software.intel.com/sites/default/files/article/401008/perc-humaninterfaceguidelines.pdf.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1202-1210 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Effect of Cold Drawing Pass Schedule on Mechanical Properties and Microstructure of ST 52 during Cold Drawing of Seamless Tubes and its Influence on Springback

Dadabhau Baban Karanjule*[1], Sunil Sukhdeo Bhamare[2], Thota Hanumant Rao[3]

[1]*Research Scholar, Sinhgad College of Engineering, Vadgaon, Pune, M.S., 411041, India*

[2]*Registrar, Dr.Babasaheb Ambedkar Technological University (BATU), Lonere, Dist.Raigad, M.S., 402103, India*

[3]*Former Director, Research and Development Department, Indian Seamless Metal Tubes Limited, Ahmednagar, M.S., 414001, India*

A R T I C L E I N F O

Article history:
Received: 30 April, 2017
Accepted: 15 June, 2017
Online: 20 July, 2017

Keywords:
Cold drawing
Pre-strain
Mechanical testing
Micro hardness
Young's Modulus

A B S T R A C T

In-elastic recovery behavior of seamless tube material has been investigated by uniaxial tensile tests. Unloading Stress–Strain curves obtained under different passes of cold drawing process shows that the percentage of in-elastic recovery to the total recovery increased with plastic deformation. This paper is an experimental study that shows Young's Modulus decreases with plastic strain for ST 52 material. It is found that with increase in plastic strain, Young's Modulus reduces rapidly initially then reduces more slowly and finally settles to stable value due to increase in plastic deformation and ultimately increased residual stresses. This variation of Young's' Modulus is related to internal stresses, residual stresses, micro cracks, dislocations during plastic deformation. Similarly, Scan Electron Microscopy (SEM) and Micro-hardness testing reveals that mechanical properties are better in the first pass sample of multiple cold drawing passes. The results of this study reveals that 10-20% degradation occurs in Young's Modulus for 5-7% plastic strain and better mechanical properties are achieved in the first pass sample.

1. Introduction

This paper is an extension of work that was originally presented in 2016 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM) [1].When a seamless tube is drawn from drawing bench after drawing, the springback phenomenon occurs due to the residual stresses leading to an undesired deformation. This magnitude of springback effect is nearly proportional to the Young's Modulus and strain hardening behavior of material. This phenomenon is particularly critical for high strength steels, which have been increasingly used in the automotive industries in regard to the performance improvement and light-weight design. Generally, the springback strain is an elastic strain.

It is well known that the elastic strain recovery is present during loading and unloading a steel sample and it leads to a hysteresis behavior or energy loss of the material. Thus, description of the elastic behavior strongly affects the springback prediction of a forming simulation. One of the most troublesome problems facing the tubing production industry is springback in the tube making process. With the ongoing miniaturization of products, springback is a dominant effect because material behavior greatly varies in this process.

The mechanism behind springback is the release of residual stresses in a work piece after the forming process. A success of the springback prediction strongly depends on the applied materials model. Among various parameters of the model, Young's Modulus is one of the most important factors influencing the calculation accuracy that should be considered as function of strain. It also influences the plastic-forming quality of the formed components, especially springback. Young's Modulus actually is having correlation with plastic strain and its accurate characterization is necessary to improve springback in the cold

*Corresponding Author: Dadabhau Baban Karanjule, Research Scholar, Sinhgad College of Engineering, Vadgaon, Pune, M.S., 411041, India
Email: karanjule.dada@gmail.com

drawing process. The purpose of the investigation described here is to study the influences of pass schedule on the mechanical properties and microstructure, since it is necessary, for the development of ultrahigh strength seamless tubes, to characterize variation of the microstructures and to clarify the mechanism how pass schedule affects the mechanical properties.

2. Literature Review

Many researchers have investigated the effect of Young's Modulus, material behavior and microstructure on springback in different manufacturing processes.

The change of Young's Modulus with increasing plastic strain was firstly investigated by [2]. In this research, he found that the actual springback was greater than that calculated with a constant Young's Modulus and independent of plastic deformation. However, a constant Young's Modulus was often used in many commercial Finite Element Methods (FEM) codes because of its simplicity. A simple analytical model was used to describe the relationship between Young's Modulus and plastic deformation based on experimental results. Great efforts have been made to study Young's Modulus and its effect on spring-back in the recent years [3].Researchers concluded that the correct choice of the yield criterion and consideration of the change in Young's Modulus are essential for improving the prediction accuracy of springback in the numerical simulation [4].Young's Modulus decreases with the increase of plastic strain during cyclic tension– compression deformation and proposed an empirical expression to describe the variation of Young's Modulus with respect to plastic strain [5].The material response is calculated under the assumption of fully elastic deformations in the complete model [6, 7, and 8].An alternative method was explained as inverse forming, during which the loads were reversed and the tools gradually retracted. Because of more computational cost, it is least used. At the same time it is more realistic because the contact forces are present during the unloading step [9, 10, 11,12 and 13].The work reported that the major improvements in the critical properties of a double corrode panel can be achieved using variable blank holder process. Springback was significantly reduced and dent resistance increased [14].Retained austenite exits in many hardenable steels and its presence induced both positive changes in mechanical engineering properties[15].Mathematical expression describing the relation of Young's Modulus with plastic strain was proposed and this expressions was implemented into LS-DYNA for simulation of U-Channel springback[16].The variation of Young's Modulus during loading and unloading was considered to identify different variables affecting ratcheting simulation [17].The work reveals the influence of pre deformation on the springback of steel AHSS sheets under consideration of the Young's Modulus change and the effect of this change as a function of plastic strength[18].An elasto-plastic model using a kinematic hardening model was made for springback analysis in sheet metal forming. The change of Young's Modulus versus plastic strain was proposed by [19]. Strain recovery after plastic pre-strain and associated elastic and inelastic behavior during unloading and unloading of DQSK steel sheet were measured. Slope of the elastic modulus variation was different during loading and unloading step due to the reciprocal dislocation movement. The amount of springback was calculated by using the process of elastic recovery as well as established values of Young's Modulus and poisson's ratio. Results concluded that inelastic strain released from the formed state was major source of additional strain recovery whose magnitude depends on forming stress state [20].

In spite of ample research conducted on springback dependence on Young's Modulus in different manufacturing processes like sheet metal bending, rolling, forming, drawing etc., still there is scope for the prediction of springback in cold drawing of seamless tubes and its dependence on Young's Modulus.

3. Materials and Methods

Material selection process has utmost importance in engineering applications. The failures arising from bad material selection are not uncommon in many industries.

Materials:

The materials under study for dies-plugs used for cold drawing and seamless tubes are:

3.1. Die and Plug material (D3):

The material for dies and plugs is D3 (high carbon high chromium cold-work tool steel) having chemical composition as shown in Table 1.

Table 1 Chemical composition of D3

Component	Min %	Max %
Carbon, C	2	2.35
Chromium, Cr	11	13.5
Manganese, Mn	0	0.6
Silicon, Si	0	0.6
Phosphorous, P	0	0.03
Sulphur, S	0	1
Tungsten, W	0	1
Vanadium, V	0	1

The physical and mechanical properties for die and plug material are tabulated as shown in Table 2.

Table 2 Properties of D3

Properties	**Value**
Physical	
Density (g/cc)	7.87
Mechanical	
Hardness (RC)	64
Modulus of elasticity (N/mm^2)	207
Poisson's ratio	0.29

3.2. Tube material

The material taken in this work is an uncoated cold rolled ST 52 steel with thickness 5.0 mm. The chemical composition of the tested tube sample is shown in Table 3 and the mechanical properties of mother or hollow tube of 75 mm X 5 mm are listed below in Table 4.

Table 3 Chemical composition of ST 52

Component	Min %	Max %
Carbon, C	0.16	0.2
Iron, Fe	-	-
Manganese, Mn	1.4	1.6
Silicon, Si	0.3	0.5
Phosphorous, P	0	0.025
Sulphur, S	0.01	0.025

Table 4 Properties of ST 52

Physical Properties	Value
Density (g/cc)	7.87
Mechanical Properties	
Hardness (Brinell)	207
Hardness (Rockwell B)	94
Hardness (Rockwell C)	17
Ultimate tensile strength (N/mm^2)	602
Yield strength (N/mm^2)	522
Elongation at break (%)	32.6
Modulus of elasticity(KN/mm^2)	210
Bulk modulus(KN/mm^2)	140
Poisson's ratio	0.3
Shear modulus(KN/mm^2)	80

The actual chemical composition on spectrometer is shown in Table 5.

Table 5 Actual chemical composition of ST 52

C	Mn	Si	P	S	Cr	Mo
0.17	0.63	0.2	0.008	0.023	0.05	0.16

4. Experimentation

An experiment is a procedure carried out to support, refute or validate a hypothesis. Experimentation is carried to draw a mother or hollow tube through three passes to get the final desired dimensions.

4.1. Cold Draw Reduction

Since a high area reduction is required, multiple drawing passes are generally necessary to avoid fracture of the tubes during the drawing process. To study the effect of Young's Modulus, reduction passes as shown in Table 6 are considered.

Table 6 Reduction passes

Pass Schedule	Outer diameter (mm)	Thick-ness (mm)	Inner diameter (mm)	Reduction on outer diameter(%)
Mother tube	75	5	65	--
First pass	68	5	58	9.3
Second pass	60	5	50	11.8
Third pass	50	5	40	16.7

4.2. Universal Testing Machine (Tensile Test)

Uniaxial tension specimens were cut from the as received tube in the longitudinal direction. The dimensions of each specimen were in accordance with ASTM E370 with the gauge length of 50 mm as shown in Figure 1.In order to minimize the influence of machining forces, linear cutting of the specimens was used. Uniaxial tensile test was carried out on UTM with a precise extensometer. For tensile properties, tensile specimens were loaded in UTM TUE-C-600.Load-elongation data were recorded and converted into stress-strain graphs. Yield strength, Ultimate Tensile strength, Young's Modulus and ductility (% elongation and reduction) were determined in accordance with ASTM standard test procedures (ASTM E18, ASTM E23, ASTM E8).

Figure 1 Test specimens for Tensile Test (Sample photograph)

Tensile test was carried out at each stage of cold draw pass to study the varying mechanical properties. The different mechanical properties measured are Yield Strength (YS), Ultimate Tensile Strength (UTS), Elongation (% E) and Hardness.

4.3. Hardness Testing

For hardness testing, the oxide layers formed during heat treatment were removed by stage-grinding and then polished. The average Brinell Hardness Number (BHN) readings were determined by taking two hardness readings at different positions on the samples, using a Brinell hardness tester. Micro hardness examination was carried out on Vickers hardness tester-MacroMann technologies with load 200 gm on scale HV 0.2 with dwell time of 20 seconds.

4.4. Microstructure Examination

Microstructure examination of all the samples were carried out in Olympus metallurgical microscope. Each sample was carefully grounded progressively on emery paper in decreasing coarseness from grit size 120, 320, 400, 600, 800.The grinding surface of the samples were polished using diamond paste. The crystalline structure of the specimens were made visible by etching using solution containing 3% nitric acid and 97% methanol on the polished surfaces. Microscopic examination of the etched surface of various specimens was undertaken using a metallurgical microscope with an inbuilt camera through which the resulting microstructure of all the samples were photographically recorded with magnification of 100 X and 500 X.

4.5. Scan Electron Microscopy (SEM)

A scanning electron microscope (SEM) is a type of electron microscope that produces images of a sample by scanning the surface with a focused beam of electrons. The electrons interact with atoms in the sample, producing various signals that contain information about the sample's surface topography and composition. For SEM, sample preparation was carried out by polishing and etching the sample as in case of metallurgical microscope. Machine used was Field Emission – SEM Model – Sigma.

5. Results and Discussions

5.1. Tensile test:

The results obtained by tensile test on universal testing machine (UTM) are provided in Table 7 for mother or hollow tube, intermediate passes and final pass.

Table 7 Properties obtained on UTM

Stage	Yield Strength (N/mm^2)	Ultimate Tensile Strength (N/mm^2)	Elongation (%)	YS/ UTS	Strain	Young's Modulus (N/mm^2)
Mother tube	451.20	520.41	18	0.87	0.180	199
First pass	484.10	550.00	17	0.88	0.175	207
Second pass	564.20	620.12	16	0.91	0.160	214
Third pass	630.51	685.21	14	0.92	0.14	222

From this it is clear that mechanical properties are found increasing with the rise in degree of cold working. Also, there is an effect of pre-existing strain of tube on Young's Modulus after cold drawing. It is found that there is inverse proportion between pre-strain and Young's Modulus. Springback prediction and compensation are the major challenges in the modern tube manufacturing industry. In spite of plenty of developments and a certain amount of progress in the field of simulation of tube making processes, the industrial needs for accurate numerical prediction of springback are not yet being met. Springback, a phenomenon that is related to the elastic strain recovery after removal of forming loads, is physically governed by the stress state achieved at the end of forming process and by elastic response of formed parts. The elastic response of the parts is associated with the removal of forming tools. Despite all that, several times it was reported that Young's Modulus drops with the increase in plastic pre-strain in a material which is found in this research work. Table 8 shows the same.

Table 8 Readings of pre-strain against Young's Modulus.

Sr. No.	Pre-strain (mm)	Young's Modulus(N/mm^2)
1	0.160	222
2	0.175	214
3	0.180	207

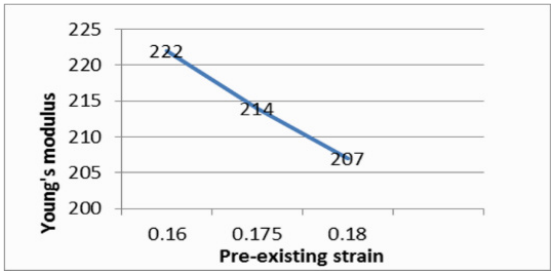


Figure 2 Graph of pre-existing strain vs. Young's Modulus

From Figure 2, it is found that degradation of Young's Modulus was found with increase in degree of plastic pre-strain in a material. Ten readings are taken at 5 different locations (Table 9) for outer diameter of the drawn tubes with digital micrometer of 1 micron accuracy to find the springback. Springback is the deviation from targeted dimension. The results are taken to co-relate the relation of Young's Modulus and springback in cold drawn tubes as shown in Figure 3 and Table 10.

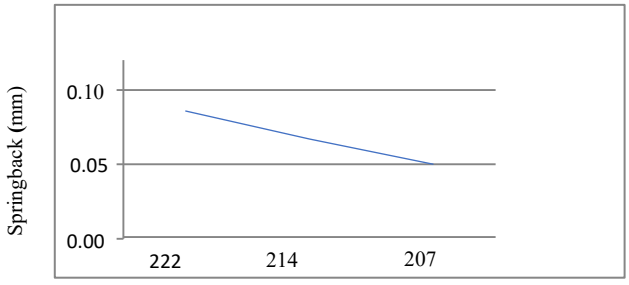


Figure 3 Graph of Young's Modulus vs Springback

Table 9 Tube outer dimensional inspection after each pass of drawing

Tube Size (mm)	Pass	Die size (mm)	Tube outer diameter dimensions (mm)						Spring back (mm)
			1	2	3	4	5	Average	
70 X 5	Mother tube	--	--	--	--	--	--	--	--
68 X 5	First pass	68	68.070	68.087	68.093	68.089	68.091	68.086	0.086
60 X 5	Second pass	60	60.071	60.060	60.063	60.071	60.068	60.067	0.067
50 X 5	Third pass	50	50.045	50.051	50.048	50.055	50.049	50.0496	0.050

Table 10 Young's Modulus against Springback

Sr. No.	Young's Modulus(N/mm^2)	Springback(mm)
1	222	0.086
2	214	0.067
3	207	0.050

From the results; the degradation of Young's Modulus is directly correlated to the degree of plastic pre-strain in tubes. Further, the actual springback is directly co-related to the degradation of Young's Modulus. Based on the studies, the degradation of Young's Modulus have significant effect on final elastic strain recovery and final shape of the product**.** Hence, a control over close dimensional tolerances can be achieved by optimizing the degradation on Young's Modulus which can be optimized by selecting degree of pre strain in tube and reduction ratio.

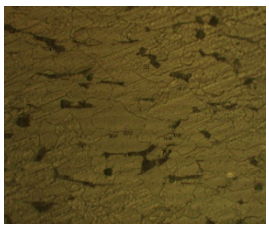
Table 11 Results of micro hardness test

Pass	Micro hardness readings (200 mg,20 sec dwell)			Average
Mother tube	131	128	130	129.66
First pass	140	140	134	138.00
Second pass	184	185	184	184.33
Third pass	184	198	200	194.00

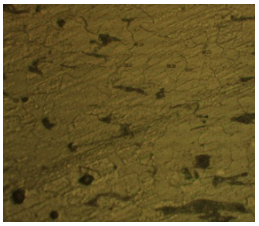
Table 11 shows that the hardness values increases as number of passes increases.

5.2. Microstructures and grain growth

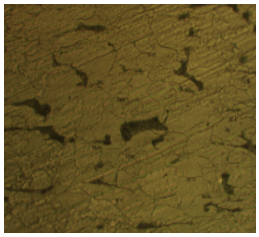
Microstructures of the tested ST 52 steel samples were examined with Metallurgical Microscope. The specimens for optical microscopy were etched using 3% Nital. The microstructures obtained under microscope are shown in Figures 4-7.



Grain size No.: 9.75,
Grain size in µm :12.98 (a)



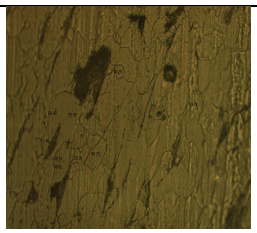
Grain size No.: 7.97,
Grain size in µm : 20.30 (b)



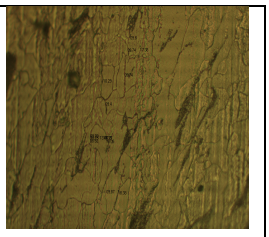
Average grain size No.:9.10
Average Grain Size in µm: 15.31

Grain size No.: 9.59,
Grain size in µm : 12.66 ©

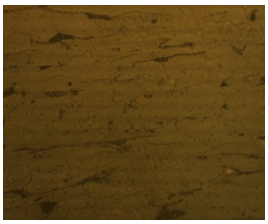
Figure 4 (a), (b), (c) Microstructures of mother tube



Grain size No.: 9.33
Grain size in µm : 12.87 (d)



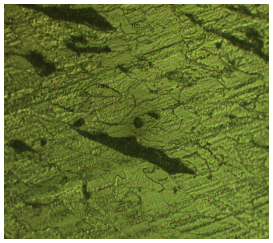
Grain size No.: 9.62
Grain size in µm :11.51 (e)



Average grain size No.: 9.34
Average grain size in µm : 12.77

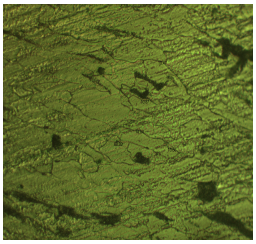
Grain size No.: 9.08
Grain size in µm: 13.95 (f)

Figure 5 (d), (e), (f) Microstructures of first pass tube



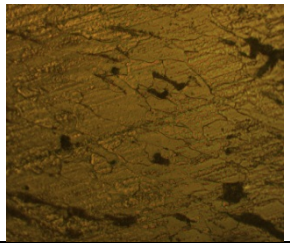
Grain size No.: 10.50,

Grain size in µm : 8.77 (g)



Grain size No.:10.47

Grain size in µm : 8.55 (h)

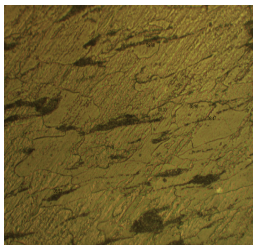


Average grain size No.: 10.35
Average grain size in µm : 8.74

Grain size No.: 10.10

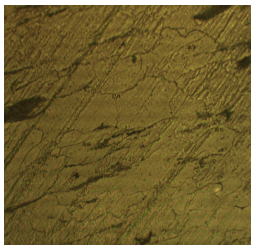
Grain size in µm : 8.90 (i)

Figure 6 (g), (h), (i) Microstructures of second pass tube



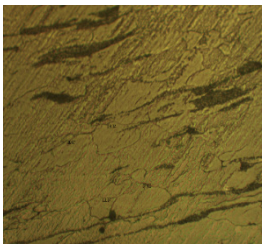
Grain size No:8.33

Grain size in µm :14.248 (j)



Grain size No.: 8.16

Grain size in µm-18.90 (k)



Average grain size No.: 8.36
Average grain size in µm : 16.96

Grain size No.: 8.61

Grain size in µm : 17.74 (l)

Figure 7 (j), (k), (l) Microstructures of third pass tube

From the observed microstructure it is found that grain size is decreasing with the rise in reduction. This is due to the increased deformation with simultaneous passes. Further the effect of grain deformation before cold drawing on springback cannot be neglected. Grain refinement is an effective means for improving the strength and lowering the ductile-brittle transition of structural alloys. The improvement can often be expressed in an equation of the Hall-Patch form. However, the appropriate use of grain refinement requires an understanding of the effective grain size that actually governs the mechanism of yielding or failure. Grain size has a measurable effect on most mechanical properties. At room temperature hardness, yield strength, tensile strength, fatigue strength and impact strength all increases with decreasing grain size. Machinability is also affected; rough machining favors coarse grain size while finish machining favors fine grain size. The effect of grain size is greatest on properties that are related to the early stages of deformation. Yield stress is more dependent on grain size than tensile strength. Fine-grain steels do not harden as deeply and have less tendency to crack than coarse-grain steels of similar analysis. There is also an equation expressing the relationship between grain size and yield strength.

5.3. Scan Electron Microscopy (SEM):

SEM scans a high-energy electron beam across the surface of a specimen and measures one of a number of signals resulting from the interaction between the beam and specimen.

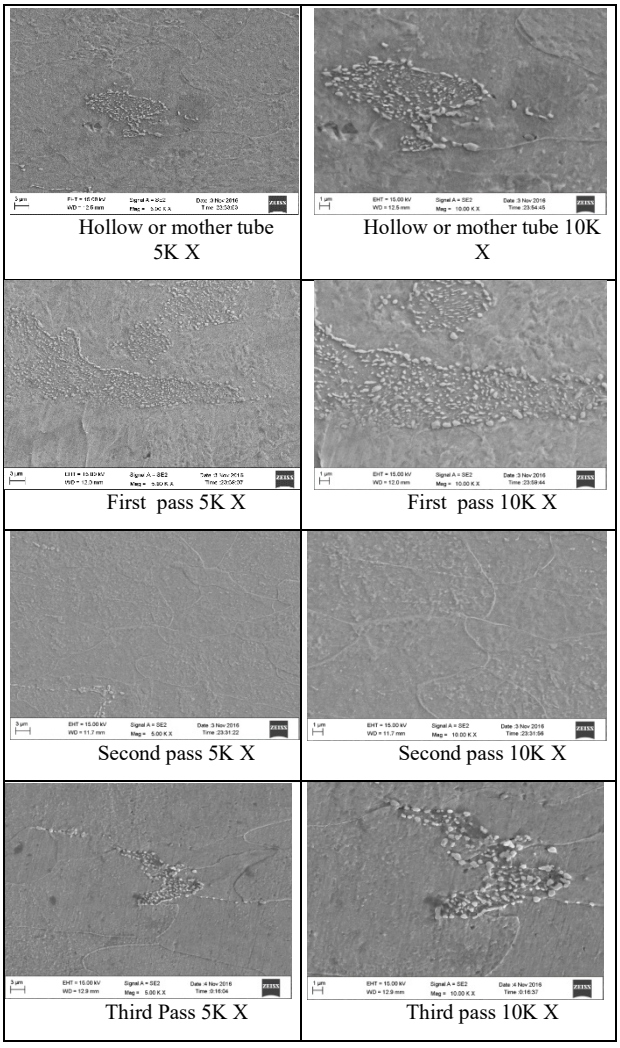


Figure 7 SEM images of all four samples

SEM microstructure shows elongated lamellar boundary structure and equiaxed grain structure. Microstructure reveals size of deformation microstructure gradually increases with increase in pass schedule. The results indicate that slight coarsening of microstructure happens at initial stage due to recovery. Few more samples are tested for mechanical properties and the results are tabulated as shown in Table 12 and Figures 8-12.

Table 12 Results of samples

Pass	Tensile strength (N/mm^2)	Hardness (BHN)	Percentage elongation (%)	Percentage reduction (%)	Young's Modulus (N/mm^2)	Yield strength (N/mm^2)
Mother tube	415.226	129.66	22.80	59.23	393225	208.774
First pass	610.530	136	8.53	32.89	248678	568
Second pass	555.659	184.66	10.94	39.78	165781	547.206
Third pass	378.219	194	37.55	59.53	131342	203.035

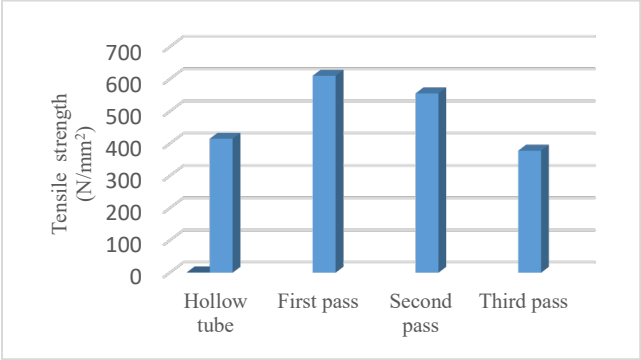


Figure 8 Tensile strength of four tube samples

Tensile strength is more in first pass sample. Greater plastic deformation induced by cold drawing introduces greater strain hardening.

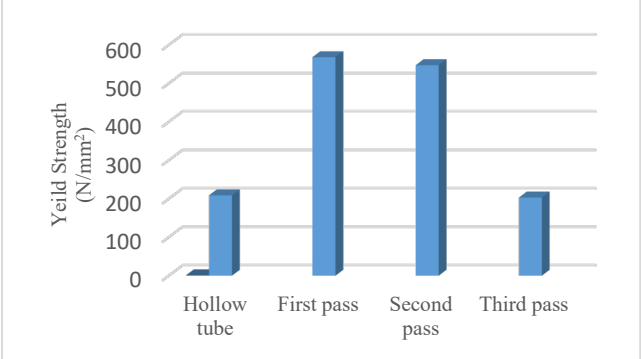


Figure 9 Yield strength of four tube samples

Yield strength is more in first pass sample as shown in Table 9.The reason is that plastic deformation occurs due to the motion of crystal lattice dislocations through the material. The result at large scales is work-hardening: as strain increases, stress needed to cause continued flow increases. Additionally, strain-to-failure decreases.

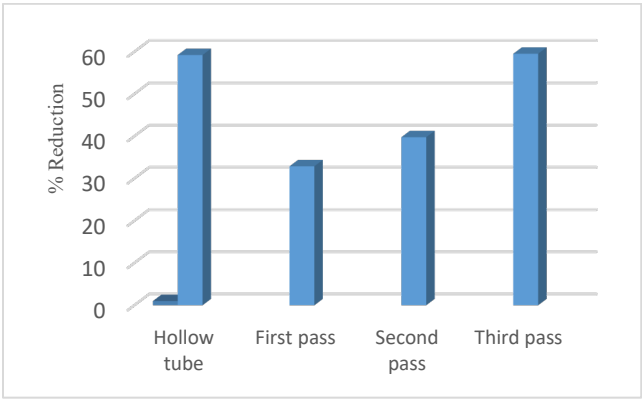


Figure 10 Percentage reduction of four tube samples

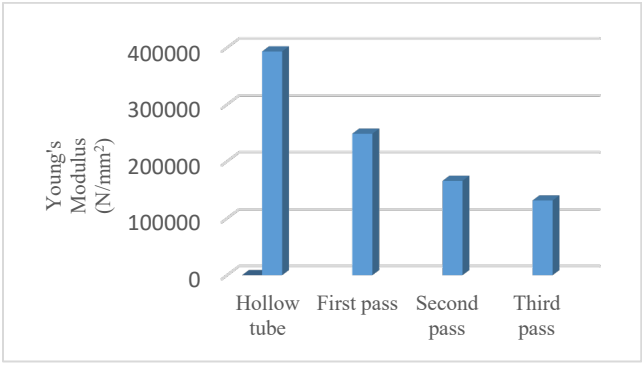


Figure 11 Young's Modulus of four tube samples

Figure 11 shows that Young's Modulus is found to be decreasing as the number of passes increases.

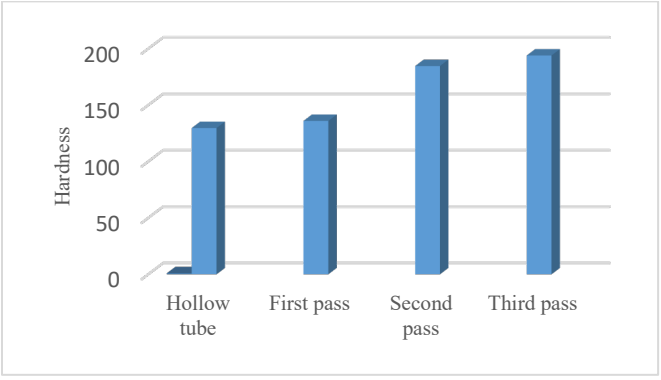


Figure 12 Hardness of four tube samples

From Figure 12, hardness is found highest in last sample due to large plastic deformation.

6. Conclusions:

The influence of pass schedule on mechanical properties, microstructures and micro hardness of cold drawn seamless tubes were studied by preparing samples viz. mother or hollow, first pass, second pass and third pass. The results indicates that both mechanical properties and microstructures were altered during deformation. Based on the experimental results and discussions in the present work, following conclusions can be drawn:

1. The mechanical properties of cold drawn seamless tubes are sensitive to pre-strain deformation, which was attributed to the reduction of boundary strengthening accompanied by the fragmentation of cementite lamellae which is introduced by pre-torsion.
2. Plastic deformation during cold drawing not only result in the variation of the microhardness but also an inhomogeneous microstructure which is responsible for the difference in microhardness behavior between the surface layer and the center layer.
3. Tensile strength, yield strength, hardness increases with plastic deformation.
4. The mechanical properties can be altered by varying heat treatment schedule.

To measure effective Young's Modulus degradation as a function of equivalent plastic strain which is for uniaxial stress case, tubes of various sizes were cold drawn with varying degree of cold working and plastic pre-strain. Further mechanical properties and Young's Modulus were checked for those samples from varying degree of cold working and plastic pre-strain.

From the results of investigation on the effect of pass schedule on mechanical properties and microstructure of ST52 steel, the following conclusions can be made:

a) Tensile strength, yield strength and hardness of ST 52 steel increased with plastic deformation. Maximum hardness is found 86-87 HRB in first pass .Maximum UTS is found in the same sample.
b) Microstructure reveals ferrite + pearlite in all phases. Elongated grains observed with the increase in number of passes.SEM photographs also validated these observations with grain size evaluations.
c) Micro hardness value is found maximum in final pass with value 162-165 and the grains are found highly stressed. Variation in mechanical properties (UTS, hardness) depends on % reduction in cold drawing. Also the heat treatment temperature and soaking time of stress relieving plays a vital role in determining the properties.
d) Vickers hardness test may vary from surface to core. Increase in reduction sample will have high micro hardness on the surface as compared to the core. The micro hardness evaluation is phase specific. Ferrite will reveal lower hardness as compared to pearlite. Finally we can choose the processing path with varying % reduction and stress relieving temperature and soaking time to obtain the required mechanical properties as per specifications.
e) Initial value of Young's Modulus is anisotropic which depends on the material in plane direction and the percentage reduction during col drawing process.In addition to it, the variation of the Young's Modulus during the deformation process is another important reason of usual underestimation of springback. It was found through this research that elastic Modulus of a material decreases by 10−20% during the plastic deformation reaching the saturation value after 5-7% of plastic strain.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

Authors thanks to Yashashree Tubes Private Limited,F-48, M.I.D.C.,Ahmednagar,M.S.,India for sponsorship and experimentation.

References

[1] Karanjule,D.B.,Bhamare,S.S.,Rao,T.H.,"Variation of Elastic Modulus during cold drawing of seamless tubes and its influence on springback" in Proceedings of the 2016 IEEE IEEM, Bali, Indonesia on 4-6 December 2016 (pp.1025- 1029). https://doi.org/10.1109/IEEM.2016.7798033

[2] Lems,W.,"The change of Young's Modulus after deformation at low temperature and its recovery*", Doctoral dissertation, TU Delft, Delft University of Technology.*1963

[3] Vin,L.J.,Streppl,A.H.,and Singh,U.P.,"A process model for air bending"*,Journal of Materials Processing Technology*,**56,**48–54,1996.

[4] Eggertsen,P.A.,andMattiasson,K.,"On constitutive modeling for springback analysis",*International Journal of Mechanical Sciences*,***52*(**6),804-818,2010. https://doi.org/10.1016/j.ijmecsci.2010.01.008.

[5] Yoshida,F.,Uemori,T.,Fujiwara,K.,"Elastioplastic behavior of steel sheets under in-plane cyclic tension-compression at large strain",*International Journal of Plasticity*,**18**,633–59,2002.https://doi.org/10.1016/S0749-6419(01)00049-3.

[6] Meinders,T.,Konter,A.W.A.,Meijers,S.E.,Atzema,E.H.,Kappert,H.,"A sensitivity analysis on the springback behaviour of the unconstrained bending problem", *Journal of Materials Processing Technology,***9**(3),365-402,2006.http://dx.doi.org/10.1063/1.2011232.

[7] Hu,Y.,and Du,C.,"Quasi static finite element algorithms for sheet metal stamping springback simulation",*Proceedings of NUMISHEET* ,**99**,71-76,1999.

[8] Karafillis,A.P.and Boyce,M.C.,"Tooling and binder design for sheet metal forming processes compensating springback error", *International Journal of Machine Tools & Manufacture*,**36**(4),503–526,1996.https://doi.org/10.1016/0890-6955(95)00023-2.

[9] Zhang,Z.T., and Hu,S.J.,"Stress and residual stress distributions in plane strain bending",*International Journal of Mechanical Sciences*,**40**(6),533–543,1998.https://doi.org/10.1016/S0020-7403(97)00075-1

[10] Li K.P., Carden W.P., and Wagoner R.H.,"Simulation of springback",*International Journal of Mechanical Sciences,***44**(1),103–122,2002.https://doi.org/10.1016/S0020-7403(01)00083-2.

[11] Meinders,T.,Konter,A.W.A.,Meijers,S.E.,Atzema,E.H.,and Kappert,H.,"A sensitivity analysis on the springback behavior of the unconstrained bending problem", *International Journal of FormingProcesses*,**9**(3),365-402,2006.http://dx.doi.org/10.1063/1.2011232.

[12] Papeleux,L.,and Ponthot,J.,"Finite element simulation of springback in tube drawing"*,Journal of Materials Processing Technology*,**125-126**,785–791,2002.https://doi.org/10.1016/S0924-0136(02)00393-X.

[13] Joannic D., and Gelin, J.C.,"Shape defects in tube making operations after springback",*The 4th Inter- national Conference and Workshop on Numerical Simulation of 3D tube making Processes*,29-34, Besancon, France,1999.

[14] Gunnarsson,L.,and Schedin,E.,"Improving the properties of exterior body panels in automobiles using variable blank holder force",*Journal of Materials Processing Technology*,**114**(2),168–173, 2001. https://doi.org/10.1016/S0924-0136(01)00727-0.

[15] Jatczak Chester F.,"Retained austenite and its measurement by X-ray diffraction" *In:SAE800426*, 1657–75,1980.doi.:10.4271/800426.

[16] Yu,H.Y.,"Variation of Elastic Modulus during plastic deformation and its influence on springback",*Materials and Design*,**30**,846-850,2009.https://doi.org/10.1016/j.matdes.2008.05.064.

[17] Abdel-Karim, M.,"Effect of Elastic Modulus variation during plastic deformation on uniaxial and multi axial ratchetting simulations",*European Journal of mechanics-A/Solids,***30**,11-21, 2011. https://doi.org/10.1016/j.euromechsol.2010.08.002.

[18] Chongthairungruang, B., Uthaisangsuk, V., Suranuntchai, S.,and Jirathearanat, S.,(2012),"Experimental and numerical investigation of springback effect for advanced high strength dual phase steel",*Materials and Design*,**39**,318-328,2012.https://doi.org/10.1016/j.matdes.2012.02.055.

[19] Morestin, F.,and Boivin, M.,(1996),"On the necessity of taking into account the variation in Young's Modulus with plastic strain in elastic–plastic software",*Nuclear Engineering and Design*,**162**,107–116, 1996. https://doi.org/10.1016/0029-5493(95)001123-4.

[20] Cleveland,R.,and Ghosh,A.,(2002),"Inelastic effects on springback in metals", *International Journal of Plasticity*,**18**,769-785, 2002. https://doi.org/10.1016/S0749-6419(01)00054-7.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1211-1217 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

IoT: Their Conveniences, Security Challenges and Possible Solutions

Davar Pishva*

Ritsumeikan Asia Pacific University, Faculty of Asia Pacific Studies, 874-8577, Japan

ARTICLE INFO

Article history:
Received: 16 May, 2017
Accepted: 17 July, 2017
Online: 21 July, 2017

Keywords:
Internet of Things
IoT
Collaborative Internet of Things
C-IoT
Mirai IoT botnet
DDoS-for-hire
Thingbot
Security
Privacy
UHG

ABSTRACT

The build-in internet-controlled functions of smart devices such as smart phone, smart television, home healthcare gadget, etc., have made them quite attractive to many segments of consumers. In recent years mankind has witnessed an upsurge usage of such devices for numerous purposes. In this paper, the author is going to show how previously forecasted security challenges of these devices are becoming realities in the present day life. The paper initially provides some introductory information about the topic, mostly by means of survey and citations of previous work. It then highlights the devastating effects of October 21, 2016 DDoS attack which mainly utilized IoT devices. It emphasizes the danger of recently revealed Mirai IoT botnet which serves as the basis for the DDoS-for-hire 'booter'/ 'stresser' service. In terms of counter measures, after highlighting IoT security implementation challenges, numerous approaches are presented. As a long-term solution, an architecture wherein security issues are managed through universal home gateway by network operators in a product based fashion is emphasized. The author shows its technical feasibility and demonstrates its partial materialization in proprietary manners. It then explains why and how numerous stake holders are needed to get together for its wide range commercial implementation. Some immediate necessary safeguard actions and intermediate schemes which include soft infrastructures are also presented for the purpose of risk reduction.

1. Introduction

Advancement in technology has been changing our way of life and digital information has become a social infrastructure. Since the expansion of the Internet in 1990s, network infrastructure has become an indispensable part of social life and industrial activity for mankind.

We are now surrounded by Internet enabled smart devices and there are computer technologies in our cars, phones, watches, entertainment systems, and home appliances. The idea of making use of their existing electronics devices and connecting them to the Internet in conjunction with some specialized software have been leading mankind to a new era of technology known as the "Internet of Things" and commonly referred to as IoT. Their build-in internet-controlled function has made them quite attractive to many segments of consumers. Adoptions of cloud computing, mobile applications and virtualized enterprise architectures have led to a tremendous expansion of applications that are connected to internet resources [1].

Japanese audio visual equipment has been Internet enabled for over a decade now. This has enabled people to enjoy network based services, such as Video on Demand (VOD), Music on Demand (MOD), remote update, e-commerce, remote control, and other similar services. Samsung's 'Family Hub' fridge can order food, play films, and even let you see inside of it remotely [2]. Researchers around the world have come up with an abundance of resourceful ideas on how to effectively use microprocessors and the Internet in other everyday household appliances. According to a study conducted by International Data Corporation, 212 billion "things" will be installed based on IoT with an estimated market value of $8.9 trillion by 2020 [3]. Those "things" will be nothing special but daily used appliances ranging from watch, light bulb to smart television, refrigerator and so on.

Commercial advertising has also greatly benefitted from Internet services and online advertising can even be considered as the foundation of web economy. Unlike conventional forms of advertising, the system of online advertising even allows its target to receive something in return for viewing the advertisement [4]. In other words, our daily life, social activity, industrial and

*Corresponding Author: Davar Pishva, Address: APU, 1-1 Jumonjibaru, Beppu, Oita 874-8577 Japan, Tel: +81-977-78-1261, Email: dpishva@apu.ac.jp

business governance highly depend on information systems. In an industrialized country like Japan, most enterprises use information technology to establish their management governance and every enterprise has its own information for its business. 'IT governance' enables them to improve their efficiency and cost performance. The impacts of information systems on their operations are quite significant. It can be said that information assets have become valuable commodities for business and information systems are the key factors to ensure the growths of enterprises.

The Internet has created new markets around the globe by means of breaking physical barriers and connecting people and organizations of similar interests together. We are now progressing towards collaborative intelligence, something which will impact our connected life and business. The era of Collaborative Internet of Things (C-IoT) is approaching wherein improvement of life quality will have a direct impact on business efficiency enhancement. Furthermore, the process of sensors generating data, data producing knowledge and knowledge driving actions is creating a new direction. The introduction of smart phones has generated wide acceptance and adaption by business, consumer and general population. This is mainly because of their conveniences and many added values like navigation, location based services, etc. Today, we see higher growth in mobile traffic than landline. Introduction of tablets and growth of ecosystem for mobile applications will cause gradual decline in the growth of PCs and rapid end of desktop equipment. People are now equipped with remote access and control capabilities to manage their home environment (energy, safety and security). They can have access to gadgets that help track their physical condition and wellbeing and generate necessary proactive course of action. For instance, there are devices that can monitor driving behaviors, children's actions and elderly people's routine life and generate multiple alerts when deemed necessary.

In other words, smart connected digital life which is composed of smart homes, offices, factories, hospitals, transports, etc., will contribute to better quality of life, generate business efficiency and additional source of significant revenues. Through cyber-physical and social data, we can better understand events and changes in our surrounding environment. Such information can enable us to monitor and control buildings, homes and city infrastructures. They can also provide better healthcare and elderly care services among many other things. However, in order to make effective use of cyber-physical and social data, integration and processing are necessary since their data come from various sources. IoT includes every device that is connected to the Internet, ranging from home automation products like smart thermostats, security cameras, refrigerators, microwaves, to home entertainment devices such as TVs and game machines, to smart retail shelves that know when they need replenishment, to industrial machinery and many more.

Nonetheless, as the value of connectivity and information continue to increase, so does the management complexity, vulnerability and attractiveness to malicious attacks. Considering that traditional approach security mechanism does not work on IoT and consumers have little knowledge or incentive to make them more secured, cybercriminals can make use of IoT for their distributed attacks. It is therefore essential to consider security in their design process, development cycle and in the effective usage of their information systems [5]. This paper is a follow up of recently published work [5] and the author would like to highlight sequence of events which have happened since then. The aim is to emphasize the danger of devastating attacks via IoT devices and numerous challenges that exit in equipping them with appropriate security. It then underlines the importance of previously proposed universal home gateway based security approach and shows some progress in its implementation process. Finally it explains why and how numerous stake holders are needed to get together for its wide range commercial implementation and urges policy makers and big players to take the security issues of IoT devices more seriously.

The rest of the paper is organized as follows: Section 2 discusses vulnerability of Internet connectivity, including those due to the operating nature of Internet Protocol (IP) and the ones originating from various Internet services. Section 3 presents some details on IoT, their vulnerability, including devastating effects of October 21, 2016 DDoS attack which mainly utilized IoT devices, and the danger of recently revealed Mirai IoT botnet. Section 4 explains why traditional security approaches do not work on IoT and highlights the main causes that make IoT attractive weapon for professional attackers. Section 5 presents a comprehensive long-term security implementation strategy, an immediate necessary and implementable safeguard action, and some intermediate schemes. Finally, summary conclusion and practical recommendation are highlighted in section 6.

2. Vulnerability of Internet Connectivity

For various reasons, today's networks are vulnerable to numerous risks, such as information leakage, privacy infringement and data corruption. Operating nature of communication protocol used in the Internet domain, availability of many free software that can carry out numerous attacks and users' unawareness about such issues are some of the main contributing factors.

2.1. Operating Nature of Internet Protocol

Internet protocol suite which is commonly known as TCP/IP (Transmission Control Protocol and Internet Protocol), is used for most Internet applications. Its IP serves as the primary component for carrying out the task of delivering packets from a source to a destination using the IP addresses contained in the packet header. Proper operation of such transaction worldwide requires source and destination to have unique IP address and included in the packet header of their information packets. The fact that each IP address gets associated with a unique entity, enables attackers to trace IP address of each holder through the packet headers.

What makes security implementation more challenging is the fact that Internet has 256 protocols and TCP is just one of them. Other commonly known Internet protocols are UCP (Universal Computer Protocol), and ICMP (Internet Control Message Protocol). Most experts who try to prevent attacks just consider these three protocols in their implementations. Many skilled attackers, however, use less known protocols in their attacks to bypass system security and such trends have been increasing during the past decade. In most cases, their attacks initially succeed until the defenders could figure out what was going on. What makes it worse is the fact that skilled attackers experiment with their new attack methodologies for years before they weaponize them via automation in order to create high volume impacts.

2.2. Vulnerability of Various Internet Services

As mentioned earlier, Internet services have their associated risks and a few them are intentionally created by service providers. Although at superficial level, they are supposed to be for better

service purposes, but are often abused for the sake of business expansion at the expense of their users' victimization. The rest are carried out by attackers, the purpose of whom may range from adventurism at individual level to financial gain at individual/group/corporate level, all the way to socio-political competition at much wider scale. This section will highlight vulnerability of some services which are used by a wide public, most of whom being technology-unaware people.

As we all know, most Internet services, e.g., web browsing, email, social networking, navigation and location services, etc., are provided for free. But most of such service providers are private companies and for their source of income, they mainly rely on online advertisement. Under the pretext of better service, most of these service providers use what is called tracking cookies, to keep track of their users' activities. For example, if they know your location, they can show you what are available in your surroundings. If they know your eating habit, they can guide you to pertinent restaurants. If they know your shopping habit, they can show you similar products, etc. When a user visits multiple websites with the same ad provider, since the same cookies are employed, the ad provider can track the user's activity in numerous sites just by compiling the information via tracking the cookies without the user's knowledge. The bottom line is to identify users, deliver targeted advertisement and persuade them do things which they may not have done under normal condition. Furthermore, since payments of most online advertisement is established on per click basis, just persuading the user to click the ad, automatically generates income for its host.

Other Internet service vulnerability can include e-commerce and social networking. We all know the convenience of online shopping and social networking. The fact that we can buy anything from any part of the world without leaving our home or connect to our networks and anyone in the world free of charge from our pc/tablet/smart phone etc., is a good evidence of such realities. Today, nations are using social networking for their election campaigns and super powers are using it to influence elections or create revolutions, etc.

It is, however, a well-known fact that privacy is implicated in e-Commerce because it requires us to disclose our personal information, such as email address, credit card information, etc., to complete the transactions. After the transaction, the retailer can use the info for their next targeted advertisement and bombard us with spam emails. What is more dangerous is the scenario of data transfer (e.g., when customer database information is sold to third parties or stolen) since it results in identity or credit card theft [6]. Furthermore, when such transactions are done via less secured networks, professional attackers can use numerous techniques to steal our credit card information and use it up to its maximum limit before we realize it. During the past decade, this approach has also penetrated to e-banking and huge amount of money have been stolen from peoples' accounts. [7-8].

At a bigger scale, security experts claim that North Korean cyber attackers have targeted banks in 18 countries for the purpose of Pyongyang using the money to boost its nuclear program [9-10]. Using cyber-attacks on nuclear power plant is even more concerning. On October 10, 2016, Reuters reported that according to Yukiya Amano, the International Atomic Energy Agency director, a nuclear power plant had been disrupted by a cyber-attack in the past two or three years [11]. Although additional details, including where the incident took place were not provided, Amano had said: "This issue of cyber-attacks on nuclear-related facilities or activities should be taken very seriously. We never know if we know everything or if it's the tip of the iceberg."

Vulnerability of social networking services is another example that the author would like to mention in this paper since it is mostly used by non-technical people. It is a scenario similar to IoT and combined utilization of different social networking services has dramatically increased, surpassing nine billion users as of April 2017 [12]. As can be guessed, this has lead cyber threats originating from social-engineering technique to also significantly increase. A study conducted by Verizon Enterprise in 2013 showed that it increased by 4 folds within the single year that they carried out the investigation [13]. Considering its rapid development, its adoption for online advertising and marketing, and its utilization by big powers even in political games, it can be foreseen that intrusions through social networking services will continue to increase in the coming years.

3. IoT and their Vulnerability

As mentioned in the introduction section, IoT includes every device that is connected to the Internet, including those ranging from home automation products like smart thermostats, security cameras, refrigerators, microwaves, to home entertainment devices such as TVs and game machines, smart retail shelves that know when they need replenishment, to industrial machinery and many more.

3.1. Some Elaborations on IoT

The term IoT was invented by a British entrepreneur Kevin Ashton in 1999 and was initially used to refer to a global network of Radio-frequency identification (RFID) connected devices [14]. Although the usage of the term IoT in its present context is less than a decade old, most of the present day IoT devices have existed for decades but they were called under different names such as smart devices, smart systems, smart home appliances, etc. Smart has been a common keyword for such devices before the invention of the term IoT.

Recent rapid expansion of IoT has been due to miniaturization of integrated circuit (IC) chips, tremendous increase in their processing/storage capabilities and huge drop in their production cost. Readily available fast, reliable and free/cheap Internet connection around the globe can be considered another major factor in the rapid expansion. Such developments have made it possible to embed various devices with electronics, software, sensors, and network connectivity and enable them to collect and exchange data. The process of sensors generating data, data producing knowledge and knowledge driving actions has enabled automation, remote sensing and remote control in many areas.

3.2. Vulnerability of IoT

As mentioned earlier, convenience of connecting to the Internet has its associated risks and IoT are no exception in this context. However, threat likelihood level of IoT for a particular type of attack would depend on its function. For example, a healthcare monitor device will be less vulnerable to data alteration attack than a security camera device.

On the other hand, huge number of IoT combined with their weak/no security, make them quite attractive for distributed denial of service (DDoS) attack, regardless of their specific functionality. Although no serious DDoS attack originating from IoT network had been reported until recently, the next section will highlight some devastating recent attacks. Nonetheless, its possibility and

upward trend had been predicted based on simple calculation and projection many years ago. If we assume that only 0.01% of the IoT network gets compromised by 2020, this could lead around 20 million appliances vulnerable to cyber-attacks. Even granting that most of the IoT will only transmit relatively small amounts of data, considering their enormous size, the attack can easily bring down a domain name system (DNS) server, or any another host. This is mainly because DDoS attack uses an integrated effect of its compromised devices. Furthermore, because each compromised element has its own unique IP address, blocking a DDoS attack becomes extremely difficult after it takes place.

3.3. Emergence of Thingbot

Traditional DDoS attack took place via a huge network of compromised PCs. Each PC in the network is transformed into a slave by means of malware. Such network is called botnet which is an abbreviation for words "ro**bot**" and "**net**work". This type of attacks occur without awareness of Internet users. Thingbot is an abbreviation similar to the word botnet, but comprised of the words "**thing**" and "ro**bot**". It indicates a huge network of compromised IoT for launching cyber-attacks. The first large scale thingbot based attack was discovered by security researchers from Proofpoint in early 2014. It consisted of more than 750,000 phishing and spam emails which were launched from thingbot. The network contained more than 100,000 hacked IoT devices, ranging from smart TVs, refrigerators and other smart household appliances [15].

Presence of huge vulnerable IoT devices and introduction of DDoS-for-hire 'booter'/'stresser' service have significantly changed and increased the risk. It enables attackers to launch DDoS attacks against target(s) of their choices in exchange for monetary compensation which usually comes in form of Bitcoin. On October 21, 2016 a DDoS attack utilized at least 150,000 hacked IoT devices and created devastating effects. It resulted in a 1 Tbps traffic, 40 to 50 times higher than normal, and brought down much of the America's internet for few hours [16]. The compromised IoT consisted of digital video recorders (DVRs), surveillance cameras and other smart devices that had weak default passwords. The created huge amount of bogus traffic targeted a major DNS service provider (Dyn) and others. This is the largest of its kind in the history as of today and its origin is traced to Mirai-based thingbot from where a significant volume of the attack traffic was originated.

The Mirai IoT botnet which was revealed in August 2016 has been launching multiple high-profile, high-impact DDoS attacks against numerous Internet properties and services worldwide. Even 2016 Rio Olympics and their associated organization were targeted by sustained and large-scale DDoS attacks, but due to their advance preparedness, they were able to minimize the impact. The botnet serves as the basis for the so called 'booter'/'stresser' service and its nodes are scattered around the world. Their concentrations are, however, higher in Spain, Brazil, Indonesia, Thailand, South Korea, Taiwan, Macau, Hong Kong, and China [17-18]. The botnet has been expanding by incorporating vulnerable IoT devices through automated continuous scanning using well-known, hardcoded administrative credentials that are present in the IoT devices.

4. IoT Security Implementation Challenges

Like other Internet enabled devices, IoT are also prone to numerous risks. Although threat likelihood level of IoT for a particular type of attack depends on its functionality, none of such threats are new. There are established security measures for most of them. However, peculiar characteristics of IoT make implementation of such security measures in a traditional way quite challenging and this section highlights some of its major causes.

4.1. Limited Resources

Unlike traditional computers, IoT are designed to serve only a specific purpose and marketability factors such as low cost, portability, tinier size, etc., prevent them from having powerful processing capabilities. It does not make sense to incorporate several hundred dollars' worth of security and processing capability in IoT devices like smart watch, smart LED, smart toaster, etc., the sale value of which are few tens of dollars. As such, most IoT end up having toy CPUs that cannot handle computationally expensive cryptographic computations and with battery power that prohibits long-lasting or high-peak computations. Hence, built-in full cryptography capability which is common for most computers becomes infeasible for most IoT devices.

4.2. Technology-Unaware Users

Even in those IoT devices, such as smart TV, smart refrigerator, healthcare monitor, etc., into which incorporation of powerful processing capabilities are feasible (cost wise, size wise, etc.), such incorporation does not produce positive results in most cases. This is mainly because most of their users are technology-unaware people, unable to take advantage of such functionality, consider their usage and requirements user unfriendly and additional burden both in terms of their cost and utilization. Hence, our competitive market makes vulnerable IoT devices more attractive in terms of their low cost and user friendliness. It also discourages makers to build secured devices in the production stage.

4.3. Around-the-clock Availability

Traditional computer users, in addition to being more technology aware people in comparison with typical IoT users, turn off their computer after they accomplished their tasks. Furthermore, they stay at their computer most of the time while it is on. In other words, Internet connectivity of traditional computers is enabled while their users actively use them and are disabled when they finish their tasks. In contrast to human-controlled computers, most IoT e.g., smart refrigerators, security cameras, gas/fire sensors, etc., are connected to the Internet 24 hours. Hence, their around-the-clock availability on the Internet, weak/no security make them quite attractive for attackers to continuously experiment their attack techniques. As mentioned earlier, skilled attackers experiment with their new attack techniques for years before they employ them at large scale via automation for high impact achievement.

5. Possible IoT Security Solutions

As we can see, despite the necessity and high importance of equipping IoT with appropriate security measures, its achievement is not that easy. The author originally got involved with investigation of security of smart home appliances over a decade ago and has recently extended it to the broader field of IoT. In this section, the author would like to share his findings in form of a comprehensive long-term security implementation strategy, an immediate necessary and implementable safeguard action, and some intermediate schemes.

Considering the existence of theoretical security measures for most threats against IoT, impracticability of incorporating such measures within each individual IoT inform of powerful security processing capability, the author would like to emphasize the importance of securing the local area network to which they are connected. The fact that most IoT rely on locally available Internet in their vicinity, e.g., home, office, etc., existence of more than one IoT in every home or office and stationary nature of most IoT, the approach can be considered as the most practical one.

5.1. Appropriate Security Model

A decade ago, the author after thoroughly investigating vulnerability of smart home appliances and existence of numerous challenges in equipping individual appliance with appropriate security, came up with the idea of "universal home gateway" (UHG) approach. The idea was to connect all Internet enabled gadgets of a house through a single gateway, build the necessary security processing capability into the gateway and let network operators and service providers to remotely manage the required security via the gateway [19-20]. The purpose was to avoid built-in security features for each gadget and embed all the security related processing capability into the gateway instead. The operation involved channeling of all local and remote access to a family area network (FAN) via the gateway by means some user-friendly authentication schemes. Its required security management was envisioned to be carried out remotely by service related third party experts in exchange for a reasonable overhead cost. The logic was, since vulnerability comes from the Internet connection, if we channel all of our devices' Internet access through a single gateway and secure the gateway, everything gets secured without requiring any of our devices to have its own built-in security related processing capability or us to perform technical security management tasks. This approach will definitely work for IoT devices like smart thermostats, security cameras, refrigerators, microwaves, etc. which are used at a fixed location. However, a different security scheme will still be needed for non-stationary IoT like smart watch, smart phone, tablet, etc. when they utilize Internet from outside the FAN.

The problem was that existing gateways were not compatible with all smart home appliances since they required different wireless technologies and protocols. The author, however, gave detailed functional requirements of such UHG, how to initiate its construction, as well as pertinent responsibilities and commitments of its numerous stakeholders. The proposed idea may have sounded like a far reachable dream and some people even wondered about the viability of being able to control many different devices by means of a single gateway. Nonetheless, even at that time, there were some vendors whose products were marketed with a central controller that could link their smart home appliances together in a FAN environment and offered some limited security and exclusive digital services [21-22]. History has shown that propriety approaches are prone to fail and the author was hopeful that in the near future, a universal controller could control various smart gadgets that are manufactured by many different companies.

5.2. Current State of the Envisioned Security Model

At present, commonly used communication methods for IoT devices include Z-Wave, Zigbee, Powerline, Bluetooth 4.0 and other radio frequency (RF) protocols. Among these, Z-Wave, Zigbee and Powerline are the most common ones for home automation devices. Z-Wave protocol seems to also have better backwards and forwards compatibility, widely adopted as it is easier to find Z-Wave compatible devices from different manufacturers. Figure 1 shows a commercial product of Z-Wave which is compatible with 1,500 products that are produced by 375 companies [23].

Figure 1 Z-Wave Wireless Technology (Z-Wave Alliance, May 2017).

5.3. Long Term Security Strategy

For a comprehensive long-term security implementation strategy, the author would still recommend the originally proposed UHG approach since no other scheme can handle the numerous challenges that exit in equipping IoT with appropriate security measures in a better way. The fact that in less than a decade, compatibility from a few single company based products has grown to 1,500 products that are produced by 375 different companies, shows that it is not a far reachable dream and we are indeed moving in that direction. Unfortunately, however, Z-Wave protocol is still a proprietary standard.

The remaining gaps are to get rid of proprietary standards, involve cooperation of manufacturers, network operators and service providers more widely, arrive at a consensus on minimum security requirements among main stakeholders and involve policy makers on the legal implementation. The size and impact of October 21, 2016 shows the urgency of the matter and an unsuccessful 2016 Rio Olympics attack should not be allowed to turn into a perfect success in a foreseen 2020 Tokyo Olympics attack. The most effective way to achieve this is to:

- Engage a network operator to build dedicated but nonproprietary universal home gateways and become the preferred trusted third party.
- Motivate IoT manufacturers to develop device drivers and application software that can run on such UHG for their control and operation.

This will, however, require support for security-driven business models and involvement of policy makers on the legal aspects of utilizing IoT. It is worth mentioning that thingbots are used against third parties rather than their IoT owners. Furthermore, consensus on minimum security requirements among main stakeholders (e.g., IoT manufacturers, third-party developers, service and solution providers and electronic communications providers) will be essential.

Therefore, design and development of an open system and involvement of big and experienced players like network operators

are vital to the successful adoption of the technologies. This way, their users can enjoy security without having to be aware of the underlying mechanisms. UHG may even be provided free of charge, like the way that cell phone sets and broadband modems are being given away by their respective service providers in Japan. Security pricing may also be implemented on a product (functional group) based scheme.

There are standards bodies which specify how to build these devices and meet their various requirements, e.g., OSGi Alliance [24-25]. It is essential to conform to these standards when building, managing and providing services for IoT in the future. Such an approach will encourage more vendors/manufacturers to conform to these standards, and the standards themselves will evolve as needs arise. The best way to proceed is to develop the security model around IoT and network components that conform to certain standards.

5.4. Immediate Necessary Safeguard Actions

Considering our globalization trends and increase in competition, the levels of bureaucracy that exist in many governments when making new policies, and their difficulties when trying to implement created policies, immediate implementation of the proposed security scheme may be difficult. Nonetheless, through some immediate safeguard actions, one can at least reduce vulnerabilities of IoT for their involvement in possible attacks. The following such actions are therefore recommended to be carried by IoT owners, their sellers or service providers:

- Change their default passwords since attackers usually penetrate through such settings.
- Disable their Universal Plug-and-Play features since they create security loopholes and are enabled by default.
- Disable their Telnet based remote management as Telnet can be used to control them remotely.
- Keep their software up-to-date so as to protect them against the known attacks.

5.5. Some Intermediate Schemes

Let us not forget that achievement of ultimate security goal requires both hard and soft infrastructures. The author has so far mainly focused on the hard infrastructure but soft infrastructures inform of appropriate education, training, guidelines, policies and governance systems are equally important. The fact that thingbot adversely affects third parties rather than their respective IoT owners, implies moral, social and "hopefully in the near future" legal obligation for their possessors. Now that the Internet is an indispensable part of our social life and business activity, security awareness and pertinent education for the common people is also quite important. Furthermore, no security will work when a user does not properly select/protect his/her passwords for various Internet services or does not know what Phishing/Pharming is, clicks such links and provide the requested information. In some collaborative work, the author has explained some relevant incidents which have taken place in online advertising and social networking services [4, 26]. Unfortunately, even in a highly industrialized country like Japan, information security awareness is not part of general education even at university level!

Considering slow progress in the UHG approach and dramatic increase in the use of social networking platforms by many non-technical people, its adoption for online advertising and marketing and victimization of many users, the author also investigated the employment of anonymous communication and its adoption to IoT network. Implementation of TOR (the Onion Router)-based anonymous communication into the IoT network as an effective alternative way to help smart home appliance users protect their privacy and make the smart home appliance system more secure from aforementioned cyber-attacks was the main objective of the work [27]. Such approach can also be effective, particularly for technology-aware users and could even be utilized for non-stationary IoT devices.

6. Conclusion

This paper explained numerous conveniences of IoT and its projected huge market in the near future. It also showed that like other Internet enabled devices, IoT are also prone to numerous risks but there exist theoretical security measures for most of them. Nonetheless, peculiar characteristics of IoT make implementation of such security measures in a traditional way quite challenging and their huge number make them quite attractive for DDoS attack. It highlighted the devastating effects of October 21, 2016 DDoS attack and emphasized the danger of Mirai IoT botnet which serves as the basis for the DDoS-for-hire 'booter'/'stresser' service and enables attackers to carry out DDoS attacks against targets of their choice in exchange for monetary compensation that usually comes in the form of Bitcoin payments. An architecture wherein security issues are managed through universal home gateway by network operators in a product based fashion was cited. It emphasized support for security-driven business models, consensus on minimum security requirements among main stakeholders and involvement of policy makers on the legal aspects of IoT environments. Some immediate necessary safeguard actions and intermediate schemes which included soft infrastructures were also recommended for the purpose of risk reduction. It is hoped that policy makers and big players take the security issues of IoT devices quite seriously and maintain the October 21, 2016 DDoS attack as the largest of its kind in the history forever.

Conflict of Interest

The author declare no conflict of interest.

References

[1] C. Drake, "FireHost detects surge in SQL injection for Q3 2013 and cross-site scripting is rising", retrieved from http://www.firehost.com/company/-newsroom/press-releases/firehost-detectssurge-in-sql-injection-for-q3-2013-with-cross-site-scripting-also-rising/, October 2013

[2] Samsung Group, "Home has a new hub", retrieved from http://www.samsung.com/us/explore/family-hub-refrigerator/, July 2016.

[3] ZDNet, "Internet of Things: $8.9 trillion market in 2020, 212 billion connected things", retrieved from http://www.zdnet.com/ article/internet-of-things-8-9-trillion-market-in-2020-212-billion-connected-things/, May 2017.

[4] Angelia, D. Pishva, "Online advertising and its security and privacy concerns" in proceeding of the 15th International Conference on Advanced Communication Technology (vol. 1, pp. 372-377), Seoul Korea, 2013.

[5] D. Pishva, "Internet of Things: security and privacy issues and possible solution" ICACT Transactions on Advanced Communications Technology (TACT), **5**(2), 797-808, 2016.

[6] M.J. Metzger, "Communication privacy management in electronic commerce" Journal of Computer-Mediated Communication, **12**(2), 335–361, January 2007.

[7] Mybroadband, "How criminals steal money from your online bank account", retrieved from https://mybroadband.co.za/news/banking/125652 -how-criminals-steal-money-from-your-online-bank-account.html, May 2017.

[8] N. Kaur, "A survey on online banking system attacks and its countermeasures" International Journal of Computer Science and Network Security (IJCSNS), **15**(3), 57-61, 2015.

[9] MailOnline, "Researchers say group linked to North Korea is attacking banks around the world", retrieved from http://www.dailymail.co.uk/news/article-4379248/North-Korea-hackers-target-banks-18-countries.html, May 2017.

[10] Reuters, "North Korean hacking group behind recent attacks on banks: Symantec", retrieved from http://www.reuters.com/article/us-cyber-northkorea-symantec-idUSKBN16M37J, May 2017.

[11] Wired, "International Atomic Energy Agency director says nuclear power plant was impacted by malicious hack", retrieved from https://www.wired.com/2016/10/security-news-week-hackers-hit-nuclear-plant/, May 2017.

[12] Statista, "Most famous social network sites worldwide as of April 2017, ranked by number of active users (in millions)", retrieved from http://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/, May 2017.

[13] Verizon Enterprise, "The 2013 Data Breach Investigations Report", retrieved from http://www.verizonenterprise.com/DBIR/2013, 2013.

[14] Wikipedia, "Internet of things", retrieved from https://en.wikipedia.org/wiki/Internet_of_things, May 2017.

[15] Cyber Defense Magazine, "Proofpoint discovered more than 750,000 phishing and spam emails launched from thingbots including televisions, fridge", retrieved from http://www.cyberdefensemagazine.com/ iot-discovered-first-internet-of-things-cyberattack-on-large-scale/, May 2017.

[16] P. Paganini, "150,000 IoT devices behind the 1Tbps DDoS attack on OVH", retrieved from http://securityaffairs.co/wordpress/51726/ cyber-crime/ovh-hit-botnet-iot.html, November 2016.

[17] Arbor Networks, "DDoS attacks from IoT botnets don't have to mean game over", retrieved from https://www.arbornetworks.com/ blog/asert/rio-olympics-take-gold-40gbsec-sustained-ddos-attacks/, November 2016.

[18] Arbor Networks, "The lizard brain of lizard stressor", retrieved from https://www.arbornetworks.com/blog/asert/lizard-brain-lizardstresser/, November 2016.

[19] D. Pishva, K. Takeda, "A product based security model for smart home appliances" in proceeding of 40th Annual IEEE International Carnahan Conferences on Security Technology (vol. 1, pp. 234-242), Lexington USA, 2006.

[20] D. Pishva, K. Takeda, "A product based security model for smart home appliances" IEEE Aerospace and Electronics System Magazine, **23**(10), 32-41, 2008.

[21] K. Yoshimi, "Addressing security requirements of network enabled home appliances, next generation IP infrastructure group report (WG3-1)", retrieved (in Japanese) from Ministry of Internal Affairs and Communication (MIC):http://www.soumu.go.jp/joho_tsusin/policyreports/chousa/jise_ip/pdf/050217_1_s1.pdf, December 2005.

[22] T. Satoru, "Information appliance system authentication technology, next generation IP infrastructure group report (WG3-2)", retrieved (in Japanese) from Ministry of Internal Affairs and Communication (MIC): http://www.soumu.go.jp/joho_tsusin/policyreports/chousa/jise_ip/pdf/050217_1_s2.pdf, January, 2006.

[23] Z-Wave Alliance, "The Internet of Things is powered by Z-Wave", retrieved from: http://z-wavealliance.org/z-wave_for_consumers/, May 2017.

[24] OSGi Alliance, "Internet of Things", retrieved from https://www.osgi.org/business/markets-and-solutions/markets-and-solutions%E2%80%8Einternet-of-things/, May 2017.

[25] OSGi Service Platform Products, "OSGi Markets and Solutions", retrieved from http://www.osgi.org/products/products.asp?section=3, July 2015.

[26] N.P. Hoang, D. Pishva, "Anonymous communication and its importance in social networking" in proceeding of the 16th International Conference on Advanced Communication Technology (vol. 1, pp. 34-39), Seoul Korea, 2014.

[27] N.P. Hoang, D. Pishva, "A TOR-based anonymous communication approach to secure smart home appliances" ICACT Transactions on Advanced Communications Technology (TACT), **3**(5), 517-525, 2014.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1218-1226 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Recognition and Position Estimation for Multiple Labware Transportation Using Kinect V2 and Mobile Robots

Mohammed Myasar Ali[*,1], Hui Liu[1], Norbert Stoll[2], Kerstin Thurow[1]

[1]*Center for Life Science Automation (celisca), Rostock University, 18119 Rostock, Germany*

[2]*Institute of Automation, Rostock University, 18119 Rostock, Germany*

ARTICLE INFO

Article history:
Received: 24 May, 2017
Accepted: 03 July, 2017
Online: 21 July, 2017

Keywords :
Labware recognition and position estimation
Mobile robot transportation
Kinect sensor V2

ABSTRACT

Mobile robots can be used to perform transportation tasks for different objects. These tasks have to be implemented carefully. Therefore, an accurate approach for object recognition and position estimation is required. This work presents a concept for identification and position estimation of multiple labware. These labware, which contain chemical and biological components, have to be manipulated and transported in life science laboratories using H20 mobile robots. The H20 robot has dual 6-DOF arms with 2-DOF grippers. Different marks are used to be attached with the labware lid for identification process. The Kinect sensor V2 is used to recognize and localize the mark of the required labware on a wide workstation. The difference of performance between the Kinect V1 and V2 is illustrated. SURF algorithm (Speeded-Up Robust Features) is used to recognize the target according to its local features. Some preprocessing steps are applied to the RGB frame to enhance the image features. The effects of strong lighting condition are eliminated by using polarization and intensity filters which are attached to the Kinect camera. The position estimation step is performed by applying a mapping process form the color frame to the depth frame of Kinect. The communication procedure between the Kinect platform and other robot platforms is done using client-server model. An efficient performance with high success rate is obtained under different lighting conditions.

1. Introduction

This paper is an extension of work originally presented in the international symposium on computational intelligence and informatics (CINTI 2016) [1]. This work shows an approach to identify required labware for mobile robot transportation in life science laboratories. In general, the realization of objects transportation using mobile robots leads to an increase of the productivity and saving human resources in the working environment. This requires several prerequisites like object recognition with position estimation, arm control, and a robot navigation system. The navigation system includes the path planning, mapping, and robot localization in the working environment. Related to the object manipulation, the robotic arm has to be guided to the target pose. The object pose can be acquired visually using a suitable sensor with a proper recognition algorithm. The Kinect V2 sensor fixed on the H20 robot is used in this work to identify and localize multiple labware. The H20 robot is a wireless networked autonomous humanoid mobile robot. It has a PC tablet, dual arms, and an indoor GPS navigation system. A Kinect holder is installed on the H20 body in a way that the labware on the workstation can be visualized clearly. Some technical achievements have been developed at the Center for Life Science Automation (celisca, University of Rostock) to improve the transportation system of the H20 mobile robots [2]-[4]. The H20 mobile robot with different labware and tube racks is shown in Figure.1. The Kinect sensor provides color frame and depth frame of the view. The RGB frame has to be processed to find the required target. This process includes the extraction of meaningful features from the image that lead to the identification of the object. Different target features can be used for this procedure such as color, shape, edges, size, and local features. Some algorithms use

[*]Corresponding Author: Mohammed Myasar Ali, Center for Life Science Automation (celisca), 18119 Rostock, Germany
Email: mohammed.myasar.ali@celisca.de

multiple features related to the target to create a robust strategy for recognition process.

Figure 1. H20 mobile robot with different tube racks.

In order to identify the required object using visual sensors, it is necessary to apply different techniques to the captured image. There are several features which can be extracted from the image to find the target. In general, these features are divided into two categories. The first is the appearance features of objects such as the color and intensity. The second is local features of the target itself. The local features have to be extracted and matched with the features in the database related to the object of interest. Image segmentation using color feature is one of the easiest methods to find the required target in the view. This method is suitable for real-time applications because it does not require a detailed prior information about the target. Color detection can be performed using different representations of color system such as RGB, HSV, and YCbCr. The HSV (hue, saturation, value) color system is more robust in dynamic lightness conditions because it is insusceptible to illumination. Sanchez-Lopez et al. used the HSV color system to track the target for service robot applications [5]. The edges detection can also be used to extract the edges of the target. Some filters are commonly used for this purpose such as Canny, Sobel, and Laplace filters. Furthermore, a set of rules of shape primitives can be applied for shape recognition such as circles, triangles, and rectangles. Yamazaki et al. used edges and shape detection with HSV color segmentation to recognize some foods and kitchen tools [6]. The success rate of object recognition can be improved by using multiple features related to the target at the same time.

In order to use specific local textures of the target as an identification reference, feature matching algorithms can be used. SIFT (scale invariant feature transform), SURF (Speeded-Up robust features), and FAST (Features from Accelerated Segment Test) are the most common algorithms for local feature descriptors to perform object recognition [7]-[9]. Theses algorithms are somehow independent to the changes in scale, illumination, and orientation. Collet et al. extracted the local descriptors for the registration and recognition of learned metric 3D models of house-hold objects using a multi-view system [10]. Marchand et al. used the features like corners and contours of known objects to perform a visual tracking method [11]. The speeded-up robust features (SURF) algorithm can be considered as an efficient object recognition method with a fast scale- and rotation-invariant detector and descriptor. The object tracker decreases the required time by providing the full region which is occupied by the target in the image at every frame. The target region is jointly estimated by iteratively updating the region information and location obtained from the previous frame. The region of interest (ROI) which contains the target is obtained using the object position in the image coordinates. In the successive images, only the interest points in this updated ROI are extracted and matched to find the object. Anh et al. proposed an object tracking method based on SURF for safe grasping tasks [12]. Zickler et al. used humanoid robots to achieve detection and localization of multiple objects on the kitchen desk [13]. Katsuki et al. attached specific marks on the required objects to handle various objects in home or office environments using robot system [14]. For the localization process of the object in the real environment, 3D visual sensor like Microsoft Kinect is more preferable. Chung et al. used the Kinect sensor installed on a service robot to help humans in object transportation [15]. Ramisa et al. used the Kinect for cloth grasping by finding the manipulation points in the depth frames [16].

In this work, the approach of recognition and position estimation of different labware is presented. Several concepts and challenges are taken into the consideration to realize an efficient performance. The visual feedback is necessary to perform a high precision manipulation of the labware that contains chemical and biological components. The strong and glossy lighting conditions in addition to the appearance of the labware affect the identification process. Therefore, a specific mark is attached to each labware to be to be distinguished from each other on a wide workstation. Speeded-Up robust features algorithm (SURF) and HSV color segmentation are used to recognize different marks. Some preprocessing steps are applied to the image to enhance its features for the recognition step. The Kinect V2 is used for this work. The difference of performance between the Kinect V1 and V2 is illustrated. Polarization and intensity filters are attached to the Kinect to reduce the effects of strong and glossy light. The connection between the Kinect platform and other robot platforms is established through a client-server model.

This paper is organized as follows: in section 2, the problem definition is presented. The proposed methodology is given in section 3. Section 4 shows the labware identification and localization process which will be followed by the system integration. Finally, the results are concluded and discussed.

2. Problem Definition and Restriction

The future of life sciences laboratories depends significantly on the innovations of automated solutions in the entire scope. Different scientific tasks, like biological testing and sample preparation, are performed in the laboratories by using automation equipments. This leads to realize high throughput, workflow optimization and reliable measurement results. Robots and especially mobile robots are very important in the life science field. Mobile robots increase the productivity and save human resources by connecting all workstations in different laboratories. For labware transportation, an intelligent procedure to grasp the target object and placing it at the right position on the workstation is required. The success of this procedure depends significantly on the success of recognition and position estimation for the required

labware. Fig. 2 shows the wide workstation which consists of 8 positions of labware containers. The required labware and its position on the workstation have to be identified and the H20 robot has to change its position to manipulate it. The laboratories and the workstations have strong lighting conditions especially with sunny weather. The glossy ceiling light blinds the visual sensor which affects the identification process since it changes the labware appearance. Also, the workstation can be affected by the sunlight if it is close to a window or is surrounded with objects which reflect some light.

Figure 2. Eight locations of labware containers on the workstation.

According to this situation, a proper visual sensor has to be located in a suitable pose on the robot with a specific height to provide an adequate view for the workbench. Moreover, the view angle of the visual sensor related to the workbench plays a role in the identification process by reducing the light effects. The H20 robot has stereo cameras in the head. The stereo vision system is fixed on a pan-tilt joints module for object tracking. The head cameras are not appropriate for labware identification task because of the limited height and distance between the head and the workstation. The cameras don't provide a clear view for the whole workbench especially at the ends of left and right sides of it. Moreover, several complex steps have to be performed in stereo vision systems to obtain the depth data. The images of stereo cameras have to be processed by steps such as undistortion, rectification, correspondence, and reprojection. It is important to correspond the video quality, resolution, and contrast in the both cameras. The two cameras have to be mounted accurately to guarantee that all their axes are parallel. Also, the kinematic solution of the head joints has to be derived and the target has to be tracked to be seen by the two cameras. To avoid these complex issues, the Kinect sensor can be considered as a proper choice for this work. It can be used to provide a sufficient and clear view for the whole workstation after fixing it in a suitable pose over a holder. Also, the Kinect provides the depth data directly without the need of performing all the complex steps which are previously mentioned for the stereo vision system.

3. Proposed Methodology

In the Center for Life Science Automation, different automation islands are connected with each other by the cooperation of stationary and mobile robots. This cooperation requires an appropriate hierarchical management systems. The mobile robots are used for maneuvering between the adjacent laboratories for transporting multiple labware and tube racks. The overall workflow starts with the user or with the hierarchical workflow management system (HWMS) to decide which target has to be manipulated and transported [17]. The manipulation system can be splitted into two parts, the target localization and the arm controller. The target localization software with the visual sensor is utilized to detect the target and to estimate its pose. The pose information is sent to the arm controller software to guide the robotic arm. The information exchange between the two parts is performed using a TCP/IP socket through client-server communication model. Fig. 3 shows the block diagram of the manipulation process.

The target pose has to be calculated related to the arm base. Then, an accurate kinematic model has to be used to control the arm joints. The kinematic analysis is how to describe the arm links motion without considering its forces. There are two terms related to the kinematic analysis, forward kinematics (FK) and inverse kinematics (IK). Using the forward kinematics model, the end-effector pose relative to the arm base can be found according to the given joint angles. On the other hand, the inverse kinematics model describes how to find the required joint angles for the given end-effector pose. The analytic solution of IK problem has been found and applied physically on the H20 arms to guide them to the target [18][19][20].

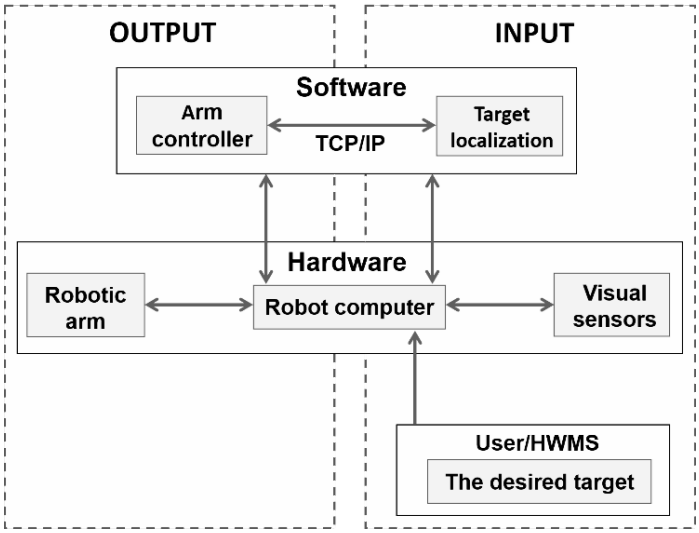


Figure 3. Block diagram of manipulation process.

The Kinect sensor can be considered as an appropriate choice for the labware manipulation approach. The Kinect provides a high quality color and depth information which is directly obtained without applying some complicated steps on the image as in the stereo vision. Both versions V1 and V2 of the Kinect sensor have been used for this work. The Kinect V2 has a wider horizontal and vertical view than V1. The RGB and depth cameras have a higher resolution and the depth measurements are more accurate. Also, the Kinect V2 uses the time-of-flight principle, whereas the Kinect V1 is based on structured light to provide the depth data [21]. Fig. 4 shows the two versions of Kinect and Table 1 shows the differences between them.

(*A*) (*B*)

Figure 4. Microsoft Kinect sensor, A: V1, B: V2.

Table 1: Differences of features between Kinect Sensor V1 and V2.

Features	Kinect V1	Kinect V2
RGB camera	640 X 480	1920 X 1080
Depth camera	320 X 240	512 X 424
Max depth	~ 4.5 M	~ 4.5 M
Min depth	40 cm in near mode	50 cm
Horizontal FOV	57 degrees	70 degrees
Vertical FOV	43 degrees	60 degrees
Tilt motor	Yes	No
USB standard	2.0	3.0
Supported OS	Win 7, Win 8	Win 8, Win 10

A holder has to be installed on the H20 body for fixing the Kinect sensor. The height and tilt angle of the holder have to be appropriate to realize a wide and clear view for all the labwares on the workstation. The distance between the Kinect and the workstation has to be suitable to obtain the target position where the minimum depth value of Kinect V2 is 50cm. Also, the FOV of the RGB camera and depth camera are not compatible. This leads to the fact that not every point in the RGB frame has a related 3D position value since the FOV of the RGB camera is wider than the FOV of the depth camera. The Kinect holder should not obstruct the head movement. In addition, the FOV of stargazer, which is located behind the H20 head, has not to be affected by the holder or the Kinect (see Fig. 5). The stargazer module is used to detect passive ceiling landmarks for mobile robot localization [22]. Moreover, any kind of changes in the Kinect position and orientation has to be avoided during the robot movements. The Kinect sensor has been supplied with the required power using a 12V battery with stabilizer. Furthermore, the interfacing between Kinect V2 and H20 robot is set through a USB-3 port of the H20 tablet. SURF algorithm for local features description and recognition is used to identify multiple labware and tube racks. This can be performed by extracting local features from the reference image to be identified with the current image. C-sharp programming language has been used to develop the manipulation system of the H20 robots. Fig. 6 shows the structure of the labware manipulation system.

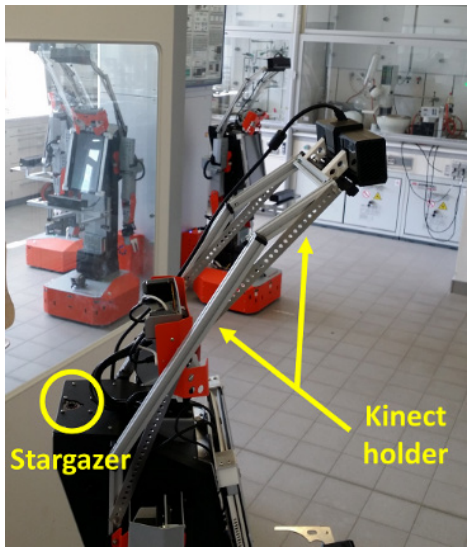


Figure 5. Frame of Kinect holder.

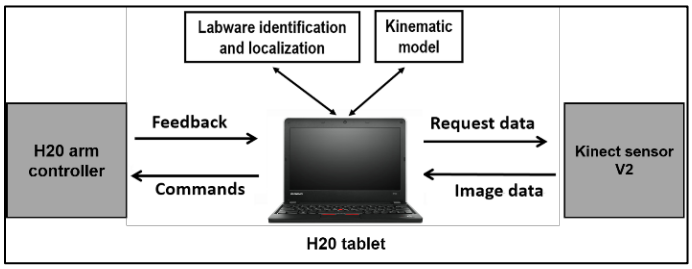


Figure 6. Structure of labware manipulation system.

Arm gripper design plays an important role in the realization of secure manipulation tasks. Several models of arm grippers and labware containers have been designed for arm manipulation [18]-[20]. Fig. 7 shows the final 3D model of the gripper with container. The goal of this design is to handle heavy labware by decreasing the lever arm of the wrist joint. This brings the labware weight center closer to the wrist. In this case, less torque is required from the wrist joint for lifting the labware. The maximum payload which can be handled with this design is 700g. To implement the visual manipulation, the labware itself has to be recognized and localized. Different object recognition algorithms can be used such as SIFT, SURF, and FAST, which are widely used for object recognition applications. The next section illustrates the process of labware recognition and position estimation using SURF algorithm with Kinect V2 sensor.

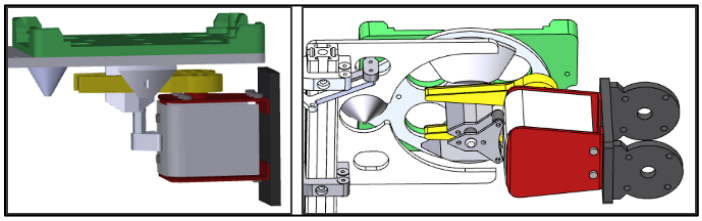


Figure 7. 3D design of labware container with gripper.

4. Labware Identification and Localization

Since the H20 robots deal with different kinds of labware, a reliable technique is required to recognize them. Different tests have been performed to show the performance of identification and localization for labwares at different positions on the workstation and under different lighting conditions. The identification tests have been performed using Microsoft Visual Studio 2015 with C# programming language. SURF algorithm is the most appropriate method that can be implemented using this programming language. The project is running on a Windows 10 platform in the H20 tablet. The process starts with an offline step by capturing an image of the target to be saved in the database as a matching reference. Strong and glossy light as well as sunlight may affect the recognition process. Therefore, to start with the online process and to decrease the light effects, some pre-processing steps have to be applied on the live image to ensure a better performance with SURF algorithm. Different procedures have been applied such as contrast with brightness correction, histogram equalization, HSV conversion, grayscale conversion, etc. The average required time for recognition process is about 3 seconds, which is long according to the application perspective. This long time is related to the features extraction and identification from a high resolution image of Kinect (1920X1080). To cope with this issue, the region of interest (ROI) can be extracted from the image using a cropping technique. The cropping process for the area of workstation with labwares is performed in the Y-axis of the image only as shown in

Fig. 8. The cropping area can be estimated according to the error range of the distance between the robot and the workbench which is about ±3cm. The resolution of the image after the cropping step is (1920X400). The recognition process requires about (1.1~1.5 sec.) using the cropped image. Thus, the required time has been decreased to the half after using the cropping image.

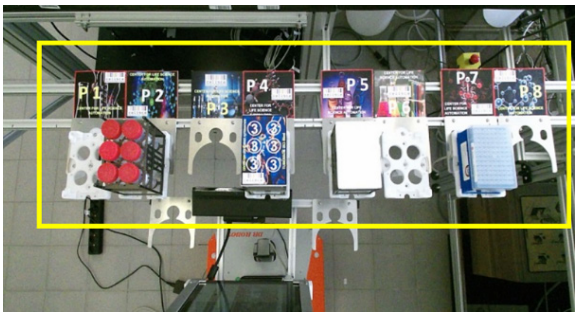


Figure 8. ROI area after cropping step.

4.1. Labware Identification Process

The recognition procedure has been performed for some tube racks without lid. In general, the lids are entirely either white or transparent. These lids are necessary to protect the tubes components from cross contamination and evaporation. Fig. 9 shows the 6-tubes rack which has different top views at different positions on the workstation. The variation in the top view are related to the 3D features of the tubes and also related to the angle of Kinect view. The lighting conditions play an effective role in the appearance of tubes rack at different positions.

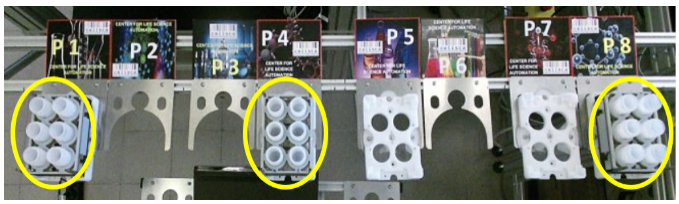


Figure 9. 6-tubes rack at different positions on the workstation.

The 6-tubes rack has been placed at 8 different positions on the workstation for the recognition process. The 8 positions are ordered from left to right in the sequence of P1-to-P8. Fig. 10 shows the recognition outcome of the 6-tubes rack by drawing a polygon around it with cross to specify the center point.

Figure 10. Identification of 6-tubes rack on the workstation.

Fig. 11 shows the success rate of the recognition process which has been tested 20 times at each position using different preprocessing steps and under different lighting conditions. The first test has been done without turning on the glossy ceiling light. The image acquired from Kinect has been used directly without applying any preprocessing steps. The identification results of the first test at each position are titled with (N) in Fig. 11. The success rate at P4 is 100% (20/20) since the top view of the tubes rack is very clear at this position. At P3, the success rate is 65% (13/20) while it is zero at the rest positions. The overall success rate for the first test is~20% (33/160). This rate is related to the 3D features of the tubes top view which gives different appearance at each position according to the angle of Kinect view. The second test has been done after turning on the ceiling glossy light. The overall success rate of the second test (N_L) is~19% (31/160) as shown in Fig. 11. The third and fourth tests have been performed after applying the brightness and contrast corrections for the Kinect image. The overall success rates are~9% and~22% respectively. The fifth and sixth tests have been performed after converting the Kinect image to grayscale and applying the histogram equalization. The overall success rates are 20% and~ 17% respectively.

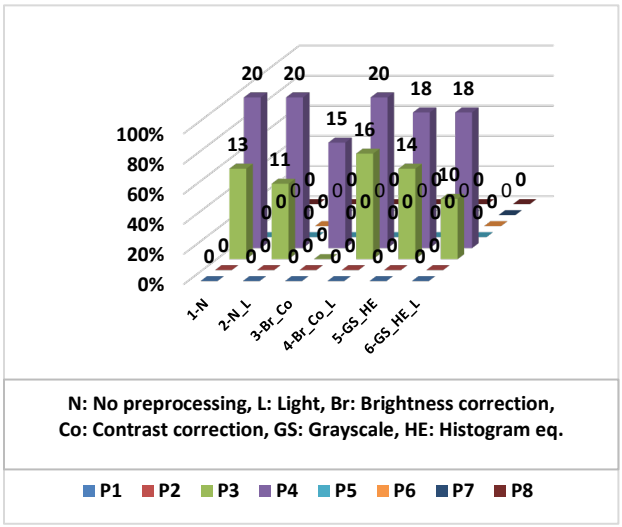


Figure 11. Success rate of 6-tubes rack identification.

The best overall success rate (~22%) has been obtained after applying the brightness with contrast correction and in lighted environment. The false positive rate for this case is about~ 13% (21/160). The false positive condition causes identifying a wrong target in the view which has to be avoided because it affects the manipulation process. As a conclusion from the overall results, it can be realized that the labware without lid has to be located at a fixed position on the workstation to be identified. This can be noticed clearly according to the overall success rate at P4 which is about ~92% (110/120). Table 2 shows the success rates summary of the identification tests for the 6-tubes rack where the term O.S.R.T means the overall success rate for each test and the term O.S.R.P means the overall success rate at each position.

Table 2: Results summary of 6-tubes rack identification.

Test	P1,2	P3	P4	P5,6	P7,8	O.S.R.T
N	0%	65%	100%	0%	0%	~20%
N_L	0%	55%	100%	0%	0%	~19%
Br_Co	0%	0%	70%	0%	0%	~9%
Br_Co_L	0%	80%	100%	0%	0%	~22%
GS_HE	0%	70%	90%	0%	0%	20%
GS_HE_L	0%	50%	90%	0%	0%	~17%
O.S.R.P	0%	~52%	~92%	0%	0%	

Specific marks have been attached to each labware lid to improve the identification process at each position on the workbench. Fig. 12 shows different lid marks which can be used to distinguish different labwares. The mark contains an information like labware name and its type with specified number. The number is used for differentiation and classification purposes. This information with the background picture gives sufficient features to differentiate multiple labware. The more features the mark has, the higher success rate of identification will be realized. Also, it is necessary to select the suitable colors of the background mark which reduce the effects of strong lighting conditions. It is known that the dark colors are much preferable to avoid the light reflection. The Kinect V2 has been used to recognize the required tube rack (6 microwaves tubes) according to the related mark with number 8 as shown in Fig. 13. The marks have been printed using coarse paper since it is more robust against light reflection.

Figure 12. Different designs of marks to distinguish multiple labware.

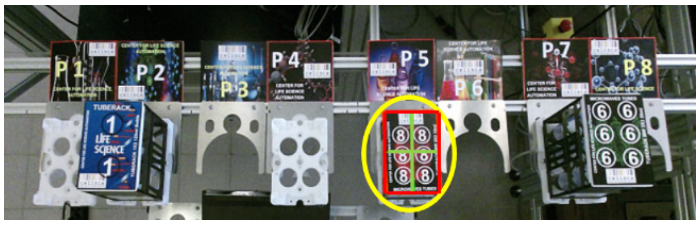


Figure 13. Labware identification according to the related mark.

The identification process has been performed for the lid mark at different positions on the workbench and under different lighting conditions. The test has been repeated 20 times at each position. Different preprocessing steps have been used before applying SURF algorithm on the Kinect image. Fig. 14 shows the recognition success rate of the mark with number 8 at each position. The improvement in the success rate results from the 2D property of mark which is not easily to be influenced by the angle of Kinect view. Also, the adequate features in the mark help to find the required one easily. In comparison with Fig. 11, it is clear that the success rate has been improved at the positions that are located at the horizontal ends of image like P1, P7, and P8. This is related to the reflected light from the mark and to the angle of view at these positions. It can be noticed also that the brightness with contrast correction improves the recognition process that is slightly affected under strong lighting conditions. The overall success rates for this case at all the 8 positions without and with glossy ceiling light are~ 98% and~ 97% respectively. On the other hand, the grayscale conversion with histogram equalization reduced the success rate of mark recognition. Furthermore, the valuable improvement that has been obtained is zero false positive rate. Tables 3 shows the success rates summary of the identification tests for the lid mark.

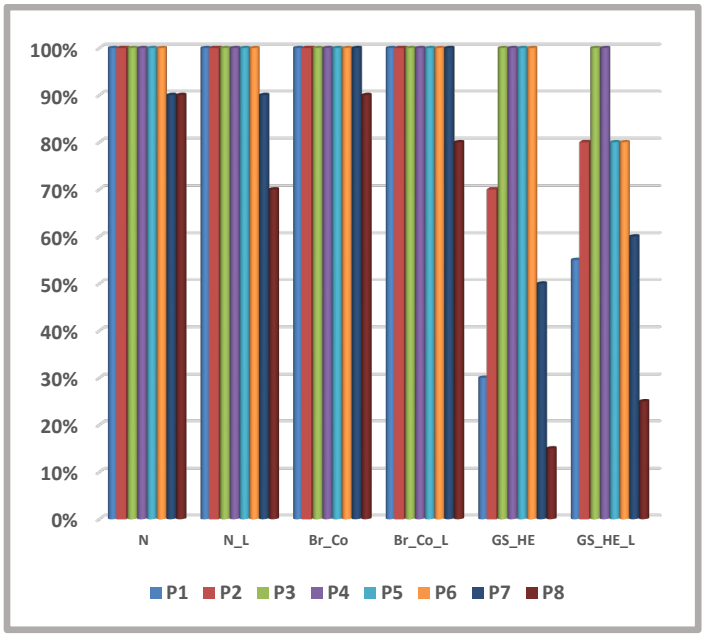


Figure 14. Success rate of lid mark identification.

Table 3: Results summary of lid mark identification.

Test	P1,2	P3,4	P5,6	P7,8	O.S.R.T
N	100%	100%	100%	90%	~97%
N_L	100%	100%	100%	80%	~97%
Br_Co	100%	100%	100%	95%	~98%
Br_Co_L	100%	100%	100%	90%	~97%
GS_HE	50%	100%	100%	~32%	~70%
GS_HE_L	~67%	100%	80%	~42%	~72%
O.S.R.P	~86%	100%	~97%	72%	

The realization of 100% success rate is the goal which has to be achieved especially under strong lighting conditions. The most appropriate solution to deal with lighting effects is to change the camera exposure time. Decreasing the exposure time leads to a reduction of the light effects on the image acquired from the camera. But Microsoft Company has blocked the camera setting of Kinect V2 which prevents the possibility of changing the exposure time or any other camera settings. Therefore, polarization and intensity filters have been affixed on the Kinect V2 camera to decrease the lighting effects as shown in Fig. 15. The intensity filters are used to decrease the brightness in the image and sharpen its edges and features. On the other hand, the polarization filters are used to increase the color saturation and decrease the reflections from glass, metals or other shiny surfaces. The success rate of mark identification under strong and glossy lighting condition is 100% at the all 8 positions on the workstation. Table 4 shows the success rates summary of the identification tests for the lid mark using filters.

Figure 15. Kinect V2 with camera filter.

Table 4: Results summary of mark identification (with filters).

Test	P1	P2,3	P4,5	P6,7	P8	O.S.R.T
N	95%	100%	100%	100%	100%	99%
N_L	100%	100%	100%	100%	100%	100%
O.S.R.P	~98%	100%	100%	100%	100%	

The recognition process requires between 1-1.5 second to be performed after cropping the ROI area. To decrease this time, the unwanted and unimportant features have to be removed from the cropped image. For this purpose, lid marks have been designed with red color features as shown in Fig. 16. Color filters can be used to extract the required color from the image after removing all the unwanted colors. This has been performed by using the HSV color filtering as a preprocessing step for the cropped image. The process of red mark identification starts with applying brightness correction with histogram equalization to increases the color saturation. Then, the color system is converted from RGB to HSV because it is more robust against strong lighting conditions. The next step is to extract the red color by removing all the unwanted colors in the image using color segmentation. The final image is converted to a binary image (black and white) to be prepared for the SURF process. Fig. 16 shows the result of this method which requires about 0.5 sec. The best success rate for this method is ~96% at all 8 positions on the workstation with using the filters in the lighted environment. Table 5 shows the results summary of this test.

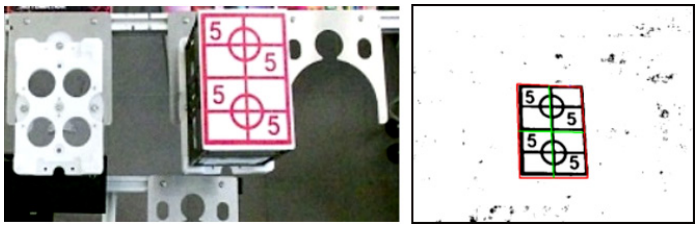


Figure 16. Identification of red-white mark.

Table 5: Results summary of red mark identification.

Test	P1	P2,…6	P7	P8	O.S.R.T
Br_H.E.	85%	100%	85%	80%	~94%
H.E._F	90%	100%	90%	85%	~96%
O.S.R.P	~88%	100%	~88%	~83%	

Table II summarizes the best results of the overall tests after applying the brightness and contrast correction on the raw image and under strong lighting condition.

Table 6: Overall summary of the Best Results.

Test case	Success rate	False pos. rate	Time
Without lid	~22%	14%	1.4sec
With lid mark	~97%	0%	1.3sec
Mark with filters	100%	0%	1.1sec
Red mark with filters	~96%	0%	0.5sec

The using of marks for labware identification and manipulation is very useful. These marks can be recognized even when they are partially occluded by some object as shown in Fig. 17. This can be considered as one of the advantages of using this approach. The mobile robot can still find and grasp the required labware even if the related mark is partially seen by the Kinect.

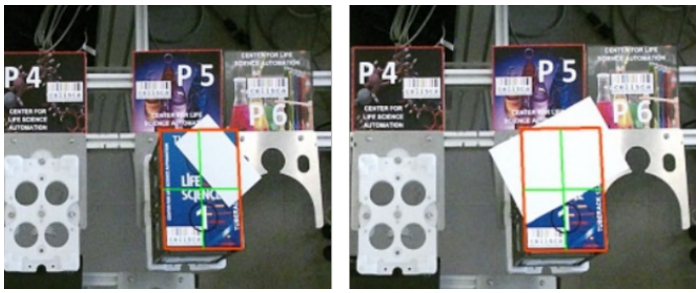


Figure 17. Mark identification with partial occlusion.

4.2. Holder Identification

The required holder has also to be identified and localized for performing the placing tasks visually. To realize that, placing mark has been designed and fixed in front of each holder as shown in Fig. 18. Each placing mark has a number and it is used for recognition and position estimation of the required holder to guide the robotic arm to the right place. In case of existing labware on the workbench, these labwares do not block the placing marks to be seen by the Kinect sensor.

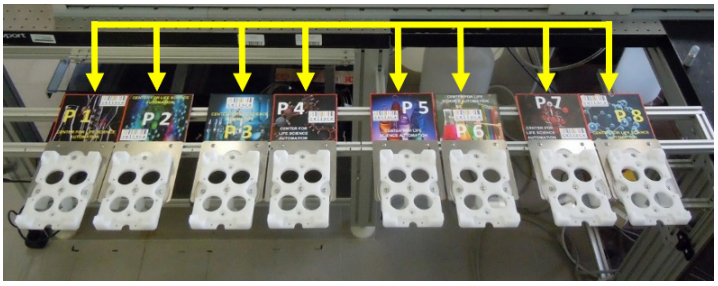


Figure 18. Holders' marks for placing tasks.

4.3. Localization Process

Since the Kinect sensor provides the depth data directly, it is simple to find the position of any point in the view. For labware or holder localization, the related mark is recognized and its center point is found in the image coordinate. The center point can be determined by finding the centroid of the bounding box drawn around the object using the box corners' positions. To find the

position of the center point related to the Kinect, mapping steps are required. The first mapping step for the related point is performed from the image frame to the depth frame. This step is essential because there is a mismatch in the resolution between the depth frame and image frame. The second mapping step for the center point is performed from depth frame to the Kinect space coordinates. The center point position of the mark related to the Kinect is used as reference to locate the position of the manipulating point where the robotic arm has to reach. At the end, the position information has to be transformed to be related to the arm shoulder space. Then, the kinematic model is used to control the arm joints and guide the arm to the target.

5. System Integration

The integration of multiple control software, which are developed using similar or different programing language, into a single one can be complicated. To cope with this issue, an interface can be developed to interact with all coding platforms simultaneously. It is useful to separate the system into multiple platforms to simplify the coding tasks and to facilitate finding the coding bugs. These platforms can communicate with each other in the system using client-server model. The labware recognition and localization part has been integrated and communicates with the multifloor navigation platform [22] and the arm manipulation platform. Using this communication model, these parts can exchange the orders and information between each other. The communication process between the client and the server is described using a sequential diagram shown in Fig. 19. Initially, sockets are created on the server and the client. The client requests a socket connection from the server using a specific port number and IP address. If the requested port is free to use, the server establishes a connection to communicate with the client. Once the connection is established, the client program sends/receives information to/from the server program. Both sockets are closed when the data transfer is successful. The information is exchanged between the client and the server in the form of strings.

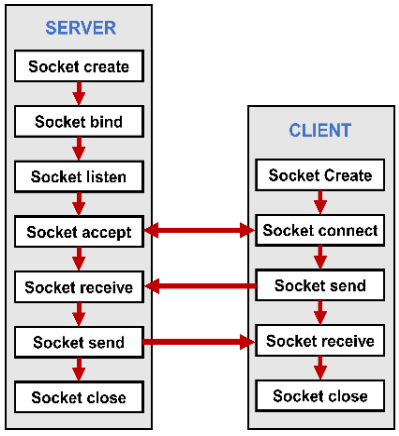


Figure 19. Diagram of client-server model.

In the described system, the multifloor navigation platform sends the work information to the arm manipulation platform whenever the robot reaches the workstation. This information includes the required task (grasping/placing) with the required number of labware or holder. This information is sent to the identification and localization platform (Kinect control). The Kinect platform finds the position of the target and transfers this information to the arm manipulation platform to perform the required task [18]. The process flowchart and the GUI of the identification and localization system are shown in Fig. 20 and Fig. 21 respectively. In the GUI window it can be clearly noticed how the required labware mark has been recognized after cropping the image.The upper part of the GUI represents the server socket which shows the used IP address with the port number. It shows also the received command related to the required target which has been sent from the arm manipulation platform. The bottom of the GUI shows the position information of the target related to the Kinect with the recognition time in milliseconds.

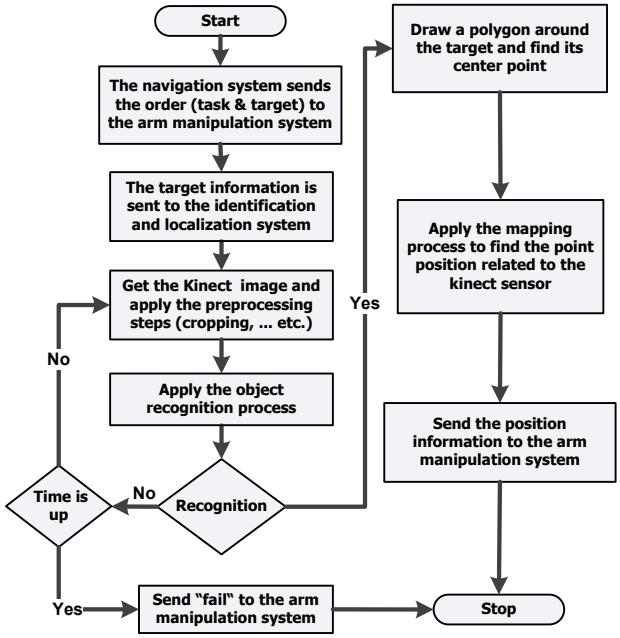


Figure 20. Flowchart of object identification and localization system.

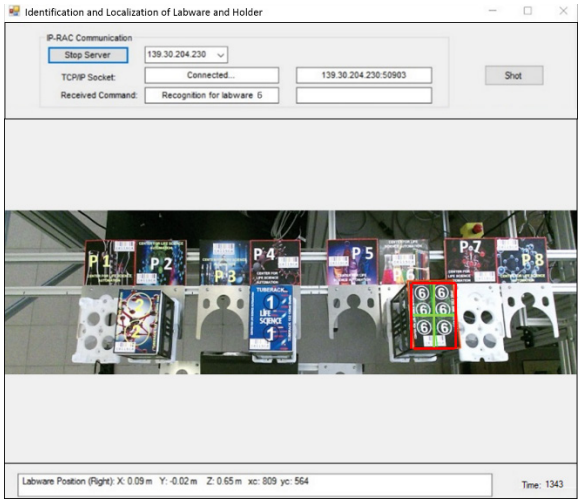


Figure 21. GUI of object recognition and localization.

6. Conclusion

Visual based manipulation plays an important role in mobile robot transportation systems. The ability to identify and localize the required object visually is very essential to guarantee successful tasks. In this paper, an approach for multiple labware and holders' identification and localization for manipulation in life science laboratories has been presented. A suitable design of grippers and labware containers is very important to perform a secure transportation for the required labware. The Kinect sensor

V2 with SURF algorithm is used for this work. The Kinect is fixed on the mobile robot using a holder which has been designed carefully. Specific marks have been designed for labware and holder identification. The marks improve the success rate of the identification process for the labware and holders at any position on the workstation. The image acquired from the Kinect has been processed initially to enhance its features and to decrease the execution time. This step improves the identification success rate using SURF algorithm. HSV color filtering has been used with red marks to decrease the identification time. Polarization and intensity filters are used with the Kinect V2 to reduce the glossy lighting effects in the working environment. Kinect V2 provides high resolution image, wide FOV, and accurate position data directly that makes it very desirable for such tasks. The position of the mark related to the robot guides the arm to perform the required task. The client server model has been used to integrate and connect the identification and localization part with the arm manipulation and Multifloor transportation systems.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This work was funded by the German Federal Ministry of Education and Research (FKZ: 03Z1KN11, 03Z1KI1). The study is supported by the German Academic Exchange Service (Ph.D. stipend M. M. Ali). The authors would also like to thank the Canadian DrRobot Company for the support of the H20 mobile robots, and Mr. Lars Woinar for his contribution in the 3D modelling and printing designs.

References

[1] M. M. Ali, H. Liu, N. Stoll, and K. Thurow, “An Identification and Localization Approach of Different Labware for Mobile Robot Transportation in Life Science laboratories”, in IEEE International Symposium on Computational Intelligence and Informatics (CINTI), Budapest, Hungary, 2016, pp. 353-358.

[2] H. Liu, N. Stoll, S. Junginger, and K. Thurow, “Mobile Robot for Life Science Automation”, International Journal of Advanced Robotic Systems, Vol. 10, pp. 1-14, 2013.

[3] M. M. Ali, H. Liu, R. Stoll, and K. Thurow, “Arm Grasping for Mobile Robot Transportation using Kinect sensor and Kinematic Analysis”, in IEEE International Conference on Instrumentation and Measurement Technology (I2MTC), Pisa, Italy, 2015, pp. 516–521.

[4] A. A. Abdulla, H. Liu, N. Stoll, and K. Thurow, “A Robust Method for Elevator Operation in Semioutdoor Environment for Mobile Robot Transportation System in Life Science Laboratories”, in IEEE International Conference on Intelligent Engineering Systems (INES), Budapest, Hungary, 2016, pp. 45–50.

[5] J. R. Sanchez-Lopez, A. Marin-Hernandez, and E. R. Palacios-Hernandez, “Visual Detection, Tracking and Pose Estimation of a Robotic Arm End Effector”, in the Proceeding of the Robotics Summer Meeting, Veracruz, Mexico, 2011, pp. 41–48.

[6] K. Yamazaki, Y. Watanabe, K. Nagahama, K. Okada, and M. Inaba, “Recognition and Manipulation Integration for a Daily Assistive Robot Working on Kitchen Environments”, in IEEE International Conference on Robotics and Biomimetics, Tianjin, China, 2010, pp. 196–201.

[7] G. Lowe, "Object Recognition from Local Scale-Invariant Features", in IEEE International Conference on Computer Vision, Corfu, Greece, 1999, pp. 1150-1157.

[8] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "SURF: Speeded Up Robust Features", Journal of Computer Vision and Image Understanding (CVIU), Vol. 110, No. 3, pp. 346–359, 2008.

[9] E. Rosten and T. Drummond, “Machine Learning for High-Speed Corner Detection,” in Computer Vision–ECCV 2006, Springer, pp. 430–443.

[10] A. Collet and S. S. Srinivasa, “Efficient Multi-View Object Recognition and Full Pose Estimation”, in IEEE International Conference on Robotics and Automation (ICRA), Anchorage, Alaska, 2010, pp. 2050–2055.

[11] E. Marchand and F. Chaumette, “Feature Tracking for Visual Servoing purposes”, Journal of Robotics and Autonomous Systems, Elsevier, vol. 52, no. 1, pp. 53-70, 2005.

[12] L. T. Anh and J. B. Song, "Object Tracking and Visual Servoing using Features Computed from Local Feature Descriptor", in International Conference on Control Automation and Systems (ICCAS), Gyeonggi, South Korea, 2010, pp. 1044-1048.

[13] S. Zickler and M. M. Veloso, “Detection and Localization of Multiple Objects”, in IEEE-RAS International Conference on Humanoid Robots, Genoa, Italy, 2006, pp. 20–25.

[14] R. Katsuki, J. Ota, Y. Tamura, T. Mizuta, T. Kito, T. Arai, T. Ueyama, and T. Nishiyama, “Handling of Objects with Marks by a Robot”, in IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Las Vegas, Nevada, 2003, vol. 1, pp. 130–135..

[15] H.-Y. Chung, C.-C. Hou, Y.-S. Chen, and C.-L. Chao, “An Intelligent Service Robot for Transporting Object”, in IEEE International Symposium on Industrial Electronics, Taipei, Taiwan, 2013, pp. 1–6.

[16] A. Ramisa, G. Alenya, F. Moreno-Noguer, and C. Torras, “Using Depth and Appearance Features for Informed Robot Grasping of Highly Wrinkled Clothes”, in IEEE International Conference on Robotics and Automation (ICRA), St. Paul, Minnesota, USA, 2012, pp. 1703-1708.

[17] X. Gu, S. Neubert, N. Stoll and K. Thurow, "Intelligent Scheduling Method for Life Science Automation Systems", in IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI), Baden-Baden, Germany, 2016, pp. 156-161.

[18] M. M. Ali, H. Liu, N. Stoll, and K. Thurow, “Multiple Lab Ware Manipulation in Life Science Laboratories using Mobile Robots”, in IEEE International Conference on Mechatronics (MEHATRONIKA), Prague, Czech Republic, 2016, pp. 415-421.

[19] M. M. Ali, H. Liu, N. Stoll, and K. Thurow, “Intelligent Arm Manipulation System in Life Science Labs Using H20 Mobile Robot and Kinect Sensor”, in IEEE International Conference on Intelligent Systems (IS'16), Sofia, Bulgaria, 2016, pp. 382-387.

[20] M. M. Ali, H. Liu, N. Stoll, and K. Thurow, “Kinematic Analysis OF 6-DOF Arms for H20 Mobile Robots and Labware Manipulation for Transportation in Life Science Labs”, Journal of Automation, Mobile Robotics & Intelligent Systems, vol. 10, no. 4, pp. 40–52, 2016.

[21] E. Lachat, H. Macher, M.-A. Mittet, T. Landes, P. Grussenmeyer, “First Experiences with Kinect V2 Sensor for Close Range 3d Modelling”, in International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS), Volume XL-5/W4, 2015, pp. 93-100.

[22] A. A. Abdulla, H. Liu, N. Stoll, and K. Thurow, “A New Robust Method for Mobile Robot Multifloor Navigation in Distributed Life Science Laboratories”, J. Control Sci. Eng., vol. 2016, Jul. 2016.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1227-1237 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Grasping and Placing Operation for Labware Transportation in Life Science Laboratories using Mobile Robots

Mohammed Myasar Ali[*,1], Hui Liu[1], Norbert Stoll[2], Kerstin Thurow[1]

[1]*Center for Life Science Automation (celisca), Rostock University, 18119 Rostock, Germany*

[2]*Institute of Automation, Rostock University, 18119 Rostock, Germany*

ARTICLE INFO

Article history:
Received: 06 June, 2017
Accepted: 07 July, 2017
Online: 21 July, 2017

Keywords:
Robotic arm manipulation
Labware transportation
Kinect sensor V2
Design of grippers and labware containers

ABSTRACT

In automated working environments, mobile robots can be used for different purposes such as material handling, domestic services, and objects transportation. This work presents a grasping and placing operation for multiple labware and tube racks in life science laboratories using the H20 mobile robots. The H20 robot has dual arms where each arm consists of 6 revolute joints with 6-DOF and 2-DOF grippers. The labware, which have to be manipulated and transported, contain chemical and biological components. Therefore, an accurate approach for object recognition and position estimation is required. The recognition and pose estimation of the desired objects are very essential to guide the robotic arm in the manipulation tasks. In this work, the problem statement of H20 transportation system with the proposed methodology are presented. Different strategies (visual and non-visual) of labware manipulation using mobile robots are described. The H20 robot is equipped with a Kinect V2 sensor to identify and estimate the position of the target. The local features recognition based on SURF algorithm (Speeded-Up Robust Features) is used. The recognition process is performed for the required labware and holder to perform the grasping and placing operation. A strategy is proposed to find the required holder and to check its emptiness for the placing tasks. Different styles of grippers and labware containers are used to manipulate different weights of labware and to realize a safe transportation. The parts of mobile robot transportation system are communicated with each other using Asynchronous socket Channels.

1. Introduction

This paper is an extension of work originally presented in the international conference on Mechatronics (Mechatronika 2016) [1]. This work shows grasping and placing operations for labware and tube racks transportation in life science laboratories using mobile robots. Different purposes can be performed using mobile robots such as product transportation [2], teleoperation for the tasks with power tools [3], domestic services [4]-[6], or material handling [7]. The development of laboratories in the field of life science depends significantly on the new techniques of automated solutions in all related areas. Different tasks such as biological testing, samples preparation, and samples analysis are performed in the laboratories using automated equipments. The robots are very necessary in the field of life sciences. All the workstations in the same or in different laboratories can be connected using stationary and mobile robots. This connection ensures a 24/7 operation and reduces tedious and routine work for the employees. Object transportation using mobile robots is one of the most important tasks in the automation field since it increases productivity and saves human resources. For safe transportation, the mobile robots have to perform the object manipulation processes in addition to the path planning, mapping, and localization. H20 robots and Kinect V2 are used in this work to perform a visual manipulation for multiple labware and tube racks. The H20 robot is a wireless networked autonomous humanoid mobile robot. It has a PC tablet, dual arms, and an indoor GPS navigation system. Some technical procedures have been

[*]Corresponding Author: Mohammed Myasar Ali, Center for Life Science Automation (celisca), 18119 Rostock, Germany
Email: mohammed.myasar.ali@celisca.de

implemented at the Center for Life Science Automation (celisca) to develop the transportation system of the H20 mobile robots [8]-[10]. The H20 robot and different labware with tube racks are shown in Fig. 1 and Fig. 2 respectively.

Figure 1. H20 mobile robot.

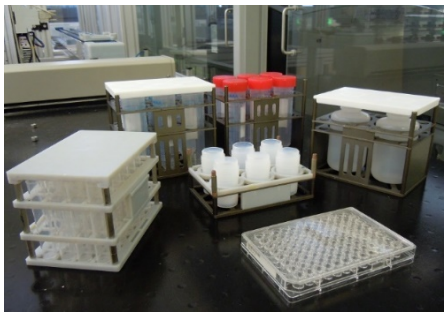


Figure 2. Different labware and tube racks.

For grasping and placing operation of multiple labware, the object identification and localization are very necessary to guide the robotic arm to the target. This requires an arm control based kinematic model to find the joints' values which move the robotic arm to perform the task. The kinematic analysis is the way of describing the motion of arm links without considering the forces causing this motion. The kinematic analysis is classified into two parts: forward kinematics (FK) and inverse kinematics (IK). Using the forward kinematics equations, the end-effector pose relative to the arm base is found according to the given joints' angles. On the other hand, the inverse kinematics equations are used to find the required joints' angles for the given pose of the end effector with respect to the arm base. The joints' limits and workspace of the robotic arm have to be taken into consideration in the manipulation process. In general, the FK solution is easier to be found than the IK solution especially for serial arms. The IK problem can be solved using two approaches: analytic and numeric. The analytical solution is more preferable to be applied in real time applications because it is computationally faster than the numerical approach. Also, all possible solutions can be found using the analytical approach. But, the IK analytical solution can only be derived for manipulators with a particular structure. For example, if 3 consecutive joints' axes are parallel, or intersect at a single point, then the analytical solution can be found. This solution often cannot be generalized to other robotic arms. Ali et al. proposed a reverse decoupling mechanism method to solve the IK problem analytically for a 6-DOF (degree of freedom) robotic arm [11]. O'Flaherty et al. developed the same method to find the closed-form solution of IK problem for a HUBO2+ humanoid robot arms [12]. The reverse decoupling mechanism method has been used to find the analytical solution of the IK problem for the arms of the H20 robot [13].

The visual approach of grasping and placing requires the use of a suitable visual sensor with a suitable object identification algorithm to realize a successful operation. Various features can be extracted from the image to find the target in the view. Color, edges, corners, shape, and local features can be considered as target characteristics for detection task. The detection of multiple characteristics related to the target improves the success rate of the process. Some applications depend on the object color as a feature to be found in the image. Different color systems can be used for this task such as RGB, HSV (hue, saturation, value), HSL (hue, saturation, lightness), etc. HSV color system can be considered as the more reliable in color segmentation in comparison with other systems [14], [15]. The HSV system is more powerful against dynamic lightness and brightness conditions. Additional information has to be used such as the target shape and area in case of presenting different objects with the same color in the view. Sanchez-Lopez et al. proposed a fast method for object tracking using HSV color segmentation and service robot [16]. Yamazaki et al. used HSV color with edges and shape for manipulation of foods and kitchen tools with daily assistive robot [17]. To recognize the required object which has a specific textures, feature matching algorithms have to be used. Generally, the local features recognition requires an off-line step to save the images of the targets in a data base as references. Then, the matching algorithm is performed by extracting local features from the reference image related to the target to be identified with the current live image. Visual matching based on local feature descriptors such as SIFT (scale invariant feature transform), SURF (Speeded-Up robust features), and FAST (Features from Accelerated Segment Test) are very popular algorithms in this field [18]-[20]. These algorithms are somehow independent to changes in orientation, scale, and illumination. Chen et al. presented an identification approach for textured objects using SIFT algorithm [21]. Grundmann et al. recognized household items using object classification based on SIFT algorithm [22]. Anh et al. proposed an object tracking method based on SURF algorithm for safe grasping operations [23]. To estimate the target position, stereo vision and 3D cameras are considered the most appropriate solutions. A 3D camera such as the Kinect sensor is more preferable for this task. Kinect provides directly the depth data without implementing any steps in image processing as in the case of stereo vision. Chung et al. proposed an intelligent service robot equipped with Kinect sensor to help humans in object transportation [2]. Stueckler et al. described a strategy for objects segmentation and manipulation using the depth frame of the Kinect sensor [24]. The design of arm grippers and labware containers is one of the most important issues which plays a main role to achieve a safe grasping and placing operation. Some factors have to be taken into consideration to design a suitable style of grippers like the task quality, the end effector structure, and the characteristics of the target like its weight, shape, etc.

In this work, a grasping and placing operation is presented to achieve labware transportation in life science laboratories. Five H20 mobile robots are used for maneuvering between the adjacent labs and workstations for transporting multiple labware and tube racks which contain chemical and biological components. Local features matching based on SURF algorithm is used to identify

multiple labware. The navigation and arm control systems of the H20 robots are developed using C-sharp programming language. SURF algorithm is the most appropriate method that can be implemented using C-sharp. The inverse kinematics solutions for the head and arms of H20 robot are listed in this work. This paper is organized as follows: in section 2, the problem statement and proposed methodology are discussed. The head and arm structures of the H20 robot with IK solutions are given in section 3. Different designs of grippers and labware containers are described in section 4. Visual (using Kinect) and blind (using sonar sensor) manipulation for the labware are presented in sections 5 and 6 respectively. Section 7 shows the target localization followed by the labware transportation process. Finally, the paper concludes the current results.

2. Problem Statement and Proposed Methodology

The future of life science laboratories depends significantly on the innovations of automated solutions in the entire scope. The stationary and mobile robots play an important role in the automated environment. In the Center for Life Science Automation, different automation islands are connected with each other by the cooperation of stationary and mobile robots. For transportation task, mobile robots usually follow a predefined path from the starting station to the ending station using a guidance control system. The H20 mobile robots use the Stargazer sensor with ceiling landmarks (see Fig. 3) for maneuvering between the adjacent labs (Hagisonic Company, Korea) [25]. This guidance system is not accurate enough and causes positioning errors in the range of ±3cm and ±2cm in the X and Y axes respectively in front of the workstation. The glossy lighting and sunlight blind the Stargazer to recognize the ceiling landmarks which leads to positioning errors. Also, the odometry system, that updates the robot pose, accumulates errors for different reasons such as different wheel diameters, wheel-slippage, and wheel misalignment. Therefore, the mobile robot transportation requires a reliable grasping and placing operation for the labware which contain chemical and biological components. This operation depends on 3 main factors as shown in Fig. 4. The identification and localization process for the desired labware has to be performed. The other essential issue is to calculate the arm joints to guide the end effector to the desired pose by solving the inverse Kinematics problem. Also, the design of arm gripper and labware containers has to be taken into consideration to realize a secure manipulation.

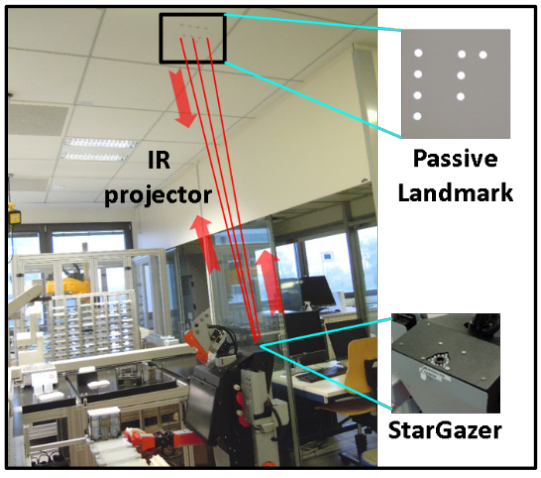


Figure 3. Stargazer sensor with ceiling landmarks.

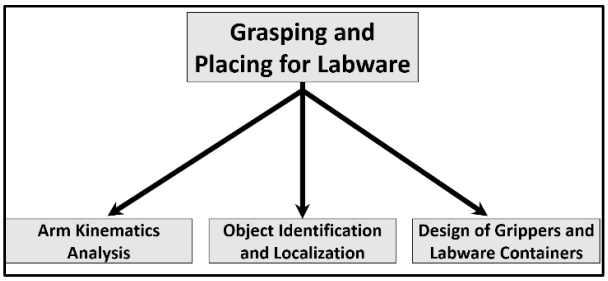


Figure 4. Requirements of grasping and placing operation.

The transportation of multiple labware between different workstations and labs requires an appropriate management systems. This is important to perform the automation of individual areas in such a way that comprehensive life science processes are realized. The hierarchical workflow management system (HWMS) controls the workflow between the stationary and mobile robots [26]. Related to the mobile robots, the management system sends the plan to the transportation system as shown in Fig. 5. The plan information includes the starting position, end position, and the required labware to be transported. The transportation system of the mobile robots includes 3 main parts: the robot remote center, the navigation system, and the grasping/placing system. The grasping/placing system is splitted into two parts, object identification and localization and the arm control. The object identification and localization software with the visual sensor is utilized to recognize the target and to estimate its pose. The pose information is sent to the arm kinematic control to guide the robotic arm.

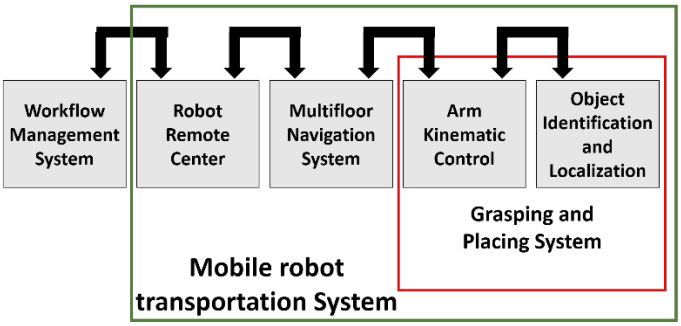


Figure 5. Overall structure of the mobile robot transportation.

3. Kinematic Analysis of H20 Robot

The H20 mobile robot has dual 6-DOF arms with 2-DOF grippers. It has also a head with 2 revolute joints (2-DOF). Fig. 6 shows the joints structure of the head and arms. According to Fig. 6, the values of d3 and d5 are 0.236m and 0.232m respectively. Also, there is a distance of de=0.069m between the wrist joint and the end-effector (E). The Denavit-Hartenberg (D-H) representation is used to describe the translation and rotation relationship between the arm links. According to this D-H representation, there are four parameters to analyze the manipulator: the link length (ai-1), the link twist (αi-1), the link offset (di), and the joint angle (θi) where (i) refers to the joint number [27]. By following the D-H rules, the homogeneous transformations between adjacent links are defined. The D-H parameters and the rotational limit for each joint of the H20 arms are described in Table 1.

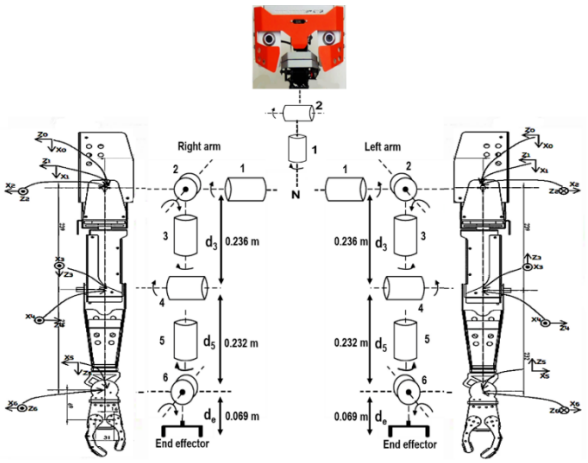


Figure 6. Arms and head structures of H20 robot.

Table 1: D-H parameters and joints limit.

Left and Right Arms						
θ_i	$\alpha_{(i-1)}$ (L)	$\alpha_{(i-1)}$ (R)	$a_{(i-1)}$ (LR)	d_i (m) (L)	d_i (m) (R)	Joints limit (LR)
θ_1	0°	0°	0	0	0	-20°~192°
θ_2	90°	-90°	0	0	0	-200°~-85°
θ_3	90°	-90°	0	-0.236	0.236	-195°~15°
θ_4	-90°	90°	0	0	0	-129°~0°
θ_5	90°	-90°	0	-0.232	0.232	0°~180°
θ_6	-90°	90°	0	0	0	-60°~85°

The solution of the IK problem for the H20 arms has been found using the reverse decoupling mechanism method [11]. The inverse matrix of the FK solution has been found which is considered the first step to solve the IK problem. Eq. (1) shows the inverse matrix of the FK model where (in_x , in_y , in_z) is the orthogonal vector, (io_x, io_y, io_z) is the orientation vector, (ia_x , ia_y , ia_z) is the approach vector, and (ip_x , ip_y , ip_z) is the position vector.

$$ {}^{E}_{0}T = {}^{E}_{6}T \cdot {}^{6}_{5}T \cdot {}^{5}_{4}T \cdot {}^{4}_{3}T \cdot {}^{3}_{2}T \cdot {}^{2}_{1}T \cdot {}^{1}_{0}T = \begin{bmatrix} in_x & io_x & ia_x & ip_x \\ in_y & io_y & ia_y & ip_y \\ in_z & io_z & ia_z & ip_z \\ 0 & 0 & 0 & 1 \end{bmatrix}. \quad (1) $$

The equations below are the solution of the IK problem where (*C*) and (*S*) are the abbreviation of angle cosine and sine respectively, (R) and (L) are related to the right and left arms respectively, and (atan2) is the two argument arc tangent function. The equation of joint 4 is the first step that has been found. The other joints' equations are followed sequentially as follows [13] [28]:

$$ \theta_{4(LR)} = \text{atan2}\left(\pm real\left(\sqrt{1 - C_4^2} \right), C_4 \right), \quad (2) $$

$$ C_{4(LR)} = \frac{(ip_y + d_e)^2 + (ip_x)^2 + ip_z{}^2 - d_3{}^2 - d_5{}^2}{2 d_3 d_5}. $$

In case that the target is outside the arm workspace, complex numbers are generated in the solution. Therefore, a (*real*) function is used to keep the real part and ignore the imaginary part for realizing the closest solution to the target position [12].

$$ \theta_{5(LR)} = \text{atan2}\left(S_5, \pm real\left(\sqrt{1 - S_5^2} \right) \right), \quad (3) $$

$$ S_{5(L)} = \frac{ip_z}{S_4 d_3}, \; S_{5(R)} = \frac{-ip_z}{S_4 d_3}. $$

$$ \theta_{6(LR)} = \text{wrapToPi}(\text{atan2}(d_3 C_4 + d_5 \; , -d_3 S_4 C_5 \;) - \gamma), \quad (4) $$

where γ = (atan2 (ip_y+d_e,ip_x)), and (wrapToPi) function is used to wrap the angle to the interval between –π and π [12].

$$ \theta_{2(LR)} = \text{atan2}\left(\pm real\left(\sqrt{1 - C_2^2} \right), C_2 \right), \quad (5) $$

$$ C_{2(L)} = ia_y (C_4 C_6 - C_5 S_4 S_6) + ia_x (C_4 S_6 + C_5 C_6 S_4) - ia_z S_4 S_5, $$

$$ C_{2(R)} = -ia_y (C_4 C_6 - C_5 S_4 S_6) - ia_x (C_4 S_6 + C_5 C_6 S_4) - ia_z S_4 S_5. $$

$$ \theta_{3(LR)} = \text{atan2}\left(S_3, C_3 \right), \quad (6) $$

$$ S_{3(L)} = -ia_y S_5 S_6 + ia_x C_6 S_5 + ia_z C_5, $$

$$ S_{3(R)} = ia_y S_5 S_6 - ia_x C_6 S_5 + ia_z C_5, $$

$$ C_{3(L)} = ia_y (C_6 S_4 + C_4 C_5 S_6) + ia_x (S_4 S_6 - C_4 C_5 C_6) + ia_z C_4 S_5, $$

$$ C_{3(R)} = -ia_y (C_6 S_4 + C_4 C_5 S_6) - ia_x (S_4 S_6 - C_4 C_5 C_6) + ia_z C_4 S_5. $$

$$ \theta_{1(LR)} = \text{atan2}\left(S_1, C_1 \right), \quad (7) $$

$$ C_{1(L)} = in_y (C_4 C_6 - C_5 S_4 S_6) + in_x (C_4 S_6 + C_5 C_6 S_4) - in_z S_4 S_5, $$

$$ C_{1(R)} = in_y (C_4 C_6 - C_5 S_4 S_6) + in_x (C_4 S_6 + C_5 C_6 S_4) + in_z S_4 S_5, $$

$$ S_{1(L)} = io_y (C_4 C_6 - C_5 S_4 S_6) + io_x (C_4 S_6 + C_5 C_6 S_4) - io_z S_4 S_5, $$

$$ S_{1(R)} = io_y (C_4 C_6 - C_5 S_4 S_6) + io_x (C_4 S_6 + C_5 C_6 S_4) + io_z S_4 S_5. $$

The IK solutions of singularity cases have been found too. Three cases of singularity have been determined within the joints limits. The singularity cases occur when one or more degrees of freedom are eliminated because some joint axes are aligned with each other. This leads to infinite number of solutions. The singularity cases of H20 arms are as follows:

A) When θ4 = 0 and θ2 ≠ -π: The axis of the third joint is aligned with the fifth joint axis

B) When θ2 = -π and θ4 ≠ 0: The axis of the first joint is aligned with the third joint axis

C) When θ4 = 0 and θ2 = -π: The joints 1, 3, and 5 are collinear

According to Eq. 2, 3, and 5, there are 8 solutions for the required pose inside the reachable workspace. To choose the suitable solution, a selecting algorithm has to be used [13]. The forward and inverse kinematics solutions have been validated and simulated using MATLAB with robotics toolbox. For the integration of the kinematic model with H20 system, the required angle value of each joint has to be converted to its related servo motors position. This step is very crucial since the H20 arms' joints are weak and they can be easily bent by the effects of gravity and payload.

Related to the H20 head, a stereo vision system is fixed on a pan-tilt module. This module consists of two revolute joints which are used for moving the head to track the desired object. The transformation matrix from the neck to the head is as follows:

$$ {}^{N}_{H}T = \begin{bmatrix} C_1C_2 & -S_1 & S_2C_1 & l_{H2}S_2C_1 \\ C_2S_1 & C_1 & S_1S_2 & l_{H2}S_1S_2 \\ -S_2 & 0 & C_1 & l_{H1}+l_{H2}C_1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (8) $$

Eq. 8 represents the forward kinematic solution of the head for the given joint angles. Related to the inverse kinematic problem, the given ${}^{N}_{H}T$ is in the form:

$$ {}^{N}_{H}T = \begin{bmatrix} n & o & a & p \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (9) $$

Thus, the solution of IK for the 2 joints of head is as follows:

$$ \theta_1 \text{ (H)} = \text{atan2}\left(o_y, -o_x\right) \quad (10) $$

$$ \theta_2 \text{ (H)} = \text{atan2}\left(n_x, a_x\right) \quad (11) $$

4. Design of Grippers and Labware Container

For labware grasping and placing with the mobile robots, different grippers with labware containers have been designed. Since the labware contain chemical and / or biological components, any kind of spilling has to be avoided. Therefore, a specific design of grippers and labware containers is required to guarantee secure grasping and placing operations. Fig. 7 shows the initial designs of the grippers and labware containers which are attached with handles.

(*A*) (*B*) (*C*)

Figure 7. Initial designs of grippers and labware containers [13].

Related to Fig. 7.A, the handle was designed to be detected using RGB color detection [29]. Each handle has a specific single color to be identified from the others. The Kinect sensor V1 was used to detect and localize the required handle to manipulate the required labware. In Fig. 4.B, the gripper and handle designs were improved. Flat panel on the upper side of the handle was resolved for fixing different colored or pictorial marks to distinguish multiple handles. Proper gripper has been designed also to fit this handle. SURF algorithm and HSV color segmentation with area and shape detection were used for identification tasks [28]. The required handle was localized using Kinect sensor V2. 350g is the maximum payload that can be manipulated using the designs in Fig. 7.A, and 7.B. The weak wrist joint of the H20 arms as well as the labware container design play a role to limit the manipulated weight. The manipulation of the handle attached with the container requires high torque due to the long lever arm of the wrist joint. The lever arm here represents the distance between the center of the labware weight and the wrist joint. A vertical handle has been designed to manipulate heavier payload as shown in Fig. 7.C. In this case, the wrist joint is rotated to have a vertical axis on the ground. Thus, the lifting process depends on the elbow joint which has more powerful torque [13]. 500g is the possible payload that can be handled using this design. This manner is still not sufficient due to the tube racks which are heavier than 500g. To cope with this issue, a new manipulation style has been created (see Fig. 8). A new gripper has been designed and the handle has been removed from the container. This brings the end-effector closer to the labware weight center. Thus, less torque is required to handle heavier objects since the lever arm of the wrist joint has been decreased. A rubber has been attached to the gripper to increase the friction for secure manipulation. The maximum payload which can be handled with this style is about 700g. Fig. 8 shows the designs of the gripper, labware container, and its holder.

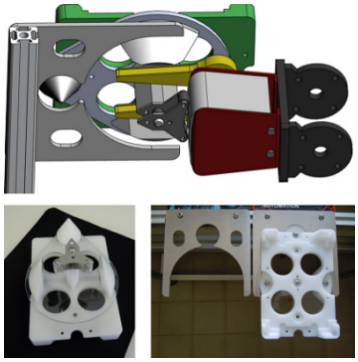


Figure 8. Final design of gripper and labware container.

5. Blind Manipulation using Sonar Sensor

A manipulation strategy has been performed using the developed kinematic model with two sonar sensors [13]. The front base of H20 robot has 3 built-in DUR5200 sonar sensors. One sensor is in the middle and the other two are on the left and right sides. These sensors can detect the range information from 4 cm to 340 cm which can be used for collision avoidance and distance detection. For the grasping and placing operation, the mobile robot has to be straight in front of the workstation. The labware container has a specific posture on the workstation. The pitch and roll orientation of the container related to the robot are fixed. The yaw orientation has to be secured by correcting the robot orientation. This can be done using the two sonar sensors on the base sides (see Fig. 9). The distance (Z) from each channel to the plastic board is checked and the robot rotates till the values of both sensors are equalized. Thus, the distance (Z) between the labware on the workstation and the robot arm base is known. Also, the height (Y) of the workstation related to the robot is fixed and known. Related to the (X) information of the labware position, this can be decided by the workflow management system [27]. Then, the X positioning feedback has to be mapped to be related to the arm base. The X-axis error of the robot position in front of the workstation is about ±2cm. This can be compensated by the design tolerance of gripper and labware container [13].

This strategy is not reliable enough due to some reasons. The positions of holders and boards, which reflect the sonar signals, have to be identical for all workstations. It is not possible in some stations to install this board due to the environment structure. Existence of obstacles leads to wrong estimation from the sonar sensors for the distance and orientation of robot in front of the

workstation. Also, there is no accurate value related to the positon of the required labware in the X-axis due to the lack of positioning feedback in this axis. To cope with all these issues, a visual sensor can be used to identify and localize the target wherever it is located on the workstation.

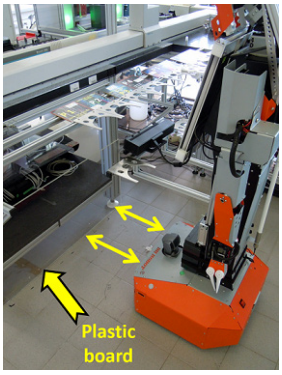


Figure 9. 2 sonar sensors for labware manipulation.

6. Vision based manipulation using Kinect Sensor

It is very crucial to have an intelligent concept to perform a visual grasping for the required labware which then will be transported to the required place. Stereo vision is one of the common methods to detect the target and estimate its position. But, it is not feasible to use the head stereo cameras of the H20 robot for this task because of the head position related to the workstation. This leads to unclear views to the wide workstation which consists of multiple positions for labware containers. Therefore, the Kinect sensor can be considered as an appropriate choice for the visual approach of labware manipulation. It provides a high quality color and depth information which is directly obtained without applying complicated steps on the image as required in the stereo vision. The Kinect V2 has been fixed on the H20 body using a holder with a suitable height and tilt angle to guarantee a clear and wide view for the whole workstation (see Fig. 1). The Microsoft Kinect sensor V2 uses "time of flight" technology to provide the depth information. It relies upon a novel image sensor that indirectly measures the time for pulses of laser light to travel from a laser projector to a target surface, and back to an image sensor. By using the "time of flight" technology, the Kinect V2 can see just as well in a completely dark room as in a well-lit room. The Kinect sensor with its features are shown in Table 2 and Fig. 10.

Table 2: Features of Kinect Sensor V2.

Features	Kinect V2
RGB camera resolution	1920 X 1080
Depth camera resolution	512 X 424
Maximum depth	~ 4.5 M
Minimum depth	50 cm
Horizontal FOV	70 degrees
Vertical FOV	60 degrees
Tilt motor	No
USB standard	3.0
Supported OS	Win 8, Win 10

Figure 10. Kinect sensor V2.

The success of visual manipulation and transportation depends significantly on the success of identification and localization for the required target. The workstation consists of 8 positions of labware holders. According to the arms workspace, the robot can deal with 4 holders (2 for each arm) for grasping and placing operations. Therefore, the robot needs 2 positions in front of the workstation to manipulate all the 8 holders. It is complicated to differentiate the holders according to their appearances. Also, it is not possible to identify the required holder for the grasping task because it will be covered with a labware with its container. To cope with this issue, an initial strategy of grasping and placing operation has been performed. This strategy includes the utilization of 2 marks as a reference for each robot position in front of the workbench. Each mark is centrally located in front of the related robot position to be used as a visual feedback for 4 holders as shown in Fig. 11. Whenever the robot arrives at the required position, the related mark is recognized and localized using Kinect V2 with SURF algorithm. Then, the mark position is modified to find the positions of the 4 holders related to the robot. The workflow management system sends the plan information to the mobile robot transportation system. This information includes the required task (grasp/place) and the holder number. According to the holder number (1-8), the robot moves to the required position (1 or 2) for performing the task. In this case, each mark is considered as a reference for 4 related holder where the holders 1, 2, 5, and 6 belong to a particular robot position. Whereas, the holders 3, 4, 7, and 8 are related to the other robot position. For this strategy, all the holders' positions related to the marks have to be fixed and identical in the other workstations in life science laboratories. To cope with this limitation, a specific mark has been attached to each holder for visual grasping and placing operation as shown in Fig. 12. The required mark is identified and the position of its center point is derived to guide the robot in the grasping and placing operation using the kinematic model.

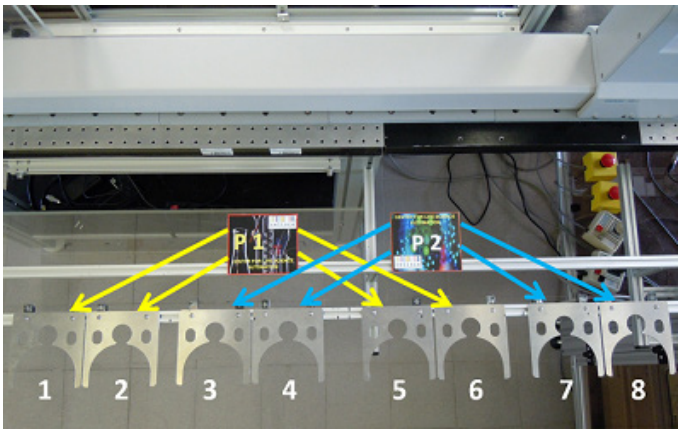


Figure 11. Specific mark for each robot position.

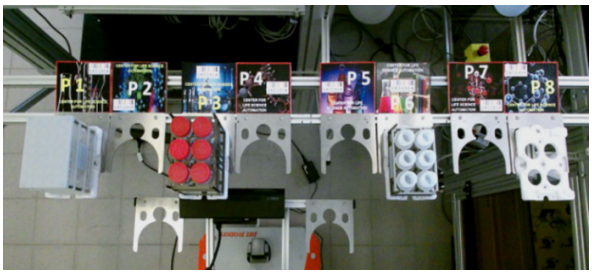


Figure 12. Workbench holders with their related marks.

For grasping the required labware, probable wrong information related to the holder number is sent to the robot transportation system. Also, there is a possibility that the required labware has been moved to another holder by an employee before the mobile robot arrives the workstation. In this case, it is very crucial to identify and localize the required labware wherever it is located on the workstation. Since the labware and tube racks are covered with transparent or white lids to protect the components from cross contamination, it is not applicable to distinguish them on the workbench. Therefore, a specific mark has been fixed on each labware lid for identification process. Fig. 13 shows two tube racks where one of them is covered with a white lid and the other with a mark fixed on its lid. Different marks have been used to identify multiple labwares. The mark contains the labware information with a particular number for classification purposes. The labware information with a background picture in the mark gives adequate features to differentiate multiple labware.

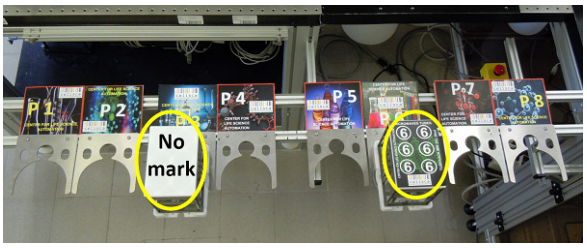


Figure 13. Labware lid with and without mark.

Several methods have been applied to test the performance of object identification and position estimation under different lighting conditions [30]. The tests have been performed for labware identification at the 8 different positions of the workstation. Various procedures have been applied to the Kinect image such as cropping, contrast with brightness correction, and histogram equalization as preprocessing steps before applying SURF algorithm. Histogram equalization process increases the global contrast of the image while the cropping process is used to extract the region of interest (ROI) from the image. The required ROI in this work is the workstation with their labware, holders, and related marks. The cropping step simplifies the recognition process and decreases its required time. The identification of the required labware is assigned by drawing a polygon around its lid mark with cross to specify the center point as shown in Fig. 14.

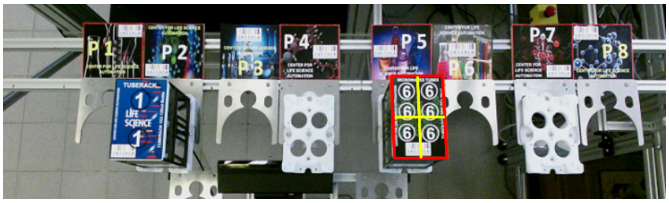


Figure 14. Recognition of the required labware.

For the grasping task, the labware orientation related to the robot can be found using coplanar POSIT algorithm which stands for POSe ITerations. It requires image coordinates of the desired object points (minimum 4 points). It also requires the real model coordinates of these points; the target has to be previously known. Furthermore, the effective focal length of the used camera has to be known [31]. The orientation angle is very important to correct the arm motion in case that the robot is not straight in front of the workstation. This has been done by finding the corners' positions of the lid mark in the image coordinates. Also, the real physical positions of the corners related to the mark center point have to be calculated. This information is inserted to the POSIT algorithm to calculate the difference in orientation between the labware and the Kinect camera or the robot. The orientation angle represents the yaw angle of the robotic arm. Fig. 15 shows the corners of the lid mark which are used to calculate the orientation angle.

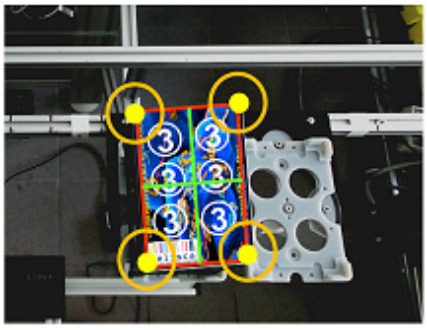
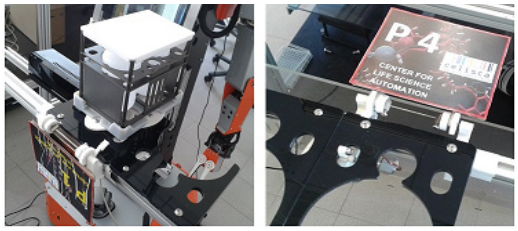
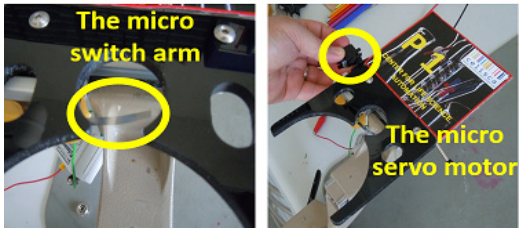


Figure 15. Corners detection of the lid mark.

The holders' marks are used for identification and position estimation purposes to perform the placing task. But, it is very necessary for the robot to check whether the holder is already occupied by a labware or not before performing the placing task. Therefore, a checking method has been implemented by fixing the holder mark on a mechanical part. A rotational joint is used to show and hide the mark from the Kinect view to determine whether the holder is empty or not. In other words, if the Kinect cannot identify the mark of the required holder, this holder is already occupied as shown in Fig. 16.A. A second method can be applied using a micro switch, micro servo motor, and servo controller to rotate the holder mark as shown in Fig. 16.B.



(*A*)



(*B*)

Figure 16. Checking the availability of the required holder.

7. Target Localization and Extrinsic Calibration

The identification step enables the identification of the required labware or holder according to the related mark. Then, the position of the center point of the mark in the image coordinate is derived. This can be done using the corners' position of the polygon drawn around the detected mark. In order to obtain the real or physical position of this center point related to the Kinect, mapping processes have to be performed. Since the RGB frame of Kinect is not identical with the depth frame, the center point has to be mapped from the image frame to its related point in the depth frame. Then, another mapping step is performed for the point in depth frame to be related to the Kinect space coordinates. The result of these mapping steps is the real position of the mark center point related to the Kinect sensor. The position information of X-axis determines which arm (left or right) can be used for the required operation. The position information of the center point has to be calibrated to calculate the position of the manipulation point in the labware container. The manipulation point represents the position where the end effector of the robotic arm has to reach during the grasping and placing operation. Fig. 17 shows the effective point of the labware container on the workstation. For the grasping task, there are different kinds of labwares which have different height. Therefore, it is not feasible to depend on the labware height to find the manipulation point position. To cope with this issue, the workstation height, which is fixed and identical for all workstations, can be used as a reference to find the manipulation point height.

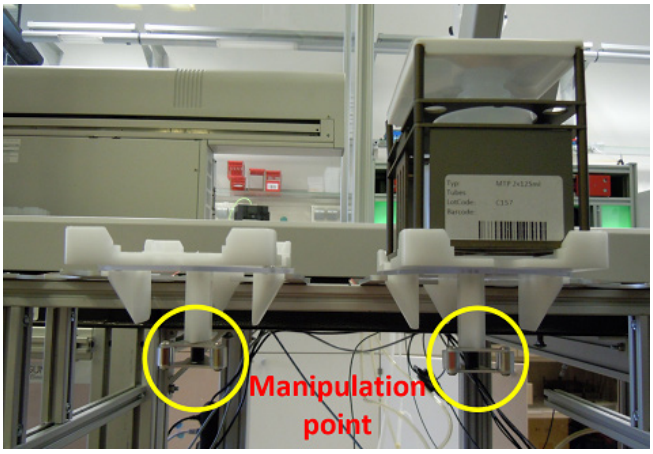


Figure 17. Manipulation point of the labware container.

The position of the manipulation point related to Kinect has to be calibrated to guide the robotic arm to the required point. This can be done by applying the extrinsic calibration step. The purpose of this step is to transform the position information from the Kinect space to be related to the arm shoulder space. Then, the inverse kinematic model is used to find the joints' values of the arm for guiding its end effector to the target. The calibration from Kinect space to arm shoulder space includes the transformation in translation and in rotation. This belongs to the difference in the position and the tilt angle (t) between the Kinect coordinates and shoulder coordinates according to the Kinect holder as shown in Fig. 18. The Kinect holder has been designed very carefully to be suitable for the grasping and placing operation. The tilt angle of the Kinect plays a role in reducing the light effects for the identification process. The holder has to be fixed very firmly to guarantee that there are no changes in the Kinect position and orientation during the robot movements.

Figure 18. Kinect holder.

8. Labware Transportation Process

The transportation process of multiple labware using the mobile robots is based on the following scheme. In the 1st step, the navigation system receives the transportation plan from the workflow management system through the robot remote center. In the 2nd step, the robot moves to the required laboratory and workstation. Then, the robot checks around the workbench to identify the lid mark related to the desired labware which has to be transported to the destination lab. In the 3rd step, the position of the target manipulation point related to the arm shoulder is found. In the 4th step, the robot uses the arm kinematic model to calculate the required joints' values depending on the position information. In the 5th step, the arm moves to the required target and grasps the labware container. After grasping operation, the robot moves to the desired destination to place the labware on the required holder. The mobile robot transportation system has been realized using three main coding platforms which are the multifloor navigation system (MFNS) [25], the arm control system (ACS), and the object identification and localization system (OILS). These 3 systems are connected with each other through a asynchronous socket. The 3 coding platforms exchange the orders and informations between each other using 2 client-server communication models as shown in Fig. 19.

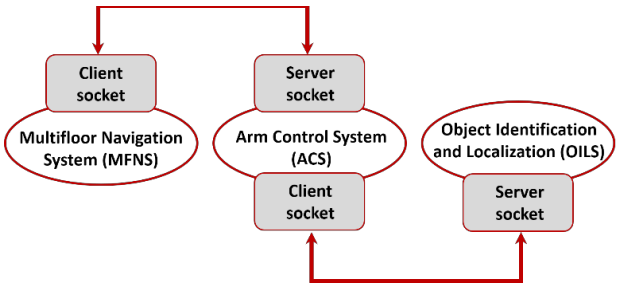


Figure 19. Client-server models for labware transportation.

One of the important information, which has to be sent to the transportation system, is the approximate weight of the manipulated labware. Since the arms of H20 robot are unstable and have weak wrist joint, the weight information helps to calibrate the wrist joint. This step is very crucial to keep the labware straight during the grasping/placing operation to avoid the spilling of its contents. Also, two labware holders have been fixed on the H20 body for the left and right arms as shown in Fig. 20. These holders support to keep the labware in a safe horizontal posture during the transportation tasks. To supply the Kinect sensor with the required

electrical power during the robot movements, a 12V battery with a voltage and current stabilizer has been installed on the robot base. If the grasping operation is performed with a particular arm (right or left), the placing operation for the grasped labware has to be achieved using the same arm. It is not possible to perform the placing operation with the other arm since the labware is positioned on the H20 holder according to the used arm. Therefore, it is very essential to perform the placing task at a position where the required arm can reach. In case that the required labware is not existing in the Kinect view or it is outside the arms workspace, a feedback information is sent to the navigation control. This information includes several decisions such as changing the robot position or skipping to the next task. Fig. 21 shows the overall flowchart of labware manipulation process (grasping/placing). Fig. 22 shows the arm movement steps for grasping the required labware and how the arm places it on the H20 holder.

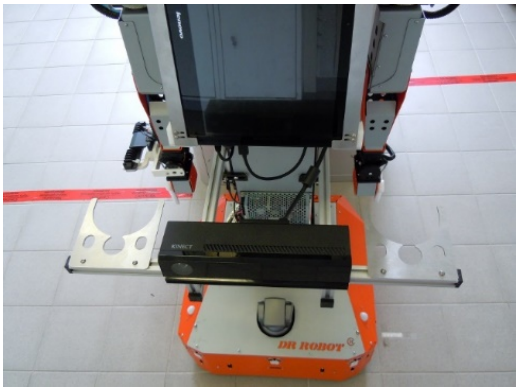


Figure 20. H20 holder frame for labware transportation.

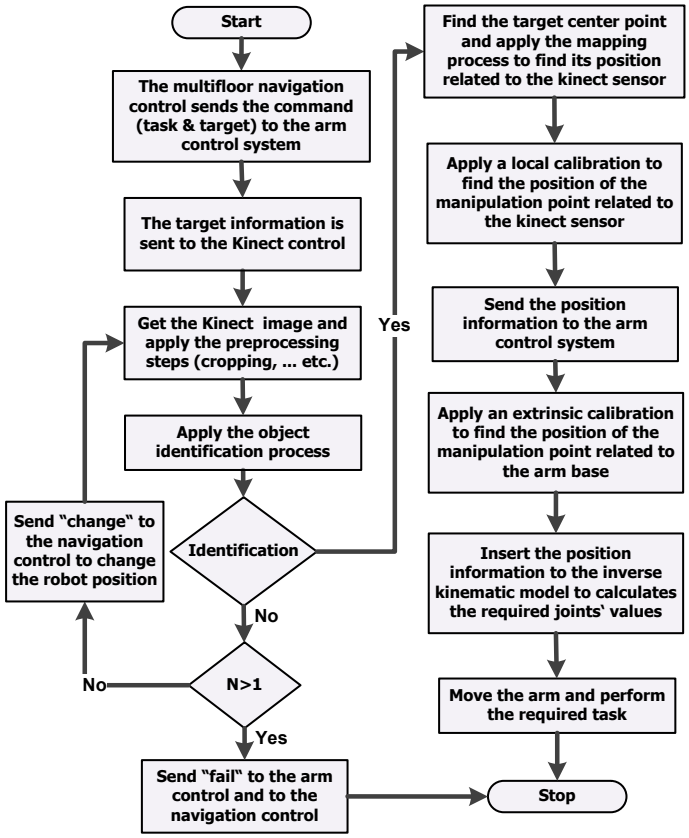


Figure 21. Flowchart of labware manipulation process.

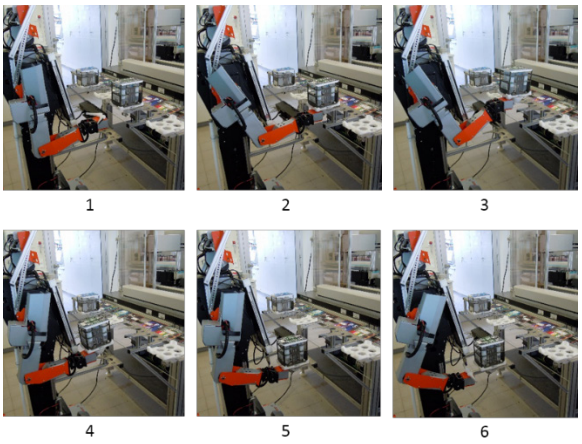


Figure 22. Arm movement steps for grasping operation.

In the grasping operation, the robotic arm moves from the rest configuration to the manipulation point on the labware container. At the rest configuration, the position of the end effector related to the arm shoulder is (X= 0.56m, Y= 0 m, Z= 0m). This position information is according to the shoulder coordinates which is shown in Fig. 6 where X represents the arm length in rest configuration. Fig. 23 shows a chart for the changes in the end effector position during the arm movement from the rest configuration to the manipulation point. In this example, the position of the manipulation point related to the arm shoulder is (X= 0.18m, Y= 0.46m, Z= 0.03m). The approximate time required to reach the manipulation point is about 30 seconds. Fig. 24 shows the path of the end effector in XY plane for the same example. Furthermore, Table 3 shows the changes of arm joints' values in degrees from the rest configuration to the manipulation point.

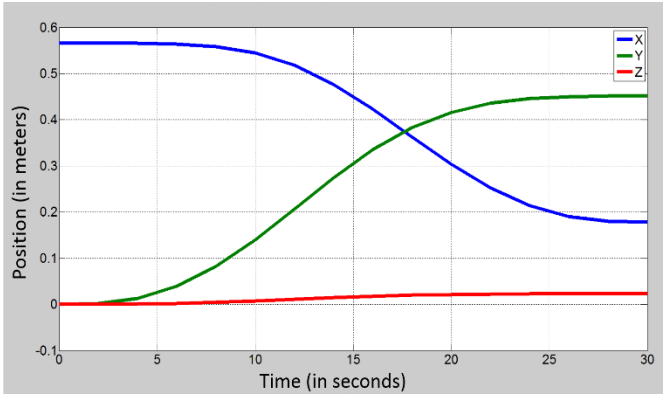


Figure 23. End effector position versus time.

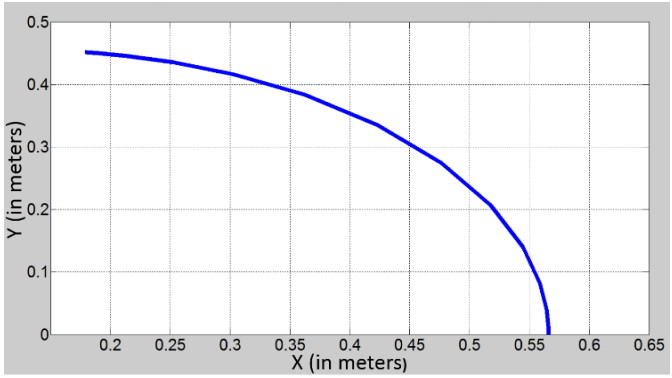


Figure 24. End effector path in XY plane.

Table 3: Joints values in degrees for grasping task.

Configuration	J1	J2	J3	J4	J5	J6
Rest	0°	-90°	-90°	0°	90°	0°
Manipulation point	32°	-95°	-92°	-64°	5°	6°

The required time for performing the grasping operation is about 69 seconds while 59 seconds are required for the placing operation. The work has been developed using Microsoft Visual Studio 2015 with C# programming language. The project is running on a Windows 10 platform in the H20 tablet. Table 4 shows the overall success rate of the grasping and placing operation for 50 attempts of transportation. Fig. 25 shows the GUI of the arm control system.

Table 4: Labware manipulation tests.

Attempts	Successful grasp	Successful place
50	92%	90%

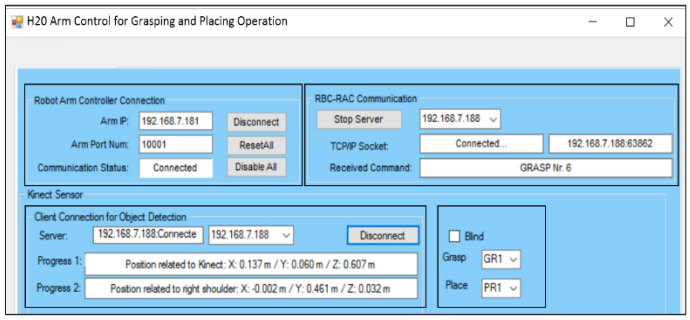


Figure 25. GUI of arm control system.

9. Conclusion

The ability to identify and calculate the position of the target in real-time plays an important role in the manipulation process of mobile robots. In this paper, a grasping and placing strategy has been presented to perform multiple labware transportation in life science laboratories. The problem statement with the proposed methodology for the mobile robot transportation system has been discussed. Different labware manipulation strategies have been described which depends on the sonar sensors and Kinect. The identification process of labware and holder, which is based on SURF method, has been integrated to the transportation system. The pose estimation of the labware related to the robot is very significant to guide the robotic arm in the grasping and placing operation. The Kinect V2 can be considered one of the powerful 3D cameras which provide the position information in a fast way. A strategy has been proposed to check whether the required holder is occupied or not for the placing operation. The design of the arm gripper and labware container have been improved to handle heavier payloads and to guarantee a secure manipulation for the labware.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The study is supported by the Federal Ministry of Education and Research (FKZ: 03Z1KN11, 03Z1KI1) and the German Academic Exchange Service (Ph.D. stipend M. M. Ali). The authors would also like to thank the Canadian DrRobot Company for the support of the H20 mobile robots, Mr. Ali A. Abdullah for controlling the multifloor navigation system with the technical collaboration, and Mr. Lars Woinar for his contribution in the 3D modelling and printing designs.

References

[1] M. M. Ali, H. Liu, N. Stoll, and K. Thurow, "Multiple Lab Ware Manipulation in Life Science Laboratories using Mobile Robots", in IEEE International Conference on Mechatronics (MECHATRONIKA), Prague, Czech Republic, 2016, pp. 415-421.

[2] H. Chung, C. Hou, Y. Chen, and C. Chao, "An Intelligent Service Robot for Transporting Object", in IEEE International Symposium on Industrial Electronics (ISIE), Taipei, Taiwan, 2013, pp. 1–6.

[3] R. O'Flaherty, P. Vieira, M. X. Grey, P. Oh, A. Bobick, M. Egerstedt, and M. Stilman, "Humanoid Robot Teleoperation for Tasks with Power Tools", in IEEE International Conference on Technologies for Practical Robot Applications, Woburn, MA, 2013, pp. 1–6.

[4] M. Ciocarlie, K. Hsiao, E. G. Jones, S. Chitta, R. B. Rusu, and I. A. Şucan, "Towards Reliable Grasping and Manipulation in Household Environments", in 12th International Symposium on Experimental Robotics (ISER), Springer Berlin Heidelberg, 2014, pp. 241–252.

[5] B. Graf, U. Reiser, M. Hägele, K. Mauz, and P. Klein, "Robotic Home assistant Care-O-bot® 3-Product Vision and Innovation Platform", in IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO), Tokyo, Japan, 2009, pp. 139–144.

[6] N. Vahrenkamp, D. Berenson, T. Asfour, J. Kuffner, and R. Dillmann, "Humanoid Motion Planning for Dual-arm Manipulation and Re-Grasping Tasks", in IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), St. Louis, USA, 2009, pp. 2464–2470.

[7] T. J. Tsay, M. S. Hsu, and R. X. Lin, "Development of a Mobile Robot for Visually Guided Handling of Material", in IEEE International Conference on Robotics and Automation (ICRA), Taipei, Taiwan, 2003, pp. 3397–3402.

[8] H. Liu, N. Stoll, S. Junginger, and K. Thurow, "A Common Wireless Remote Control System for Mobile Robots in Laboratory", in IEEE Instrumentation and Measurement Technology Conference (I2MTC), Graz, Austria, 2012, pp. 688–693.

[9] H. Liu, N. Stoll, S. Junginger, and K. Thurow, "Mobile Robot for Life Science Automation," International Journal of Advanced Robotic Systems, Vol. 10, pp. 1-14, 2013.

[10] A. A. Abdulla, H. Liu, N. Stoll, and K. Thurow, "A Backbone-Floyd Hybrid Path Planning Method for Mobile Robot Transportation in Multi-Floor Life Science Laboratories", in IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI), Baden-Baden, Germany, 2016, pp. 406-411.

[11] M. A. Ali, H. A. Park, and C. G. Lee, "Closed-form Inverse Kinematic Joint Solution for Humanoid Robots", in IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Taipei, Taiwan, 2010, pp. 704–709.

[12] R. O'Flaherty, P. Vieira, M. Grey, P. Oh, A. Bobick, M. Egerstedt, and M. Stilman, "Kinematics and Inverse Kinematics for the Humanoid Robot HUBO2", Georgia Institute of Technology, Atlanta, GA, USA, Technical Report, 2013.

[13] M. M. Ali, H. Liu, N. Stoll, and K. Thurow, "Kinematic Analysis OF 6-DOF Arms for H20 Mobile Robots and Labware Manipulation for Transportation in Life Science Labs", Journal of Automation, Mobile Robotics & Intelligent Systems, vol. 10, no. 4, 2016, pp. 40–52.

[14] W. Chen, Y. Q. Shi, and G. Xuan, "Identifying Computer Graphics using HSV Color Model and Statistical Moments of Characteristic Functions", in IEEE International Conference on Multimedia and Expo, Beijing, China, 2007, pp. 1123–1126.

[15] Z.-K. Huang and D.-H. Liu, "Segmentation of Color Image using EM Algorithm in HSV Color Space", in IEEE International Conference on Information Acquisition, Jeju, Korea, 2007, pp. 316–319.

[16] J. R. Sanchez-Lopez, A. Marin-Hernandez, and E. R. Palacios-Hernandez, "Visual Detection, Tracking and Pose Estimation of a Robotic Arm End Effector", in the Proceeding of the Robotics Summer Meeting, Veracruz, Mexico, 2011, pp. 41–48.

[17] K. Yamazaki, Y. Watanabe, K. Nagahama, K. Okada, and M. Inaba, "Recognition and Manipulation Integration for a Daily Assistive Robot Working on Kitchen Environments", in IEEE International Conference on Robotics and Biomimetics, Tianjin, China, 2010, pp. 196–201.

[18] G. Lowe, "Object Recognition from Local Scale-Invariant Features", in IEEE International Conference on Computer Vision, Corfu, Greece, 1999, pp. 1150-1157.

[19] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "SURF: Speeded Up Robust Features", Journal of Computer Vision and Image Understanding (CVIU), Vol. 110, No. 3, pp. 346–359, 2008.

[20] E. Rosten and T. Drummond, "Machine Learning for High-Speed Corner Detection", in Computer Vision–ECCV 2006, Springer, pp. 430–443.

[21] C. H. Chen, H. P. Huang and S. Y. Lo, "Stereo-Based 3D Localization for Grasping Known Objects with a Robotic Arm System", in the World Congress on Intelligent Control and Automation (WCICA), Taipei, Taiwan, 2011, pp. 309–314.

[22] T. Grundmann, R. Eidenberger, M. Schneider, M. Fiegert, and G. v Wichert, "Robust High Precision 6D Pose Determination in Complex Environments for Robotic Manipulation", in IEEE International Conference on Robotics and Automation, Alaska, 2010, pp. 1–6.

[23] L. T. Anh and J. B. Song, "Object Tracking and Visual Servoing using Features Computed from Local Feature Descriptor", in International Conference on Control Automation and Systems (ICCAS), Gyeonggi, South Korea, 2010, pp. 1044-1048.

[24] J. Stueckler, R. Steffens, D. Holz, and S. Behnke, "Real-Time 3D Perception and Efficient Grasp Planning for Everyday Manipulation Tasks", In Proceedings of 5th European Conference on Mobile Robots (ECMR), Örebro, Sweden, 2011, pp. 177-182.

[25] A. A. Abdulla, H. Liu, N. Stoll, and K. Thurow, "A New Robust Method for Mobile Robot Multifloor Navigation in Distributed Life Science Laboratories", J. Control Sci. Eng., vol. 2016, Jul. 2016.

[26] X. Gu, S. Neubert, N. Stoll and K. Thurow, "Intelligent Scheduling Method for Life Science Automation Systems", in IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI), Baden-Baden, Germany, 2016, pp. 156-161.

[27] J. Denavit and R. S. Hartenberg, "A Kinematic Notation for Lower-Pair Mechanisms Based on Matrices", ASME Journal of Applied Mechanics, vol. 22, 1955, pp. 215 -221.

[28] M. M. Ali, H. Liu, N. Stoll, and K. Thurow, "Intelligent Arm Manipulation System in Life Science Labs Using H20 Mobile Robot and Kinect Sensor", in IEEE International Conference on Intelligent Systems (IS'16), Sofia, Bulgaria, 2016, pp. 382-387.

[29] M. M. Ali, H. Liu, R. Stoll, and K. Thurow, "Arm Grasping for Mobile Robot Transportation using Kinect Sensor and Kinematic Analysis", in IEEE International Conference on Instrumentation and Measurement Technology (I2MTC), Pisa, Italy, 2015, pp. 516–521.

[30] M. M. Ali, H. Liu, N. Stoll, and K. Thurow, "An Identification and Localization Approach of Different Labware for Mobile Robot Transportation in Life Science Laboratories", in IEEE International Symposium on Computational Intelligence and Informatics (CINTI), Budapest, Hungary, 2016, pp. 353-358.

[31] D. Oberkampf, D. F. DeMenthon, and L. S. Davis, "Iterative Pose Estimation Using Coplanar Feature Points", Computer Vision and Image Understanding, vol. 63, 1996, pp. 495–511.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 2, No. 3, 1238-1244 (2017)
www.astesj.com
Special Issue on Recent Advances in Engineering Systems

ASTES Journal
ISSN: 2415-6698

Face Recognition and Tracking in Videos

Swapnil Vitthal Tathe[*,1], Abhilasha Sandipan Narote [2], Sandipan Pralhad Narote[3]

[1] *Department of Electronics & Telecommunication, Sinhgad College of Engineering, Pune, India*

[2]*Department of Information Technology, SKN College of Engineering, Pune, India*

[3]*Department of Electronics & Telecommunication, Government Residence Women Polytechnic, Tasgaon, India*

ARTICLEINFO

Article history:
Received: 03 April, 2017
Accepted: 01 June, 2017
Online: 23 July, 2017

Keywords :
Face Detection
Face Recognition
Tracking

ABSTRACT

Advancement in computer vision technology and availability of video capturing devices such as surveillance cameras has evoked new video processing applications. The research in video face recognition is mostly biased towards law enforcement applications. Applications involves human recognition based on face and iris, human computer interaction, behavior analysis, video surveillance etc. This paper presents face tracking framework that is capable of face detection using Haar fea-tures, recognition using Gabor feature extraction, matching using correlation score and tracking using Kalman filter. The method has good recognition rate for real-life videos and robust performance to changes due to illumination, environmental factors, scale, pose and orientations.

1 Introduction

In recent years, there has been rapid growth in research and development of video sensors and analyzing technologies. Cameras are vital part of security systems used to acquire information from environment. It includes video, audio, thermal, vibration and other sensors for civilian and military applications. The intelligent video acquisition system is a convergence technology including detecting and tracking objects, analyzing their movements and responding to them.

Face recognition system is used to automatically identify a person from an image or a video source. The recognition task is performed by obtaining facial features from an image of the subject's face. The main objective of video-based face recognition and tracking is to identify a video face-track of unknown individuals. It identifies the facial features by extracting features from image of subjects face and analyzing the relative position, size, shape etc. These obtained features are utilized to search corresponding matching features in other images or subsequent frames.

The face recognition system can be divided into three main parts:detection, representation and classification. The face detection from video is challenging task as it needs compensation to distortions arising from motion. The most popular algorithm for last decade is Viola Jones algorithm based on cascaded Haar feature detector [1]. Due to its high computational speed and robustness it is widely used to detect face in still images as well as video frames. Skin color of human being has limited range of chrominance that can be modeled to represent skin color pixels [2]. To overcome the limitation of complex background with skin color like objects it detects feature like shape and appearance to validate face detection [3]. The representation stage extracts distinct and discriminative features that represents face texture and appearance for face classification. The feature extraction stage is an important as the final matching decision is biased to the quality of features. Recognition stage consists of generating reference template by averaging multiple faces and matching with the detected faces. Eigenface method is simple and popular method widely used in this field [4]. Other methods like LBP [5], LDA [6], Gabor Filter [7], etc. Neural Network is efficient method used for classification [8]. It requires large training samples suitable for face recognition in images. The computation time of recognition system depends on the size of feature vector and number of database images. The state-of-art methods are mostly used for still image face recognition. Research is in progress to develop robust methods for face recognition in real-time.

This paper is an extension of work originally presented in IEEE Annual India Conference [9]. The system proposed here is intended to detect and identify the person. The first stage captures the video segments. Pre-processing algorithms like histogram

[*]Swapnil Vitthal Tathe, Sinhgad College of Engineering, Pune, India & swapniltathe7@gmail.com

https://dx.doi.org/10.25046/aj0203156

equalization are employed to remove the unwanted artifacts present in the image [10]. The second stage detects faces in complex backgrounds using Viola-Jones [1] face detector. Next stage is representation of face features using Gabor filter [11] that inculcates directional features. The Gabor features are compared with the database images using correlation coefficients [12]. The later stage tracks the detected face in subsequent frames. Tracking utilizes popularly used Kalman filter [13] that makes use of statistical analysis to predict the location of face. The proposed system identify a person in high security premises to restrict the movements of unauthorized person.

The paper is organized as follows: Section 2 presents state-of-art technologies of the field. Section 3 describes methodologies used. Section 4 presents Results and Section 5 gives Conclusions.

2 Related Work

Recognition of individuals is of high importance for various reasons. The real-time algorithms has hard limitations to overcome as any delay in processing may result in loss of important information. Facial features are used for describing individuals as it has unique nature i.e. no two persons can have same faces. The facial features commonly used in many of the studies are skin color [14], spatio-temporal [15], geometry[16] and texture pattern [17]. Face detection algorithms are computationally intensive, which makes it is difficult to perform face detection task in real-time. As human face changes with factors like facial expression, mustache, beard, glasses, size, lightening conditions and orientation, detection algorithms must be adaptive and robust[18].

The simplest approach to detect faces is to find skin pixels present in frames and then apply geometrical analysis to locate exact position of face. The conventional images or frames in videos are modeled with RGB color space. Skin pixel values for normalized RGB plane are - r plane pixel values are in the range 0.36 to 0.456 and g values in the range 0.28 to 0.363. The second model in representing images is HSV color space. In HSV model pixel is classified as skin pixel if it has values $0 \leq H \leq 50$ and $0.20 \leq S \leq 0.68$ [19]. The third model in color representation is YC_bC_r color space. Skin color pixel has C_r value around 100 and C_b value around 150 [20]. Dhivakar et al. [21] used YC_bC_r space for skin pixel detection and Viola-Jones [1] method to verify correct detection.Skin color based method is sensitive to illumination variation and fails if background contains skin color like objects.

The other complex methods used for face detection are based on geometry and texture pattern that uses depth features to detect face. Face detection using edge maps and skin color segmentation is presented in [22]. Mehta et al. [6] proposed LPA and LBP methods to extract textural features with LDA dimensionality reduction and SVM classification. Lai et al. [23] used logarithmic difference edge maps to overcome illumination variation with face verification ability of 95% true positive rate. Contour based face detection is proposed in [24]. Xiao et al. [25] used eigen faces for face detection using texture and shape information in bi view face recognition system. Viola and Jones [1] proposed haar feature based real-time method for face detection. Many researchers popularly used haar face detection with adaboost classification [8] in face detection.

Face recognition systems are popular in biometric authentication as they do not require the user intervention. Face recognition systems can be classified as sub-class of pattern recognition systems. Eigenfaces and Fisher faces were proposed in the early 1990s by Turk et al. [26]. Bayesian face recognition has better performance compared to eigenfaces [27]. Rawlinson et al. [28] presents eigen and fisher face for face recognition. Feature based methods are robust to illumination and pose variations. SIFT and SURF has robust performance in unconstrained scenarios [29]. Gangwar et al. [30] introduced Local Gabor Rank Pattern method that uses both magnitude and phase response of filters to recognize face. Neural Network is most popular tool of recent days that is used for classification which has outperformed many methods [31]. Meshgini et al. [7] proposed a method with combination of Gabor wavelets for feature extraction and support vector machine for classification. Recognition methods are time consuming as it requires accumulates training and testing phases.

Tracking is of prime importance to monitor activities of individuals in various environments [32]. The location of object in next frame is predicted based on location of object in current or preceding frames. More details for tracking methods can be found in [33]. Elafi et al [34] proposed an approach that can detect and track multiple moving objects automatically without any learning phase or prior knowledge about the size, nature or initial position. Gyorgy et al. [35] proposed Extended Kalman filter for non linear systems. Kumar et al. [36] extracted moving objects using background subtraction and applied Kalman filter for successive tracking. The other popular tracker is mean shift that employs mean shift analysis to identify a target candidate region, which has most similar appearance to target model in terms of intensity distribution. Adaptive mean shift (Camshift) algorithm is popular OpenCV algorithm for tracking [37]. In [38] combined mean shift and kalman filter is used for tracking. Particle filter tracker uses probabilistic approach to estimate position of object in next frame. The particle filter searches for histogram that best matches with reference [39]. Tracking using sparse representation [40] is gaining momentum for fast and accurate tracking.

3 Methodology

Automatic face detection is the very first step in many applications like face recognition, face retrieval, face tracking, surveillance, HCI, etc. Face is vital part of human being representing most information about the individual. Developing a representation model is difficult because face is complex, multi-dimensional visual stimuli. General architecture of proposed system is shown in Figure 1.

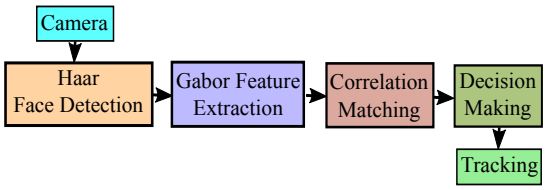


Figure 1: Human detection and tracking system

3.1 Haar Face Detection

This method is most popular method for face detection in real-time applications. Most of mobile platform uses the technique to locate faces in camera applications. Viola and Jones [1] proposed real-time AdaBoost algorithm for classification of rectangular features. The detection technique is based on the idea of the wavelet template that defines the shape of an object in terms of a subset of wavelet coefficients. The different Haar features shown in Figure 2 are used to analyze the intensity variation patterns of face regions. For example the eye region appears darker than the surrounding region that can be modeled using two rectangle feature (black part corresponding to eye region and white part corresponding to surrounding region). Combination of four, six, eight and more features is used in practical face detection to make detection robust. For a given random variable X, the variance value of X is as follows:

$$Var(X) = E(X^2) - \mu^2 \quad (1)$$

where $E(X^2)$ is expected value of squared of X and μ is expected value of X.

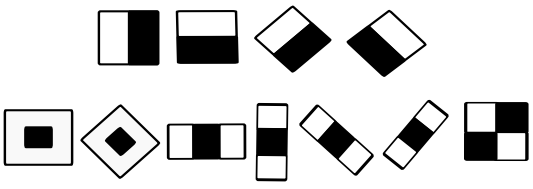


Figure 2: Haar features

Feature value is computed by subtracting the sum of variance of black region from sum of variance of white regions. To increase the computation speed, feature values can be computed using integral image [41]. Each feature is classified by a Cascaded Haar feature classifier. The Haar feature classifier multiplies the weight of each rectangle by its area and results are added together. The cascaded structure helps to speed up the detection process by eliminating non-face regions at every stage of cascade. Figure 3 shows face detection results from a video stream capture using webcam.

Figure 3: Face detection using Haar Features

Haar based method is capable of detecting multiple faces in near real-time and robust to illumination variations.

3.2 Gabor Filter

Feature extraction is one of the most important steps in designing a pattern recognition system. The features must have features have a small within class variation and strong discriminating ability among classes. Gabor filters have band pass performance that emphasizes features in face region that are analogous to scale and orientation of filter. The characteristics of Gabor filters are appropriate for texture representation and discrimination [7]. This method is robust to face pose and illumination changes.

The Gabor filter extracts face features from gray-level images. A two-dimensional Gabor filter is a Gaussian kernel function modulated by a complex sinusoidal plane wave represented as Equation 2.

$$\Psi_{\omega,\theta}(x,y) = \frac{1}{2\pi\sigma^2} exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) exp\,(j\omega x') \quad (2)$$

$$x' = xcos\theta + ysin\theta, y = -xsin\theta + ycos\theta \quad (3)$$

where (x,y) is the pixel position in the spatial domain, x is the central angular frequency of the complex sinusoidal plane wave, h is the anti-clockwise rotation(orientation) of the Gabor function, r represents the sharpness of the Gaussian function along both x and y directions and $\sigma \approx \frac{\pi}{\omega}$.

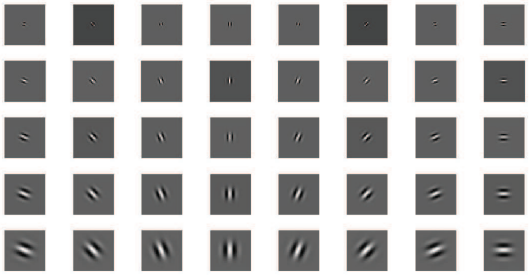


Figure 4: Gabor filters at 5 scales and 8 orientations

To extract useful features from an face image a set of Gabor filters with five frequencies (scales) and eight

orientations is used. Figure 4 shows the Gabor filter bank. The scale and orientation values are obtained using Equations 4 and 5.

$$\omega_u = \frac{\pi}{2} \times \sqrt{2}^u, u = 0, 1, \ldots, 4 \quad (4)$$

$$\theta_v = \frac{\pi}{8} \times v, v = 0, 1, \ldots, 7 \quad (5)$$

The Gabor representation of a face image $I(x, y)$ is obtained as:

$$G_{u,v}(x, y) = I(x, y) * \Psi_{\omega_u, \theta_v}(x, y) \quad (6)$$

where $G_{u,v}(x, y)$ denotes the two-dimensional convolution result corresponding to the Gabor filter at scale u and orientation v. This process can speed-up with the use of frequency domain transformation. Figure 6 and 7 shows the convolution results of a face image shown in Figure 5) with Gabor filters.

Figure 5: Face Image

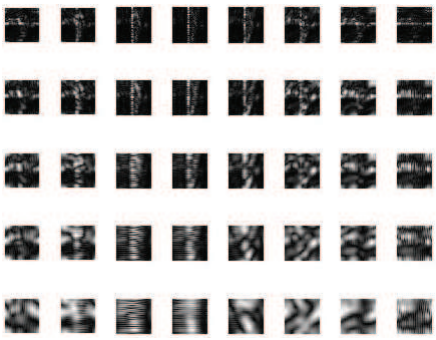


Figure 6: Magnitude of Face

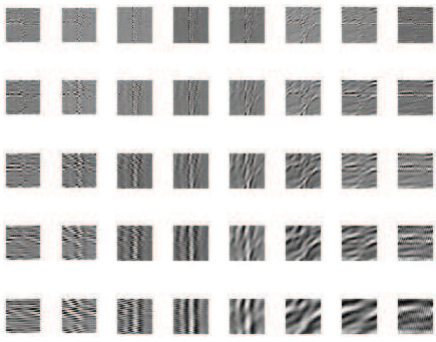


Figure 7: Real Part of Face

Each image $I(x, y)$ is represented as set of Gabor wavelet coefficients $G_{u,v}(x, y)|u = 0, 1, \ldots, 4; v = 0, 1, \ldots, 7$. The magnitude $G_{u,v}(x, y)$ is normalized to zero mean and unit variance and represented as a vector $z_{u,v}$ given by Equation 7 [7].

$$z = \left[(z_{0,0})^T (z_{0,1})^T \ldots (z_{4,7})^T\right]^T \quad (7)$$

Gabor based recognition is robust to pose and illumination variations with additional overhead of processing time and complexity.

3.3 Correlation Score Matching

Matching is a search algorithm that compares the features from query image to the database images to find best match. The comparison is done using correlation coefficients [32] between the Gabor features of detected face and the images present in database. Prior to matching stage all images in database are applied with Gabor filter for efficient feature matching. The 2D images are represented by 1D feature vector that speeds up the matching process. The correlation score is computed as :

$$Score = \sum_{i=1}^{m} f_q(i).f_{db}(-i) \quad (8)$$

where m is size of feature vector, f_q represents query feature vector and f_{db} represents database feature vector. The query image best matches with the database image that results into high correlation score.

3.4 Kalman Filter Tracking

Tracking objects in video sequences is one of the most important tasks in computer vision. The approach is to search target within a region centered on last position of the object. The important facts that should be taken care for effective tracking are: speed of moving object, frame rate of video and search region. These factors are dependent on each other, higher frame rate is used to get hold of faster moving target which on other hand requires faster processing. Tracking is defined as localization and association of features across series of frames. The two major components of tracking system are object representation and data association. The first component describes object and second utilizes the information for tracking.

The location of face from face detection stage is used as apriori estimate or sometimes called as initialization for tracking. Based on this location Kalman filter estimates and updates the prediction behavior of this stage. Initialization is an important stage for the tracker.

Kalman filter predicts the next time state vector based on the movement information in the previous stages. Kalman filter equations are categorized as - time update equations and measurement update equations. The time update equations projects forward the current state and error covariance estimates to obtain the apriori estimates for the next time step. The measurement update equations are the feedback i.e. it incorporates a new measurement into the a priori estimate to obtain an improved aposteriori estimate. The time update equations are predictor equations, while the measurement update equations are

corrector equations. The process for Kalman filter estimation and update is given in Algorithm 1.

Algorithm 1 Kalman Filter

1: Initial Estimates for $\hat{X}_{k-1}$ and P_{k-1}
2: Time Update

- Project the state

$$\hat{X}_k^- = A\hat{X}_{k-1} + Bw_k$$

- Project the error covariance

$$P_k^- = AP_{k-1}A^T + Q$$

3: Measurement Update

- Compute Kalman gain

$$K_k = \frac{P_k^- H^T}{HP_k^- H^T + R}$$

- Update estimate with measurement Z_k

$$\hat{X}_k = \hat{X}_k^- + K_k(z_k - H\hat{X}_k^-)$$

- Update error covariance

$$P_k = (1 - K_k H)P_k^-$$

The state equation is

$$X_k = AX_{k-1} + Bw_k \tag{9}$$

The measurement model is

$$Z_k = HX_k + v_k \tag{10}$$

A and B are the system parameters and are matrix in the multi-model system, H is a parameter in a measure system, which is matrix in the multi-measurement system. w_k and v_k represents the process and measurement noise respectively.

The state vector $X_k = [x_k y_k x'_k y'_k]^T$, x_k, y_k are the object center in the x, y-axis location, x'_k, y'_k are the x, y axis speed. $Z_k = [x_{ck} y_{ck}]^T$ is the observation vector. P_k^- is apriori estimate error covariance and P_k is aposteriori estimate error covariance.

The search for object is carried in a region selected based on the estimated observation matrix Z_k. It fails if image is not properly illuminated. Kalman filter overcomes problem of Kanade-Lucas-Tomasi(KLT) [42]. It can track multiple objects with variations in pose and occlusion. The drawback is that it cannot handle rapid changes in motion of target.

Table 1: Result Comparison

Video	Frames	Face Det.	True Recog.		False Recog.	
			Eigen	Gabor	Eigen	Gabor
1	237	71	47	37	24	34
2	329	99	97	31	2	68
3	257	223	52	98	171	125
4	339	225	65	138	160	87
5	448	305	225	7	80	298
6	438	283	127	4	156	279
7	353	176	140	96	36	187
8	404	274	242	195	32	88
9	198	191	101	2	90	189
10	248	206	46	19	160	187
11	78	78	21	65	57	13
12	128	124	36	74	88	50
13	324	252	122	224	130	28
14	353	250	112	216	138	36
15	258	176	107	82	69	94
16	328	191	125	95	66	96
17	346	238	65	108	173	130
18	426	392	148	289	244	103
19	318	302	144	48	158	344
20	388	257	146	16	111	376

4 Results

The algorithms are applied on video sequences obtained from NRC-IIT Facial Video Database and our own videos recorded with webcam. Table 1 gives the comparison of proposed method with eigen face method. The values presented in table are the number of frames with true result. Table 2 gives summary of results obtained on the videos. Haar face detector used has robust performance for frontal faces. It cannot handle larger pose variations. Face detection fails if the face is very small in size compared to size of frame and in random pose changes. The second part of algorithm is face recognition that makes use of gabor filters. Gabor filter has robust performance in pose and illumination variations. Once person is recognized Kalman filter keeps track of movements the face in subsequent frames. Kalman filter has good tracking performance in partial occlusions but fails in case of sudden movements. Recognition result is shown in Figure 8. The values in table shows the playback time in sections for videos under test with face detection and recognition stage.

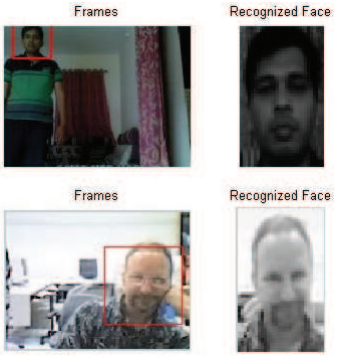


Figure 8: Face Recognition Results

Table 2: Results for Tracking

Sample Video	No. of Frames	Face Detect. (Time (sec))	Face Recog. (Time (sec))
1	237	15.5880	54.5289
2	329	20.9837	52.9032
3	257	16.4133	41.3256
4	339	21.4623	86.5154
5	448	31.4490	63.3743
6	438	30.2323	65.4514
7	353	25.3845	56.76
8	404	28.7957	113.2256
9	198	13.7490	31.8384
10	248	16.5841	39.8784
11	78	5.4555	5.97056
12	128	8.6597	20.5824
13	324	22.5488	59.4159
14	353	24.4202	66.3573
15	258	17.0961	41.4864
16	328	20.7186	52.7424
17	346	22.1367	55.6368
18	426	28.5429	68.5008
19	318	20.9056	51.1344
20	388	25.4525	35.1624

5 Conclusions

The proposed system detects the facial features of human being and verify identity of a person. Haar based face detection is popular and efficient face detection algorithm for real-time applications. Haar based face detector has high accuracy in frontal face detection. Recognition performance depends on the type of images in database. Larger the database more is recognition time. Gabor face recognition is better than many other methods as it compares features at different scale and orientations. Disadvantage of recognition stage is that comparison of large number of features requires more time for recognition that affects real-time performance. Reducing feature set reduces recognition time but affects accuracy of system. Kalman filter gives good performance in complex environment. Future directions can be optimizing algorithms to minimize processing time and achieve real-time performance.

References

[1] P. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, pp. 137–154, May 2004.

[2] C. Dong, T. Lin, X. Wang, and L. Mei-cheng, "Face detection under particular environment based on skin color model and radial basis function network," in *IEEE Fifth International Conference on Big Data and Cloud Computing*, pp. 256–259, Aug 2015.

[3] G. Lee, K. H. B. Ghazali, J. Ma, R. Xiao, and S. A. lubis, "An innovative face detection based on ycgcr color space," *Physics Procedia*, vol. 25, pp. 2116 – 2124, 2012.

[4] F. Crenna, G. Rossi, and L. Bovio, "Perceived similarity in face measurement," *Journal of Measurement*, vol. 50, pp. 397–406, 2014.

[5] G. Lian, J.-H. Lai, C. Y. Suen, and P. Chen, "Matching of tracked pedestrians across disjoint camera views using ci-dlbp," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, pp. 1087–1099, July 2012.

[6] R. Mehta, J. Yuan, and K. Egiazarian, "Face recognition using scale-adaptive directional and textural features," *Pattern Recognition*, vol. 47, pp. 1846–1858, May 2014.

[7] S. Meshgini, A. Aghagolzadeh, and H. Seyedarabi, "Face recognition using gabor-based direct linear discriminant analysis and support vector machine," *Computers and Electrical Engineering*, vol. 39, no. 910, p. 727745, 2013.

[8] M. Da'san, A. Alqudah, and O. Debeir, "Face detection using viola and jones method and neural networks," in *International Conference on Information and Communication Technology Research (ICTRC)*, pp. 40–43, May 2015.

[9] S. V. Tathe, A. S. Narote, and S. P. Narote, "Face detection and recognition in videos," in *IEEE Annual India Conference (INDICON)*, pp. 1–6, Dec 2016.

[10] C. Y. Lin, J. T. Fu, S. H. Wang, and C. L. Huang, "New face detection method based on multi-scale histograms," in *2016 IEEE Second International Conference on Multimedia Big Data (BigMM)*, pp. 229–232, April 2016.

[11] F. Jiang, M. Fischer, H. K. Ekenel, and B. E. Shi, "Combining texture and stereo disparity cues for real-time face detection," *Signal Processing: Image Communication*, vol. 28, no. 9, pp. 1100 – 1113, 2013.

[12] M. B. Hisham, S. N. Yaakob, R. A. A. Raof, A. B. A. Nazren, and N. M. W. Embedded, "Template matching using sum of squared difference and normalized cross correlation," in *IEEE Student Conference on Research and Development (SCOReD)*, pp. 100–104, Dec 2015.

[13] S. Vasuhi, M. Vijayakumar, and V. Vaidehi, "Real time multiple human tracking using kalman filter," in *3rd International Conference on Signal Processing, Communication and Networking*, pp. 1–6, March 2015.

[14] B. Dahal, A. Alsadoon, P. W. C. Prasad, and A. Elchouemi, "Incorporating skin color for improved face detection and tracking system," in *IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI)*, pp. 173–176, March 2016.

[15] H. Yang, L. Shao, F. Zheng, L. Wang, and Z. Song, "Recent advances and trends in visual tracking: A review," *Neuro Computing*, vol. 74, pp. 1–9, November 2011.

[16] J. Yan, X. Zhang, Z. Lei, and S. Z. Li, "Face detection by structural models," *Image and Vision Computing*, vol. 32, no. 10, pp. 790 – 799, 2014. Best of Automatic Face and Gesture Recognition 2013.

[17] D. T. Nguyen, W. Li, and P. O. Ogunbona, "Human detection from images and videos: A survey," *Pattern Recognition*, vol. 51, pp. 148–175, March 2016.

[18] S. Liao, A. K. Jain, and S. Z. Li, "A fast and accurate unconstrained face detector," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, pp. 211–223, Feb 2016.

[19] I. Hemdan, S. Karungaru, and K. Terada, "Facial features-based method for human tracking," in *17th Korea-Japan Joint Workshop on Frontiers of Computer Vision*, pp. 1–4, 2011.

[20] M. Zhao, P. Li, and L. Wang, "A novel complete face detection method in color images," in *3rd International Congress on Image and Signal Processing*, pp. 1763–1766, 2010.

[21] B. Dhivakar, C. Sridevi, S. Selvakumar, and P. Guhan, "Face detection and recognition using skin color," in *3rd International Conference on Signal Processing, Communication and Networking*, pp. 1–7, March 2015.

[22] R. Sarkar, S. Bakshi, and P. K. Sa, "A real-time model for multiple human face tracking from low-resolution surveillance videos," *Procedia Technology*, vol. 6, pp. 1004 – 1010, 2012.

[23] Z.-R. Lai, D.-Q. Dai, C.-X. Ren, and K.-K. Huang, "Multiscale logarithm difference edgemaps for face recognition against varying lighting conditions," *IEEE Transactions on Image Processing*, vol. 24, pp. 1735–1747, June 2015.

[24] A. Dey, "A contour based procedure for face detection and tracking from video," in *3rd International Conference on Recent Advances in Information Technology (RAIT)*, pp. 483–488, March 2016.

[25] B. Xiao, X. Gaoa, D. Tao, and X. Li, "Biview face recognition in the shapetexture domain," *Pattern Recognition*, vol. 46, pp. 1906 – 1919, July 2013.

[26] M. A. Turk and A. P. Pentland, "Face recognition using eigenfaces," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 586–591, 1991.

[27] B. Moghaddam, T. Jebara, and A. Pentland, "Bayesian face recognition," *Pattern Recognition*, vol. 33, no. 11, pp. 1771 – 1782, 2000.

[28] T. Rawlinson, A. Bhalerao, and L. Wang, *Handbook of Research on Computational Forensics, Digital Crime and Investigation: Methods and Solutions*, pp. 53–78. Ed. C-T. Li, 2010.

[29] A. Vinay, D. Hebbar, V. S. Shekhar, K. N. B. Murthy, and S. Natarajan, "Two novel detector-descriptor based approaches for face recognition using {SIFT} and {SURF}," *Procedia Computer Science*, vol. 70, pp. 185 – 197, 2015. Proceedings of the 4th International Conference on Eco-friendly Computing and Communication Systems.

[30] A. Gangwar and A. Joshi, "Local gabor rank pattern (lgrp): A novel descriptor for face representation and recognition," in *IEEE International Workshop on Information Forensics and Security (WIFS)*, pp. 1–6, Nov 2015.

[31] N. Jindal and V. Kumar, "Enhanced face recognition algorithm using pca with artificial neural networks," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 3, pp. 864–872, June 2013.

[32] G. Adhikari, S. K. Sahani, M. S. Chauhan, and B. K. Das, "Fast real time object tracking based on normalized cross correlation and importance of thresholding segmentation," in *International Conference on Recent Trends in Information Technology (ICRTIT)*, pp. 1–5, April 2016.

[33] A. W. M. Smeulders, D. M. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah, "Visual tracking: An experimental survey," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 36, pp. 1442–1468, July 2014.

[34] I. Elafi, M. Jedra, and N. Zahid, "Unsupervised detection and tracking of moving objects for video surveillance applications," *Pattern Recognition Letters*, vol. 84, pp. 70 – 77, 2016.

[35] K. Gyrgy, A. Kelemen, and L. Dvid, "Unscented kalman filters and particle filter methods for nonlinear state estimation," in *The 7th International Conference Interdisciplinarity in Engineering*, vol. 12, pp. 65 – 74, 2014.

[36] S. Kumar and J. S. Yadav, "Video object extraction and its tracking using background subtraction in complex environments," *Perspectives in Science*, vol. 8, pp. 317 – 322, 2016. Recent Trends in Engineering and Material Sciences.

[37] M. Coskun and S. Unal, "Implementation of tracking of a moving object based on camshift approach with a uav," *Procedia Technology*, vol. 22, pp. 556 – 561, 2016.

[38] S. V. Tathe and S. P. Narote, "Mean shift and kalman filter based human face tracking," in *Communications in Computer and Information Science, SPC*, pp. 317–324, June 2013.

[39] M. Z. Islam, C.-M. Oh, and C.-W. Lee, "Video based moving object tracking by particle filter," *International Journal of Signal Processing, Image Processing and Pattern*, vol. 2, pp. 119 – 132, March 2009.

[40] D. Wang, H. Lu, and C. Bo, "Online tracking via two view sparse representation," *IEEE Signal Processing Letters*, vol. 21, pp. 1031–1034, September 2014.

[41] J. Jin, B. Xu, X. Liu, Y. Wang, L. Cao, L. Han, B. Zhou, and M. Li, "A face detection and location method based on feature binding," *Signal Processing: Image Communication*, vol. 36, pp. 179 – 189, 2015.

[42] V. M. Arceda, K. F. Fabian, P. L. Laura, J. R. Tito, and J. G. Caceres, "Fast face detection in violent video scenes," *Electronic Notes in Theoretical Computer Science*, vol. 329, pp. 5 – 26, 2016.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1245-1251 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Efficiency of Photovoltaic Maximum Power Point Tracking Controller Based on a Fuzzy Logic

Ammar Al-Gizi[1,2*], Sarab Al-Chlaihawi[1,3], Aurelian Craciunescu[1]

[1]*Electrical Engineering Faculty, University Politehnica of Bucharest, 060042, Romania*

[2]*Al-Mustansiriyah University, Faculty of Engineering, 10001, Baghdad, Iraq*

[3]*Al-Furat Al-Awsat Technical University, Najaf Technical Institute, 31001, Najaf, Iraq*

ARTICLE INFO

Article history:
Received: 18 May, 2017
Accepted: 29 June, 2017
Online: 23 July, 2017

Keywords:
Photovoltaic systems
Maximum Power Point Tracking
Fuzzy Logic Controller
Perturb and Observe
Climate conditions

ABSTRACT

This paper examines the efficiency of a fuzzy logic control (FLC) based maximum power point tracking (MPPT) of a photovoltaic (PV) system under variable climate conditions and connected load requirements. The PV system including a PV module BP SX150S, buck-boost DC-DC converter, MPPT, and a resistive load is modeled and simulated using Matlab/Simulink package. In order to compare the performance of FLC-based MPPT controller with the conventional perturb and observe (P&O) method at different irradiation (G), temperature (T) and connected load (R_L) variations - rising time (t_r), recovering time, total average power and MPPT efficiency topics are calculated. The simulation results show that the FLC-based MPPT method can quickly track the maximum power point (MPP) of the PV module at the transient state and effectively eliminates the power oscillation around the MPP of the PV module at steady state, hence more average power can be extracted, in comparison with the conventional P&O method.

1. Introduction

The extracted output power from the photovoltaic (PV) system is affected by the climate conditions represented by a solar irradiation (*G*) and a temperature (*T*). Hence, an effective maximum power point tracking technique is required improve the efficiency of the overall PV system.

In the last years, many MPPT techniques have been proposed to track the MPP of the PV module [1–6]. In [1], the FLC based MPPT is proposed to track the MPP at standard technical condition (STC) and variable resistive load. Where, the change in duty cycle (ΔD) was used as an FLC's output to adjust the duty cycle (D) of the DC-DC converter, in order to meet the load matching between the PV module and the connected resistive load.

This work is an extension of work originally presented in 2016 International Conference on Applied and Theoretical Electricity (ICATE) [1]. In this paper, the efficiency of FLC-based MPPT for the PV system is examined under variable climate conditions (irradiation and temperature) and variable values of connected resistive load. The utilized PV system includes a BP SX150S PV module, buck-boost DC-DC converter, MPPT controller, and resistive load (R_L). The circuit diagram of the PV system is shown in Figure 1.

The rest of this paper is arranged as follows: Section 2 explains the basic modeling of the utilized PV module. Section 3 tackles the MPPT control and the mechanism of load matching through the DC-DC converter. Section 4 describes the details of considered MPPT algorithms: conventional perturb and observe (P&O) and the proposed FLC. Simulation and comparison results obtained by using Matlab software are presented in Section 5. Finally, conclusions of the work are revealed in Section 6.

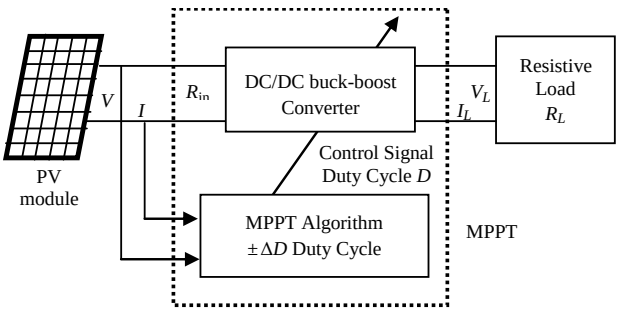


Figure 1. The PV system with MPPT control.

*Ammar AL-GIZI, Ph.D. student, University Politehnica of Bucharest, 060042, Romania, +40732481220, ammar.ghalib@yahoo.com

2. Modeling of the PV Module

In this paper, a single-diode equivalent circuit model is used. This model consists of two resistances (series R_s and parallel R_{sh}) as depicted in Figure 2 [7].

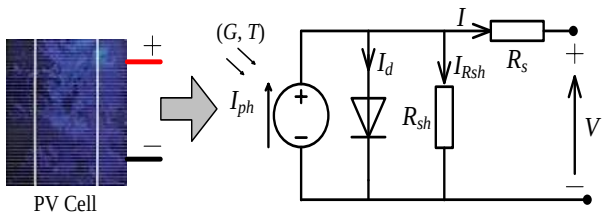


Figure 2. The equivalent circuit of the PV model.

According to the Kirchhoff's law, the characteristic equation of the PV model can be expressed by [8]:

$$I = I_{ph} - I_d - I_{Rsh} \tag{1}$$

By developing the terms of I_d and I_{Rsh}, the PV current can be written as,

$$I = I_{ph} - I_o\left(\exp\left(\frac{q(V + I R_s)}{N_s n K T}\right) - 1\right) - \left(\frac{V + I R_s}{R_{sh}}\right) \tag{2}$$

Where I_{ph} is the light current that is equal to short-circuit current I_{sc}. I_d and I_{Rsh} are the parallel currents through the diode and parallel resistance R_{sh}, respectively. I_o is the cell's reverse saturation current, also known as the "dark current", q is the electron charge (1.602×10^{-19} C), n is the diode ideality factor of a value is between 1 and 2, K is the Boltzmann's constant (1.381×10^{-23} J/K), T is the PV cell temperature (K), and N_s is the number of series-connected solar cells to organize the utilized PV module [1,8,9]. In this paper, the BP SX150S PV of N_s =72 cells is utilized to produce a peak power output of 150 W at standard technical condition (STC). Under STC, the cell temperature T_r is 298 °K which equivalent to 25 °C and the solar irradiation G_r is 1000 W/m^2 at air mass (AM=1.5). Table 1 illustrates the utilized PV module's electric parameters [10].

Table 1. Electrical parameters of BP SX150S PV module

Parameter	Value
Maximum Power (P_{max})	150 W
Voltage at P_{max} (V_{mpp})	34.5 V
Current at P_{max} (I_{mpp})	4.35 A
Short-circuit current (I_{sc})	4.75 A
Open-circuit voltage (V_{oc})	43.5 V
Temperature coefficient of I_{sc}	(0.065 ± 0.015) %/ °C
Temperature coefficient of V_{oc}	– (160 ± 20) mV/ °C

The reverse saturation current (I_o) is dependent on the temperature (T) and can be expressed by:

$$I_o = I_{or} \times \left(\frac{T}{T_r}\right)^3 \times \exp\left(\frac{q E_g}{n K}\left(\frac{1}{T_r} - \frac{1}{T}\right)\right) \tag{3}$$

Where E_g is the band-gap energy of the semiconductor used in the cell.

To determine the module's open-circuit voltage (V_{oc}), we set (I=0 and V=V_{oc}) and assume R_{sh}= ∞ in (2), which results,

$$V_{oc} = N_s n \frac{KT}{q} \ln\left(\frac{I_{sc}}{I_o} + 1\right) \tag{4}$$

The PV module's short-circuit current (I_{sc}) depends on the cell temperature (T) and solar irradiation (G) as,

$$I_{sc} = \frac{G}{G_r}\left(I_{scr} + \alpha(T - T_r)\right) \tag{5}$$

Where α is the short circuit current temperature coefficient.

The PV module's maximum power point (MPP) is the point at which the peak product of V and I is obtained, hence V=V_{mpp}, I=I_{mpp}, and P=P_{max} at MPP. This point changes with the variation of climate conditions as shown in Figure 3 and Figure 4. It can be seen that the current and power of the PV module increase linearly as irradiation increases, while the voltage approximately unchanged as shown in Figure 3. In contrast, for an increasing temperature, the voltage and power decrease while the current approximately unchanged as shown in Figure 4. Different MPPT methods can be used to maintain the PV module's operating point at or near the MPP, thereby, extracting a maximum available output power from the PV module.

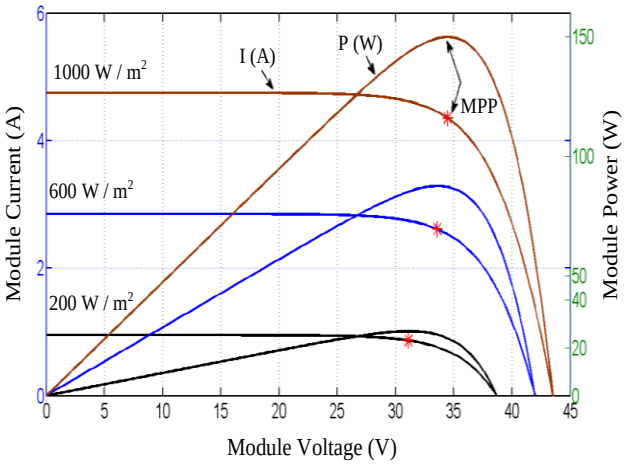


Figure 3. *I-V* and *P-V* Characteristics of the PV module at various irradiations and 25 °C of temperature.

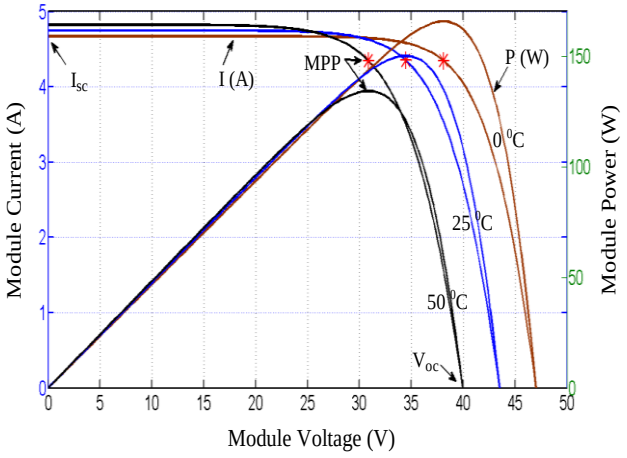


Figure 4. *I-V* and *P-V* Characteristics of the PV module at various temperatures and 1000 W/m^2 of irradiation.

3. MPPT Control and Load Matching

The switch mode DC-DC converter represents a main part of the PV MPPT. The DC-DC converter is used for two purposes; firstly, to regulate the PV operating point at the MPP. Secondly, to maintain the matching between the PV optimal impedance (R_{opt}) and the connected load impedance (R_L), consequently, a maximum power transfer between the PV module and the load can be obtained. The DC-DC converter uses the duty cycle control signal produced by the MPPT algorithm [1].

According to an ideal (lossless) DC-DC converter, the input-output voltage and current relationships can be described by:

$$P_{in} = P_L , \tag{6}$$

$$V \times I = V_L \times I_L , \tag{7}$$

$$\frac{V_L}{V} = \frac{I}{I_L} = \frac{D}{1-D} \tag{8}$$

Where, V_L and I_L are the load's voltage and current, respectively. Whereas V and I are the PV module's voltage and current, respectively. The relationship between the voltage and duty cycle can be expressed as:

- If $0 < D < 0.5$, the converter output voltage does not exceed the input voltage (Buck type).
- If $D = 0.5$, the converter's input and output voltages are identical.
- If $0.5 < D < 1$, the converter output voltage exceeds the input voltage (Boost type).

The DC-DC converter's input impedance (i.e. input impedance seen by the PV module) can be written as:

$$R_{in} = \frac{V}{I} \tag{9}$$

By substituting (8) in (9), R_{in} can be expressed as:

$$R_{in} = \frac{(1-D)^2}{D^2} \times \frac{V_L}{I_L} = \frac{(1-D)^2}{D^2} \times R_L , \tag{10}$$

$$R_L = \frac{V_L}{I_L} \tag{11}$$

From (10), it can be seen that by changing the duty cycle, the value of R_{in} can be matched with the optimal input impedance (R_{opt}) of the PV module at the MPP, where,

$$R_{opt} = \frac{V_{mpp}}{I_{mpp}} \tag{12}$$

At STC, the relationship between the duty cycle and extracted power from the PV module at various values of resistive load (R_L) is shown in Figure 5 [1]. As it is seen in this figure, when R_L is increased, the optimal duty cycle (D_{opt}) is also increased to reach the unique MPP. In this paper, the goal of the two utilized MPPT methods is to maintain D_{opt} under various climate (G and T) conditions and resistive load (R_L).

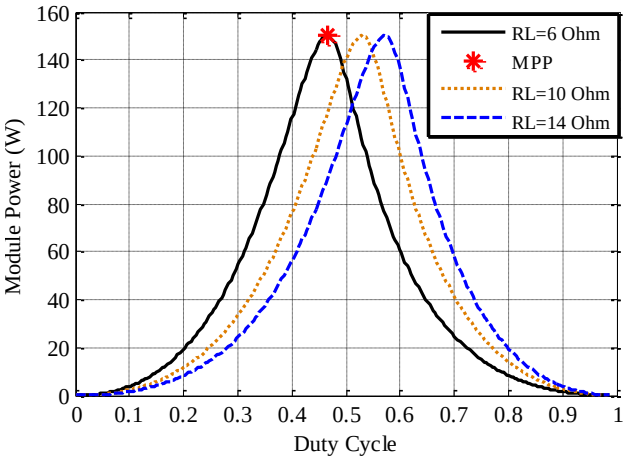


Figure 5. Module power vs. duty cycle at STC and various resistive loads.

4. MPPT Algorithms

4.1. Perturb and Observe (P&O) Method

The perturb and observe (P&O) method, also known as the "hill climbing" method is widely used as an MPPT algorithm for PV systems applications due to its simplicity and ease of implementation and simplicity [11]. The algorithm perturbs the duty cycle by increasing or decreasing it using a constant duty step-size (ΔD) and observes resulting variations in the PV output power (ΔP). If the $\Delta P>0$, that means the new operating point has moved closer towards the MPP, hence the duty cycle is further perturbed in the same direction; otherwise, the direction will be reversed. This process is repeated periodically until MPP is maintained. Unfortunately, by this algorithm, the operating point oscillates around the MPP at the steady state. Consequently, the PV efficiency is slightly reduced. Although the steady state oscillation can be decreased by decreasing ΔD, however, the dynamic response will be slowed. Hence, the tradeoff between these two requires that ΔD is finely tuned. The flow chart of the P&O algorithm is shown in Figure 6.

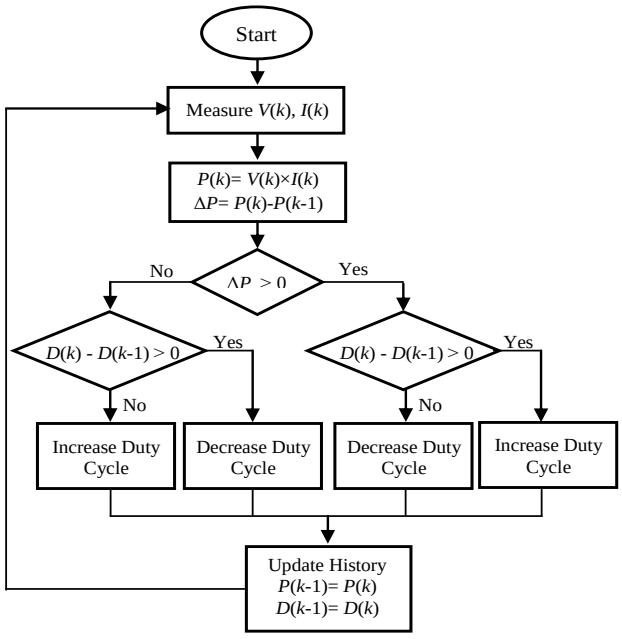


Figure 6. Flow chart of the P&O algorithm.

4.2. Fuzzy Logic Control (FLC) Method

The fuzzy logic was developed by Zadeh in 1965. The FLC is used to convert the human information to the rule-based model that can control a plant with linguistic explanations [12,13].

In this paper, the utilized FLC has two inputs and one output. The inputs are an error (E) and its variation (ΔE), whereas the output is the change in duty cycle (ΔD), as shown in Figure 7.

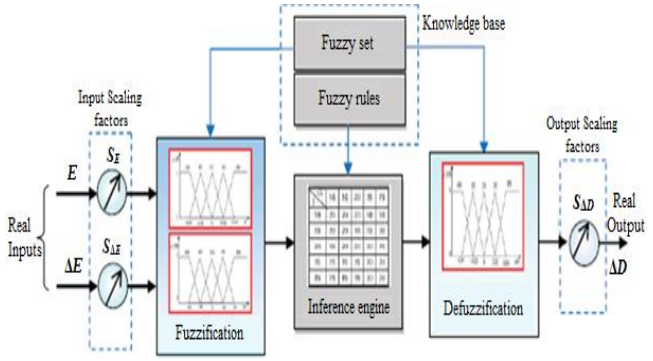


Figure 7. The Structure of a fuzzy logic controller.

The inputs and output, at a sampling instant k are expressed as follows:

$$E(K)=\frac{P(K)-P(K-1)}{V(K)-V(K-1)}, \quad (13)$$

$$\Delta E(K)=E(K)-E(K-1), \quad (14)$$

$$D(K)=D(K-1)+\Delta D(K) \quad (15)$$

Where $P(k)$ and $V(k)$ are the output power and voltage of the PV module at sampling k. $\Delta D(k)$ is the change in duty cycle used as the FLC's output to calculate the DC-DC converter's actual duty cycle $D(k)$ at sampling k.

$E(k)$ is the P-V curve's slope. Consequently, the sign of $E(k)$ indicates the operating point's location at instant k, either on the left or on the right of the MPP on the P-V curve of PV module as illustrated in Figure 8. However, the input $\Delta E(k)$ expresses the moving direction of this point. The typical FLC includes three main components: fuzzification, inference engine, and defuzzification, as shown in Figure 7 [14].

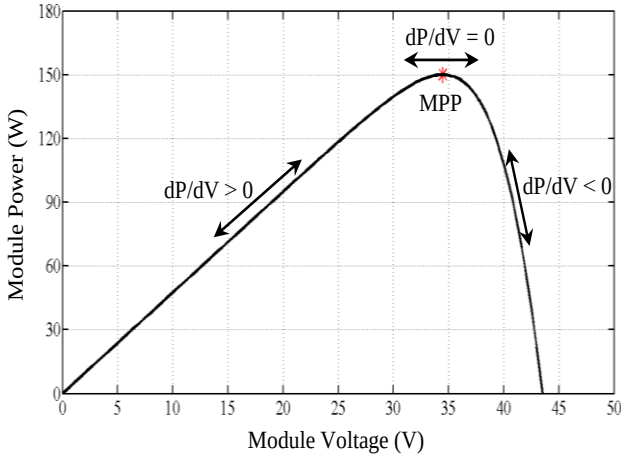


Figure 8. The slope of P-V characteristics of PV module at STC.

I. Fuzzification

After normalizing the actual (crisp) FLC's inputs by scaling factor (S_E) and ($S_{\Delta E}$), as shown in Figure 7, the inputs are fuzzified. The fuzzification process converts the input and output variables from real crisp variables to fuzzy variables which are expressed by linguistic terms such as negative big (*NB*), negative small (*NS*), zero (*Z*), positive small (*PS*), and positive big (*PB*). Figure 9 shows the fuzzy sets for the inputs and output variables used in this paper with triangular and trapezoidal membership functions.

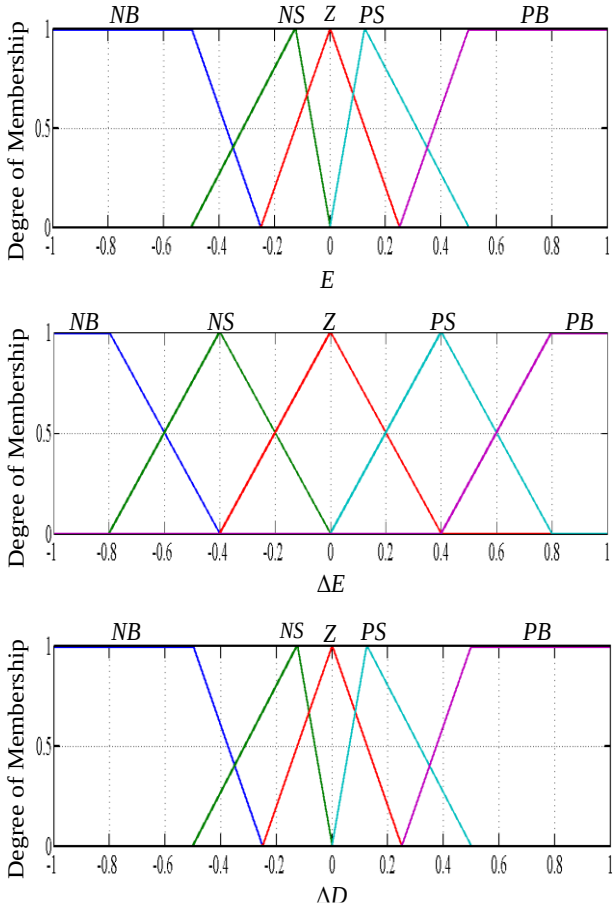


Figure 9. Membership functions: E and ΔE are the inputs; ΔD is the output.

II. Fuzzy Rules and Inference Engine

The inference engine applies the rules to the fuzzy inputs which are produced by the fuzzification process to generate n aggregated fuzzy outputs. In this paper, since each of the input and output has five fuzzy subsets, 25 fuzzy IF-THEN rules shown in Table 2 are used. The rules are used to control the DC-DC converter by a suitable ΔD.

The main concept of the rules is related to the location of the operating point from the MPP. If the operating point moves closer toward the MPP, the ΔD will be increased or decreased to a small extent and vice versa if the operating point diverges away from the MPP. For example, in the shaded rule in Table 2: If E is *NB* AND ΔE is *Z* THEN ΔD is *PB*. This means that if the operating point is located at a distance to the right of the MPP and there is no change in the P-V slope, then the duty ratio is substantially increased.

Table 2. Fuzzy logic controller rules

ΔD		ΔE				
		NB	*NS*	*Z*	*PS*	*PB*
E	*NB*	*Z*	*Z*	*PB*	*PB*	*PB*
	NS	*Z*	*Z*	*PS*	*PS*	*PS*
	Z	*PS*	*Z*	*Z*	*Z*	*NS*
	PS	*NS*	*NS*	*NS*	*Z*	*Z*
	PB	*NB*	*NB*	*NB*	*Z*	*Z*

III. Defuzzification

The defuzzification represents the inverse of the fuzzification process. It is used to transform the output from the fuzzy domain into the single real (crisp) domain, using the well-known center of gravity (COG) defuzzification methods. This method computes the centroid of the final composite area representing the output fuzzy term, which is resulted by the union of all 25 rule output fuzzy sets [15,16], to produce the real value of ΔD as:

$$\Delta D = \frac{\sum_{i=1}^{n} \Delta D_i \times \mu(\Delta D_i)}{\sum_{i=1}^{n} \mu(\Delta D_i)} \quad (16)$$

Finally, this crisp output is denormalized using the output scaling factor ($S_{\Delta D}$) then substituted in (15) to generate the actual duty cycle $D(k)$.

5. Simulation Results

In this paper, the load, irradiation, and temperature variations are presented in Figures 10–12, respectively.

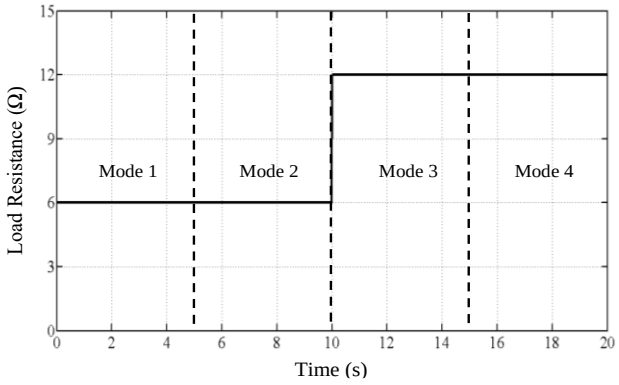


Figure 10. Resistive load variation.

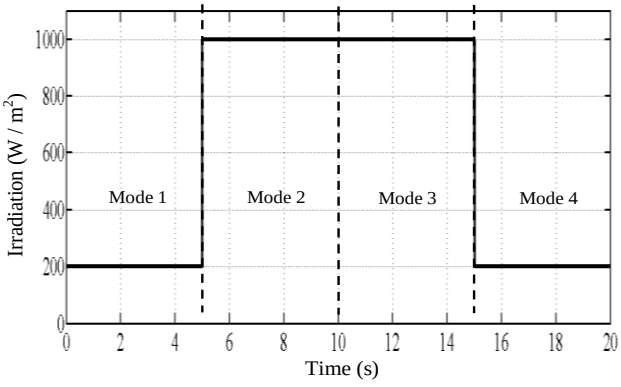


Figure 11. Irradiation variation.

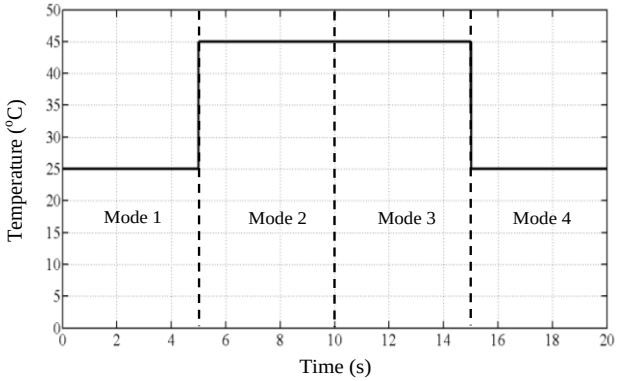


Figure 12. Temperature variation.

The overall operating period can be classified into four different modes:

- Mode 1 is the period where G=200 W/m^2, T=25 °C, and R_L=6 Ω.
- Mode 2 is the period where R_L is kept constant whereas G and T are suddenly increased to 1000 W/m^2 and 45 °C, respectively.
- Mode 3 is the period where the load is changed from 6 Ω to 12 Ω at constant G and T of the previous mode.
- Mode 4 is the period where the load is kept constant whereas G and T are suddenly decreased to the same its values in mode 1.

During the time period of mode 1 shown in Figures 10–12, the MPPT methods increase the duty cycle from initial value (D=0.1) to (D_{opt1}=0.29), as shown in Figure 13 and decrease the converter input resistance (R_{in}) by using (10) to match (R_{opt1}=35.92 Ω), moving the operating point of the PV module from starting point ***A*** to reach MPP at point ***B***, as shown in Figure 14, thereby extracting maximum power (P_{max1}=27 W) from the module as shown in Figures 15–17.

Due to a large increasing in the irradiation at the starting mode 2, the operating point will suddenly jump to point ***C***, as shown in Figure 14 with a small increasing in power to (P=40 W), as shown in Figures 15–17. During this mode, the MPPT methods increase D to (D_{opt2}=0.476), as shown in Figure 13 and decrease R_{in} to match the new (R_{opt2}=7.27 Ω), moving the operating point from point ***C*** to the new second MPP at point ***D***, as shown in Figure 14, thereby increasing the power to (P_{max2}=137.3 W), as shown in Figures 15–17.

At the starting of mode 3, when the load is suddenly doubled, R_{in} is also doubled from 7.27 Ω to 14.54 Ω, thus the operating point will suddenly jump away in the right direction from point ***D*** to point ***E***, as shown in Figure 14 and the power drops from P_{max2} to 95 W, as shown in Figures 15–17. During this mode, the tracking methods try again to recover the operating point from current point ***E*** to the MPP at point ***F*** which is the same location of optimal point ***D***, by moving it in the left direction, by increasing D from D_{opt2} to new (D_{opt3}=0.5624), as shown in Figure 13 and decreasing R_{in} to match R_{opt2} again, as shown in Figure 14, to reach P_{max2} again, as shown in Figures 15–17.

Finally, at the starting of mode 4, when the load is kept constant and the climate conditions drops as the inverse of starting condition of mode 2, the operating point will suddenly jump to point ***G***

shown in Figure 14 with a large decreasing in power from P_{max2} to 7.5 W, as shown in Figures 15–17. During this period, the MPPT methods decrease D to (D_{opt4}=0.3663), as shown in Figure 13 and increase R_{in} to match (R_{opt1}=7.27 Ω) again, moving the operating point in the right direction from point ***G*** to the MPP at point ***H*** which is the same location of optimal point ***B***, as shown in Figure 14, thereby increasing the power and reaching P_{max1} again, as shown in Figures 15–17.

From the different operation modes, it is clear that the FLC provides a fast and effective transient response in terms of tracking and recovering speed and this method also provides an effective steady response with less oscillation as compared with P&O tracking method, according to the responses of duty cycle and extracted power from the PV module shown in Figures 13, 16, respectively.

At variable climate and load operation modes explained in Figures 10–12, Table 3 lists the load matching results of voltage, current, and power of PV module, duty cycle and input resistance of the DC-DC converter, and the corresponding load voltage and current using MPPT.

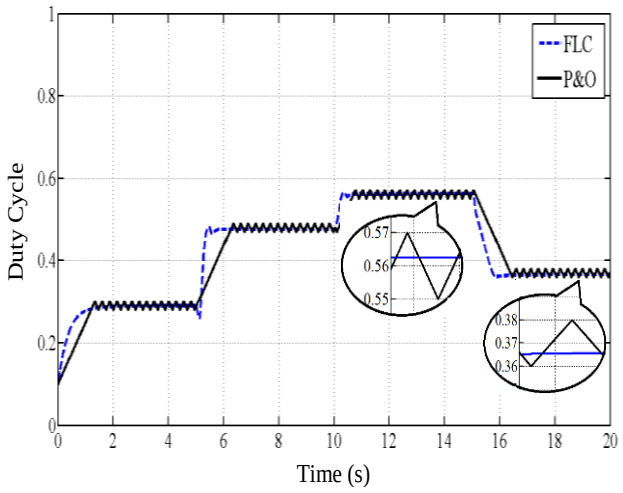


Figure 13. Duty cycle performance using MPPT methods under variable climate and load conditions.

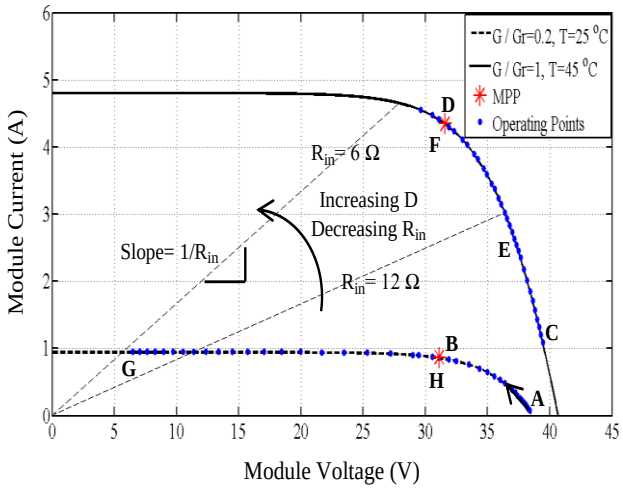


Figure 14. *I-V* Characteristics of PV module and traces of operating point under variable climate and load conditions.

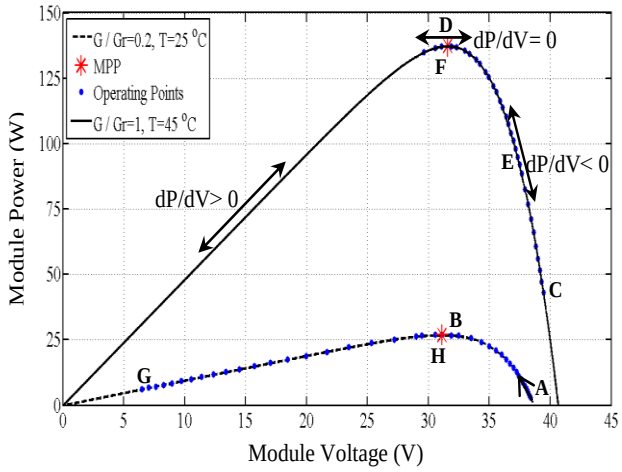


Figure 15. *P-V* Characteristics of PV module and traces of operating point under variable climate and load conditions.

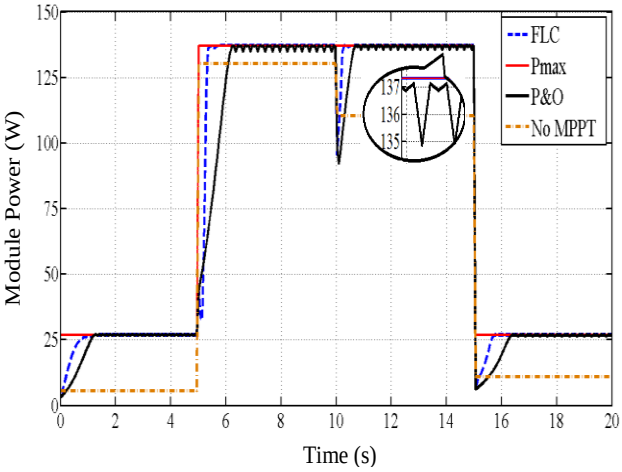


Figure 16. Module power performances using MPPT methods under variable climate and load conditions.

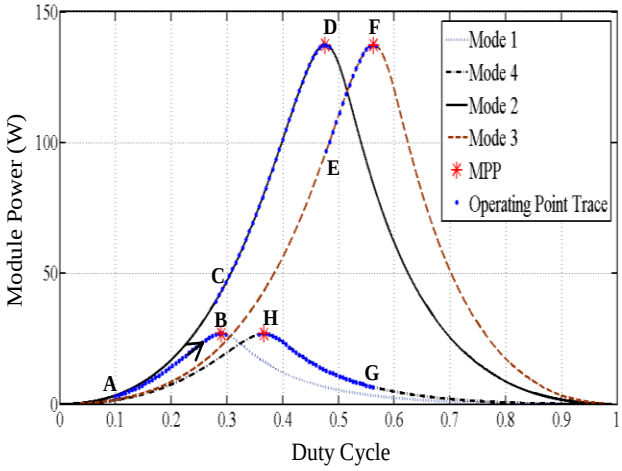


Figure 17. Module power vs duty cycle and traces of operating point under variable climate and load conditions.

Whereas, Table 4 lists the comparison results of MPPT methods in terms of rising time, recovery time, total average power, and MPPT efficiency during the overall operating period.

Table 3. Load matching results under different operating modes

Operating Modes	V_{mpp} (V)	I_{mpp} (A)	P_{max} (W)	R_{opt} (Ω)	D_{opt}	V_o (V)	I_o (A)
Mode 1	31.1	0.86	26.92	35.92	0.29	12.71	2.12
Mode 2	31.6	4.34	137.3	7.27	0.47	28.70	4.34
Mode 3	31.6	4.34	137.3	7.27	0.56	40.59	3.38
Mode 4	31.1	0.86	26.92	35.92	0.36	17.97	1.49

Table 4. Comparison of MPPT methods under different operating modes

MPPT Method	t_r (s)	Recovering Time (s)	Ideal P_{av} (W)	Total P_{av} (W)	Efficiency %
Without (D=0.5)	–	–	82.1137	64.153	78.12
P&O	1.25	0.75	82.1137	76.3	92.92
FLC	0.75	0.25	82.1137	79.66	97.01

6. Conclusions

In this paper, the effectiveness of the FLC and classical P&O based MPPT methods are analyzed by simulation and compared using different operation modes of irradiation, temperature, and load requirement.

In the case of rapid changes in the climate conditions and constant connected load, as represented by starting periods of modes 2 and 4, the simulation results show that the FLC tracking method can quickly reach the new MPP with less rising time (t_r) at transient response and eliminates the power oscillation around the MPP at steady state; thus providing a higher power output efficiency, as shown in Figures 13, 16 and Table 4.

According to a rapid fluctuation of the load, as represented by starting periods of mode 3, also the FLC method has better performances in comparison with P&O method – less recovering time and better steady responses of less oscillation as shown in Figures 13, 16 and Table 4. Consequently, the PV system with FLC-based MPPT control is more efficient due to the fact that allows harvesting more solar power in comparison with the classical P&O tracking methods.

Acknowledgment

The first author thanks the Iraqi government represented by the Ministry of higher education and scientific research, Al-Mustansiriya University, Baghdad, Iraq for its financial support.

References

[1] A. G. Al-Gizi, "Comparative study of MPPT algorithms under variable resistive load" in Proc. of the IEEE International Conference on Applied and Theoretical Electricity (ICATE 2016), Craiova, Romania, 2016. doi: 10.1109/ICATE.2016.7754611

[2] V. M. Elena and Papadopoulou, Photovoltaic Industrial Systems: An Environmental Approach, Green Energy and Technology, Springer Heidelberg Dordrecht, London-New York, 2011.

[3] F. L. Tofoli, D. Pereira, W. Josias, "Comparative study of maximum power point tracking techniques for photovoltaic systems" International Journal of Photoenergy, **2015**, 1–10, Jan. 2015. doi: 10.1155/2015/812582

[4] S. Z. Hassan, H. Li, Y. H. Liu, T. Kamal, U. Arifoglu, S. Mumtaz, L. Khan, "Neuro-fuzzy wavelet based adaptive MPPT algorithm for photovoltaic systems" Energies, **10**(3), 1–16, 2017. doi: 10.3390/en10030394

[5] P. C. Cheng, B. R. Peng, Y. H. Liu, Y. S. Cheng, J. W. Huang, "Optimization of a fuzzy-logic-control-based MPPT algorithm using the particle swarm optimization technique" Energies, **8**(6), 5338–5360, 2015. doi: 10.3390/en8065338

[6] C. L. Liu, J. H. Chen, Y. H. Liu, Z. Z. Yang, "An asymmetrical fuzzy-logic-control-based MPPT algorithm for photovoltaic systems" Energies, **7**(4), 2177–2193, 2014. doi: 10.3390/en7042177

[7] T. Salmi, M. Bouzguenda, "Matlab/Simulink based modeling of solar photovoltaic cell" International Journal of Renewable Energy Research, **2**(2), 2012.

[8] V. V. Naidu, T. M. Mohan, "Modeling and simulation of photovoltaic system" Golden Research Thoughts, **3**(6), 1–6, 2013.

[9] M. Spasoje, M. Nedeljkovic, "The solar photovoltaic panel simulator" Rev. Roum. Sci. Techn. – Électrotechn. Et Énerg, **60**(3), 273-281, 2015.

[10] http://www.abcsolar.com/pdf/bpsx150.pdf

[11] M. T. Makhloufi, M. S. Khireddine, Y. Abdessemed, "Maximum power point tracking of a photovoltaic system using a fuzzy logic controller on dc-dc boost converter" IJCSI International Journal of Computer Science, **11**(3), 1694–0784, 2014.

[12] A. V. Pavan, A. M. Parimi, K. U. Rao, "Implementation of MPPT control using fuzzy logic in solar-wind hybrid power system" in Proc. of the IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES), Kozhikode, India, 2015. doi: 10.1109/SPICES.2015.7091364

[13] A. M. Othman, A. Fathy, A. Ghitas, "Realworld maximum power point tracking simulation of PV system based on fuzzy logic control" NRIAG Journal of Astronomy and Geophysics, 186–194, 2012. doi: 10.1016/j.nrjag.2012.12.016

[14] B. Bendib, F. Krim, H. Belmili, M. F. Almi, S. Boulouma, "Advanced fuzzy MPPT controller for a stand-alone PV system" Energy Procedia, 50, 383–392, Dec. 2014. doi: 10.1016/j.egypro.2014.06.046

[15] A. G. Al-Gizi, A. Craciunescu, S. J. Al-Chlaihawi, "The use of ANN to supervise the PV MPPT based on FLC" in Proc. of the 10th IEEE International Symposium on Advanced Topics in Electrical Engineering (ATEE), Bucharest, Romania, 703–708, 2017. doi: 10.1109/ATEE.2017.7905128

[16] S. Z. Hassan, H. Li, T. Kamal, M. Nadarajah, F. Mehmood, "Fuzzy embedded MPPT modeling and control of PV system in a hybrid power system." In *Emerging Technologies (ICET), 2016 International Conference on*, pp. 1-6. IEEE, 2016. doi: 10.1109/ICET.2016.7813236

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1252-1259 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Model-based Development of Safety-critical Functions and ISO 26262 Work Products using modified EAST-ADL

Bülent Sari[*1], Hans-Christian Reuss[2]

[1]*Electronics Powertrain Technology, ZF Friedrichshafen AG, Friedrichshafen, Germany*

[2]*Forschungsinstitut für Kraftfahrwesen und Fahrzeugmotoren Stuttgart (FKFS), Stuttgart, Germany*

ARTICLE INFO

Article history:
Received: 22 March, 2017
Accepted: 06 June, 2017
Online: 23 July, 2017

Keywords:
model-driven approach
safety
ISO 26262
EAST-ADL

ABSTRACT

Safety is becoming more and more important with the ever increasing level of safety related E/E Systems built into the cars. Increasing functionality of vehicle systems through electrification of power train, in future even more by autonomous driving, leads to complexity in designing system, software and safety architecture. ISO 26262 aims to reduce the complexity and to approve the traceability of the different safety activities. This paper presents an approach about model-based development of system, software and safety architecture using Electronics Architecture and Software Technology - Architecture Description Language (EAST-ADL), being in line with the relevant standard ISO 26262. In particular, we briefly discuss how the main safety related activities, such as hazard analysis and risk assessment, developing functional and technical safety concepts and performing safety analysis can be performed model-based and how the activities can be related with system and software development. The state-of-art is also provided and compared with the proposed approach.

1. Introduction

This paper is an extension of work originally presented in Electrical Systems for Aircraft, Railway, Ship propulsion and Road Vehicles & International Transportation Electrification Conference (ESARS-ITEC 2016) [1]. Nowadays, in a premium vehicle up to 100 processing units (ECUs) are installed, which are capable of computing complex algorithms. The embedded software of a premium automobile contains up to 100 million lines of source code. On the contrary, the new “Boeing Dreamliner 787” needs for all onboard systems around 6.5 million lines of code [2].

This comparison shows how complex the software is already in today's vehicles. Complexity and the scale of the software will continue growing [2]. Major reasons for the large amount of software are the electrification of the automobile and already available advanced driver assistance systems. It is believed that the ratio of the cost of electric and electronic components to the total production cost of a vehicle could rise up to 35% by 2020 and up to 50% by 2030 [3]. With the increase of electrification, the proportion of safety-critical systems is also growing. Malfunctions like "unintended blocking of the drive axle while driving", "power assisted steering acts in the wrong direction", or "false tripping of the airbag while driving" are a few examples that can lead to life-threatening injuries [4].

Increasing functionality of existing or new vehicle systems both result in increasing proportion of E/E systems and an increase in complexity of the system, software and safety architecture. Engineers from different fields need to be involved. The system and software architectures have to satisfy the different requirements in any phase of development. At this point the question arises, how the industry is responding to the increase in safety and non-safety related functions without sacrificing the quality of the software.

An approach, how the engineers deal with these challenges, uses model-driven system, software and safety development. This research presents an approach for model-based development of safety critical functions and model-based development of ISO work products. EAST-ADL is used as the basis of this approach. However EAST-ADL should be modified in a way to bring the system architecture and safety architecture together, because ISO 26262 requires consistency and traceability by the development of safety critical systems. But the current EAST-ADL specification is not enough to implement these requirements, because currently the

[*]Corresponding Author: Bülent Sari, Electronics Powertrain Technology, ZF Friedrichshafen AG, Friedrichshafen, Germany | Email: buelent.sari@zf.com

system architecture and safety architecture are separated in different models and there is not direct relation between system architecture models and safety models. The system architecture is developed within the abstraction levels of EAST-ADL and the safety models are realized within the dependability model. It is a big challenge and also requirement of ISO 26262 (Part 4 – Figure 2, Part 10 - Figure 8 and Figure 9) to show the dependencies between hazards and risks, safety goals, functional and technical safety requirements and safety functions. By safety assessments, the developers should be able to show which safety goals are implemented by which safety functions and it should be proven that the safety functions and safety goals have the same Automotive Safety Integrity Level (ASIL). Therefor it was essential to modify EAST-ADL in a way to bring these models together in order to provide consistency check, traceability of safety functions and also generate the ISO26262 work products such as safety concept automatically. The extensions enable to show the relationship between the safety goals and the safety functions in order to prove easily which function is implemented for which safety goal. So it is easier to prove the completeness of the safety activities.

The second main contribution of the paper is to verify and to validate the technical safety concept and functional safety concept as required from ISO 26262-4:2018, Clause 7.4.8 within simulation of safety requirements in the earlier development phase of project. So it is possible to improve the safety concepts earlier and also to find out the systematic errors in the earlier phase of development. The modeling of a complete system is an extensive project and requires domain specific knowledge. Furthermore, knowledge is needed about various tools, because the architecture is described using several tools. Current approach can be replaced within developed approach based on domain-specific language with ADL [6, 7]. By this it is possible now within the extensions and improvements that all information is created within a system model that describes the complete system, safety and software architecture and offers the consistency, traceability between system and safety architecture and enables to verify the safety concepts in the earlier development phase.

Section 2 shows how EAST-ADL is modified for this purpose.

2. Description of the Approach

The developed approach in Figure 1 shows how the system, software, and safety development are combined and thus the complexity of the system is reduced by a coherent architecture.

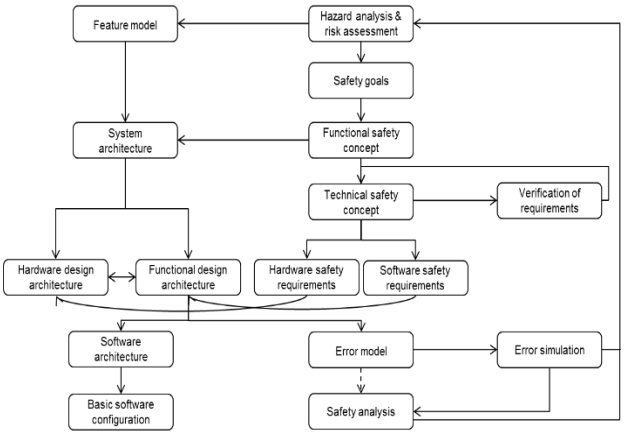


Figure 1 Model driven approach for system, safety and software development

The developed method consists of five main parts, which are shown in Figure 2:

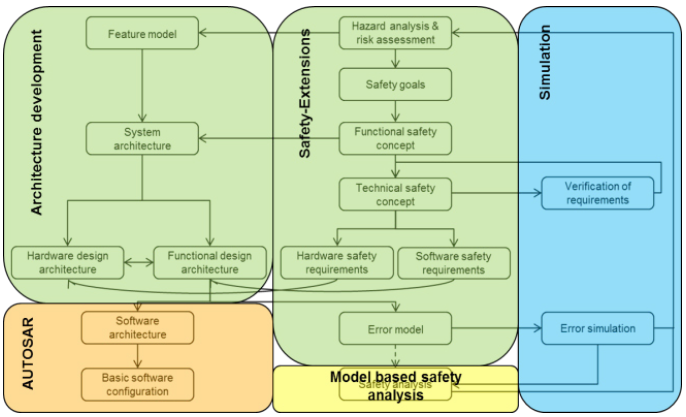


Figure 2 Main parts of the approach

First part is the architecture development [7]. This part is about the creation of feature model, system architecture, functional design architecture (FDA) and hardware design architecture (HDA). It is also possible to allocate the functions from FDA model to the corresponding hardware elements of HDA model. This part is extended within additional safety attributes in order to combine the system architecture and safety architecture and thus it is possible to realize the relationship between system architecture and safety architecture in order to prove traceability and consistency. The further details will be explained by chapter 2.1.

Second part is safety extensions [8, 9]. This part deals with the model-based creation of ISO 26262 work products as an extension to architecture development. Firstly, hazard analysis and risk assessment is performed. Secondly, the safety goals are derived from hazard analysis and risk assessment. In order to fulfill the safety goals, functional safety concept and technical safety concept are created in the following step. This part is also extended with additional safety attributes to realize a relationship between safety model, system model and requirements model. Safety model enables to realize the safety work products model based such as Hazard Analysis and Risk Assessment (HARA), functional and technical safety concept. The extensions enable now additionally to generate these safety work products automatically from the models using the developed scripts. The further details will be explained by chapter 2.2.

Feature model can contain both non-safety-critical and safety-critical properties of the system. After hazard analysis and risk assessment, the safety-critical aspects of the system will be considered as being part of the feature model.

In the system architecture, the safety goals and corresponding safety functions of the system are taken into account. In the functional and hardware design architecture, the safety-critical functions are detailed.

Third part is AUTOSAR [10]. This is about the software architecture and basic software configuration. The software architecture can be created from FDA model, which contains the necessary software features.

Fourth part is model-based safety analysis. In this step, the fault trees (FTA) of fault models are automatically generated from error model of ADL or simulation model with external tools [11] [12].

The last part is simulation and verification. This part is developed within this research. In this step, error simulation and verification of the requirements are carried out and this becomes possible already in the earlier phases of development. On the one hand, the safety requirements are verified with usage of a simulation environment. On the other hand, it is possible to determine the system impacts of causes with error simulation. Thus, the defined top events of the system from hazard analysis and risk assessment can be approved. The further details will be explained by chapter 2.4.

The developed approach allows for consistency and traceability of individual steps. The method also permits the efficient tracing from the software architecture to the feature model and from the safety analysis to the hazard analysis and risk assessment.

In the following subsections the details of the main parts will be described.

2.1. Architecture development and AUTOSAR

The architecture development can be realized with UML, SysML or EAST-ADL. For our tasks, the architecture description language EAST-ADL (Electronics Architecture and Software Technology - Architecture Description Language) is investigated.

The description language EAST-ADL represents an ADL, which was specified initially in research project EAST-EEA (Electronic Architecture and Software Technology – Embedded Electronic Architecture) and developed further to EAST-ADL2 in other research projects ATESST1 (Advancing Traffic Efficiency and Safety through Software Technology), ATESST2 [7] and MAENAD (Model-based Analysis & Engineering of Novel Architectures for Dependable Electric Vehicles) [8], SAFE (Safe Automotive Software Architecture) [9] and Synligare [21]. In addition the language was adapted to AUTOSAR [10], such that AUTOSAR can be used for the detailed description and the implementation of the software architecture. It is used to describe electronic systems in vehicles to facilitate the development of vehicle electronics [5]. Thus EAST-ADL is a domain-specific language.

The purpose of EAST-ADL is to provide the engineers facilitation for the representation and description of electronic systems in vehicles in a standardized form [5]. It can be used during different activities for the modeling of functional requirements, safety work products, as well as for analysis and design purposes [13].

The metamodel of EAST-ADL is organized in four different abstraction levels. Each of them fulfills specific roles. Each level considers a "different phase of vehicle development" [14], represents the complete system and provides different perspectives of the whole EEA [5].

The 4 levels “vehicle level, analysis level, design level and implementation level” are described in detail in the following Fig. 3:

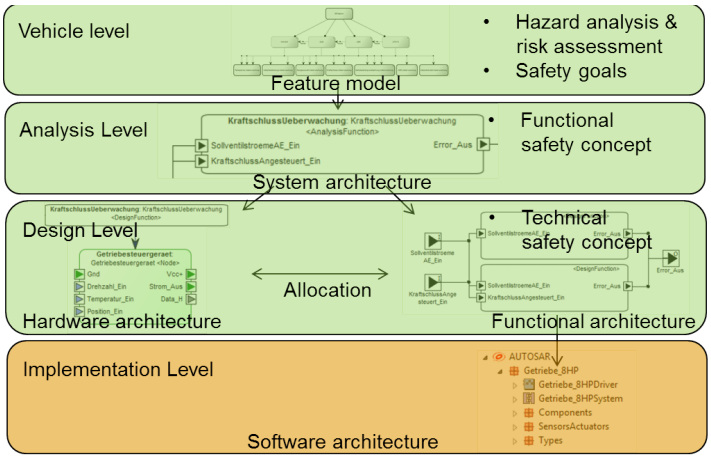


Figure 3 The EAST-ADL levels and safety activities of the approach

As mentioned before, the library elements of these abstraction levels were extended further as follow in Fig. 4. The library elements of EAST-ADL abstraction layer contain the additional information about corresponding safety goal, safety requirements and ASIL classification which enable the relationship between safety model and system model. Additionally the attribute “verified” shows whether the function is already verified in the earlier development phase. This information is very useful for safety case generation and also required from ISO 26262-4:2018, Clause 7.4.8.1.

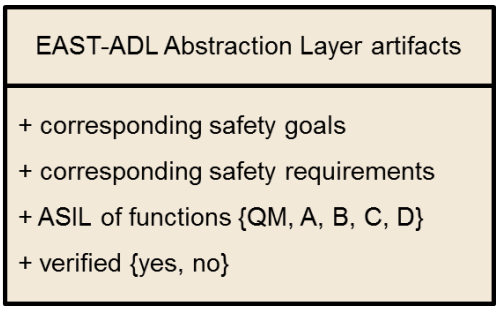


Figure 4 The extensions of EAST-ADL Abstraction layer

At vehicle level, the vehicle or system characteristic is described with a lot of features. In this step, the question 'what' must be realized with the architecture development, but not the question "how". To detect and to eliminate risks as early as possible, it makes sense to perform the hazard analysis and risk assessment in parallel on the same level [15].

The analysis level describes the realization of the features in the functional analysis architecture (FAA) in form of functions. Here we find a top-down decomposition of features on functions and on abstract sensors and actuators [16].

The design level consists of FDA and HDA. The FDA comprises the functions of the FAA, which can be further detailed and decomposed. In addition, an allocation graph is modeled, in which the function blocks of the FDA are allocated to the hardware components of the HDA [18].

The implementation level describes one-to-one the software architecture of the FDA in AUTOSAR [17].

2.2. Safety Extensions

EAST-ADL offers the dependability package with safety extensions, which enable to create the work products of safety development process such as hazard analysis and risk assessment or safety goals, as it is shown in Figure 6.

The dependability model is additionally designed to support the developer in creating safety requirements and performing safety analysis, as well as in the modeling of systematic errors and their propagation, which leads to failures.

As mentioned before, within this research the library elements of dependability model were extended further as follow in Fig 5.

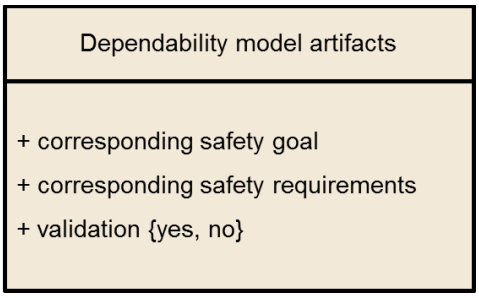


Figure 5 The extensions of EAST-ADL dependability model

The developed extensions within this research enable the generation of safety concepts using the developed automation script. They are created automatically from modified requirements model and dependability model.

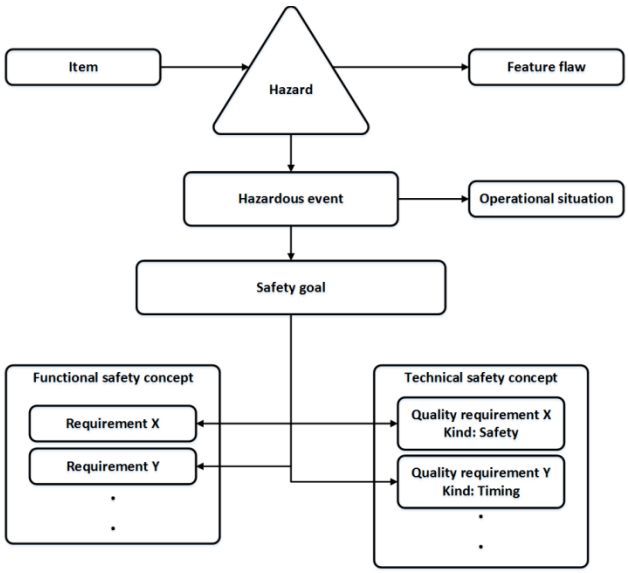


Figure 6 Dependability model

During the development of a system, requirements should be defined, which can be modeled in a separate requirement diagram as shown in Fig 5. Requirements model may include both non-safety-critical and safety-critical requirements. Therefore, it is recommended to begin with the requirements modeling on the system architecture level.

The requirements model in EAST-ADL is based on the SysML (Systems Modeling Language), adapted to the metamodel of EAST-ADL [5]. User-defined features can be added to the requirements, as it is also possible in DOORS (Dynamic Object Oriented Requirements System). Information such as status, author and responsible person are added, in order to achieve better traceability of the changes. Requirements model can be exported in the format ReqIF (Requirements Interchange Format) to facilitate exchanges with other requirements tools. As mentioned before, the requirements model is extended further as in Fig. 7 and 8 with safety related features such as ASIL classification, corresponding safety goal or decomposition relevance.

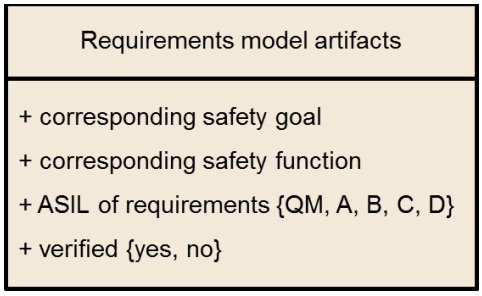


Figure 7 The extensions of EAST-ADL requirements model

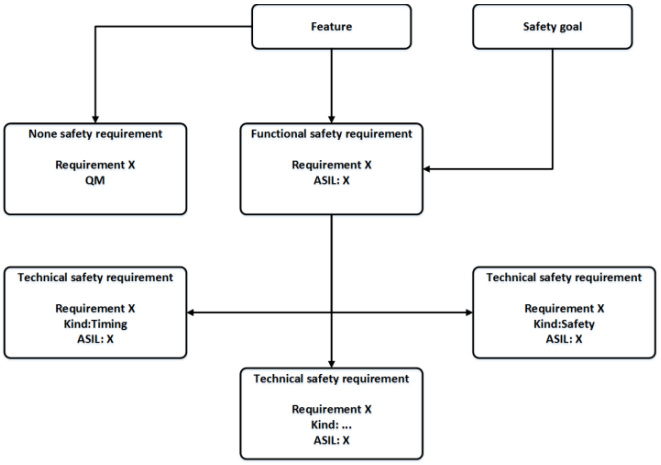


Figure 8 Requirements model

The extensions within this research enable the generation of safety concepts using the developed automating script as follow in Fig. 9.

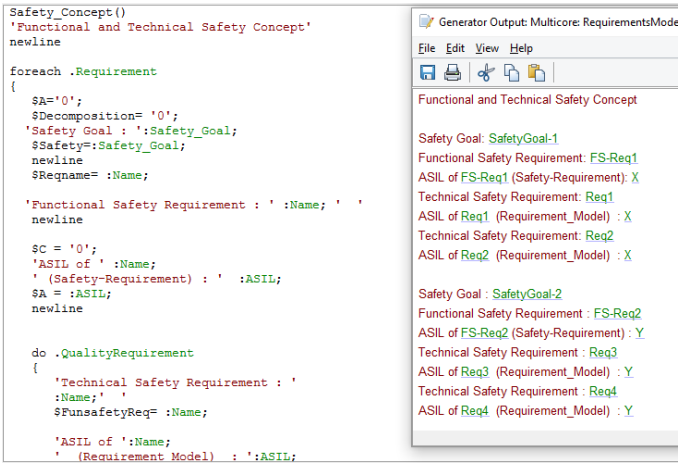


Figure 9 Safety concept script and generated document

It is possible to generate error models automatically from functional design architecture model (FDA). Error description and error logic to the individual subsystems are modeled afterwards.

Because vehicles in general and especially electrical vehicles consist of complex systems, it is very difficult to determine all measures to prevent a hazard. Of course, Original Equipment Manufacturers (OEMs) and their system suppliers are committed to reduce all risks to an acceptable level. This means that for all hazards and risks relevant safety goals, safety concept, safety requirements and safety measures have to be defined. For this reason, ISO 26262 requires to carry out the safety analysis based on fault models in the automotive industry. For example with the help of the fault tree analysis it is possible to find the error causes for a particular hazard [19].

In EAST-ADL, it is possible to model errors and their propagation in an error model to recreate a failure behavior [20]. On the basis of error models, additional tools such as HiP-HOPS [11], ALTARICA [12] can be used to perform model-based safety analysis.

Figure 6 shows how the error model and error logic can be created.

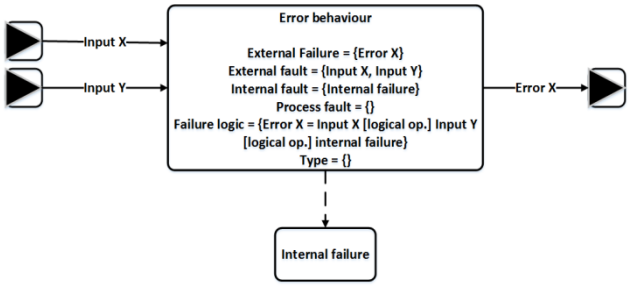


Figure 10 Error Model

2.3. Model based safety analysis

It is possible to use commercial tools such as HiP-HOPS [11] to perform the safety analysis. These tools are capable of generating fault tree analysis (FTA) and failure modes and effects analysis (FMEA) automatically from an error model [11].

Safety analysis enables to find out possible error causes which should be detected and avoided with appropriate safety measures. For this purpose, safety measures and requirements are defined in order to introduce countermeasures. Additionally, functional and technical safety requirements can be defined or extended based on the results of the safety analysis.

2.4. Simulation

For verification of the requirements and error simulation, a simulation environment such as Simulink (Matlab), Dymola, CarMaker, etc. can be used. For instance, the power train of an electric vehicle can be simulated with the help of a simulation environment and error models can be designed and implemented. The behavior of the vehicle can be simulated with error models in a way where the error causes such as failure sensor values can be used as stimulation signal. The simulation is used for the verification of the requirements, for detection of the errors in the earlier phase of development and for detection of the system impacts of errors. If it is found that the safety requirement cannot be implemented as intended, then the requirement should be redefined. This is shown in Figure 11.

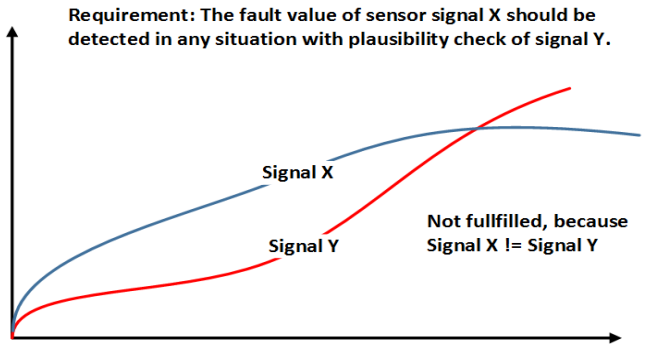


Figure 11 Verification and simulation

Fig. 12 shows the evaluation procedure of the defined hazards. The simulation is performed within the error causes in order to find out the system reaction of these failures. If the vehicle behavior is the same as defined hazards, then HARA is approved, otherwise HARA should be extended further as required from ISO 26262-4:2018, Clause 7.4.8 in order to develop the necessary safety measures to detect and prevent the possible hazards.

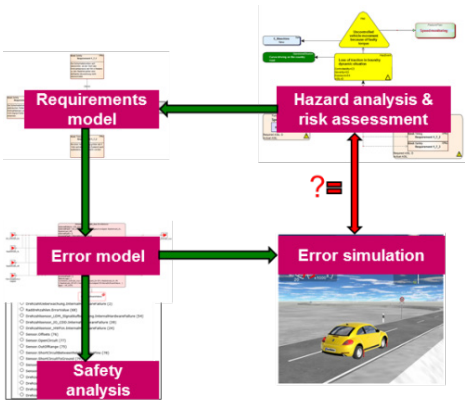


Figure 12 Error simulation

3. Use-Case

In order to show the benefits of the developed model-driven approach, an example is created which shows the safety-critical function development. It starts with hazard analysis and risk assessment and ends with verification of safety goals and safety requirements. Figure 13 shows a typical dependability model of a system. In this concrete case an electrical machine is considered as an item. Hazard is defined as “Uncontrolled vehicle movement because of faulty torque”. ASIL of hazardous event is classified based on the operational situation. In this case the hazardous event “Loss of traction in boundary dynamic situation” leads by driving a curve to a dangerous situation. Therefor the ASIL of this hazardous event is classified as ASIL D (C3, S3, E4). At the next step safety goal “No faulty torque in boundary dynamic situation” is defined with the safe state “Electric machine is torque-free, vehicle is rolling freely” in order to detect and avoid the failure. After this functional safety concept and technical safety concept are created in order to achieve the safety goal. Functional and technical safety requirements are linked from the requirements model and shouldn`t be specified again in the dependability model.

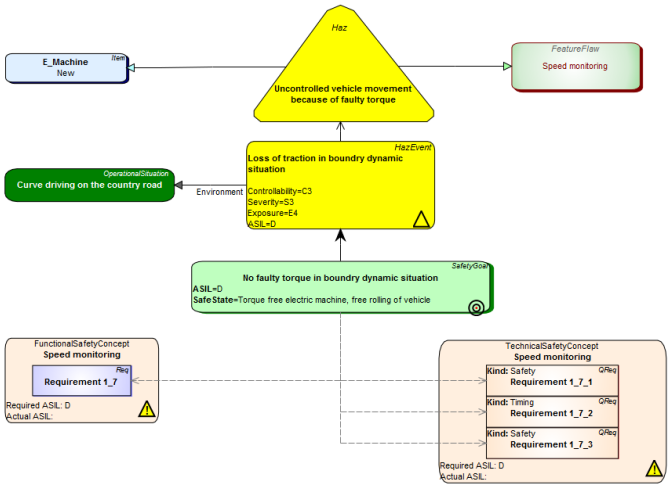


Figure 13 Dependability model – power train example

Functional and technical safety requirements are specified with associated information within the requirements model. Their allocation to architectural elements is realized just by using the satisfy connection. In this case the safety goal is realized with the function “Speed monitoring” (FAA) as shown in Figure 14.

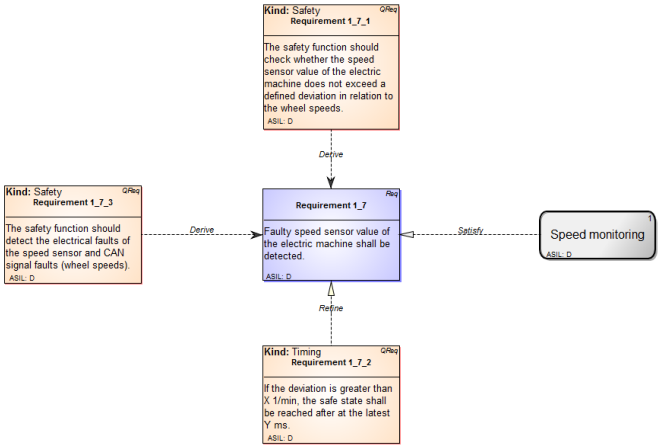


Figure 14 Requirements model – power train example

Functional and technical safety concept can be created automatically from the dependability model and requirements model using the developed script as shown in Figure 15.

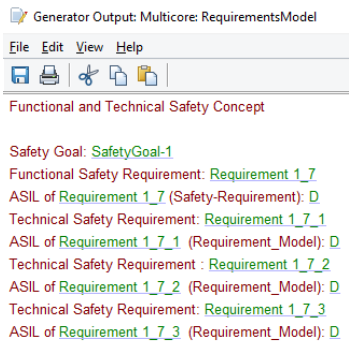


Figure 15 Functional and Technical Safety Concept – power train example

Figure 16 shows the FDA model. The safety requirements are implemented within the subfunctions of the FDA model which are extended within safety attributes to enable consistency and traceability check.

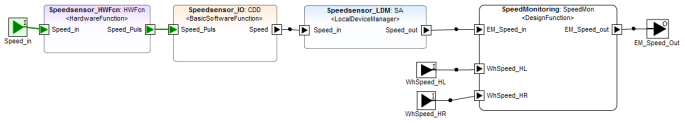


Figure 16 Functional design model – power train example

Error model of the hazard is generated automatically from FDA-Model. Afterwards the error logic of subsystems will be described individually with possible fault causes, considering the error propagation. Error logic contains all possible output failures of every submodule, with regard to internal hardware failures, internal software failures and input signal failures, as it is shown in Figure 17.

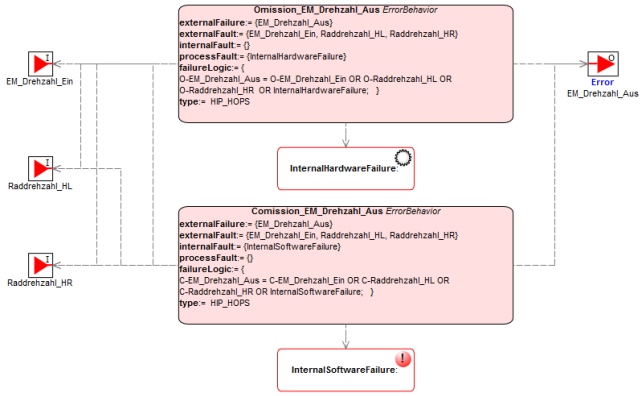


Figure 17 Error model logic – power train example

The additional tool HiP-HOPS enables to perform model based safety analysis automatically from the existing error models. This tool is capable to generate cut sets, fault tree analysis (FTA) and failure modes and effects analysis (FMEA). Figure 18 shows the generated FTA for the top event “faulty torque”. The causes such as sensor failures, internal software failures and internal hardware failures, which lead to the top event, are listed in the FTA. In this case a qualitative safety analysis is realized. The tool also enables to perform a quantitative analysis.

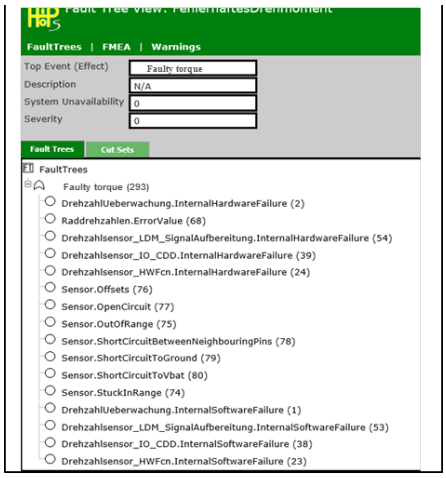


Figure 18 Model based safety analysis with HiP-HOPS – power train example

Finally, the error simulation is performed, in order to verify the system effects (top events) which were defined before within the HARA. For this reason the simulation is stimulated within the error causes (basic events) regarding fault tree analysis. The basic events should lead to the top events. After simulation the results are compared with the HARA results. If they are same, then the defined hazardous events are correct and thus verified. But if the vehicle behavior is other than defined events, the HARA should be extended. If necessary, a new safety goal should then be defined, in order to avoid new detected system impacts. In this example a concrete hazardous event (faulty torque due to E-Machine speed sensor failures) leads to loss of traction, which causes the vehicle to enter an instable state. In the boundary dynamic situations, e. g. when driving a curve, the driver cannot control the vehicle anymore. This may lead to a fatal accident. Figure 19 shows both the verification results of technical safety requirements and validation of safety goals regarding the error simulation.

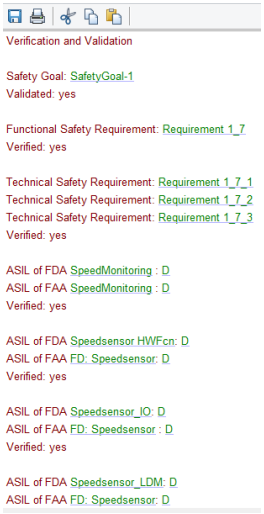


Figure 19 Verification and Validation – power train example

4. Conclusion

The presented approach supports developers in creating a model based system, software and safety architecture of E/E systems. Various tools are used currently for different special development fields. The architecture description language EAST-ADL, the main component of this approach, replaces these tools in a way that the entire development of system and safety architecture can be realized within EAST-ADL. The proposed approach has the advantage to verify the functional safety concept and technical safety concept and also to validate the hazard analysis and risk assessment within the developed simulation environment in the earlier development phase of project.

The biggest advantage of this method is to prove the traceability of the different development steps and safety work products with the help of extended library elements and developed scripts. Once the user became familiar with the method, it is easy for him to understand the overall architecture. Figure 20 highlights the relationship between the various levels and shows the traceability of different steps.

The approach enables to understand the system, software and safety architecture of an item very quickly. When applying this approach, requirement model, dependability model and library elements of EAST-ADL is extended within safety features (such as ASIL classification of requirements) in order to model the safety-critical requirements with necessary attributes and in order to create functional and technical safety concepts automatically using developed script and in order to combine system and safety architecture. The user can thus find out quickly which requirements are implemented for which functions.

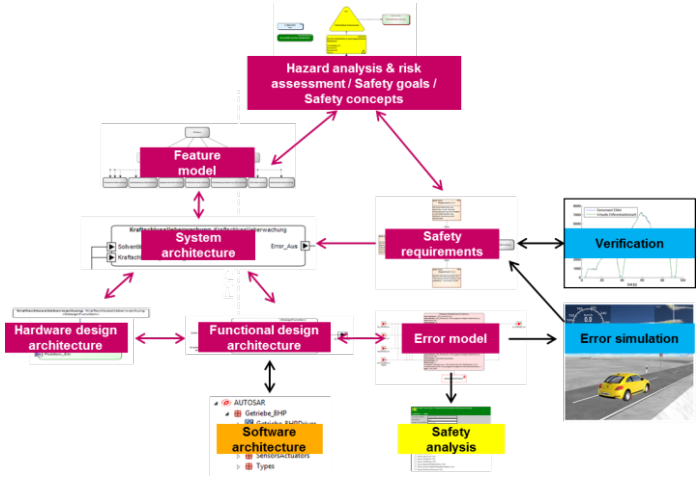


Figure 20 Traceability of the development steps

Another major advantage of the approach is the mapping of the EAST-ADL levels to the work products and requirements of ISO 26262. Already, at the initial phase the relationship between EAST-ADL and ISO 26262 is very significant. With dependability package of EAST-ADL, safety aspects of ISO 26262 can be early considered in the architecture development workflow. Thus, the hazard analysis and risk assessment is performed to find out the safety goals.

Dependability models allow the modeling of the necessary requirements for achieving the defined safety goals. Subsequently, the functional and technical safety concept can be created. So the user gets an overview, which safety concept is formulated for which safety goal. EAST-ADL offers the possibility to generate the error models automatically from FDA-models. With help of the approach and external analysis tools, safety analysis can be performed automatically from error models.

Systematic errors can be detected by the verification of the requirements and error simulation and it is also possible that safety measures are specified early in the development phase.

Finally, advantages of the presented approach can be summarized as follow:

- Modeling of safety-related functions in an architecture description language.
- Achievement of an efficient and consistent model-based development of automotive embedded system.
- Model based automated creation of ISO 26262 work products from hazard analysis and risk assessment to safety requirements using developed scripts.
- Combination of system and safety development to achieve the traceability and to show the relationship between safety goals and safety functions considering ASIL.

• Verification and validation of functional safety concept and technical safety concept in the earlier development phase of project.

• Early detection of systematic errors and system impact of errors.

References

[1] B. Sari, H.C. Reuss, "A model-driven approach for the development of safety-critical functions using modified Architecture Description Language (ADL), *Electrical Systems for Aircraft, Railway, Ship propulsion and Road Vehicles & International Transportation Electrification Conference (ESARS-ITEC 2016)*.

[2] R. N. Charette,: This Car Runs on Code. Hyperlink: http://www.real-programmer.com/interesting_things/IEEE%20SpectrumThisCarRunsOnCode.pdf, Februar 2009. Accessed January 30, 2017.

[3] PWC DEUTSCHLAND: Autoindustrie treibt Chipnachfrage an. Hyperlink: http://www.pwc.de/de/automobilindustrie/autoindustrie-treibt-chipnachfrage-an.html. Accessed January 30, 2017.

[4] AK-L_Orientation-list-V1.2_2010-11-25_DE (AA-I3/AK 16 - Functional Safety).

[5] H. Blom, H. Lönn, F. Hagl, Y. Papadopoulos et al.: EAST-ADL – An Architecture Description Language for Automotive Software-Intensive Systems – White Paper Version 2.1.12. Hyperlink: http://www.maenad.eu/public/conceptpresentations/EAST-ADL_WhitePaper_M2.1.12.pdf. Accessed January 30, 2017.

[6] P. Cuenot, D. Chen, S. Gerard, H. Lönn et al.: Managing Complexity of Automotive Electronics Using the EAST-ADL. In: Engineering Complex Computer Systems, 2007. 12th IEEE International Conference on, 2007.

[7] ATESST2: ATESST2. Hyperlink: http://www.atesst.org. Accessed January 30, 2017.

[8] MAENAD: MAENAD. Hyperlink: http://www.maenad.eu/. Accessed January 30, 2017.

[9] SAFE: SAFE. Hyperlink: http://www.safe-project.eu/. Accessed January 30, 2017.

[10] AUTOSAR: AUTOSAR. Hyperlink: http://www.autosar.org/. Accessed January 30, 2017.

[11] HIP-HOPS: HiP-HOPS, Automated Fault Tree, FMEA and Optimisation Tool. Hyperlink: http://hip-hops.eu/. Accessed January 30, 2017.

[12] ALTARICA: ALTARICA. Hyperlink: http://openaltarica.fr/. Accessed January 30, 2017.

[13] ATESST2: EAST-ADL Overview. Hyperlink: http://www.atesst.org/home/liblocal/docs/ConceptPresentations/01_EAST-ADL_OverviewandStructure.pdf, 2010. Accessed January 30, 2017.

[14] M. Hillenbrand, „Funktionale Sicherheit nach ISO 26262 in der Konzeptphase der Entwicklung von Elektrik / Elektronik Architekturen von Fahrzeugen“, Ph. D. Thesis, KIT Karlsruhe, 2011.

[15] ATESST2: EAST-ADL2 Vehicle Level. Hyperlink: http://www.atesst.org/home/liblocal/docs/ConceptPresentations/02_EAST-ADL_Vehicle_Level.pdf, 2010. Accessed January 30, 2017.

[16] ATESST2: EAST-ADL Analysis Level. Hyperlink: http://www.atesst.org/home/liblocal/docs/ConceptPresentations/03_EAST-ADL_Analysis_Level.pdf, 2010. Accessed January 30, 2017.

[17] ATESST2: EAST-ADL Overview Implementation Level. Hyperlink: http://www.atesst.org/home/liblocal/docs/ConceptPresentations/05_EAST-ADL_Implementation_Level.pdf, 2010. Accessed January 30, 2017.

[18] ATESST2: EAST-ADL Overview Design Level. Hyperlink: http://www.atesst.org/home/liblocal/docs/ConceptPresentations/04_EAST-ADL_Design_Level.pdf, 2009. Accessed January 30, 2017.

[19] ISO 26262 “Functional safety of road vehicles”, 2011.

[20] EAST-ADL ASSOCIATION: EAST-ADL – Domain Model Specification – Version V2.1.12. Hyperlink: http://east-adl.info/Specification/V2.1.12/EAST-ADL-Specification_V2.1.12.pdf, 2013. Accessed January 30, 2017.

[21] Synligare: Synligare (literally “more visible” in Swedish). Hyperlink: http://synligare.eu/HomePage.html. Accessed January 30, 2017.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1260-1265 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

On-Chip Testing Schemes of Through-Silicon-Vias (TSVs) in 3D Stacked ICs

Shadi MS. Harb[*1], William R. Eisenstadt[2]

[1]*Intel Corporation, Hillsboro, OR 97124, USA*

[2]*University of Florida, Department of Electrical & Computer Engineering, Gainesville, FL 32611, USA*

A R T I C L E I N F O

Article history:
Received: 02 June, 2017
Accepted: 09 July, 2017
Online: 23 July, 2017

Keywords:
Interconnect
Crosstalk
3D Technology
TSVs
Test
Stacked

A B S T R A C T

This paper presents on-chip testing structures to characterize and detect faulty Through Silicon Vias (TSVs) in 3D ICs technology. 3D Gunning Transceiver Logic (GTL) I/O testing is proposed to characterize the performance of 3D TSVs in high speed applications. The GTL testing circuit will fire different data patterns at different frequencies to characterize the transient performance of TSVs. In addition, Different testing schemes based on an oscillation ring testing methodology are proposed to detect TSVs faults such as stuck-at, open, slope and delay degradation, and severe crosstalk TSVs coupling. A parallel ring-based oscillator test structure is proposed and simulated based on a high performance fully tunable electrical circuit pi-model where a single and coupled TSVs with ground-signal-ground (GSG) and ground-signal-signal-ground (GSSG) 3D vias configurations are used as a test vehicle for 3D interconnect characterization and test. Simulation results are presented using the Keysight/Agilent Advance Design System (ADS) and a standard 0.25 µm CMOS process.

1. Introduction

With deep sub-micron silicon processing technology, it becomes more difficult for integrated circuits to achieve higher integration through the scaling down of the transistor feature size. 3D ICs technology stacks multiple dies together and connects them using Through-Silicon Vias (TSVs), which become a promising technology to provide a low cost and high efficient way to increase integration for modern systems [1-2]. However, there are some problems associated with this technology. One of the most important issues is the compound yield loss due to die stacking and TSV failures, which might occur either in the fabrication of the TSV themselves, in the bonding of the TSVs to the next layer, or during the life time of the 3D stack. Thus, to guarantee the stacking yield, the interconnections (TSVs) reliability must be tested. Common TSV defects in this emerging technology can be caused by TSV shorts, opens, delay faults [3], crosstalk induced-delay faults and glitches [4-7].

To insure a low probability of 3D IC failure, integrating embedded test structures in 3D stacked ICs becomes a more attractive solution to avoid traditional probe testing, which becomes an infeasible approach to validate inaccessible nodes after stacking. The embedded test circuits can be integrated during the fabrication process to test a group of TSVs or each TSV individually for different possible defects in the die before stacking. In addition, they can be used to characterize the TSV's signal integrity effects and the impact of TSVs on the 3D circuit performance after fabrication. The results of these test circuits can then be read out off-chip for post processing. In this paper, A 3D Gunning Transceiver Logic (GTL) I/O testing is proposed to characterize signal integrity of TSVs. Additionally, an oscillation ring (OR) 3D test methodology with different circuits testing schemes, based on our work [8], are presented to detect TSVs failure induced by TSVs shorts, opens, delay faults and crosstalk coupling.

The paper is organized as follows: a high performance TSV electrical pi-model, which can be used as DUT for 3D TSV testing, is discussed in Section 2. Section 3 discusses the 3D GTL I/O testing to evaluate TSV performance. Section 4 proposes the ring-based oscillation testing schemes to detect TSVs failures and their induced 3D performance degradation. Section 5 concludes the paper.

[*]Corresponding Author: Shadi MS. Harb, Intel Corporation, Hillsboro, OR 97124, USA | Email: shadiharb@ieee.org

2. An Equivalent Electrical Circuit For TSVs

To test and study the effects of TSVs on 3D circuit performance, a high frequency electrical pi-model, commonly used in literature [9-11], is applied for TSV test and characterization, which is extracted from physical configurations of a common TSV structure. Ground-signal-ground (GSG) and ground-signal-signal-ground (GSSG) vias configurations are used to test single and coupled TSVs for different 3D routing topologies. By leveraging these fully parametric TSV models, different TSV designs can be explored to study the impact of TSVs on the 3D performance. In addition, changing the model parameters gives more insight about how these parameters such as induced resistance and capacitance of TSVs affect the 3D circuit. Figure 1(a) shows a schematic drawing of the high frequency equivalent circuit model of a GSG via configuration [12].

The electrical model of the TSV is expressed as an inductor (Lvia) and a resistor (Rvia), while the capacitive coupling between the vias is modeled as coupling capacitors (Cvia_ox, Cox, and Csil). The parameter Cvia_ox denotes the capacitance of the thin oxide layer surrounding the via barrel, and Cox denotes the capacitance of the oxide layer on the silicon surface and the fringing field between the vias. The capacitance of the silicon substrate is denoted by Csil, and the loss property of the silicon substrate between the signal via and the ground via is denoted by Gsil. The model parameters (Rvia=0.22 Ω, Lvia=3 pH, Cvia_ox=1 PF, Cox=5 fF, Csil=4 fF, Gsil=2.92m/Ω) are chosen from [4] as default values for the proposed testing applications. It is worth mentioning that the Rvia and Cvia_ox seem to be the main parameters to consider in order to model the impact of TSVs on 3D circuits. Figure 1(b) shows also the pi electrical model for a GSSG structure. It is assumed that the two outer TSVs are grounded and the other two inner ones are signal TSVs.

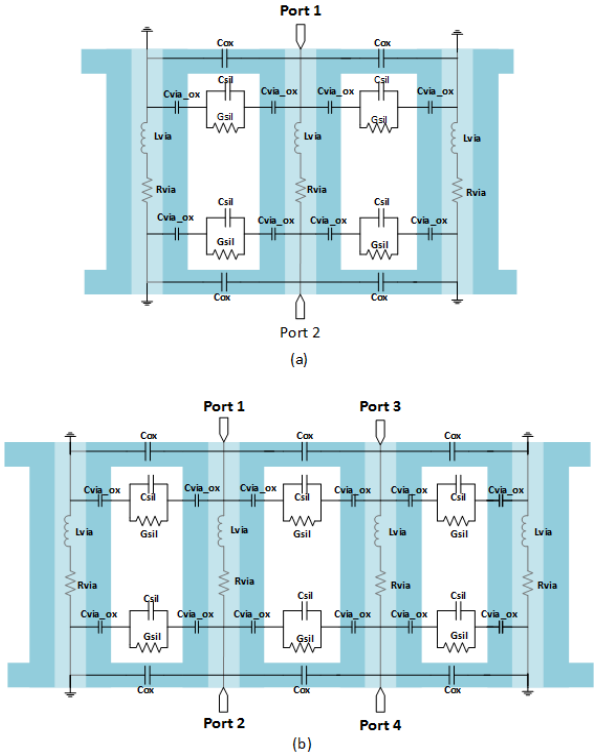


Figure 1: Equivalent Circuit Model of (a) a Standard GSG TSV and (b) a standard GSSG TSV [12]

3. 3D GTL I/O Testing

3.1. Single-Ended Gunning Transceiver Logic (GTL) I/O Design

Single-ended GTL (Gunning Transceiver logic) I/O standard is used in a wide range of applications from ASICs and processors to interface logic devices. A GTL I/O test structure [13] consists of a GTL clock and data receiver, data depths control logic, a serial-to-parallel converter, and GTL I/O drivers as shown in Figure 2. The GTL clock and data can be provided from off-chip using an external test equipment. The data depth control logic controls the depth of the fired data where the control signals come from an external stimulus system. The serial-to-parallel converter loads the serial data from off-chip and fire the parallel data out to the I/O ports. The basic cell consists of a flip-flop and a multiplixer that selects either the serial or parallel data. When the shift register goes to the parallel mode, the register starts firing the parallel data indefinitely with period specified by number of GTL cells. GTL I/O driver is basically an open drain circuit. The circuit has NMOS shunted to ground, and the other end is pulled up to the termination voltage with a 50Ω off-chip resistor.

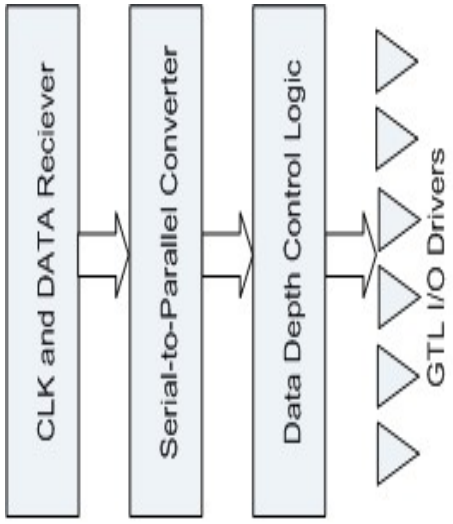


Figure 2: A GTL I/O test structure

3.2. GTL Circuit Architecture in 3D Stacked ICs

A Gunning Transceiver Logic (GTL) I/O test circuit can be utilized to characterize the performance of the 3D TSVs under high speed transient simulation in 3D stacked ICs. The 3D GTL I/O test circuit will fire different data patterns at different frequencies having different data depths across multiple tiers to characterize the impact of TSVs by observing the transient performance through different 3D routes. Figure 3 shows a block diagram of the proposed test structure, where the GTL serial-to-parallel shift registers (SIPOs) are placed in different tiers. A control MUX with Tri-state buffers and control logic can be used to switch between different I/O signals. Figure 4 shows the 3D GTL I/O circuit implementation, where each tier has a 16-bit serial IN-2-bit parallel out GTL test structure. Each I/O driver fires signals out through TSVs from its corresponding tier.

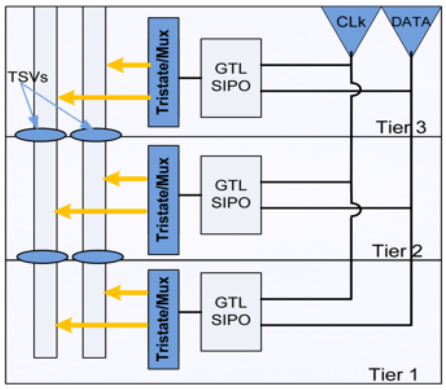


Figure 3: The GTL I/O Test Structure in 3D Stacked IC

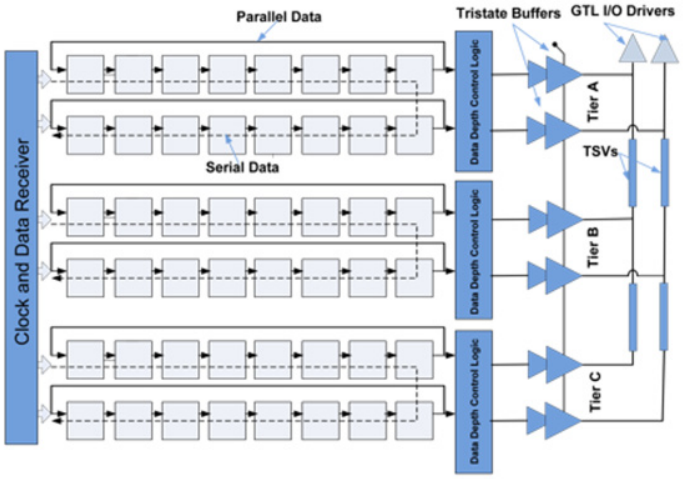


Figure 4: 3D GTL I/O Circuit Implementation

The 3D GTL I/O test structure can be also dedicated to measure the NEXT/FEXT crosstalk between vertical connections by firing high speed signals from different tiers as shown in Figure 5. FEXT and NEXT crosstalk tests can be realized by activating only the I/O signal in tier A and tier C respectively. For example; FEXT crosstalk can be observed by firing high speed signal only at the bottom layer and monitoring the coupling noise at the far-end of the victim line. Similarly, NEXT crosstalk can be observed by firing high speed signal only at the top layer and monitoring the coupling noise at the near-end of the victim line.

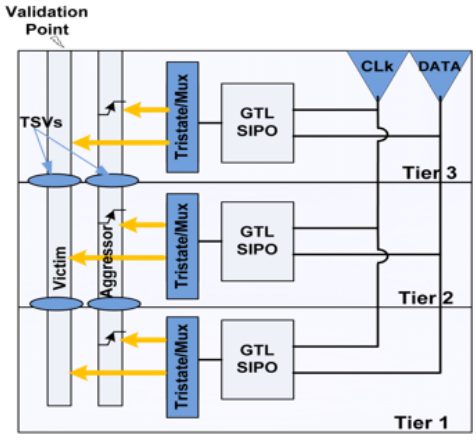


Figure 5: Measuring Crosstalk between TSVs in 3D Stacked ICs

A 2 port GTL I/O test for three tiers system was designed and simulated using the Advanced Design System (ADS) and 0.13 µm UMC CMOS process. Figure 6 shows the simulated results of the GTL test circuit at 1Gb/s data rate.

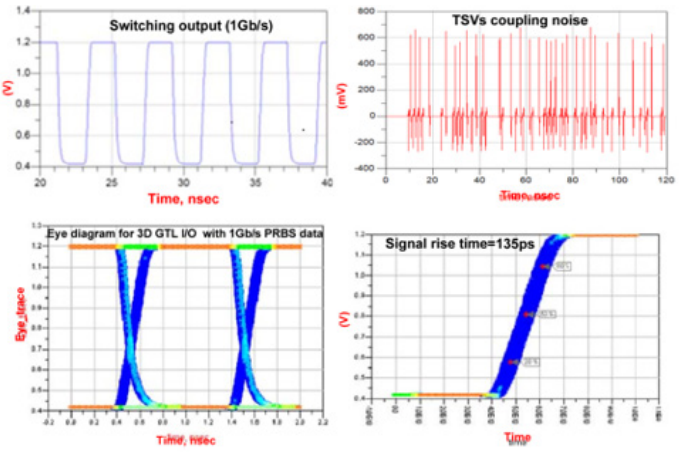


Figure 6: simulated 3D GTL Test Results at 1Gb/s data rate.

3.3. TSV Test Using the GTL I/O Shift Register as Scan Chain

A testing scheme for TSVs in 3D ICs can be realized by using the Gunning Transceiver Logic (GTL) shift register circuit. The GTL serial-to-parallel (SIPO) shift register can be designed where the TSVs are connected serially to form a scan chain structure [14]. In this case, each TSV is connected to a multiplexer and a flip-flop which can be used as scan chain as shown in Figure 7. When the GTL registers are in the data load mode, the registers start loading data serially from the top tier and shifting them down and up between tiers. After loading the data, data starts circulating up and down and vice versa. Circulating data with different patterns can test the reliability of TSVs.

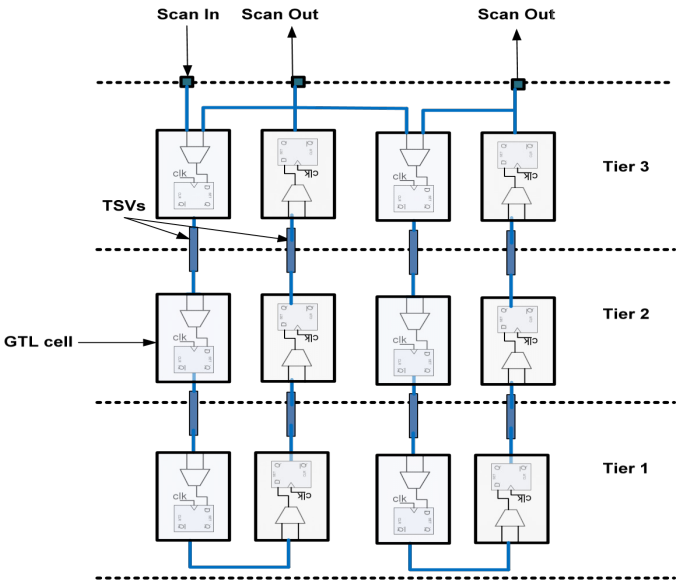


Figure 7: The GTL serial-to-parallel shifts register for 3D TSV test

The GTL scan chain test structure was simulated using ADS and 0.25 µm standard CMOS process. The TSVs connect the GTL cells serially between tiers. The TSVs are tested by injecting test vectors (i.e. data patterns) in one tier (source tier). The test vectors

are then propagated to the destination tier (e.g. the top tier), where they are captured and transmitted off-chip. Figure 8 shows a test vector with data stream ("101010.."), which is injected from the source tier and captured at the destination tier (i.e. top tier). In case of faulty TSVs, the captured data has bit errors as shown in Figure 8(b).

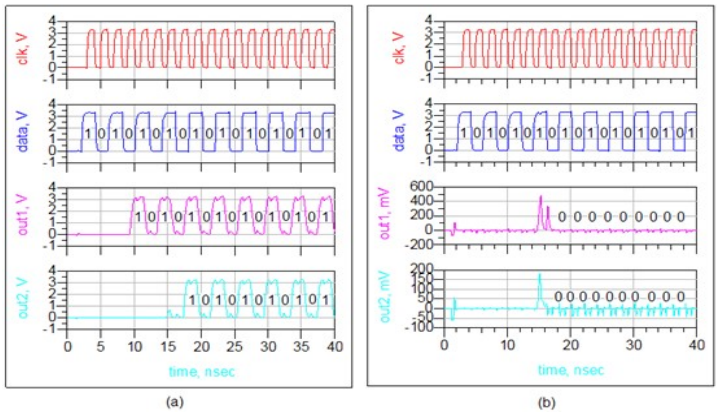


Figure 8: Outputs of the scan chain using the GTL shift register (a) with fault-free TSVs (b) with faulty TSVs.

4. Oscillation Ring-based 3D Testing Schemes

Oscillation ring (OR) test is a useful and efficient method to detect faults in functional circuits and devices [15-16]. An oscillation ring is a closed loop with an odd number of signal inversions. Intuitively, if an oscillation ring fails to oscillate, it implies that there exists stuck-at or open faults in the components of the oscillation ring. The period of the oscillation signal can also be measured to test the delay faults. For example; in a circuit with gate or path delay faults, the oscillation frequency is different from the fault–free circuit. By observing the oscillation signal at the circuit output, one can decide whether the circuit-under-test is faulty or not. To evaluate the impact of a TSV on 3D circuit performance due to the additional delay induced by a TSV, an Oscillation Ring (OR) 3D Built-in Self-Test (BIST) testing schemes, shown in Figure 9, are designed and simulated using the Advanced Design System (ADS) and 0.25 µm standard CMOS process. The target fault models of this testing architecture are TSV stuck-at, open, delay and slope degradation, and TSVs strong crosstalk coupling. If a triggered oscillator fails to oscillate, there exist stuck-at or open faults in the components of the RO. In addition, measuring the TSV delay fault can also be achieved. In this proposed testing scheme, a counter is included at the output of each triggered oscillator to detect TSVs delay faults.

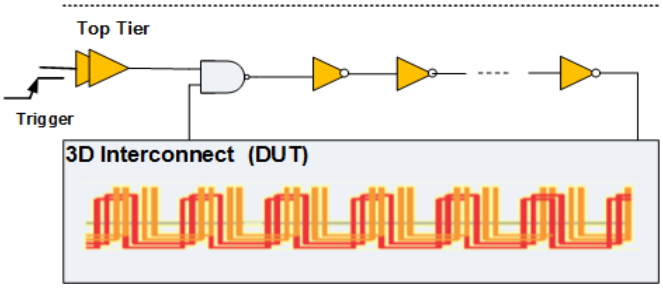


Figure 9: OR Testing Methodology for 3D Stacked ICs

4.1. TSVs Test with Different Configurations

In this testing scheme, a parallel ring oscillators (ROs) testing structure is presented, each RO has different routing TSVs configuration across different stacking tiers as shown in Figure 10. An edge detector circuitry is used to trigger ROs in parallel when the trigger enable signal (OR_TEST) is asserted and a rising edge of the reference clock is detected. This will imply that each one will have a different oscillating frequency and thus a different counting value, which can be scanned out for off-chip testing. Since the frequency of each triggered oscillator is predetermined during the design phase, a delay fault can thus be detected and measured by inspecting the contents of the counters. In this test scheme, let assume that we have m triggered oscillators for different TSV configurations and the oscillation ring test is enabled by the OR_TEST signal. Also let assume that the triggered oscillator unloaded by TSVs is allowed to count only up to n at frequency of f0 and the counter contents of the other triggered oscillators, be n1, n2, …, nm, respectively. An estimation of the ith ring's oscillation frequency fi can be approximated by:

$$f_i = f_0 * \frac{ni}{n} \quad (1)$$

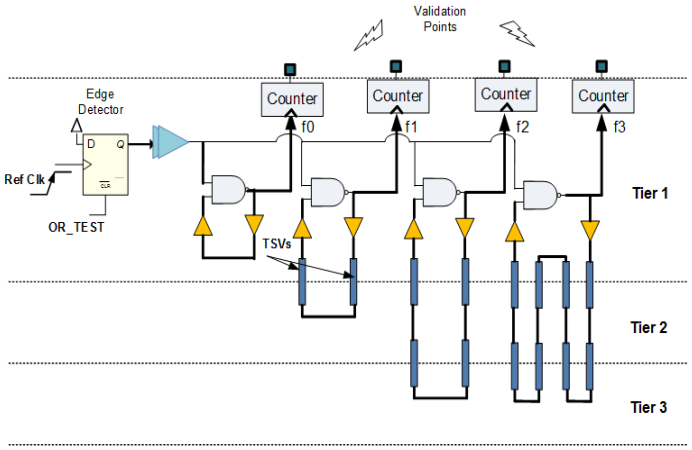


Figure 10: 3D Interconnect Test based on a Parallel triggered ROs Structure

Figure 11 shows simulation results for the triggered oscillators connected with different 3D configurations based on the pi-model which is used to electrically characterize the TSV. Each triggered oscillator has different output frequencies where the triggered oscillator with unloaded TSV has the highest frequency (1.37 GHz) compared to the ones loaded by TSVs. For example; the triggered oscillators loaded by 2 TSVs, 4 TSVs, and 8 TSVs have the following frequencies: 662 MHz, 543 MHz, and 436 MHz respectively. Each output frequency will have a different counter value. Figure 12 shows the counting values 'b111, 'b101, 'b010 for the 2 TSVs, 4 TSVs and 8 TSVs respectively. Obviously, the addition of a TSV is similar to the addition of a short interconnect, thus leading to a degradation of rise time and more induced delay in the ring oscillator and thus slower oscillating frequency.

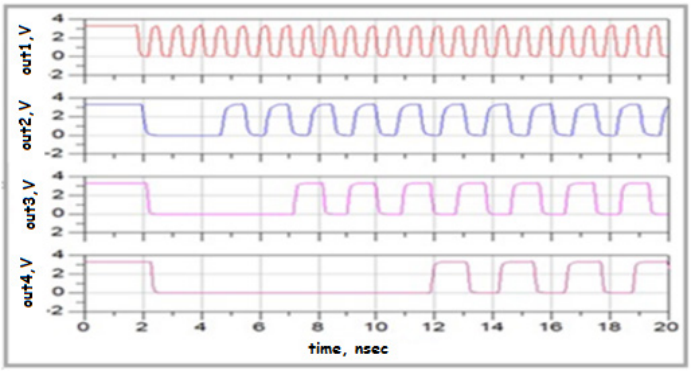


Figure 11: Output frequencies for the triggered oscillators

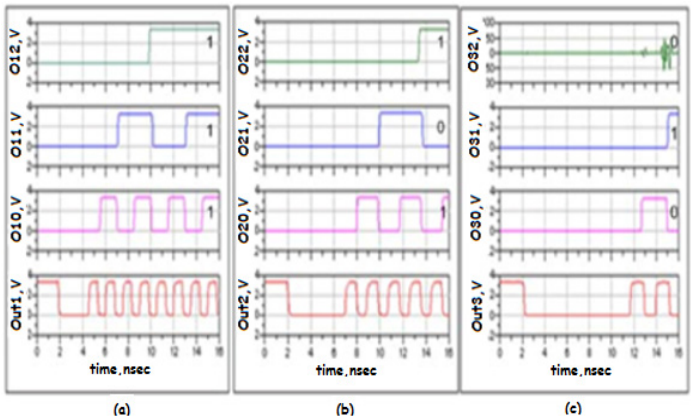


Figure 12: Output Frequencies and Counter Values for Different 3D Configuration (a) TSVs, (b) 4 TSVs, (c) 8 TSVs

4.2. TSVs Rise Time Degradation Test

To characterize the effect of the rise time degradation using the OR test, the capacitance of the thin oxide layer is changed from 10fF to 4pF, which corresponds to 45 ps, and 1.3 ns rise time respectively as shown in Figure 13. This will presents the importance of RC delay caused by the time needed to load TSV oxide capacitance. Figure 14 shows the output frequency and the counter values for triggered oscillators loaded by 2 TSVs for both cases (45 ps, 1.3 ns). The 1.15 Ghz output frequency with counter value of 'b111 corresponds to the 45ps TSV rise time, and the 833 MHz output frequency with counter value 'b010 corresponds to the 1.3 ns TSV rise time.

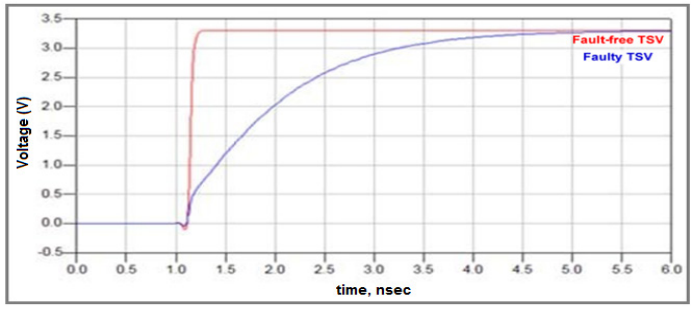


Figure 13: TSV Rise Time (red: 45ps, blue: 1.3ns)

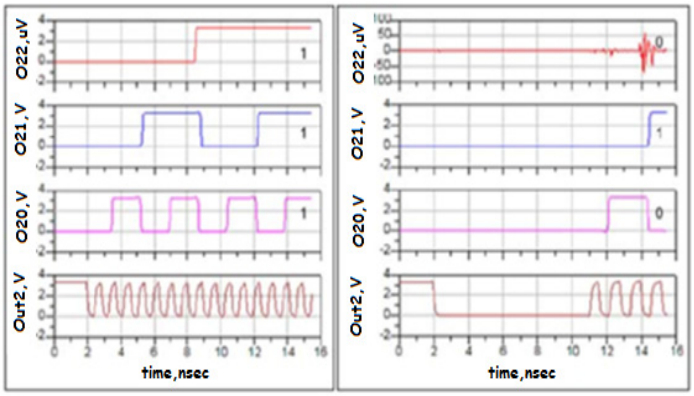


Figure 14: Output Frequencies and Counter Values for (a) 45ps TSV rise time, (b) 1.3ns TSV rise time

4.3. TSVs Crosstalk test

In this test, a consecutive triggered parallel ROs structure running same frequency is used to characterize the crosstalk effect between TSVs. Figure 15 shows four triggered oscillators; two oscillators are crosstalk-coupled and the other two are crosstalk-free. The proposed ROs parallel structure creates a delta phase shift difference between each consecutive triggered oscillators, which is equal to the time difference between the delay buffer chain and the oscillation time period of the triggered oscillators. 3D crosstalk detection can be achieved by observing the frequency of crosstalk-coupled oscillators, which is different from the frequency of the crosstalk-free oscillators.

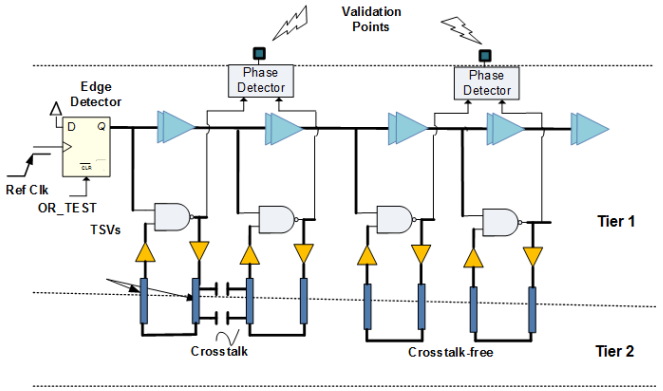


Figure 15: 3D Crosstalk Detection Scheme

Figure 16 shows the output frequency for both crosstalk-coupled (725 MHz) and crosstalk-free oscillators (667MHz).

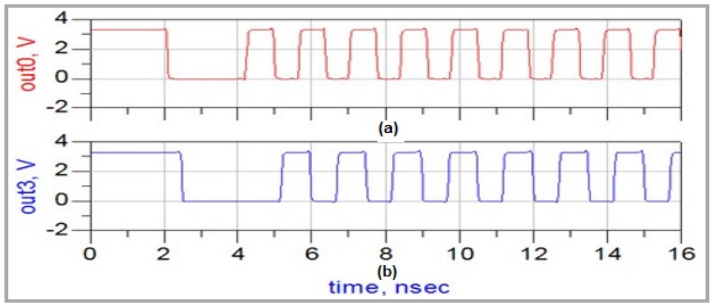


Figure 16: Output frequency for (a) Crosstalk-coupled and (b) Crosstalk-free oscillators

Interestingly, the two crosstalk-coupled triggered oscillators have less oscillation time delay (i.e. faster frequency) than the crosstalk-free oscillators, and the phase difference between the two crosstalk-coupled oscillators diminishes due to the coupling effect as shown in Figure 17. Phase detection at the output of the crosstalk-coupled triggered oscillators can be used as an indicator for strong coupling between TSVs. On the other hand, the edges of crosstalk-free oscillators are still separated by a deterministic phase shift dictated by the time difference between the delay buffer chain and the ring oscillation time period.

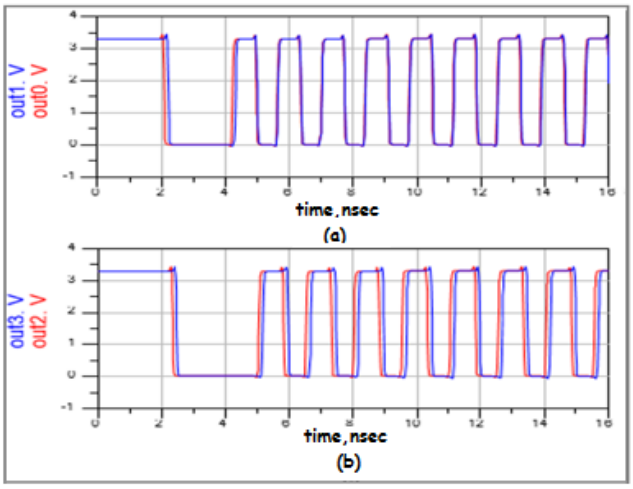


Figure 17: Simulated Results for 3D Interconnect Crosstalk (a) Crosstalk-coupled (b) Crosstalk free

5. Conclusion

In this paper, on-chip testing techniques are presented to test and characterize the signal integrity effects of TSVs after fabrication. Characterizing TSVs using an on-chip 3D testing will provide accurate information since it is closely matches real IC operating conditions, which is very important to evaluate the 3D circuit performance and validate the electrical models of these TSVs. In this paper, 3D testing techniques for TSVs are presented. A Gunning Transceiver logic (GTL) testing structure is utilized to provide a unique way of characterizing crosstalk and Inter-Symbol-Interference (ISI) effects of TSVs. This includes a proposed methodology of characterizing near-end crosstalk (NEXT) and the far-end crosstalk (FEXT) between different 3D routes. A GTL shift register can be also utilized as scan chain test to characterize the reliability of TSVs. Finally, an oscillation ring (OR) 3D testing methodology to detect and identify faulty TSVs and characterize their impact on 3D circuits performance is presented. The proposed OR testing methodology adopt a parallel ROs structure, which is dedicated to detect TSV faults such as stuck-at, open, delay faults and crosstalk coupling between TSVs. The parallel RO structure can be built in different 3D configurations connected through TSVs across different 3D stacking tiers to evaluate and characterize the reliability of TSVs. For open and short faults, the circuit outputs of this test scheme are based on pass/fail criteria. TSVs delay and rise time degradation can be detected based on the counter data for each triggered RO. In addition, the proposed testing structure provides a new way of detecting a strong noise coupling between TSVs. In this test, the outputs of the coupled ring oscillators, which is running at same frequency, will eventually lock in phase due to the coupling noise between TSVs, where a phase detection circuit can be used to indicate a strong TSVs crosstalk noise.

References

[1] J. Q. Lu, "3-D Hyperintegration and Packaging Technologies for Micro-Nano Systems," in Proc. of IEEE, vol. 97(1), 2009, pp. 18-30.

[2] W.R. Davis et al., "Demystifying 3DICs : the Pros and Cons of Going Vertical", IEEE Design and Test of Computers, vol. 22, no. 6, pp. 498 - 510, 2005.

[3] R. Wang, K. Chakrabarty, S. Bhawmik,"At-speed interconnect testing and test-path optimization for 2.5D ICs," in Proc. IEEE VLSI Test Symp. (VTS), 2014, pp. 1-6.

[4] R. Weerasekera, M. Grange, et al., "Compact modelling of throughsilicon vias TSV s in three dimensional (3D) integrated circuits," in IEEE Int'l Conf. on 3D System Integration, 2009, pp.1-8.

[5] W. Yao, S. Pan, et al., "Modeling and application of multi-port TSV networks in 3-D IC,"IEEE Trans. Comput.-Aided Design Integr. Circuits Syst., vol. 32, no. 4, Apr. 2013., pp. 487-496.

[6] Z. Xu, A. Beece, et al., "Crosstalk evaluation, suppression and modeling in 3D through-strata-via (TSV) network," in Proc. IEEE Int. Conf. 3D Syst. Integr., Nov. 2010, pp. 1-8.

[7] A. Mercha et al., "Comprehensive analysis of the impact of single and arrays of through silicon vias induced stress on high-k/ metal gate CMOS performance," in Proc. IEEE Int. Electron Devices Meeting (IEDM), San Francisco, CA, USA, 2010, pp. 2.2.1–2.2.4.

[8] Shadi MS Harb, William Eisenstadt, "Oscillation ring testing methodology of TSVs in 3D stacked ICs," IEEE Nordic Circuits and Systems Conference (NORCAS), 2016.

[9] D. M. Jang, C. Ryu, K. Y. Lee, B. H. Cho, J. Kim, T. S. Oh, W. J. Lee, and J. Yu, "Development and evaluation of 3-D SiP with vertically interconnected through silicon vias (TSV)," Proc 2007 Elect. Comp. and Tech. Conf, pp. 847-852, 2007.

[10] C. Ryu, D. Chung, J. Lee, K. Lee, T. Oh and J. Kim, "High frequency electrical circuit model of chip-to-chip vertical via interconnection for 3-D chip stacking package," IEEE 14th Topical Meeting on Electrical Performance of Electronic Packaging, pp. 151-154, 2005.

[11] Z. Xu, A. Beece, K. Rose, T. Zhang, and J. Q. Lu, "Modeling and evaluation for electrical characteristics of Through Strata Vias (TSVs) in three-dimensional integration," IEEE 3D System Integration Conference, 2009.

[12] C. Ryu, J. Lee, H. Lee, K. Lee, T. Oh, and J. Kim, "High frequency electrical model of through wafer via for 3-D stacked chip packaging," Electronics System integration Technology Conference, vol. 1, pp. 215-220, 2006.

[13] A. JongShick, and W. R. Eisenstadt, "In-situ characterization of high speed I/O chippackage systems," PhD thesis, University of Florida, 2007.

[14] I. Loi, S. Mitra, T. H. Lee, S. Fujita, and L. Benini, "A low-overhead fault tolerance scheme for TSV-based 3D network on chip links," IEEE/ACM Int. Conf. ComputerAided Design, pp. 598–602, 2008.

[15] M. Kaneko, and K. Sakaguchi, "Oscillation fault diagnosis for analog circuits based on boundary search with perturbation model," IEEE ISCAS, pp. 93-96, 1994.

[16] K. Arabi, and B. Kamiska, "(OBIST) scheme for functional and structural testing of analog and mixed-signal integrated circuits," Int'l Test Conf. pp. 786-795, 1997.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1266-1271 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Stability and nonlinear controller design of Fast-Lock Phase-Locked Loop in 0.18-µm CMOS

Fatima Al-Mutairi[*,1], Reem Al-Mutairi[2]

[1]*Department of Electrical and Computer Engineering, Texas A & M University, 77843, College Station, TX*

[2]*Department of Electrical Engineering, Kuwait University, 965, Kuwait*

A R T I C L E I N F O

Article history:
Received : 01 June, 2017
Accepted : 13 July, 2017
Online : 28 July, 2017

Keywords:
Adaptive controller
Stability
Phase Margin(PM)
Frequency divider
VCO
PLL

A B S T R A C T

This paper introduces a kind of novel backstepping phase- locked loops (PLL). The paper presents a developed theoretical model of (PLL) frequency synthesizer and provides analytical equations for calculating the desired specification such as phase margin (PM). The proposed model aims to improve the control accuracy, ensure stability and reducing the lock time. Adaptive controller algorithm has been used to design a conventional phase locked loop to operate at 2.2GHz using 0.18 µm CMOS technology. The lock time was reduced by adding nonlinear controller in addition with the low-pass filter (LPF) and the tuning variables of the adaptive control to ensure the control accuracy. The simulation results confirm that the stability of the system has been improved and increase Phase Margin (PM) from 56° to 60°. Furthermore, the simulation results indicate that the performance of adaptive PLL control has greater speed response and smaller overshoot than conventional PLL. The lock time for the conventional PLL was 2.1 µs and it is reduced to 1 µs by adding adaptive controller, showing a reduction of the lock time by 53% over the classical PLL.

1. Introduction

This paper is an extension of work originally presented in conference [1]. The phase-locked loop (PLL) considered as an important part in many applications. Phase-Locked loop (PLL) has been widely used in different applications such as electronic application, communication system etc. [2]. Moreover, it is used for data transmission and are manufactured as integrated circuit [3]. PLL considered as an electronic circuit with voltage driven oscillator that permanently adjust to correspond the frequency of an input signal. PLL has been used to generate, stabilize, filter or recover signal from noisy communication channel [4]. Most application of different system requires the oscillator to be tunable [5]. Therefore, their output frequency will be a function of a control input and frequency can be adjusted or controlled by voltage of the control signal. Stability of the system is a critical issue for many designers. Phase margin consider as one of the stability metrics for assigning system robustness. This paper includes stability analysis of the PLL to remain system unchanged over time from external condition.

[*]Corresponding Author: Fatima Al-Mutairi, Department of Electrical and Computer Engineering, Texas A & M University, 77843, College Station, TX
Email: falmutairi@tamu.edu

2. Modeling PLL

The systematic model of the phase- locked loop (PLL) nonlinear controlled system is shown in Figure 1. Moreover, the linearized model of the PLL and nonlinear controller is illustrated in Figure 2. The PLL consists of the voltage control oscillator, phase frequency detector, charge pump, divider and traditional low-pass filter. The first functional block of the PLL is the PFD (phase frequency detector) which effects on the performance of the PLL where it compares the phase of two input signal one usually comes from VCO [5]. Figure 3 and Figure 4 show the structure of PFD. Secondly, the charge pump works as a switch current source that controlled by the phase difference between oscillator and reference [6]. Figure 5 shows the charge pump. Furthermore, the loop filter and backstepping controller are considered as a one of functional block of PLL. The second order loop filter is shown in Figure 6. The voltage gains of the VCO (voltage-controlled oscillator) is a major parameter in the PLL that effects lock time [7]. An illustration of the three stages current-starved VCO (CS-VCO) is shown in Figure 7. Frequency divider consists of TSPC technique. TSPC employs a true single-phase clock that is the only clock signal needed and never inverted, which avoids the usage of two non-overlapping two-phase clocks and thus increase the speed [8]. The divider based on TSPC is shown in Figure 8.

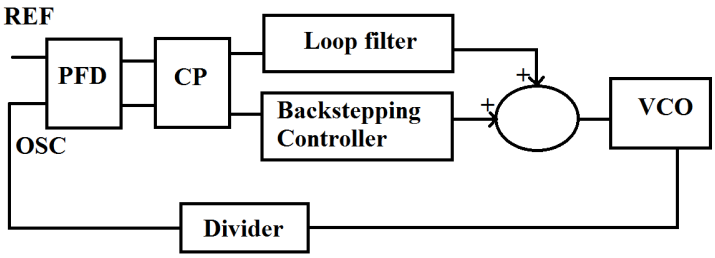


Figure 1: Systematic model of Phase Locked Loop with adaptive Control

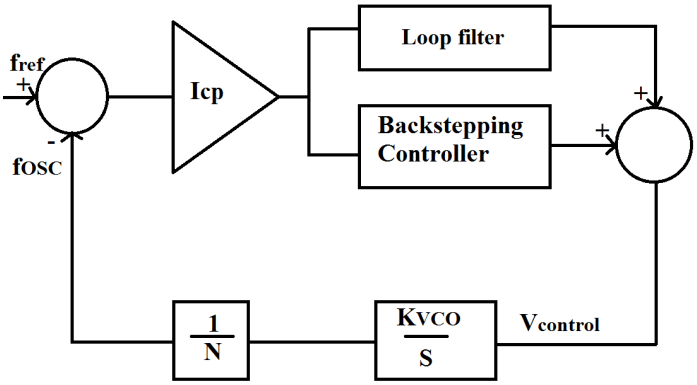


Figure 2: Linearized model of the system

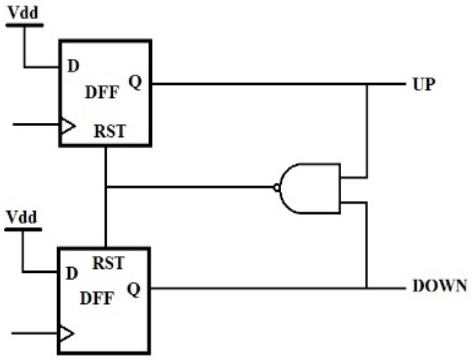


Figure 3: PFD

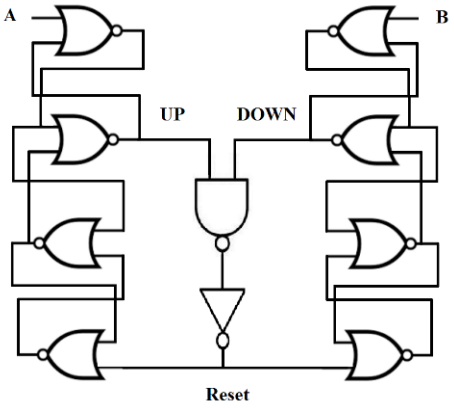


Figure 4: Structure of the PFD

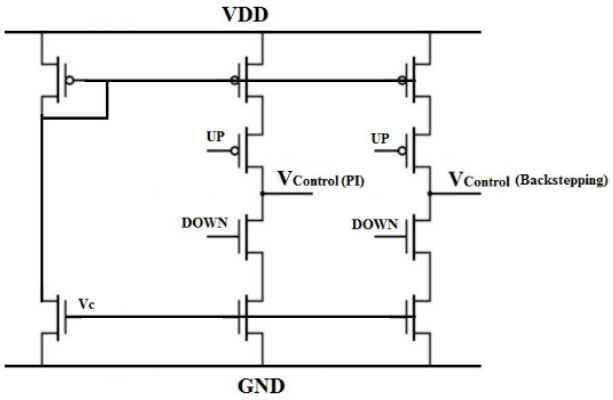


Figure 5: Structure of the CP

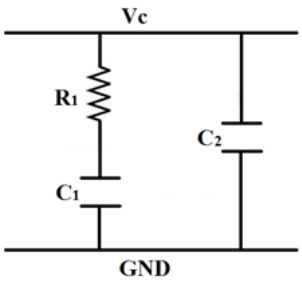


Figure 6: Loop Filter

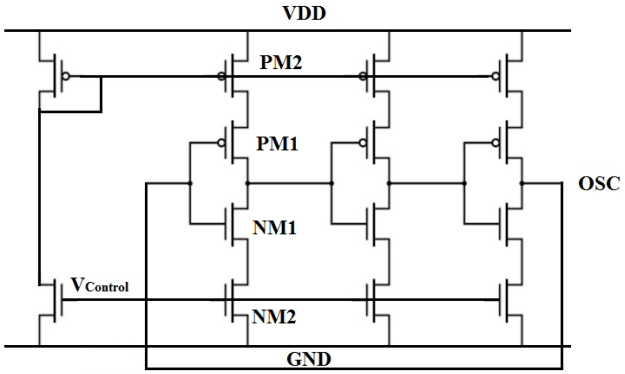


Figure 7: Structure of the current-Starved VCO

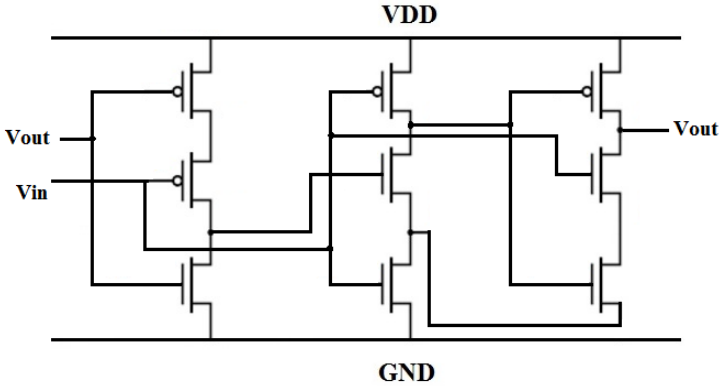


Figure 8: Structure of the TSP

3. Theoretical Analysis

Backstepping design technique is a motivating topic for many researchers, it is widely used and became as one of the featuring new methods for stabilizing strict-feedback, nonlinear systems [9]. Backstepping control is a nonlinear control that used one of the state as a control input to control the other states [10]. The technique is based on generating a family of global asymptotically stabilizing control laws starting the design procedure by defining stable system and determining controller that gradually stabilize each outer subsystem. The process terminated when the external control is attained [11]. The method basically depends on applying Lyapunov functions in which guarantees the stability of the related system [12].

4. Design Proposed

In this section, we will explain the steps needed to implement the backstepping method. The open loop transfer function of PLL can be defined in Equation (1).

$$G(s) = \frac{sk_b + k_a}{s^2\left(s + \frac{\left(1+\frac{C_1}{C_2}\right)}{R_1C_1}\right)} \quad (1)$$

Where $k_a = \left(\frac{k_{vco}I_{cp}}{NR_1C_1C_2}\right)$, $k_b = \frac{k_{vco}I_{cp}}{NC_2}$,

K_{VCO} is the VCO gain and Icp is the charge pump current. Finding the state space of the system from the transfer function Equation (1).

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & \frac{-\left(1+\frac{C_1}{C_2}\right)}{R_1C_1} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} u \quad (2)$$

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} k_a & k_b & 0 \\ 0 & \frac{k_a}{k_{vco}} & \frac{k_b}{k_{vco}} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \end{bmatrix} u \quad (3)$$

Rearrangement the system is the first procedure of the backstepping controller.

$$V_1 = \frac{k_1}{2}x_2^2 \quad (4)$$

The derivative of Equation (4) is taking.

$$\dot{V}_1 = k_1x_2(\dot{x_2}) \quad (5)$$

Substituting Equation (2) in Equation (5) yields.

$$\dot{V}_1 = (x_2)(k_1x_3) \quad (6)$$

Choosing the controller in Equation (10) to ensure the stability of the system.

$$k_1x_3 = -k_2x_2 \quad (7)$$

$$k_1x_3 + u = -k_2x_2 + \mathrm{u} \quad (8)$$

$$k_1x_3 + \dot{x_3} + \frac{\left(1+{}^{C_1}/_{C_2}\right)}{R_1C_1}x_3 = -k_2x_2 + \mathrm{u} \quad (9)$$

$$u = \dot{x_3} + \left(k_1 + \frac{\left(1+{}^{C_1}/_{C_2}\right)}{R_1C_1}\right)x_3 + k_2x_2 \quad (10)$$

$$\dot{V}_1 = -k_2x_2^2 < 0 \quad (11)$$

The schematic of backstepping controller is shown in Figure 9.

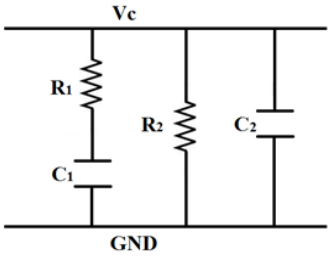


Figure 9: Backstepping Control

5. Stability Analysis

The open loop transfer function of PLL system is shown in equation (12).

$$G(s) = \frac{K_{VCO}I_{CP}}{NC_2\left(1+\frac{C_1}{C_2}\right)}\left(\frac{(s\,R_1C_1+1)(s^2(2C_2R_2R_1C_1)+s(R_1C_1+2R_2C_1)+1)}{s^2(s^2(C_2R_2R_1C_1)+s(R_1C_1+R_2C_2+R_2C_1)+1)\left(\frac{R_1C_1s}{1+\frac{C_1}{C_2}}+1\right)}\right) \quad (12)$$

The magnitude of the G(s) has been plotted with respect to frequency ω as shown in Figure 10.

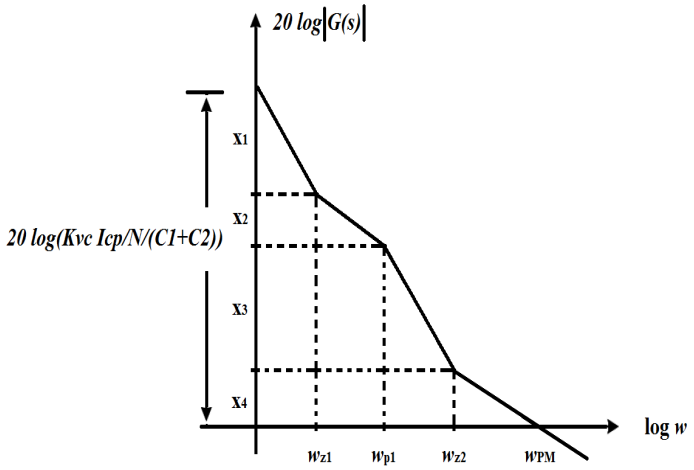


Figure 10: Bode Plot of the G(s)

The crossover frequency (ω_{PM}) has been determined [13].

$$20log\left(\frac{K_{VCO}I_{CP}}{N(C_2+C_1)}\right) = x_1 + x_2 + x_3 + x_4 \quad (13)$$

$$\frac{x_1}{log(\omega_{z1})} = 40 \quad (14)$$

$$\frac{x_2}{log(\omega_{p1})-log(\omega_{z1})} = 20 \quad (15)$$

$$\frac{x_3}{log\left(\frac{1}{R_1C_1}\right)-log(\omega_{p1})} = 40 \quad (16)$$

$$\frac{x_4}{log(\omega_{PM})-log\left(\frac{1}{R_1C_1}\right)} = 20 \quad (17)$$

$$\omega_{PM(with\ backstepping)} \approx \frac{K_{VCO}I_{CP}}{N(C_2+C_1)}\frac{\omega_{p1}}{\omega_{z3}\omega_{z1}} = \frac{\omega_{p1}}{\omega_{z1}}\omega_{PM(without\ backstepping)} \quad (18)$$

Where

$$\omega_{p1} = -\left(\frac{R_2+R_1}{2C_2R_2R_1}\right) + \sqrt{\left(\frac{R_2+R_1}{2C_2R_2R_1}\right)^2 - \left(\frac{1}{C_2R_2R_1C_1}\right)}, \quad (19)$$

$$\omega_{p2} = -\left(\frac{R_2+R_1}{2C_2R_2R_1}\right) - \sqrt{\left(\frac{R_2+R_1}{2C_2R_2R_1}\right)^2 - \left(\frac{1}{C_2R_2R_1C_1}\right)}, \quad (20)$$

$$\omega_{z1} = -\left(\frac{2R_2+R_1}{2R_1C_2R_2}\right) + \sqrt{\left(\frac{2R_2+R_1}{2R_1C_2R_2}\right)^2 - \left(\frac{2}{C_2R_2R_1C_1}\right)}, \quad (21)$$

$$\omega_{z2} = -\left(\frac{2R_2+R_1}{2R_1C_2R_2}\right) - \sqrt{\left(\frac{2R_2+R_1}{2R_1C_2R_2}\right)^2 - \left(\frac{2}{C_2R_2R_1C_1}\right)}, \quad (22)$$

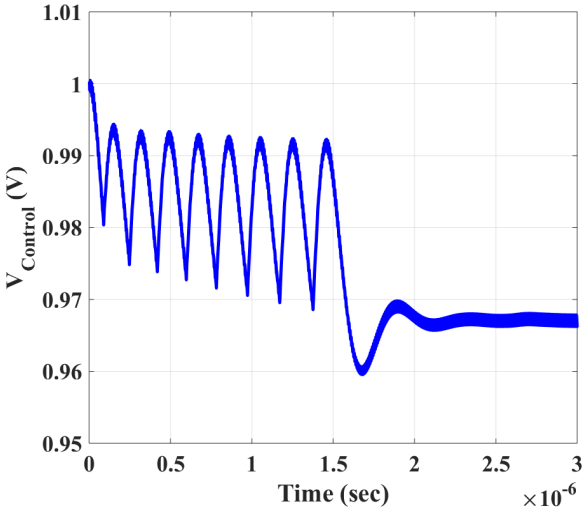
$$\omega_{z3} = -\left(\frac{1}{C_1R_1}\right), \quad (23)$$

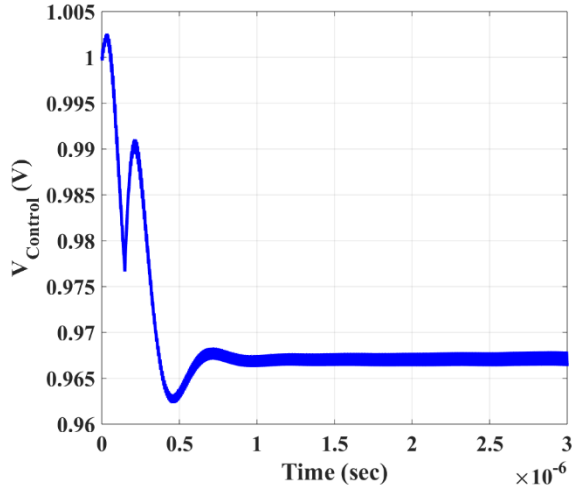
The phase margin can be calculated as shown below:

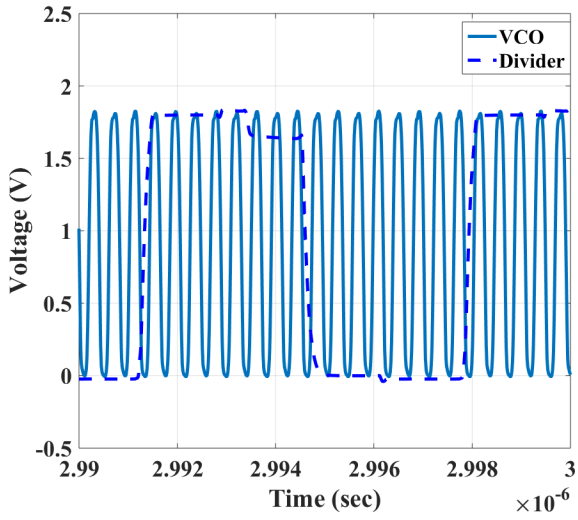
$$PM = tan^{-1}\left(\frac{\omega_{PM}}{\omega_{z1}}\right) + tan^{-1}\left(\frac{\omega_{PM}}{\omega_{z2}}\right) + tan^{-1}\left(\frac{\omega_{PM}}{\omega_{z3}}\right) - tan^{-1}\left(\frac{\omega_{PM}}{\omega_{p3}}\right) - tan^{-1}\left(\frac{\omega_{PM}}{\omega_{p2}}\right) - tan^{-1}\left(\frac{\omega_{PM}}{\omega_{p1}}\right), \quad (24)$$

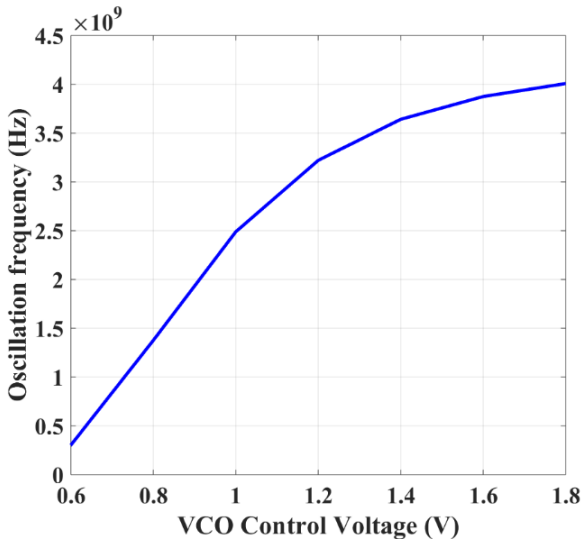
6. System Simulation

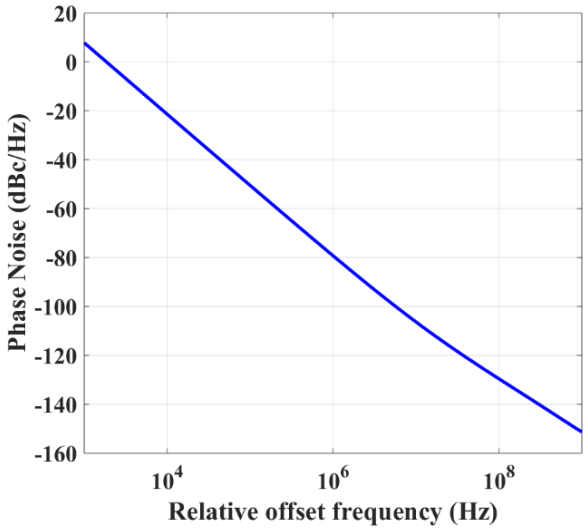
A 0.18µm CMOS technology was used to simulate the proposed design, test and verify the dynamic response of backstepping phase lock loop (PLL). The transistors size of the VCO, Divider, Gates and Charge Pump is shown in Table 1. Maximum overshoot and settling time have been considered as a performance metrics to compare the performance of conventional PLL and backstepping controller with PLL. The comparison between the actual PLL and the implementation of backstepping controller with PLL has been shown in Figure 11 and Figure 12 respectively. The results based on proposed design is showing a smaller overshoot and fast response time with lock time of 1µs where the lock time of the classical PLL was 2.1µs. The new scheme increases the speed and make the system more faster in order to settle the desired voltage where each voltage applying to the VCO represents as a frequency as shown in Figure 12. Figure13 shows Division ratio. The division ratio is 16 when the input and output frequencies are 2.3 GHz and 142 MHz. Figure 14 shows the increased of the measured tuning range of the VCO. The measured tuning range increases from 0.3 to 4 GHz, as the control voltage is increased from 0.6 to 1.8 V. Figure 15 represents that the phase noise is -79.24 dBc/Hz at 1MHz offset. The performance summary of this paper is shown in Table 2.











To insure stability, the phase margin should be at least 45 degree. Typically, the value of C2/C1 is set in the range of 1/10 to 1/6 for conventional PLL that limit phase margin and settling time. By adding backstepping the phase margin is function of R_2 and it is improved as shown Figure 16 and Figure 17. The phase margin without backstepping is 56.4˚ and with backstepping is 60.6˚.The improvement is 7.44%.

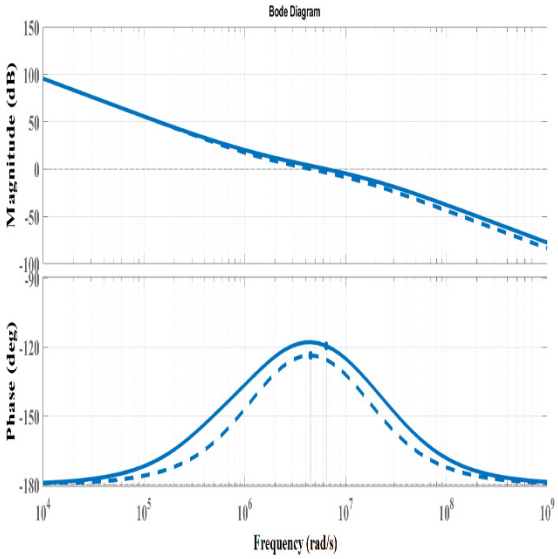


Table 1: Sizing of the Transistors

Device	Transistor	Ratio
VCO	PM1	27.778
	PM2	28.333
	NM1	11.667
	NM2	11.111
Divider	PM	16.667
	NM	11.111
Not Gate	PM	27.78
	NM	11.67
Nand Gate	PM	27.78
	NM	17.78
Nor Gate	PM	66.67
	NM	15.56
Charge Pump	PM1	17.78
	PM2	8.88
	NM2	2.22
	NM1	11.11

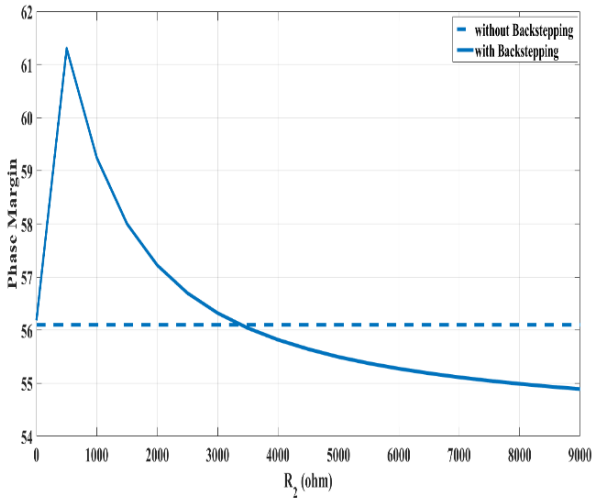


Figure 17: Phase margin as a function of R_2

Table 2: Performance Summery

Technology	CMOS 0.18µm
Voltage supply	1.8 V
Power Dissipation	11.079 mW
Phase noise	-79.24dBc/Hz @1MHz
Tuning Range	300MHz – 4GHz
Lock Time	1µs

In this paper, we present adaptive controller to PLL. The modeling, tuning and simulation of proposed design are presented in this work. In addition, stability analysis is produced by determining phase margin of open loop transfer function of PLL from Bode plot. A 0.18µm CMOS technology was used for circuit designed. In general, it is evident that the established model shows through simulations the superiority and effectiveness performance of the backstepping controlled PLL over the conventional PLL. The proposed controller proved to achieve greater speed response and smaller overshoot when it compared to conventional PLL. The results based on proposed design is showing a fast response time with lock time of 1µs where the lock time of the classical PLL was 2.1µs reducing the lock time by a 53%. Simulation results also confirm that the proposed controller works very well, in which it improved control accuracy and it stability. Furthermore, the phase margin that is considered as one of the stability metrics of system showing an increase from 56° to 60°.

References

[1] Almutairi, Fatima, and Reem Almutairi. "Fast-Lock Phase-Locked Loop With Adaptive Controller In 0.18-Mm CMOS". International Conference On Electronic Devices, Systems, And Applications. 2016. Print.

[2] R. Baker, H. Li and D. Boyce, CMOS circuit design, layout, and simulation. New York: IEEE Press, 1998.

[3] R. Best, Phase-locked loops. New York: McGraw Hill, 1997.

[4] Phase-Lock Loop Applications Using the MAX9382", Dallas Maxium, Application Note 1130, 9th May 2005, http://pdfserv.maxim-ic.com/en/an/AN1130.pdf

[5] H. Janardhan and M. Wagdy, "Design of a 1 GHz Digital PLL Using 0.18um CMOS Technology", in IEEE Third International Conference on Information Technology, 2006.

[6] Woogeun Rhee, "Design of High- Performance CMOS Charge Pumps in Phase- Locked Loops"

[7] R. Patil and V. Nasre, "A Performance Comparison of Current Starved VCO and Source Coupled VCO for PLL in 0.18um CMOS Process", International Journal of Engineering and Innovative Technology, vol. 1, no. 2, 2012.

[8] Behzard Razavi. RF Microelectronics. Prentice-Hall, 1998.

[9] Ki-Seok Kim and Youdan Kim, "Robust backstepping control for slew maneuver using nonlinear tracking function", IEEE Transactions on Control Systems Technology, vol. 11, no. 6, pp. 822-829, 2003.

[10] P. Yip and J. Hedrick, "Adaptive dynamic surface control: A simplified algorithm for adaptive backstepping control of nonlinear systems", International Journal of Control, vol. 71, no. 5, pp. 959-979, 1998.

[11] G. Li, S. Zhou and K. Yang, "Generalized projective synchronization between two different chaotic systems using active backstepping control", Physics Letters A, vol. 355, no. 4-5, pp. 326-330, 2006.

[12] Y. Li, S. Qiang, X. Zhuang and O. Kaynak, "Robust and Adaptive Backstepping Control for Nonlinear Systems Using RBF Neural Networks", IEEE Transactions on Neural Networks, vol. 15, no. 3, pp. 693-701, 2004.

[13] Adrag. "A systematic Procedure For Analysis Of Third- Order Charge Pump Phase- Locked Loop (PLL)". Academic Journals, vol. 7, no.2, pp. 190-198, 2012.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 2, No. 3, 1272-1279 (2017)
www.astesj.com
Special Issue on Recent Advances in Engineering Systems

ASTES Journal
ISSN: 2415-6698

Towards Deployment Strategies for Deception Systems

Daniel Fraunholz[*,1], Marc Zimmermann[1], Hans Dieter Schotten[1,2]

[1]*Intelligent Networks, German Research Center for Artificial Intelligence, 67663 , Germany*

[2]*Wireless Communication and Navigation, University of Kaiserslautern, 67663, Germany*

ARTICLEINFO

Article history:
Received: 22 May, 2017
Accepted: 07 July, 2017
Online: 01 August, 2017

Keywords :
Information Security
Network Security
Deception Systems
Honeypots
Deployment Strategy
Machine Learning
Artificial Intelligence

ABSTRACT

Network security is often built on perimeter defense. Sophisticated attacks are able to penetrate the perimeter and access valuable resources in the network. A more complete defense strategy also contains mechanisms to detect and mitigate perimeter breaches. Deceptive systems are a promising technology to detect, deceive and counter infiltrations. In this work we provide an insight in the basic mechanisms of deception based cyber defense and discuss in detail one of the most significant drawbacks of the technology: The deployment. We also propose a solution to enable deception systems to a broad range of users. This is achieved by a dynamic deployment strategy based on machine learning to adapt to the network context. Different methods, algorithms and combinations are evaluated to eventually build a full adaptive deployment framework. The proposed framework needs a minimal amount of configuration and maintenance.

1 Introduction

Several studies suggest that cyber crime and espionage frameworks are flourishing. In the United States of America the monetary loss due to cyber crime is amounted to $1,070,000,000 in 2015 [1]. The European Union was also in the focus of organized cyber crime. 15 reported major security breaches leaked more than 41 million records of sensitive information, such as credit card information, email addresses, passwords and private home addresses [2]. In the context of highly sophisticated cyber crime such as industrial espionage, digital repression and sabotage it is common to not only trust perimeter based network security [3]. Several cyber attacks and developped attack methods such as *AirHopper* [4] proved that even physical isolation can be circumvented. This leads to a permanent and latent threat of successful infiltrations, which are undetectable by state of the art defense mechanisms such as firewalls, antivirus, rule based intrusion detection and prevention systems (IDS/IPS), network separation and user authentication. Deception systems (DS) enable in depth network defense support for the IT security concept. They mimic productive, secret or critical resources in the target system. Intruders can not distinguish between a *DS* and the actual resource. However, defenders easily detect intrusions because no connections, traffic and activities are expected on a *DS*. Any interaction with such a system can be classified as malicious. This technology therefore comes along with no false positive classifications, from which other defense in depth technologies such as anomaly detection often suffer. Typical issues for state of the art network defense are: Inside or insider attacks, encryption, high-throughput traffic, polymorphism and highly fluctuating signatures. Deception systems do not suffer any drawbacks on these issues. More than that, technology changes such as IPv6 do not impact *DS*s. However, there are other drawbacks coming along with *DS*s. A major drawback is the deployment [5]. The *DS* needs to mimic a actual system and additionally fit in the network structure [6]. State of the art for a proper configuration, deployment and maintenance is manual effort [7]. We state that a framework consisting of a scanning engine for context observation, a back-end database for proper storage of context information in combination with an engine for machine learning based on context analysis and a *DS* dependent deployment engine can solve this issue. This enables *DS*s for a broad range of applications and companies. Especially small and medium size companies will profit from manageable *DS*s, because they cannot afford cumbersome manual configuration, de-

[*]Daniel Fraunholz, Trippstadter Str. 122, +49 (0)631 / 205 75-0 & daniel.fraunholz@dfki.de

https://dx.doi.org/10.25046/aj0203161

ployment and maintenance of network security mechanisms.

This work is structured as follows: Since the idea of machine learning and deception in network defense is around 30 years old, we first identify recent trends and related work in chapter 2. Investigated machine learning methods as well as their advantages and drawbacks are introduced in chapter 3. In chapter 4 we propose our adaptive deployment framework and discuss important modules. The proposed framework is evaluated in chapter 5. Our work is concluded in chapter 6.

2 Related Work

In strategic defense and attack the idea of deception dates back to the 5th century BC [8]. It was first described from Clifford Stoll as digital strategy [9] in 1990 and first implemented from Lance Spitzer as network defense strategy [10].

2.1 Deception Systems

Modern *DS*s provide a vast variety of fake resources to deceive intruders. The most popular concept are server side systems. These systems mimic typical server protocols such as FTP, SSH or SMB. Connecting intruders trigger alarms and are under observation while they try to exploit the server. Other concepts are client side systems, which connect to potential malicious servers and observe the servers behavior. This concept is common to investigate web based attacks such as drive by downloads. A more recent concept employs tokens as trigger for alarms. Tokens impersonate documents, credentials or accounts. Stack canaries can be interpreted as token-based *DS*. Long-term and large scale studies with deception systems enable high quality insight in recent threats and their developments [11][12].

2.2 Deployment Strategies

Except for client-side *DS*s all need to be implanted in an existing and often fluctuating context. This context can be a IP-based network, a file system or any other architecture to defend. In this work we will focus on IP-based networks. There are two major groups of deployment modes: Research and production. In research mode the *DS* is directly connected with the Internet. In this mode its main purpose is the collection of threat intelligence, botnet observation and other trends. For non IT security companies this mode is not relevant. The production mode deploys *DS*s behind the perimeter. *DS*s in this mode typically have less interaction. However, in this mode any interaction is a strong indicator for perimeter breaches or internal misuse. In the production mode, six basic deployment concepts are prevalent [13][14]: Sacrificial lamb, deception ports on production systems, proximity decoys, redirection shield, minefield, zoo. In table 1 the different concepts are described.

State of the art deployment strategies do not employ automated deployment. Our adaptive framework supports all deployment concepts except for deception ports, since access to the production machines is not natively available. Furthermore, we argue that manipulation of software on production systems is not acceptable for most operators and vendors. This restricts the usage of the deception port concept in industrial scenarios and proprietary systems. We also argue that sacrificial lamb and zoo deployment suffer from lower attraction to intruders and less knowledge about the actual network security state. Both are implications of the deployment in a different subnetwork. Minefield deployment is a good choice to detect intrusions in an early state, but if an intruder circumvents the minefield there are no more defense in depth mechanisms. We focus on proximity decoys, since we think it is the most promising deployment concept for defense in depth strategies. Please note that redirection shield is a special case of all other concepts, where the *DS*s hardware is not located in the internal network, but the malicious traffic is tunneled out to an external environment.

2.3 Artificial Intelligence for Deception based Network Security

Artificial intelligence enables context-awareness. In network security this is crucial, since modern networks are heterogeneous and entities within the network can often change. To adapt *DS*s in these scenario several researches have been conducted. These researches can be classified in two major domains: Interaction and Deployment. Context-aware interaction focuses on decision making for *DS*s [15][16][17]. The adaptive deployment domain is in an early stage compared to the first usage of *DS*s. However, this domain decreases the probability for being fingerprinted by adapting to other entities in the network and also increases the intrusion detection probability by optimizing the ratio between *DS*s and production systems within a network. Conducted works are learning mechanisms of new unknown services and protocols [18], context-awareness for *DS*s [19] and automated configuration [20]. An overview of conducted research is given by Zakaria [21][22].

3 Unsupervised Machine Learning

The data acquired from our framework is not labeled. Even the number of clusters is unknown. To determine the optimal *DS*s deployment, we employ unsupervised machine learning methods to identify clusters and derive deployment prototypes. In this chapter we introduce and investigate several methods we identified as promising. These methods are later employed in our framework.

Table 1: Deployment concepts for *DS*s in internal networks

Concept	Description
Sacrificial lamb	Single deployment isolated from any production systems
Deception ports on production systems	Deployment on the production system
Proximity decoys	Deployment near the production systems
Redirection shield	Redirection of certain traffic outside the internal network
Minefield	Deployment of a vast amount of *DS*s near the perimeter
Zoo	Deployment of a vast and versatile amount of *DS*s isolated from any production system

3.1 Methods and Algorithms

We investigated three different clustering algorithms. All three are assigned to a different class of cluster algorithms. First is the centroid based k-medoids method [23]. In difference to the well known k-means algorithm, k-medoids always sets an entity from within a cluster as centroid. This centroid is called medoid. As given in (1), we define the Jaccard-Tanimoto metric [24] as distance measurement:

$$d(x,y) = \frac{\left|x \cup y\right| - \left|x \cap y\right|}{\left|x \cup y\right|} \tag{1}$$

where x and y are either a feature set of an observation or a feature set of an aggregation of observations. We employ this distance measurement as reference for all further investigations in this paper. There are, however, several distance measurements that are also feasible such as the Manhattan, Euclidean, Simpson, Dice and Mahalanobis distance [25]. The definition of the k-medoids method is given in (2):

$$\arg\max_{S} \sum_{i=1}^{k} |S_i| \, Var S_i \tag{2}$$

where k is the number of clusters and $S = S_1, S_2, ..., S_k$ the sets of all observations.

Our evaluation is based on the partition around medoids (PAM) [23] implementation. *PAM* is a heuristic method, employed to circumvent the *NP*-hardness of k-medoids.

Second is the connectivity based single linkage clustering [26]. We also chose the Jaccard-Tanimoto distance as distance measurement to ensure comparability. The single linkage method is an agglomerative hierarchical clustering method. All observations are considered as cluster and then merged into an agglomeration of clusters based on the distance between the clusters. The distance is calculated by a linkage function, which is given in (3) for the single linkage method

$$D(S_i, S_j) = \min_{u \in S_i, v \in S_j} d(u,v) \tag{3}$$

where D is the linkage function, S_i and S_j are subsets of S, u is a observation in cluster S_i and v a observation in cluster S_j. In our experiments we found that more complex linkage functions such as WPGMA, UPGMA and WPGMC do not significantly improve the results of our application. We used the SLINK implementation [27] to decrease the time complexity from $O(n^2 log(n))$ to $O(n^2)$.

Finally, we evaluated the density based spatial clustering of applications with noise (DBSCAN) method [28]. *DBSCAN* defines a distance measurement $d(x,y)$ and a minimal number of observations $minPts$ that need to be in a certain distance ϵ of a given observation x to consider the observation x as part of the cluster. If a observation x is within the distance ϵ of less than $minPts$ observations, it is considered as cluster edge and is part of the cluster. The Jaccard-Tanimoto metric is employed as $d(x,y)$.

All three methods imply different advantages and disadvantages. A comparison is given in table 2.

It can be seen that the optimal algorithm depends on the application. Determining a suitable method requires an understanding of the data set. In our application it is not possible to assume a certain distribution of systems within a network. The diversity of clusters and the occurrence of outliers depend on the network architecture.

3.2 Convergence Criteria

The introduced algorithms require a proper parametrization to ensure reasonable results. Even methods that need no predetermination of k need parameters to calculate k.

We employed three methods to estimate the convergence criteria: The Elbow method, the GAP method and the Silhouette coefficient. An increasing number of clusters decrease the mean squared error (MSE). The *MSE* is defined as follows:

$$\sum_{i=1}^{k} \sum_{u \in S_i} \left\| u - \mu_i \right\|^2 \tag{4}$$

where k is the number of clusters, u an observation in cluster S_i and μ_i the mean value of S_i. The elbow method [29] investigates, if further incrementation of the number of clusters do significantly decrease the *MSE*. If the decrease is not significant, the optimal number of clusters is found. The GAP method [30] is based on the elbow method, but instead of $\frac{\Delta MSE}{\Delta k}$, the maximal difference between the *MSE* of the elbow function and the *MSE* of randomly distributed observations indicates the optimal number of clusters. A widely employed method to determine the number of clusters in machine learning applications is the silhouette coefficient [31]. The definition is given in (5).

Table 2: Comparison of different clustering algorithms

Feature	k-medoids	Single linkage	DBSCAN
Class	Centroid	Connection	Density
Predetermined k	Yes	No	No
Outliers	-	+	+
Efficiency	PAM: $O(k(n-k)2)$	SLINK: $O(n^2)$	$O(nlog(n))$
Divers clusters	-	0	+
Deterministic	No	Yes	No

$$s_j = \begin{cases} 0 & \text{for } dist(S_i,u) = 0 \\ \frac{dist(S_v,u)-dist(S_i,u)}{\max\{dist(S_i,u),dist(S_v,u)\}} & \text{else} \end{cases} \quad (5)$$

The distance measure for the silhouette method based on (1). For the distance between an observation and a cluster, the mean value of the cluster is employed as defined in (6) and (7).

$$dist(S_i,u) = \frac{1}{|S_i|}\sum_{x\in S_i} d(x,u) \quad (6)$$

$$dist(S_j,u) = \min_{S_y\neq S_i} \frac{1}{|S_x|}\sum_{y\in S_y} d(y,u) \quad (7)$$

The distance between S_j and u is the difference as defined in (1) to the nearest cluster $S_y \in S$. For an evaluation we will employ the three introduced convergence criteria.

4 Adaptive Deployment Framework

We developed an adaptive deployment framework consisting of a data acquisition engine (DAE), a clustering engine (CE) and a deployment engine (DE). A specific data format was also developed. In this chapter we describe our framework and the single components. The adaptive deployment consists of four consecutive processes: Context perception, context evaluation, configuration and deployment. In the first step the *DAE* collects context information such as other hosts. The acquired data is then stored in our data format. Based on this data, the *CE* statistically analyzes the stored data and determines k prototypes P. These prototypes P are *DS*s that are $\min d(P,S_i)$. The configuration process depends on the *DE*. In general, however, the required configuration file is generated in this process. Finally, the *DE* deploys the *DS*s based on the configuration file. The overall process of adaptive deployment is restartable at any time. This enables a fast adaption to changing architectures and contexts. The process is shown in Figure 1.

4.1 Data Acquisition Engine

The *DAE* captures the context and stores it in a defined data format. In our implementation we define the other hosts in the same subnetwork as context. To capture as much information as possible about the context, the *DAE* combines passive information gathering by *p0f* [32] and active information gathering by *nmap* [33] and *xprobe* [34]. For each host in the subnetwork the information sources decide by vote for an operating system. The services available from a host are determined by *nmap*.

4.2 Data Format

The data format we developed is based on the Extensible Markup Language (XML). First an unique identifier (ID) is generated for each host. These *ID*s are associated with features. There are three major sections: meta data, services and operating system. The first section contains available meta data such as up time, MAC address, IP address and a time stamp. In the second section open TCP and UDP ports are listed. We map port numbers directly to services. This is efficient and produces sufficiently reliable results. In the third section we store information about the TCP stack based fingerprint. This information is extracted from the *nmap* and *xprobe* scan.

4.3 Clustering Engine

In the *CE* the prototypes for the deployment are generated. These prototypes need to contain all information that is needed for a sufficient deployment. In our implementation we employ the same data format for context information and prototypes. The *CE* determines k clusters containing S_i hosts. The TCP stack and the available services for each P are equal to the medoid in S_i. However, meta information is generated on distributions within S_i. For example the MAC address: The first three octetes are extracted from the most prevalent vendor within S_i and the other three are chosen randomly. For the IP address we developed an algorithm to reduce impact on the distribution in subnetworks. First a random IP within the cluster is chosen then the upwards next unoccupied IP address is assigned to the prototype. By the use of this algorithm the distribution within the cluster remains the same, since a specific probability distribution is preserved if only uniformly distributed observations are added on the existing observations. Please note that IP addresses are only assigned to one host at the same time and therefore the distribution is not perfectly preserved. Uptimes for prototypes are determined based on the mean uptime within a cluster.

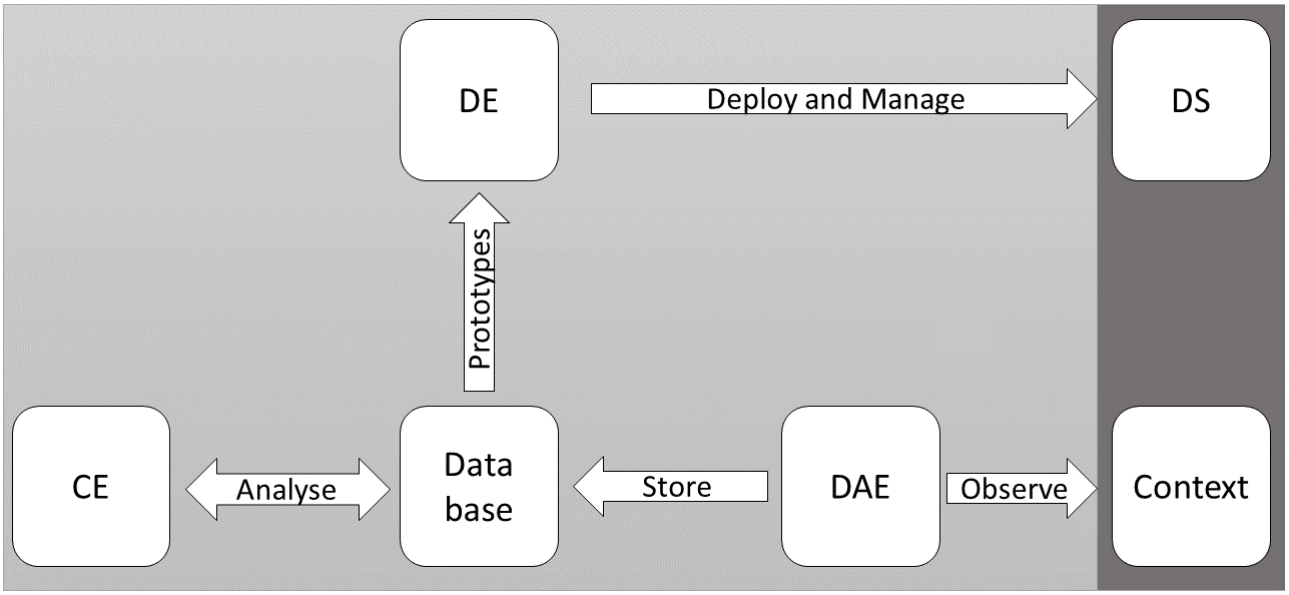


Figure 1: Overall process of the adaptive deployment framework

4.4 Deployment Engine

In a last step the actual deployment is executed. This step is most crucial to all previous steps. The required information for a proper configuration needs to be calculated or assumed. In our implementation we employ *honeyd* [35] as *DE*. *honeyd* is able to emulate a vast amount of hosts with TCP stack and offers the ability to open TCP and UDP ports as well as the execution of scripts to emulate services on the open ports. If it is needed *honeyd* is also able to emulate large network architectures including network elements such as routers, switches and tunnels [36].

5 Evaluation

In the evaluation chapter two different settings are investigated. First, an artificial scenario is evaluated. This scenario consists of several virtual machines (VMs) in an isolated network. The second scenario is an actual production network in which we deploy *DS*s by our framework.

5.1 Artificial Data Sets

As shown in table 3 eight different *VM*s are prepared for the simulation of a production network: Windows 10, Windows 7, Ubuntu 17.04, Ubuntu 12.04, Debian 8.8.0, Fedora 25, openSUSE 42.2 and Android 4.3. Two scenarios are defined in this evaluation. The first scenario mimics a network with equally distributed cluster sizes. In the second scenario the cluster sizes are different. We chose these diverse settings to not favor a specific algorithm. The deployment is realized with Virtualbox.

5.2 Real World Scenario

For scenario 3 we scanned a class C development network. The network consists of: 7 Windows 10 machines, 4 Ubuntu machines, 2 TP Link switches, 2 Cisco switches, 11 Raspberry Pis, 1 Android system and 4 other Unix systems. Unlike in the artificial scenarios the configurations of the systems are different.

5.3 Results

First we evaluated the determination of the number of clusters. In Figure 2 the comparison of combinations of different methods in scenario 1 is shown.

As it can be seen for the elbow method and the silhouette coefficient all three algorithms perform similarly. However, for GAP there are differences. We found that *DBSCAN* is not suitable when using GAP. Please note, that the determined number of clusters is six in this scenario for all algorithms. This is because Ubuntu 12.04 and Ubuntu 17.04 as well as Windows 7 and Windows 10 have closely resembling TCP-Stack implementations and similar open ports in the default configuration, reducing the number of clusters from eight to six. In Figure 3 we compare the same algorithms for scenario 2.

For *DBSCAN* the elbow method does not give a feasible result. The GAP method results only for *SLINK* in suitable results. *PAM* as well as *DBSCAN* result in a number of clusters of four. The silhouette coefficient only results in suitable values for the *PAM*. Figure 4 compares the results for the development network.

DBSCAN is not feasible with any convergence criteria in this scenario. This fits in our overall evaluation. However, it is recommend to estimate ϵ not on the number of cluster, but on the k-distance graph for *DBSCAN*. By doing so the results are probably better. *PAM* and *SLINK* both result in reliable values for the

Table 3: Definition of the investigated scenarios

	Scenario 1	Scenario 2
Windows 10	3	10
Windows 7	3	5
Ubuntu 12.04	3	0
Ubuntu 17.04	3	4
Debian	3	2
Fedora	3	1
openSUSE	3	1
Android	3	5

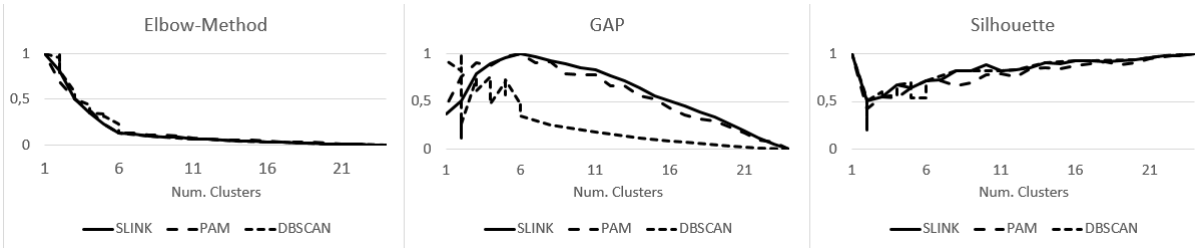


Figure 2: Evaluation of algorithms to estimate the number of clusters in Scenario 1

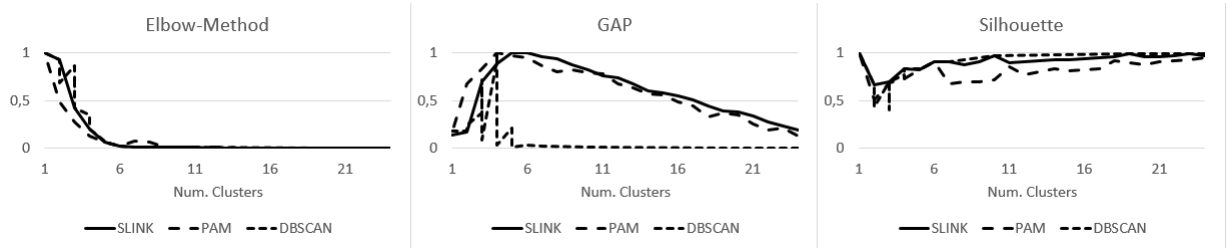


Figure 3: Evaluation of algorithms to estimate the number of clusters in Scenario 2

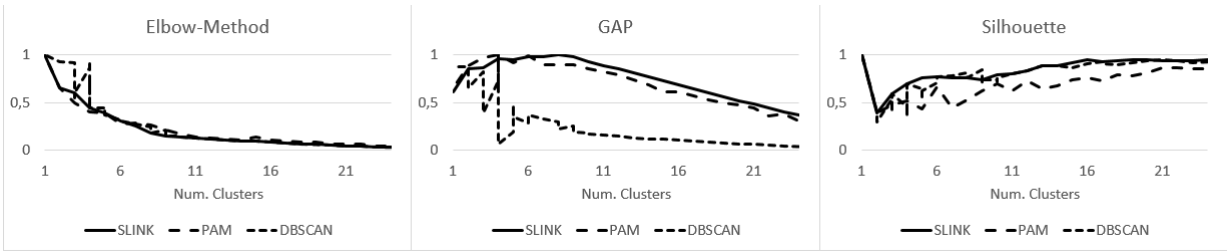


Figure 4: Evaluation of algorithms to estimate the number of clusters in Scenario 3

Table 4: Relative error for the estimation of the number of clusters

	Scenario 1			Scenario 2			
	Elbow Method	GAP	Silhouette	Elbow Method	GAP	Silhouette	Mean
SLINK	0.25	0.25	0.5	0.25	0.25	0.25	0.29
PAM	0.25	0.25	0.13	0.25	0.5	0.25	0.27
DBSCAN	0.25	0.75	0.38	0.25	0.5	0.5	0.44
Mean	0.25	0.42	0.33	0.25	0.42	0.33	

Table 5: Relative error for the estimation of the entities within the clusters in scenario 1

	C1	C2	C3	C4	C5	C6	C7	C8	Mean
SLINK	0.33	1.00	1.00	0.00	1.00	0.33	1.00	0.00	0.58
PAM	0.67	1.00	1.00	1.00	1.00	0.50	1.00	0.00	0.77
DBSCAN	0.33	1.00	1.00	0.00	1.00	0.33	1.00	0.00	0.58

Table 6: Relative error for the estimation of the entities within the clusters in scenario 2

	C1	C2	C3	C4	C5	C6	C7	Mean
SLINK	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
PAM	1.00	1.00	0.67	1.00	0.00	0.00	0.80	0.64
DBSCAN	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

elbow method and the silhouette coefficient. The overall performance evaluation is given in Table 4.

Comparing the algorithms the best results are achieved for the *SLINK* implementation. For the convergence criteria the elbow method appears to provide the best results. However, the elbow method requires an additional criterion for the detection of the *elbow*. Formally a criterion detecting significant changes for $\frac{\Delta MSE}{\Delta k}$ is required. These criteria tend to be unreliable [37]. For the silhouette coefficient it is also difficult to detect a reliable number of clusters. This is because the local maximum before a monotonic increase determines the optimal number of clusters and this maximum can be ambiguous, as shown in Figure 2.

Besides the optimal number of clusters the clustering results are of importance for the adaptive deployment. In scenario 1 eight clusters are existing, all with the same size. In Table 5 the clustering results are evaluated. As similarity measurement we employ the Jaccard index.

In scenario 1 *PAM* performed best. This is as expected since a particular strength of centroid based clustering algorithms are equally sized clusters. However, in networks an equal distribution of hard- and software cannot be assumed. To evaluate also heterogeneous environments, scenario 2 features an unequal distribution of systems.

It can be seen, that *SLINK* and *DBSCAN* outperform *PAM* clearly. This result was expected since connection and density based algorithms are better suited for unequal sized clusters. The obtained results in our experiment suggest, that *SLINK* in combination with the elbow method or *GAP* produce the best results. However, since we did not compare *SLINK* with other connection based clustering algorithms, it is possible that other algorithms outperform the single linkage algorithm. The proposed method of an adaption of the *DS* to the context by observing and scanning the network and determining prevalent systems to mimic is possible by an employment of the investigated methods.

6 Conclusion and Discussion

In this work the authors proposed an adaptive framework for the deployment of deception systems for cyber defense. The proposed framework is implemented for an evaluation. Different algorithms and convergence criteria are evaluated in different aspects such as computational time, determination of the number of clusters and the cluster accuracy. The focus of the implementation are server-side deception systems. However, the framework can easily be extended to feature also token based deception systems. We found that *SLINK* provides the best results. Even though the lowest error was achieved for the elbow convergence criteria, we recommend to consider *GAP* in this application because of its robustness and the simple determination of the global maximum. The adaptive deployment framework enables deception based security mechanisms for a broad range of users and a significant decrease in configuration, deployment and maintenance effort of such systems. It provides an enhanced security concept in a simple to use solution.

Conflict of Interest The authors declare no conflict of interest.

Acknowledgment This work has been supported by the Federal Ministry of Education and Research of the Federal Republic of Germany (Foerderkennzeichen KIS4ITS0001, IUNO). The authors alone are responsible for the content of the paper.

References

[1] Federal Bureau of Investigation. Internet crime report 2015. 2015.

[2] Europol. Internet organized crime threat assessment. 2016.

[3] Bundesamt für Sicherheit in der Informationstechnik. Die Lage der IT-Sicherheit in Deutschland 2015. 2015.

[4] M. Guri, G. Kedman, A. Kachlon, and Y. Elovici. Airhopper: Bridging the air-gap between isolated networks and mobile phones using radio frequencies. *International Conference on Malicious and Unwanted Software*, 9, 2014.

[5] L. Spitzner. Dynamic honeypots. 2003.

[6] D. Fraunholz, F. Pohl, and H. Schotten. Towards basic design principles for high- and medium-interaction honeypots. *European Conference on Cyber Warfare and Security*, 16, 2017.

[7] R. Grimes. *Honeypots for Windows: Configure and Manage Windows Honeypots*. 2005.

[8] Sun-tzu and S. Griffith. *The Art of War*. 1964.

[9] C. Stoll. *The cuckoo's egg: Tracking a spy through the maze of computer espionage*. 1990.

[10] L. Spitzner. Honeypots: catching the insider threat. *Annual Computer Security Applications Conference*, 19:170–179, 2003.

[11] D. Fraunholz, M. Zimmermann, S. Duque Antón, S. Schneider, and H. Schotten. Distributed and highly-scalable wan network attack sensing and sophisticated analysing framework based on honeypot technology. *International Conference on Cloud Computing, Data Science & Engineering*, 7, 2017.

[12] D. Fraunholz, D. Krohmer, S. Duque Antón, and H. Schotten. Investigation of cyber crime conducted by abusing weak or default passwords with a medium interaction honeypot. *International Conference On Cyber Security And Protection Of Digital Services*, 2017.

[13] D. Moran. Trapping and tracking hackers: Collective security for survival in the internet age. *Information Survivability Workshop*, 3, 2000.

[14] B. Scottberg, W. Yurcik, and D. Doss. Internet honeypots: Protection or entrapment? 2003.

[15] G. Wagener, R. State, A. Dulaunoy, and T. Engel. Heliza: talking dirty to the attackers. *Computer Virology*, 2010.

[16] G. Wagener. *Self-Adaptive Honeypots Coercing and Assessing Attacker Behaviour*. PhD thesis, Universite du Luxembourg, 2011.

[17] G. Wagener, R. State, T. Engel, and A. Dulaunoy. Adaptive and self-configurable honeypots. *IFIP/IEEE International Symposium on Integrated Network Management*, 12:345–352, 2011.

[18] V. Chowdhary, A. Tongaonkar, and T. Chiueh. Towards automatic learning of valid services for honeypots. *International Conference Distributed Computing and Internet Technology*, 1.

[19] C. Hecker, K. Nance, and B. Hay. Dynamic honeypot construction. *Colloqium for Inforation Systems Security Education*, 10:95–102, 2006.

[20] C. Leita, K. Mermoud, and M. Dacier. Scriptgen: an automated script generation tool for honeyd. *Annual Computer Security Applications Conference*, 21, 2005.

[21] W. Zakaria and L. Kiah. A review on artificial intelligence techniques for developing intelligent honeypot. 2012.

[22] W. Zakaria and L. Kiah. A review of dynamic and intelligent honeypots. *ScienceAsia*, 2013.

[23] L. Kaufmann and Rousseeuwm R. Clustering by means of medoids. 1987.

[24] P. Jaccard. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37(547-579), 1901.

[25] P. Mahalanobis. On the generalised distance in statistics. *Proceedings of the National Institute of Science of India*, pages 49–55, 1936.

[26] J. Gower and G. Ross. Minimum spanning trees and single linkage cluster analysis. *Journal of the Royal Statistical Society*, pages 54–64, 1969.

[27] R. Sibson. Slink: an optimally efficient algorithm for the single-link cluster method. *The Computer Journal*, pages 30–34, 1973.

[28] M. Ester, H. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. *International Conference on Knowledge Discovery and Data Mining*, 2, 1996.

[29] R. Thorndike. Who belongs in the family? *Psychometrika*, 18:267–276, 1953.

[30] R. Tibshirani, Waltherm G., and T. Hastie. Estimating the number of clusters in a dataset via the gap statistic. 2000.

[31] P. Rousseuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Computational and Applied Mathematics*, pages 53–65, 1987.

[32] M. Zalewski. p0f v3: Passive fingerprinter. 2012.

[33] F. Vaskovich. The art of port scanning. *Phrack Magazine*, 7, 1997.

[34] A. Ofir and F. Yarochkin. Xprobe2 - a fuzzy approach to remote active operating system fingerprinting. 2003.

[35] N. Provos. Honeyd: A virtual honeypot daemon. *DFN-CERT Workshop*, 10, 2003.

[36] R. Chandran and S. Pakala. Simulating networks with honeyd. 2003.

[37] D. Fraunholz, M. Zimmermann, and H. Schotten. An adaptive honeypot configuration, deployment and maintenance strategy. *International Conference on Advanced Communication Technology*, 19, 2017.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 2, No. 3, 1280-1289 (2017)
www.astesj.com
Special Issue on Recent Advances in Engineering Systems

ASTES Journal
ISSN: 2415-6698

Classification of patient by analyzing EEG signal using DWT and least square support vector machine

Mohd Zuhair[*,1], Sonia Thomas[2]

[1]*RMSoEE, IIT Kharagpur, 721302, India*

[2]*Department Computer Science, State University of New York (SUNY) Korea, 21985, South Korea*

ARTICLEINFO

Article history:
Received: 25 May, 2017
Accepted: 15 July, 2017
Online: 01 August, 2017

Keywords:
ApEn
LS-SVM
EEG
DWT
Kernel

ABSTRACT

Epilepsy is a neurological disorder which is most widespread in human beings after stroke. Approximately 70% of epilepsy cases can be cured if diagnosed and medicated properly. Electro-encephalogram (EEG) signals are recording of brain electrical activity that provides insight information and understanding of the mechanisms inside the brain. Since epileptic seizures occur erratically, it is essential to develop a model for automatically detecting seizure from EEG recordings. In this paper a scheme was presented to detect the epileptic seizure implementing discrete wavelet transform (DWT) on EEG signal. DWT decomposes the signal into approximation and detail coefficients, the ApEn values the coefficients were computed using pattern length (m= 2 and 3) as an input feature for the Least square support vector machine (LS-SVM). The classification is done using LS-SVM and the results were compared using RBF and linear kernels. The proposed model has used the EEG data consisting of 5 classes and compared with using the approximate and detailed coefficients combined and individually. The classification accuracy of the LS-SVM using the RBF and Linear kernel with ApEn using different cases is compared and it is found that the best accuracy percentage is 100% with RBF kernel.

1. Introduction

Epilepsy in humans is an intrinsic brain pathology and its major manifestation is epileptic seizures. Epileptic seizures may affect partial part of the brain (partial) or the whole cerebral mass (generalized), seizures are recurrent with interictal period ranging from several minutes to several days. Brain's electrical activity is measured through the electroencephalogram (EEG) signal which is an effective tool for studying functioning of brain and diagnosing epilepsy [1]. EEG signals are a non-invasive testing method that provides valuable details of distinct physiological states of the brain. The data recorded usually are of long duration and inspected by experts to analyze the huge data recorded in the form of EEG signal to detect epilepsy.

The advanced signal processing has enabled to process and store the EEG signal digitally. The processed EEG data is then feed as an input for the automatic detection system to detect the traces of epilepsy. The automatic system reduces the workload of neurologists by reducing the amount of effort and time required to detect the traces of epilepsy in the recorded EEG signal. Automatic prediction and detection of epilepsy from the EEG signal are developed using different signal processing techniques like frequency domain analysis, wavelet analysis, spike detection, and non-linear methods. Gotman [2] has presented an automatic detection system for epilepsy by decomposing the EEG signal into elementary waves. Srinivasan et al., [3] has used time domain and frequency domain analysis to detect epilepsy the author has selected five features out of which two features were from frequency-domain and three features were from time-domain. The author has used the recurrent neural network for detecting epilepsy and the neural net-

[*]Mohd Zuhair, RMSoEE, IIT Kharagpur, 9038273634 & md.zuhair.cs@gmail.com

https://dx.doi.org/10.25046/aj0203162

work was trained and tested for the epileptic EEG signal with an accuracy of 99.6%, Keshri et al., [4] has used the slope of the lines between each pair of two consecutive data points (x1, y1) and (x2, y2) and feed it into Deterministic Finite Automata and got the accuracy level as high as 95.68%. Geva et al., [5] has used wavelet analysis as both time and frequency domain view can be provided with the use of WT.

Various methods are available to detect and predict the epileptic seizure. Artificial Neural Network (ANN) has been widely used for detecting the epileptic spike [1, 6, 7, 8, 9]. Features extraction is an important aspect in the performance of ANN models as the model is trained in and tested on these extracted features. ApEn was proposed by Pincus [10] as a statistical parameter to measure the regularity of time series data. ApEn is being predominantly used in the electrocardiogram and other related heart rate data analysis [11, 12, 13] as well as in the analysis of endocrine hormone [14]. It is the measure of regularity as smaller value of ApEn depicts a high regularity and higher value of ApEn depicts low regularity in time series data [15]. Diambra et al., [16] has denoted that parameter ApEn gives the valuable temporal localization of a variety of epileptic activity. Elman, PNN and SVM network for detection of epilepsy through ApEn based feature with 100% overall accuracy. Hence, it is an acceptable feature for automated detection of epilepsy.

The Support Vector Machine (SVM) is a classier method that performs classification tasks by constructing hyperplanes in a multidimensional space which try to find a combination of samples to build a plane maximizing the margin between two classes. SVM is widely used in epileptic detection and prediction [17] has used permutation entropy as a parameter, as it drops during the seizure interval. The Burg Burg AR coefficients has been used as an input for SVM that shows accuracy of 99.56% [18].

In this paper an automated detection of the epileptic seizure was discussed using LS-SVM comparing two kernel functions RBF and linear. In this work EEG signal were decomposed into six sub-bands namely D1-D5 and A5 using DWT. The analysis of complexity in the sub-bands are done by ApEn which acts as an input feature for SVM. Experiments are done using different cases having a different combination of EEG data sets.

2. Proposed Method

In this paper there are four main tasks in sections 2.1) Clinical data 2.2) Preprocessing EEG data through DWT to decompose into several sub-bands in section 2.3) Feature extraction (ApEn) in section 2.4) Classification of EEG data using feature (ApEn). Figure 1 shows the proposed approach used in this paper.

2.1. Clinical data

The data set used in this paper is available in public domain and is accessed online from the University of Bonn, Germany which consists of five different data set of EEG data [19]. The data used is artifact free EEG time series data and is widespreadly used by the ongoing research on epilepsy [1, 3, 7, 15, 17, 20]. The complete data set contains five sets marked as (A-E) each set has 100 single channel EEG segment having duration of 23.6-sec. Each segments of data are selected and cut out after visual inspection for artifacts from continuous mechanical EEG recordings with the sampling frequency of 173.61 Hz with the band pass settings of 0.53-40 Hz. The data contains three different classes normal, epileptic background (pre-ictal), and epileptic seizure (ictal). The normal EEG data (A and B) was collected from five healthy volunteers. The pre-ictal EEG data (C and D) was recorded during the period when there were no traces of seizure from five epileptic patients. The ictal EEG data (E) was recorded during the epileptic seizure from the same five patients. EEG signals were recorded from 128- channel amplifier using an average common reference and digitized at 173.61 Hz sampling rate and 12-bit A/D resolution. Figure 2 and Figure 3 shows the specimen of an epileptic and normal signal.

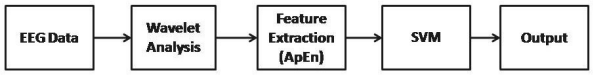


Figure 1. *Proposed Approach*

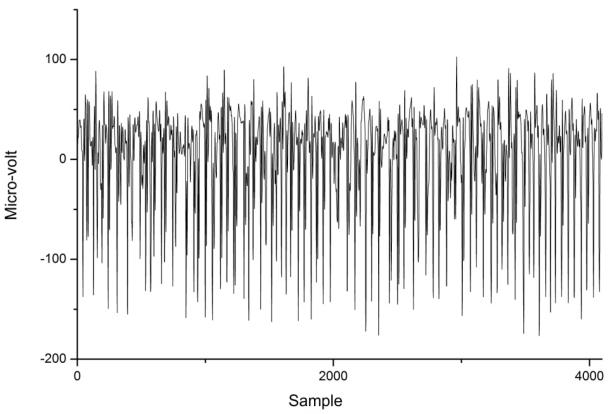


Figure 2. *Epileptic Signal*

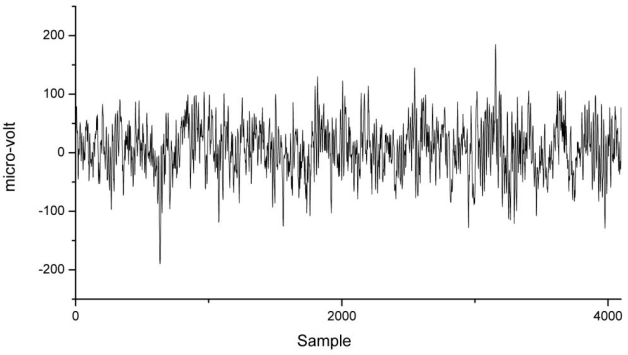


Figure 3. *Normal Signal*

2.2. Discrete Wavelet Transform

Wavelet transform (WT) uses variable window size, has well-known data compression and time-frequency filtering capabilities. It provides a good local representation of the signal in time as well as in the frequency domain which makes it as an effective tool for analyzing the signal and extracting features. The wavelet transform looks for the spatial distribution of singularities whereas fourier transform provides a description of the overall regularity of signals [5]. WT captures transient features and localizes them in time as well as frequency content and is widely used in epileptic seizure detection [21, 22]. WT uses long time windows for obtaining finer low-frequency resolution as well as the short time windows for obtaining high-frequency information. Thus, WT provides specific frequency information at low frequencies and specific time information at high frequencies [23].

Discrete wavelet transform (DWT) x(t), is defined as,

$$DWT(j,k) = \frac{1}{\sqrt{|2^j|}} \int_{-\infty}^{\infty} x(t)\psi\left(\frac{t-2^j k}{2^j}\right) dt \qquad (1)$$

Where ψ, a and b are wavelet function called as scaling and shifting parameters. The quadrature mirror filters are used to realize the scheme by passing the passing the signal through a series of low-pass (LP) and high-pass (HP) filter [24] The out put from the low pass filter is termed as . Approximation (A) and output from the high pass filter id termed as detail (D) coefficients. The figure illustrates the fifth level wavelet decomposition of a signal showing the coefficients A1, D1, A2, D2, A3, D3, A4, D4, A5, and D5 .

In this model fifth-level wavelet decomposition was performed on normal subjects data (A, B, C and D) as well as on epileptic patient data (E) using Daubechies order 4 wavelet (db4). The researchers have found that db4 wavelet is most appropriate for the analysis of epileptic EEG data [25]. The structure for the wavelet decomposition at each level with its approximation and detail coefficients are shown in Figure 4.

2.3. Approximate entropy (ApEn)

Approximate entropy is a statistical feature for quantifying regularity and complexity in a time-series data [10]. ApEn is a non-negative number and has been successfully applied in the field of pattern recognition. ApEn have potential application throughout medicine and prominently in ECG and EEG [14].The following steps determine the value of ApEn [1, 8, 26, 27, 28].

1. The data set contains N data points let it be A=[a(1), a(2), . . . , a(N-m+1)].
2. Let a(i) be a subsequence of A such that a(i) =[a(i), a(i+1), a(i+2), . . . , a(i+m-1)] for $1 \leq i \leq N-m+1$, m is the number of samples used
3. Denote the distance between A(i) and A(j) by d[A(i), A(j)]=

$$\max_{k=0,\ldots,m-1} (|\, a(i+1) - a(j+1)\,|) \qquad (2)$$

4. For a given A(i), number of j is counted as ($j = 1,\ldots,N-m+1, j \neq i$) such that $d[A(i),A(j)] \leq r$ denoted as $M^m(i)$ for $i = N-m+1$

$$C_r^m(i) = \frac{M^m(i)}{(N-m+1)}, \qquad (3)$$

for $i = 1,\ldots,N-m+1$

5. For each $C_r^m(i)$ compute natural logarithm, and average it over i

$$\Phi^m(r) = \frac{1}{N-m+1} \sum_{i=1}^{N-m+1} \ln C_r^m(i) \qquad (4)$$

6. . Repeat step (1) to step (4) by increasing the dimension m to m+1 and find $C_r^{m+1}(i)$ and $\Phi^{m+1}(r)$
7. Finally ApEn is computed as

$$ApEn(m,r,N) = \Phi^m(r) - \Phi^{m+1}(r) \qquad (5)$$

To compute the ApEn value of the signal of length N two parameters length of the compared run m tolerance window r are specified. We have taken the value of m=2 and 3 and r is in between 0.1 to 0.25 times the standard deviation of data. In this model the ApEn values of the approximate (A1-A5) and detailed coefficients (D1-D5) are computed using the length of the compared run (m), tolerance window (r) were set to m= (2,3) and r= (0.2)*standard deviation of the data to compute ApEn.

2.4. Support vector machine (SVM)

Support vector machine (SVM) has been used in several EEG signal classification problems [17, 20] and was first introduced in 1995. SVMs belong to the family of kernel-based classifiers and are extremely powerful classifiers. Linear as well as non-linear classification can be performed in SVM using different kernel functions [29]. The approach of SVMs is to implicitly map the classification data into higher dimension input space where a hyperplane separating the classes may exist. The implicit mapping is achieved through different Kernel functions. In the case of linear SVM to classify linearly separable data, the training data, $\{a_i, b_i\}$ for $i = 1,,m$ and $y_i \in \{-1,1\}$ then the following decision function is determined by [18]:

$$D(x) = w^t g(x) + y \qquad (6)$$

g(x)is a mapping function that maps x into l-dimensional space, y is a scalar and w is the l-dimensional vector. The decision function satisfies the following condition to separate the data linearly:

$$b_i\left(w^t g(a_i) + y\right) \geq 1 \quad for \quad i = 1,,M \qquad (7)$$

There are infinite number of decision functions that satisfy Eq.(8) for a linearly separable feature space. The largest between the two classes are selected between the two classes. The margin given by $D|x|/\|w\|$. Let the margin is ρ then the following condition are required to be satisfied:

$$\frac{b_i D(a_i)}{\|w\|} \geq \rho \quad for \quad i = 1,,M \tag{8}$$

The product of ρ and $\|w\|$is fixed

$$\rho\|w\| = 1 \tag{9}$$

In order to obtain the maximum margin for the optimal separating hyperplane, $\bar{w}$ with the minimum $\|w\|$ that satisfying Eq. (9) should be found from Eq. (10), This provides optimization problem as follows. Minimizing

$$\frac{1}{2}w^t w \tag{10}$$

subject to the constraints: $b_i\left(w^t g(a_i)+y\right) \geq 1 \quad for \quad i = 1,,M$ When training data are not linearly separable a slack variables is introduced $\xi_{\mathbf{i}}$ into Eq. (8)

$$b_i\left(w^t g(a_i)+y\right) \geq 1-\xi_i, \quad \xi_i \geq 0 \quad for \quad i = 1,,M \tag{11}$$

To maximize the margin and minimize the training error the optimal separating hyperplane is determined and is achieved by minimizing,

$$\frac{1}{2}w^t w + \frac{C}{2}\sum_{i=1}^{n}\xi_i^p \tag{12}$$

subject to the constraints:$b_i\left(w^t g(a_i)+b\right) \geq 1 - \xi_i, \quad \xi_i \geq 0 \quad for \quad i = 1,,M$

The tradeoff between the maximum margin and minimum classification error is determined by the parameter C.

2.5. Least squares support vector machines (LS-SVMs)

The LS-SVM are trained by minimizing

$$\frac{1}{2}w^t w + \frac{C}{2}\sum_{i=1}^{n}\xi_i^2 \tag{13}$$

subject to the equality constraints:

$$b_i\left(w^t g(a_i)+y\right) = 1-\xi_i, \quad \xi_i \geq 0 \quad for \quad i = 1,,M \tag{14}$$

The conventional SVM uses inequality constraints where as in the LS-SVM equality constraints are used. The equality constraint has reduced the complexity to obtain the optimal solution by solving a set of linear equations rather than solving a quadratic programming problem. To derive the dual problem of Eqs. (14) and (15) Lagrange multipliers are used

$$Q(w,y,\alpha,\xi) = \frac{1}{2}w^t w + \frac{C}{2}\sum_{i=1}^{n}\xi_i^2 - \sum_{i=1}^{M}\alpha_i\left\{y_i\left(w^t g(a_i)+y\right)-1+\xi_i\right\} \tag{15}$$

where $\alpha = (\alpha_i,...,\alpha_M)^t$ is Lagrange multipliers and by differentiating the above equation with respect to w, ξ_i, b, and α_i and equating the resulting equations to zero the the conditions for optimality are determined [30].

2.5.1. Kernel Function

Classically SVMs were designed to classify the data in linear space, in the nonlinear space SVMs do not preformed well to overcome this limitation on SVMs, kernel approaches were developed. The following kernels are most commonly used [31].

1. Linear Kernels: $k(a,a') = (a^t a')$;
2. Polynomial kernels: $k(a,a') = ((a^t a')+1)^y$ where d is a positive intege r
3. RBF kernels: $k\left(a,a'\right) = \exp\left(-\gamma\left\|a-a'\right\|\right)$

To train a SVM classifier, the user has to determine a suitable kernel function, optimum hyper parameters, and proper regularization parameter. in this paper we have used two kernels 1) Linear Kernel and 2) RBF kernel. The goal to achieve optimum hyper parameter and regularization parameter is accomplished by cross-validation technique. The cross-validation technique can be used to select parameters.

2.6. Cross-Validation

Cross-validation is a validation technique used to determine the quality of the classification model. It partitions a sample data into different subsets such that the analysis is initially implemented on a single subset. The remaining subset(s) are kept for validating the result of initial analysis. The data subset used for initial test is called as training set while the other subsets are called as testing or validation sets [32]. In K-Fold cross validation data is partitioned into k roughly equal size sets and each set is used once as a test set while other remaining sets are used as training sets. For each k=1,2,...N, fit the model parameter for other retained K-1 parts. The cross-validation procedure is repeated for K number of times using each of the K set exactly once as validation data. The average of the K result obtained from the folds produces a single estimation. In this paper, we have used 10-fold scheme to achieve best performance efficiency.

2.7. Performance evaluation parameters

The performance of LS-SVM is estimated by using the parametres, namely, sensitivity (SE), specificity (SP) and overall accuracy (OA) defined as:

$$SE(\%) = \frac{TN_{CP}}{TN_{AP}} * 100 \tag{16}$$

Where TN_{CP} denotes the count of correctly detected positive patterns and TN_{AN} denotes the count of actual positive pattern. The positive pattern represents a detected seizure.

$$SP(\%) = \frac{TN_{CN}}{TN_{AN}} * 100 \tag{17}$$

Where TN_{CN} denotes the count of correctly detected negative pattern and TN_{AN} denotes the actual count of negative pattern. The negative pattern represents a detected non-seizure.

$$OA(\%) = \frac{TN_{CDP}}{TN_{APP}} * 100 \quad (18)$$

Where TN_{CDP} denotes the count of correctly detected pattern and TN_{APP} denotes the count of applied patterns [1].

3. Design of Experiment

The data is processed and the ApEn values of the detailed and approximate coefficients are computed. The design of experiment is done to make different cases of the processed data-set and then compare the results of different cases the cases are as follows

- Case 1: Data set A versus Data set E
- Case 2: Data set B versus Data set E
- Case 3: Data set C versus Data set E
- Case 4: Data set D versus Data set E
- Case 5: Data set A, C and D versus Data set E
- Case 6: Data set A, B, C and D versus Data set E

These cases are then modeled combining all the coefficients A5 and D1-D5 in one set as an input to LS-SVM and using different coefficients individually as an input to LS-SVM. The results were analyzed using all different cases and combinations and compared to find the model providing the best efficiency.

4. Result and Discussion

The EEG data set is decomposed into different sub bands by applying DWT using db4 wavelet having 5 level of decomposition shown in Figure 4. The frequency ranges of these sub-bands are: A1(043.4 Hz), A2 (021.7Hz), A3 (010.85Hz), A4 (05.43 Hz), A5 (02.70Hz), D1 (43.486.8 Hz), D2 (21.743.4 Hz), D3 (10.8521.7Hz), D4(5.4310.85Hz) and D5 (2.705.43 Hz. ApEn values have been computed from the approximation and detail coefficients of each sub-bands of entire 500 EEG epochs of five data sets A-E. Figure 5 and 6 shows the decomposition of EEG signal epoch of data set A and E.

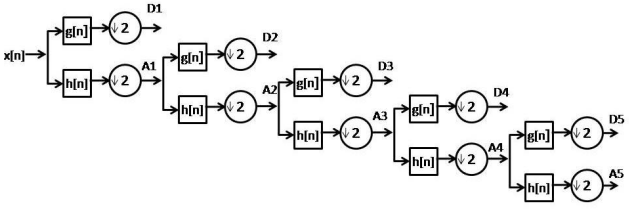


Figure 4. *Decomposition of EEG signal using fifth level decomposition*

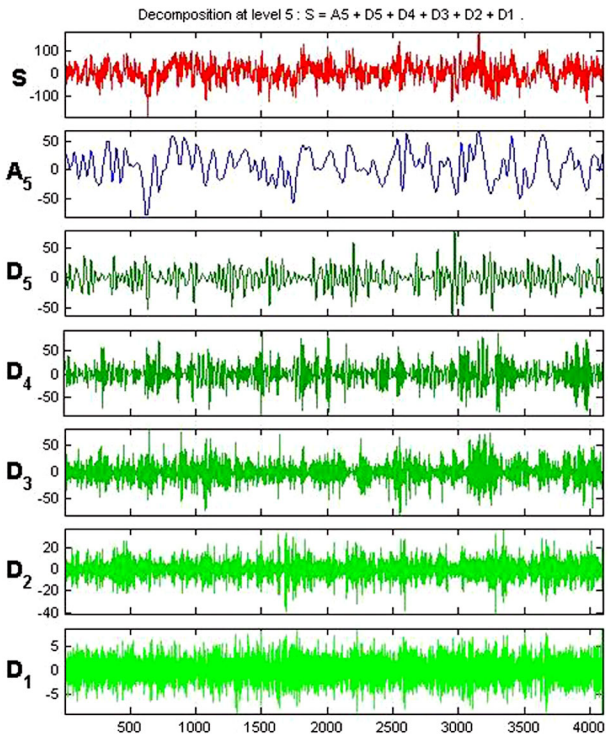


Figure 5. *Wavelet decomposition of Epileptic signal (Data Set E)*

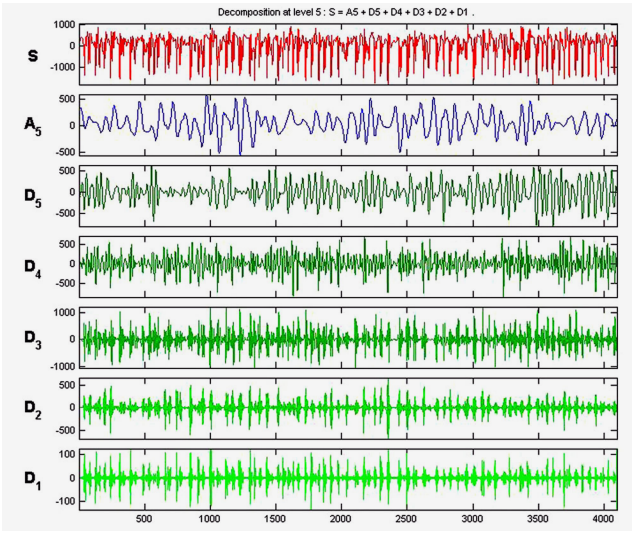


Figure 6. *Wavelet decomposition of Normal Signal (Data Set A)*

The ApEn values of the approximate and detailed coefficients are computed from the entire data set A-E consisting 100 epochs having parameters m=(2,3) and r=0.2*standard deviation of data set. ApEn value of the detailed coefficient D1 from data set A, B which were recorded from the surface of the scalp of the normal subject while they are in a relaxed and an awake state with (Data Set A) eyes open and (Data set B) eyes closed vs E are recorded from the epileptic subjects through intracranial electrodes and having embedded

dimension m=2 is shown in Figure7. The ApEn values for data set A and B are higher than that of data set E which means that the data set E is more ordered and periodic than set A and B. Similarly Figure8 shows the Apen values of detailed coefficient D1 from data set C and D v/s data set E recorded from intracranial electrodes having embedded dimension m=2. The ApEn values of data set C and D are also higher than that of data set E which that data set A,B,C,D are more complex than data set E. Figure 7 and Figure 8 shows that the complexity of the data set A recorded from normal subject and data set C which is recorded from opposite to the epileptogenic zone are almost similar. ApEn values for embedded dimension m=3 are shown in Figure 9 and Figure 10 depicting similar attributes. The calculated value of ApEn got reduced in this case than that of having embedded dimension m=2 the curves and variation of data also got reduced. Data set D and E are overlapping in both the case as the data set is acquired within the epileptogenic zone.

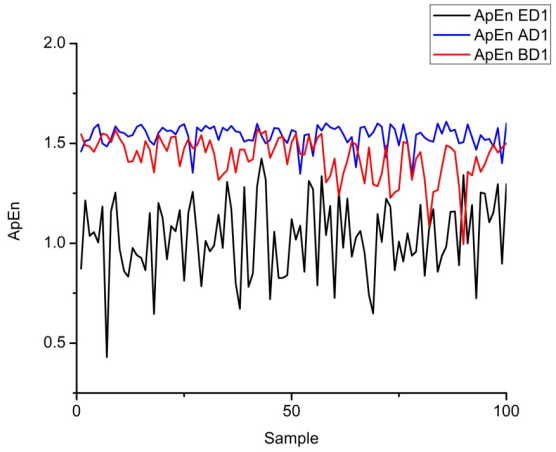


Figure 7. *Epileptic Signal*

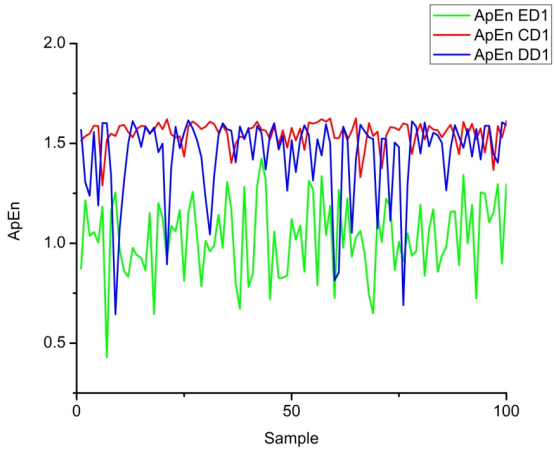


Figure 8. *Normal Signal*

In this paper SVM is implemented by using MATLAB *R2012b* and LS-SVM toolbox [33].The input feature vector ApEn is divided into two parts training data set 60% and testing 40%. The training data set is used to train the SVM while testing data set is used for verifying the accuracy of the trained SVM. For dividing the data set into training and testing part we have used holdout method of cross validation. The SVM is initialized by initlssvm function trained using tunelssvm and trainlssvm function. The performance parameters were tuned by using tunelssvm function for regularization and kernel parameter (gam, sig2) of LS-SVM [34].

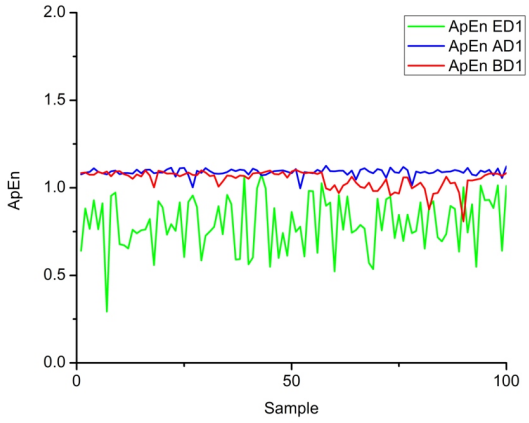


Figure 9. *Normal Signal*

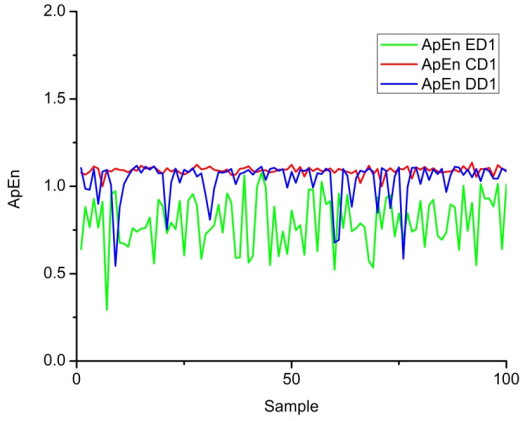


Figure 10. *Normal Signal*

The SVM algorithm is used with linear and Gaussian radial basis kernel functions. Linear kernel function require gamma parameter for training the SVM. SVM with rbf kernel function require gamma as well as sigma parameter which has to be selected based on training data. In this paper we have set gamma $\in [0-1]$ for linear Kernel and gamma $\in [0-10]$ for rbf Kernel. The sigma parameter for rbf kernel is $\in [0.7-9]$ using tunelssvm function. Each row of the input data matrix is one observation and its column is one feature.

The feature vector of each data set has 100 rows (epochs) and 6 columns (D1-D5) and A5. Case $(1-4)$ consist 200 observations, Case 5 consist of 400 observations, Case 6 consists 500 observations. The obser-

Table 1. Statistical parameters with ApEn embedded dimension m=2 linear kernel

Embedding dimension (m=2)	SVM (Linear)					
Cases for seizure detection	D1	D2	D3	D4	D5	A5
Case 1 (A-E)	0.9875	0.9375	0.9375	0.8625	0.55	0.95
Case 2 (B-E)	0.9375	0.9375	0.7375	0.7625	0.575	0.9125
Case 3 (C-E)	0.9875	0.975	0.875	0.6375	0.775	0.8625
Case 4 (D-E)	0.95	0.7875	0.7	0.5875	0.7625	0.825
Case 5 (ACD-E)	0.925	0.8875	0.8063	0.7562	0.7813	0.9063
Case 6 (ABCD-E)	0.94	0.89	0.83	0.81	0.805	0.925

Table 2. Statistical parameters with ApEn embedded dimention m=2 rbf

Embedding dimension (m=2)	SVM (RBF)					
Cases for seizure detection	D1	D2	D3	D4	D5	A5
Case 1 (A-E)	0.9875	0.95	0.975	0.875	0.5875	0.9
Case 2 (B-E)	0.9625	0.9375	0.75	0.7875	0.6	0.9
Case 3 (C-E)	0.9875	0.9625	0.925	0.675	0.85	0.875
Case 4 (D-E)	0.925	0.75	0.7375	0.5875	0.825	0.875
Case 5 (ACD-E)	0.9437	0.9125	0.825	0.7688	0.8063	0.9063
Case 6 (ABCD-E)	0.945	0.91	0.84	0.84	0.815	0.93

vations are normalized by scaling between 0 and 1 and then 60% and training and 40% of testing data is used for training and testing SVM.

- The Case when D1-D5 and A5 is used individually as an Input to LS-SVM

The highest precision accuracy of for embedded dimension $m = 2$ are found in cases 1 and 3 for detailed coefficient D1 are 98.75% using linear kernel and same for rbf kernel ie. in cases 1 and 3 are 98.75%. The lowest accuracy was found to be 55% for case 1 for detailed coefficient D5 in linear kernel and 59% for case 1 for detailed coefficient D5 in rbf kernel. The results suggests that there is marginal differences between the results of rbf and linear kernel but the overall average efficiency remains same for both in this cases as shown in the Table 1 and Table 2. The highest accuracy of for embedded dimension $m = 3$ is found in cases 1, 2 and 3 for detailed coefficient D1 are 96.25% and for RBF kernel it is 98.75% for cases 1 and 3 for detailed coefficient D1. In this modeling RBF kernel shows better results than that of linear kernel the overall result is summarized in Table 3 and Table 4.

- The Case when D1-D5 and A5 is combined as an Input to LS-SVM

The precision of the proposed system for embedded dimension m = 2 with linear kernel is 100% which is maximum and a lowest of 98:12% respectively for cases 1, 3, 4, 5 and case 6. Similarly, the precision of the proposed system for embedded dimension m = 2 with RBF kernel is 100% maximum and lowest of 99.50% respectively for cases 1-5 and case 6. The precision of the proposed system for embedded dimension m = 3 with linear kernel is 100% maximum and a lowest of 97.0% respectively for cases 1, 3, 4 and case 6. Similarly, the precision of the proposed system for embedded dimension m = 3 with RBF kernel is 100% maximum and a lowest of 99.5% respectively for cases 1, 3 – 5 and case 6. It is clearly observed in the results summarized in Table 6 and Table 7 that RBF kernel based automatic epileptic seizure detection system gives better precision than linear kernel based automatic epileptic seizure detection system. Table 5 presents the the comparison of results between the existing and the proposed method using same data set.

5. Conclusion and Future Scope

The least squares version of support vector machine classifiers is discussed in this paper. The quadratic programming problem has been eased to solving a set of linear equations with the use of equality constraint instead of inequality constraint. In this paper we have modeled LS-SVM using EEG data set with different combinations of input features as well as changing the parameters of the ApEn. The results suggest that the combination of detailed and approximate coefficients ie. D1-D5 and A5 as an input to classifier produces better results than using single feature D1 as a classification input. The kernel function also plays significant role in the classifying the EEG data set using LS-SVM and RBF kernel produces better results than the linear kernel in this modelling. The embedded dimension m for calculation the ApEn is also significant ad this model suggests that at m=2 the classification system performs better than that at m=3. The proposed approach can be deployed as a quantitative measure for monitoring EEG signal associated with epilepsy. As an extension of the proposed method, it would be challenging to scrutinize the efficacy of the proposed method for other neurological disorders which uses brain signals for analyzes such as Parkinson diseases etc [35]. Furthermore, it would be interesting to analyze the learning effectiveness of this model on other database. In this study, the

Table 3. Statistical parameters with ApEn embedded dimention m=3 lin

Embedding dimension (m=3)	SVM (Linear)					
Cases for seizure detection	D1	D2	D3	D4	D5	A5
Case 1 (A-E)	0.9625	0.9375	0.95	0.9875	0.7125	0.925
Case 2 (B-E)	0.9625	0.8875	0.775	0.7125	0.7375	0.875
Case 3 (C-E)	0.9625	0.95	0.8625	0.8	0.6375	0.875
Case 4 (D-E)	0.9125	0.8	0.7	0.7125	0.6	0.7625
Case 5 (ACD-E)	0.9	0.85	0.8	0.775	0.8063	0.875
Case 6 (ABCD-E)	0.905	0.885	0.85	0.83	0.84	0.895

Table 4. Statistical parameters with ApEn embedded dimention m=3 rbf

Embedding dimension (m=3)	SVM (RBF)					
Cases for seizure detection	D1	D2	D3	D4	D5	A5
Case 1 (A-E)	0.9875	0.9375	0.9375	0.9875	0.7	0.925
Case 2 (B-E)	0.975	0.925	0.825	0.7875	0.7375	0.85
Case 3 (C-E)	0.9875	0.9625	0.875	0.8125	0.65	0.875
Case 4 (D-E)	0.9125	0.825	0.775	0.7375	0.65	0.7875
Case 5 (ACD-E)	0.9437	0.885	0.85	0.8438	0.8187	0.8938
Case 6 (ABCD-E)	0.95	0.91	0.85	0.83	0.845	0.895

Table 5. Comparison with Existing Models

Authors	Year	Method	Cases	Maximum OA %
Srinivasan et al. [3]	2005	Time and frequency domain features using ANN	Case 1	99.60
Srinivasan et al. [1]	2007	Approximate Entropy using ANN	Case 1	100
Ocak [7]	2009	DWT and (ApEn)	Case 1 Case 9	99.60 96.65
Guo et al. [8]	2010	Discrete wavelet transform using ANN	Case 1 Case 9	99.85 98.27
Ubeyli [18]	2010	LS-SVM model-based method coefficients	Case 1	99.56
Nicolaou et al [17]	2012	Permutation entropy using SVM	Case 1 Case 2 Case 3 Case 4	93.55 82.88 88.00 79.94
Kai et al. [36]	2014	Time-frequency image using SVM	Case 1	99.125
Anindya et al. [37]	2016	Dual tree complex wavelet transform using SVM	Case 1 Case 3 Case 4 Case 9	100 100 100 100
Proposed Method		DWT based ApEn and Artificial neural network, LS-SVM	Case1 Case2 Case3 Case4 Case5 Case6	100 100 100 100 100 99.5

Table 6. Results obtained for proposed model with embedded dimension m=2

Embedding dimension (m=2)	SVM (Linear)			SVM (RBF)		
Cases for seizure detection	SE	SP	OA	SE	SP	OA
Case 1 (A-E)	1	1	1	1	1	1
Case 2 (B-E)	0.9722	1	0.9875	1	1	1
Case 3 (C-E)	1	1	1	1	1	1
Case 4 (D-E)	1	1	1	1	1	1
Case 5 (ACD-E)	1	0.9831	0.9875	1	1	1
Case 6 (ABCD-E)	0.9697	0.982	0.98	0.9756	1	0.995

Table 7. Results obtained for proposed model with embedded dimension m=3

Embedding dimension (m=2)	SVM (Linear)			SVM (RBF)		
Cases for seizure detection	SE	SP	OA	SE	SP	OA
Case 1 (A-E)	1	1	1	1	1	1
Case 2 (B-E)	0.9474	1	0.975	0.975	1	0.9875
Case 3 (C-E)	1	1	1	1	1	1
Case 4 (D-E)	1	1	1	1	1	1
Case 5 (ACD-E)	0.9524	0.9915	0.9812	1	1	1
Case 6 (ABCD-E)	0.9091	0.982	0.97	0.995	0.9939	0.995

proposed model has been tested on different datasets from the same database.

References

[1] V. Srinivasan, C. Eswaran, N. Sriraam, Approximate entropy-based epileptic eeg detection using artificial neural networks, Information Technology in Biomedicine, IEEE Transactions on 11 (3) (2007) 288–295.

[2] J. Gotman, Automatic recognition of epileptic seizures in the eeg, Electroencephalography and clinical Neurophysiology 54 (5) (1982) 530–540.

[3] V. Srinivasan, C. Eswaran, Sriraam, N, Artificial neural network based epileptic detection using time-domain and frequency-domain features, Journal of Medical Systems 29 (6) (2005) 647–660.

[4] A. K. Keshri, R. K. Sinha, R. Hatwal, B. N. Das, Epileptic spike recognition in electroencephalogram using deterministic finite automata, Journal of medical systems 33 (3) (2009) 173–179.

[5] A. B. Geva, D. H. Kerem, Forecasting generalized epileptic seizures from the eeg signal by wavelet analysis and dynamic unsupervised fuzzy clustering, Biomedical Engineering, IEEE Transactions on 45 (10) (1998) 1205–1216.

[6] A. Alkan, E. Koklukaya, A. Subasi, Automatic seizure detection in eeg using logistic regression and artificial neural network, Journal of Neuroscience Methods 148 (2) (2005) 167–176.

[7] H. Ocak, Automatic detection of epileptic seizures in eeg using discrete wavelet transform and approximate entropy, Expert Systems with Applications 36 (2) (2009) 2027–2036.

[8] L. Guo, D. Rivero, A. Pazos, Epileptic seizure detection using multiwavelet transform based approximate entropy and artificial neural networks, Journal of neuroscience methods 193 (1) (2010) 156–163.

[9] A. Subasi, J. Kevric, M. A. Canbaz, Epileptic seizure detection using hybrid machine learning methods, Neural Computing and Applications (2017) 1–9.

[10] S. M. Pincus, Approximate entropy as a measure of system complexity., Proceedings of the National Academy of Sciences 88 (6) (1991) 2297–2301.

[11] K.-J. Bär, M. K. Boettger, M. Koschke, S. Schulz, P. Chokka, V. K. Yeragani, A. Voss, Non-linear complexity measures of heart rate variability in acute schizophrenia, Clinical Neurophysiology 118 (9) (2007) 2009–2015.

[12] K. K. Ho, G. B. Moody, C.-K. Peng, J. E. Mietus, M. G. Larson, D. Levy, A. L. Goldberger, Predicting survival in heart failure case and control subjects by use of fully automated methods for deriving nonlinear and conventional indices of heart rate dynamics, Circulation 96 (3) (1997) 842–848.

[13] X. Hu, C. Miller, P. Vespa, M. Bergsneider, Adaptive computation of approximate entropy and its application in integrative analysis of irregularity of heart rate variability and intracranial pressure signals, Medical engineering & physics 30 (5) (2008) 631–639.

[14] S. M. Pincus, I. M. Gladstone, R. A. Ehrenkranz, A regularity statistic for medical data analysis, Journal of clinical monitoring 7 (4) (1991) 335–345.

[15] U. R. Acharya, F. Molinari, S. V. Sree, S. Chattopadhyay, K.-H. Ng, J. S. Suri, Automated diagnosis of epileptic eeg using entropies, Biomedical Signal Processing and Control 7 (4) (2012) 401–408.

[16] L. Diambra, J. de Figueiredo, C. P. Malta, Epileptic activity recognition in eeg recording, Physica A: Statistical Mechanics and its Applications 273 (3) (1999) 495–505.

[17] N. Nicolaou, J. Georgiou, Detection of epileptic electroencephalogram based on permutation entropy and support vector machines, Expert Systems with Applications 39 (1) (2012) 202–209.

[18] E. D. Übeyli, Least squares support vector machine employing model-based methods coefficients for analysis of eeg signals, Expert Systems with Applications 37 (1) (2010) 233–239.

[19] R. G. Andrzejak, K. Lehnertz, F. Mormann, C. Rieke, P. David, C. E. Elger, Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state, Physical Review E 64 (6) (2001) 061907.

[20] Y. Kumar, M. L. Dewal, R. S. Anand, Relative wavelet energy and wavelet entropy based epileptic brain signals classification, Biomedical Engineering Letters 2 (3) (2012) 147–157.

[21] A. Subasi, E. Erçelebi, Classification of eeg signals using neural network and logistic regression, Computer methods and programs in biomedicine 78 (2) (2005) 87–99.

[22] S. Patidar, T. Panigrahi, Detection of epileptic seizure using kraskov entropy applied on tunable-q wavelet transform of eeg signals, Biomedical Signal Processing and Control 34 (2017) 74–80.

[23] H. Ocak, Optimal classification of epileptic seizures in eeg using wavelet analysis and genetic algorithm, Signal processing 88 (7) (2008) 1858–1867.

[24] S. G. Mallat, A theory for multiresolution signal decomposition: the wavelet representation, Pattern Analysis and Machine Intelligence, IEEE Transactions on 11 (7) (1989) 674–693.

[25] H. Adeli, Z. Zhou, N. Dadmehr, Analysis of eeg records in an epileptic patient using wavelet transform, Journal of neuroscience methods 123 (1) (2003) 69–87.

[26] J. Bruhn, H. Röpcke, A. Hoeft, Approximate entropy as an electroencephalographic measure of anesthetic drug effect during desflurane anesthesia, Anesthesiology 92 (3) (2000) 715–726.

[27] J. S. Richman, J. R. Moorman, Physiological time-series analysis using approximate entropy and sample entropy, American Journal of Physiology-Heart and Circulatory Physiology 278 (6) (2000) H2039–H2049.

[28] P. Artameeyanant, W. Chiracharit, K. Chamnongthai, Spike and epileptic seizure detection using wavelet packet transform based on approximate entropy and energy with artificial neural network, in: Biomedical Engineering International Conference (BMEiCON), 2012, IEEE, 2012, pp. 1–5.

[29] C. J. Burges, A tutorial on support vector machines for pattern recognition, Data mining and knowledge discovery 2 (2) (1998) 121–167.

[30] D. Tsujinishi, S. Abe, Fuzzy least squares support vector machines for multiclass problems, Neural Networks 16 (5) (2003) 785–792.

[31] K. Polat, B. Akdemir, S. Güneş, Computer aided diagnosis of ecg data on the least square support vector machine, Digital Signal Processing 18 (1) (2008) 25–32.

[32] T. Kayikcioglu, O. Aydemir, A polynomial fitting and¡ i¿ k¡/i¿-nn based approach for improving classification of motor imagery bci data, Pattern Recognition Letters 31 (11) (2010) 1207–1215.

[33] J. A. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, J. Vandewalle, J. Suykens, T. Van Gestel, Least squares support vector machines, Vol. 4, World Scientific, 2002.

[34] K. De Brabanter, P. Karsmakers, F. Ojeda, C. Alzate, J. De Brabanter, K. Pelckmans, B. De Moor, J. Vandewalle, J. Suykens, Ls-svmlab toolbox users guide, ESAT-SISTA Technical Report (2011) 10–146.

[35] M. Nilashi, O. Ibrahim, A. Ahani, Accuracy improvement for predicting parkinsons disease progression, Scientific reports 6.

[36] K. Fu, J. Qu, Y. Chai, Y. Dong, Classification of seizure based on the time-frequency image of eeg signals using hht and svm, Biomedical Signal Processing and Control 13 (2014) 15–22.

[37] A. B. Das, M. I. H. Bhuiyan, S. S. Alam, Classification of eeg signals using normal inverse gaussian parameters in the dual-tree complex wavelet transform domain for seizure detection, Signal, Image and Video Processing 10 (2) (2016) 259–266.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 2, No. 3, 1290-1297 (2017)
www.astesj.com
Special Issue on Recent Advances in Engineering Systems

ASTES Journal
ISSN: 2415-6698

Proposal for a Visualization System of Purchase Relationship Using ID-POS Data

Ko Hashimoto[1], Kohei Otake[2], Takashi Namatame[*2]

[1]*Graduate School of Science and Engineering, Chuo University, 112-8551, Japan*

[2]*Faculty of Science and Engineering, Chuo University, 112-8551, Japan*

ARTICLEINFO

Article history:
Received: 31 May, 2017
Accepted: 06 July, 2017
Online: 01 August, 2017

Keywords :
Marketing
POS data
Visualization

ABSTRACT

In recent years, big data analysis is gaining immense credence in the fields of academics and business. Businesses such as management and marketing have demonstrated a strong inclination and interest in data analytics. However, many businesses are unable to utilize data even if they have access to it. The main reason for that is the lack of familiarity with data analytics procedures. Hence, a system needs to be developed that can perform data analytics and demonstrate its benefits. In this study, we use point of sales data obtained from a supermarket chain to analyze and show the relationship between purchase goods at the same time. A supermarket is one of the ideal places to demonstrate data analysis because retail stores have many purchase records and are always conducting various marketing activities. We propose an easy-to-handle visualization system to show the goods that are inter-related. By using our system, a store manager can obtain information about the item sales easily and interactively.

1 Introduction

In recent years, big data analytics has been receiving increasing attention due to the progress of information and communication technologies in various fields [1,2]. Big data can be defined as data characterized by the 3Vs-Volume, Variety, Velocity, i.e., a large amount of data of rich variety and accumulated at high frequencies [3]. Big data has increased drastically due to the increase in internet penetration with the development of digital devices such as smart phones and the prevalence of social networking services. More and more businesses are trying to utilize big data for business planning and corporate strategies using in-house expertise, so that they do not have to depend on data analysis experts.

Supermarkets and department stores are trying to devise an effective sales strategy to address diverse consumer needs by critically analyzing point-of-sales (POS) data. However, effective big data analysis requires extensive knowledge of data structures and programming, which can be built over a period of time. Hence, businesses and services are still not able to effectively utilize big data analytics on-site.

Retailers such as supermarkets are often required to make quick decisions. Quick decision-making requires quick data analysis to grasp the various needs of the consumer, and consequently, the store, and utilize the insights derived for strategic planning to increases sales, minimize inventory, improve efficiency, etc. Consequently, retail stores need interactive operations and real-time analysis capabilities.

This study is organized as follows: Section 2 presents the purpose of our study. Section 3 describes the study conducted to attain the objectives and the features of the proposed system. Section 4 describes the basic concepts and technology. Section 5 outlines the proposed system, its functions, and the way it can be applied. Section 6 describes the experimental evaluation of the proposed system. Finally, Section 7 concludes the paper and discusses the future topics to follow up on this study. This paper is an extension of work originally presented in *Future Technologies Conference (FTC) 2016* [4].

2 Purpose of our research

In this study, we propose a system that allows users to visualize relationships between goods while manipu-

[*]Takashi Namatame, 1-13-27 Kasuga, Bunkyo-ku, Tokyo, 112-8551 Japan, +81-3-3817-1932& nama@indsys.chuo-u.ac.jp

https://dx.doi.org/10.25046/aj0203163

lating interactively. The proposed system enables:

- Automation of the data analysis process from data extraction to visualization to reduce the burden on users.
- Development of a real-time sales strategy by constructing a platform that can freely operate under various conditions.

The proposed system aims to devise a real-time sales strategy that is not dependent upon the intuition and experience of individuals, instead relying on insights gained from the data collected on-site. Even if the data analyst does not have programming knowledge, the proposed system can help him/her understand the relationship between the goods. From the analysis of relationships between the goods, the analyst can derive the insights. Moreover, the proposed system is available as a web application, allowing users to share the results of the analysis by using a web browser. The proposed system can be utilized for the arrangement of goods on shelves and offer suggestions for their effective promotion, such as the distribution of a free recipe with hot-selling goods.

3 Related studies

In this section, we discuss other studies related to this study and describe the characteristics of the proposed system.

As mentioned above, various data are being accumulated in recent years. Under such situation, SNA(social network analysis) which is based on "graph theory", has been attracting attention as a method of analyzing a large amount of data. SNA is an analytical method that clarifies the structure of a network composed of "Vertex" and "Edge". In recent years, with the advent of social media, there are many studies targeting SNS. In these studies, the users are regarded as vertexes, join relationships defined between users are regarded as edges. For the purpose of these studies, the characteristics of the community, comparison of network structures, creation of formation model of social network are set up[5, 6, 7]. On the other hand, there is also research that systematizes technology of social network analysis[8].

Ito et al. proposed a visualization system using the conditional parts from the results of association rule mining. With this system, the user can figure out the relationship between the conditional parts in association rules by using hierarchical graphs [9].

Matsuo et al. used databases of academic theses and picked out the human relationship within a specific community from information acquired from the web. From the results, the human relationships could be understood appropriately. The information on human relationships can then be used to find people who are in the same academic field for the benefit of students and researchers. To judge the strength of the relationship, they used an index, which is based on a combination of the principles of the Dice coefficient and the Jaccard coefficient. They then performed a comparative verification of the index [10].

These studies showed that it is useful as a system using network analysis. However, such a system is not effective in reducing the burden on analysts and providing an easy-to-use platform. In this study, we propose a system that satisfies the following expectations:

1. *Real-time processing*:
   The system is able to visualize the network even if the user wants real-time data. For example, the system can help the analyst identify how the relationship between the goods will change in an hour's time in case of a limited-time sale and which relationships are strong in a particular time zone, all of this in real time.
2. *Interactive processing*:
   The system allows the analyst to interactively extract data and set the threshold, allowing him/her to customize information based on specific needs.
3. *Ease of operation*
   In a conventional analysis, the analyst needed to run the script, which was written as an R script, to prepare appropriate data for visualization. Moreover, it was necessary to write more scripts for appropriate visualization. Since it takes time to understand these scripts, it is difficult for beginners to process them. Therefore, the proposed system automates these processes to help the analyst who is not good at programming understand the results easily.

4 Basic concept and technical elements

In this section, we explain the basic concepts used in this study including association rule mining, co-occurrence indexes, and network analysis.

4.1 Association rule mining

In recent years, POS data has become easy to obtain. However, one of the significant problems here is the effective use of the enormous accumulated data. Association rule mining [11] is a popular and well-known method for analyzing useful relationships between variables in large databases. It extracts frequent combinations (association rules) of items from transaction data. The main aim of this method is to derive the pertinent rules from the data. The association rule uses Support (1), Confidence (2), and Lift (3) as the indexes that show the relationships between the goods. The rule which states that "When X is bought, Y is also bought" is transcribed concisely as {X}⇒{Y}. The sets of items X and Y are called antecedent (left-hand-side or LHS) and consequent (right-hand-side or RHS) of rule respectively. Moreover, in a few cases, one rule

is formed by more than three goods. For example, {potato, carrot}⇒{onion} means that "When potatoes and carrots are bought, onions are also bought."

$$Support(X,Y) = \frac{|X \cap Y|}{|M|} \tag{1}$$

$$Confidence(X,Y) = \frac{|X \cap Y|}{|X|} = \frac{Support(X,Y)}{Support(X)} \tag{2}$$

$$Lift(X,Y) = \frac{Confidence(X,Y)}{Support(Y)} \tag{3}$$

where X, Y denote the set of items respectively. $|X|$ is the number of transactions that includes X. $|X \cap Y|$ is the number of transactions that include X and Y. $|M|$ indicates all transactions.

4.2 Co-occurrence index

Co-occurrence indexes [12] are used in the field of language processing. These indexes are mainly used for text data analysis. Co-occurrence means the simultaneous appearance of more than one word in one sentence. The sentences are usually taken from platforms such as Twitter or search history. For example, "test" and "pass" are often used at the same time in many cases, which indicates that "test" and "pass" have a strong co-occurrence. In this way, it is possible to guess a relationship between words by extracting the strong co-occurrence from a large number of combinations. In this study, we try to extract the relationship by using the co-occurrence index to measure the strength of the relationship between goods. Then we use indexes such as Jaccard coefficient (4), Simpson coefficient (5), Dice coefficient (6), and Cosine coefficient (7). Features of these indexes are discussed below:

$$Jaccard(X,Y) = \frac{|X \cap Y|}{|X \cup Y|} \tag{4}$$

$$Simpson(X,Y) = \frac{|X \cap Y|}{\min(|X|,|Y|)} \tag{5}$$

$$Dice(X,Y) = \frac{2|X \cap Y|}{|X| + |Y|} \tag{6}$$

$$Cosine(X,Y) = \frac{|X \cap Y|}{\sqrt{|X||Y|}} \tag{7}$$

The values of these indexes range from 0.0 to 1.0. The correlation is presumed to be strong when the value is higher. The features of the indexes are indicated in Table 1.

Table 1: Features of each index

Co-occurrence index	Features
Jaccard	Something with the large number of appearances by the independence whole the value are lower.
Simpson	If the number of appearances by independence is extremely small, the value becomes high even if the relationship is not strong.
Dice	When there is a great difference between two appearances, the lower value remains lower even if the relationship is strong.
Cosine	When both appearances are 0, those are ignored.

It is necessary to use the appropriate index that expresses co-occurrences according to the information needed and based on the situation.

4.3 Basic expressions about the network analysis

In this section, we explain the concepts of centrality [13] and modularity [14], which are used in the network analysis. A network of relationships using goods of Electronic Commerce (EC) site is researched widely in recent years[15].

4.3.1 Centrality

Degree centrality:
The degree is the number of adjacencies for a node. Degree centrality is the most basic method of calculating centrality in which the nodes where the edges converge are deemed as having higher centrality. If we assume a node p_k, the degree of p_k, $Degree(p_k)$:

$$Degree(p_k) = \sum_{i=1}^{n} a(p_i, p_k) \tag{8}$$

where $a(p_i, p_k)$ is 1 if and only if p_i and p_k are connected by a line; otherwise, it is 0.

Closeness centrality:
The closeness is measured by summing the geodesic distances from one node to other nodes. Closeness centrality is a method of calculating centrality in which centrality is higher when the closer the distance (therefore other nodes can be reached with a small step) is between the nodes. If we assume a node p_k, $Closeness(p_k)$:

$$Closeness(p_k) = \left(\sum_{i=1}^{n} d(p_i, p_k) \right)^{-1} \tag{9}$$

where $d(p_i, p_k)$ is the number of edges in the geodesic linking p_i and p_k.

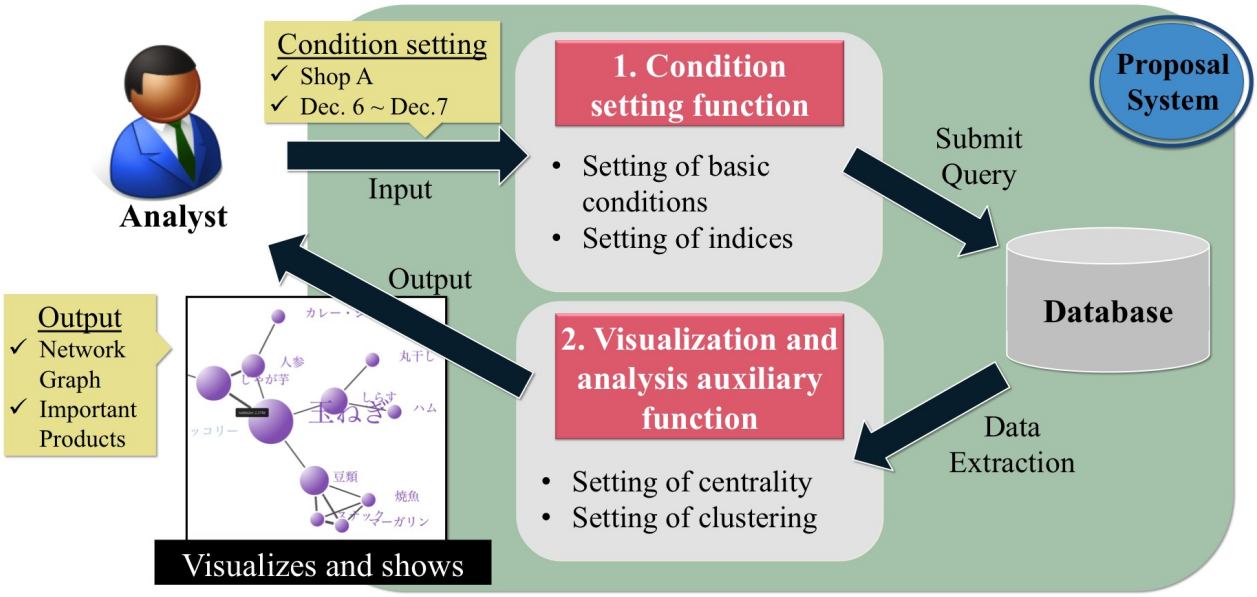


Figure 1: Image of the proposal system

Betweenness centrality:
The betweenness is based upon the frequency with which a node falls between pairs of other nodes on the shortest or geodesic paths connecting them. Betweenness centrality is a method of calculating centrality in which centrality is higher when more channels pass through a certain node. Therefore, this implies the importance of nodes acting as bridges between nodes in the network. If we assume a node p_k, the betweenness of p_k, $Betweenness(p_k)$:

$$Betweenness(p_k) = \sum_{i}^{n} \sum_{j}^{n} \frac{g_{ij}(p_k)}{g_{ij}} \quad (i < j) \qquad (10)$$

where $\frac{g_{ij}(p_k)}{g_{ij}}$ is the probability that node p_k falls on a randomly selected geodesic linking p_i with p_k.

4.3.2 Modularity

The modularity is an effects function that is used for network and graph analysis. It is a numerical index of how appropriate a particular division of communities is from the network. For a division with g groups, we define a $g \times g$ matrix E whose component e_{ij} is the fraction of edges in the original network that connects vertices in group i to those in group j. Then, the modularity Q is defined as follows.

$$Q = \sum_{i} e_{ii} - \sum_{i,j,k} e_{ij} e_{ki} \qquad (11)$$

The modularity ranges from 0 to 1. Local peaks in modularity during the progress of the community structure algorithm indicate particularly good divisions of the network.

5 Proposal and implementation of the visualization system

5.1 Outline of the proposal system

In this study, we develop the web application by using the R Language's package from Shiny [1]. We use R as a computational engine. For visualization, we use D3.js[2], which is a JavaScript library. We accumulate the analysis data, which is fixed in the database of MySQL[3]. In Figure 1, we show the analysis image of the proposed system. It will be possible to get the result through the web browser by mounting the system on a web server. We developed the system in the following environment.

- OS : Mac OS X Yosemite
- CPU : Intel Core i7 2.9GHz
- Memory : 8GB
- Language : R, JavaScript, SQL

In Figure 1, we explain the working of the proposed system. First, an analyst inputs a condition (e.g., shop name, periods) to analyze POS data. Next, a condition setting function defined in subsection 5.2 submits a query to the database based on the input

[1] http://shiny.rstudio.com/
[2] http://d3js.org/
[3] https://www.mysql.com/

conditions. Then, a visualization and analysis auxiliary function described in subsection 5.2 performs processing for visualization. This function calculates the centrality and clustering values of the network. Finally, this function delivers the network graph of the product relationships and important goods to the analysts.

5.2 Two functions of the proposal system

In this section, we are going to explain two functions of the proposed system outlined in Figure 1. The proposed system has the following two functions:

1. Condition setting function
2. Visualization and analysis auxiliary function

The condition setting function is a function to select dates, store numbers, and visualization indexes. With this function, users can interactively manipulate the data and enable free condition setting. The visualization and analysis auxiliary function is a function to set centrality and clustering and decide how to visualize. With this function, users can see the relationship between goods under the conditions set that they can set by themselves.

5.3 Details of the proposal system

For this study, we used POS data with the ID of a supermarket in Japan. POS data is the data of the sales of all goods in a store. POS data with the ID contains the customer information. By utilizing this data it is possible to grasp the customer behavior and improve the sales in a store. The data used in this study is obtained from 18 stores of a supermarket through one fiscal year in 2014. It consists of purchase data and personal attribute data.

- The purchase data
  - Purchased goods, number of goods to be purchased, purchase price, purchase stores, purchase dates, and so on
- The personal attribute data
  - Gender, age, and so on

In this study, we use purchase stores, purchase dates, days of the week, time zone, and receipt numbers for identifying the receipt of the purchased goods and purchased goods. Those are accumulated in a MySQL database.

The proposed system visualizes on the basis of the indexes of the relationship. The user can choose the index. The indexes are association rules that are used in simultaneous analysis, and co-occurrence indexes show the similarity between the sets. By using association rules and co-occurrence indexes, it is possible to obtain the rules from an enormous network. However, if the rules are also enormous, it is difficult to derive the useful rule just by visualizing. Generally, a threshold value of the index is set and the important rule is picked out. However, when the threshold is made too high, important information may not be even picked out. To address this problem, we provide the system with the function that sets the threshold freely. The system allows the analyst to freely set the dates, time zone, index, threshold value, and the number of clusters by using dynamic D3.js. Thereby, the analyst can easily inspect sets as the rule. Here, in Figure 2, we show the whole layout of the system and explain the operation of the system.

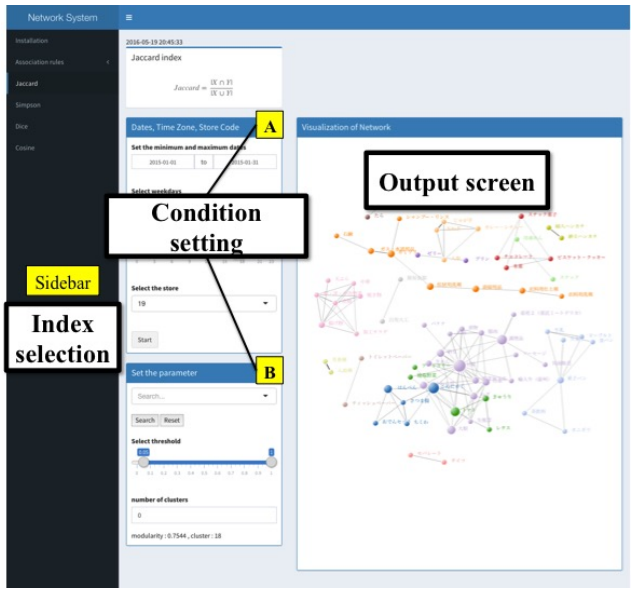


Figure 2: Screenshot of our proposal system

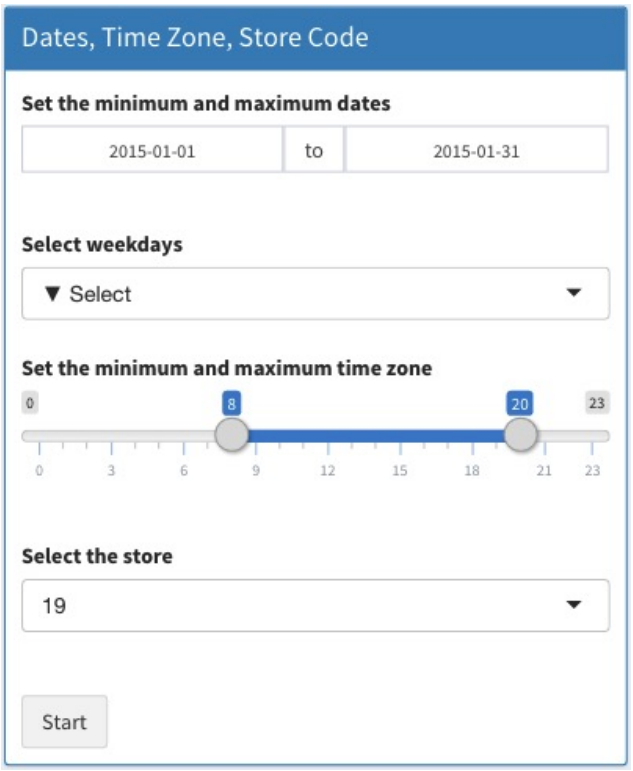


Figure 3: Screenshot of condition setting about dates, time zone, and store code

After the application starts, the user selects the index by clicking on "Sidebar," as shown in Figure 2. The layout of the application is switched to the respective indexes on selection of the index.

In Figure 3, the condition of the analysis period, the time zone and the analysis target shop is set to extract the data from the database. First, it sets the analysis period that the analyst wants through "Set the minimum and maximum dates." As shown in Figure 3, it sets the period from "2015-01-01" to "2015-01-31" as 1 month. In this way, the analyst can set it by a week unit or a month unit as well as a single date and time. Next, the day of the week can be selected when necessary in the period. In Figure 3, day of the week setting is not being used because of "▽Select." Next, it sets the analysis time zone that the analyst wants in the period using "Set the minimum and maximum time zone." In Figure 3, it sets the time zone from 8 to 20, which is the time zone from "8 a.m." to "8 p.m." in the period that is set from "2015-01-01" to "2015-01-31" as 1 month. It sets the target store through "Select the store." In this way, the analyst can set the target store that the analyst wants. Finally, the series of processing steps starting from the extraction of the data to visualization is begun by clicking the button "Start."

Figure 4 depicts the way in which the threshold of the index and the number of clusters can be set. In addition to the condition that was set in Figure 3, it is possible to set condition setting for threshold and the number of clusters. As shown in Figure 4, the visualized goods can be searched using the "Search..." button. If the goods for which the search was conducted existed in the graph, only the relevant goods are indicated. Otherwise, it displays the message "The search goods did not exist in the visualization." It sets the threshold of the selected index in "Set threshold." In this way, it is possible to adjust the enormous network by choosing rules in the case of many goods. In Figure 4, the rules that are greater than 0.05 of Jaccard coefficient value are selected. A visualized network is divided into more than one cluster in "number of clusters." The default number of clusters is 0 and this setting is calculated by using the modularity. In Figure 4, the number of clusters in the network is divided by the default setting, and the color of network is set automatically.

Here, the network is visualized based on the conditions set in Figure 3 and Figure 4. Then, the analyst can perform actions such as clicking, moving, and scaling of the network. In our proposed system, when the analyst clicks the target node, the system shows the relationship between the target node and the others that are related to the target node. When the analyst double-clicks the node, the network returns the original state. Figure 5 shows that all goods are interrelated, which is because goods that have no relationships at all are excluded beforehand. Though we excluded goods that have no inter-relationships from the system, there is a possibility that they are contributing to the sales. Figure 5 shows an example of the network visualization, which is based on the condition of the purchase history from "8:00" to "20:00" of January at a particular store. The index is the Jaccard coefficient and the threshold is 0.05. The color of the network is divided by the modularity. And we focus on two partition parts in Figure5.

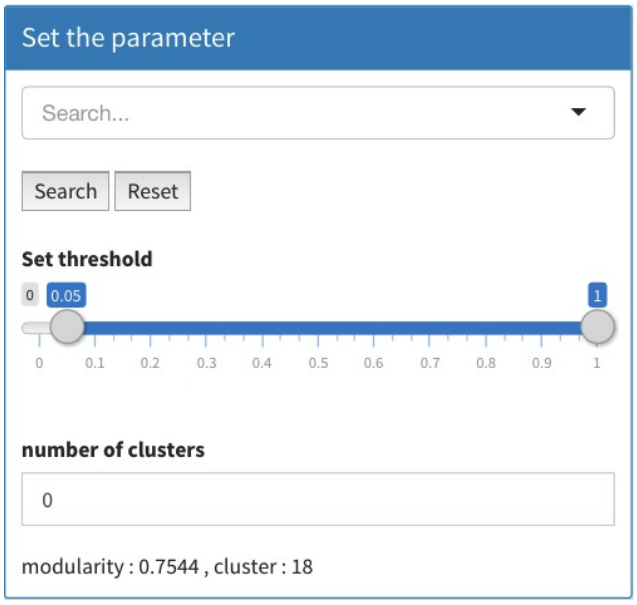


Figure 4: Screenshot of condition setting about threshold of the network index and the number of clusters

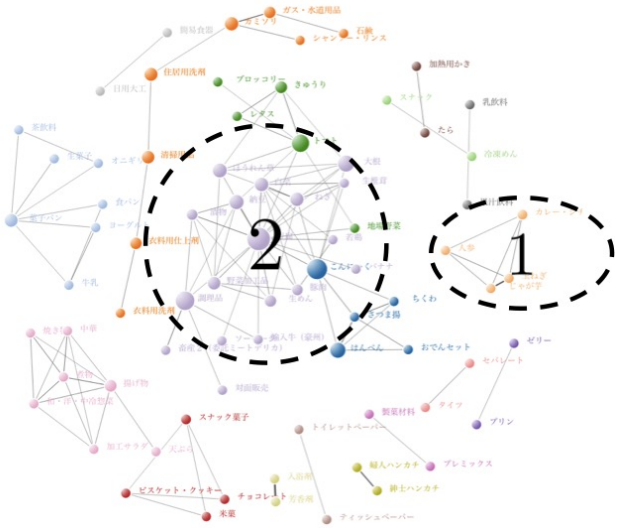


Figure 5: An example of the output screen

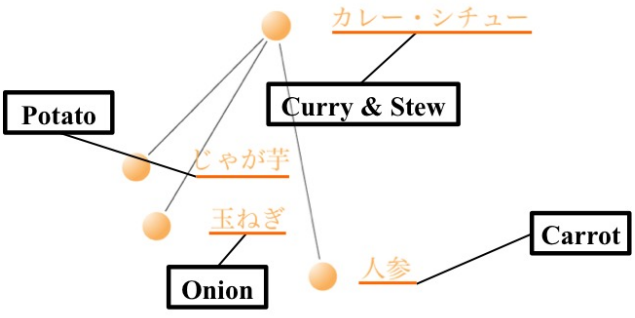


Figure 6: 1st partition part

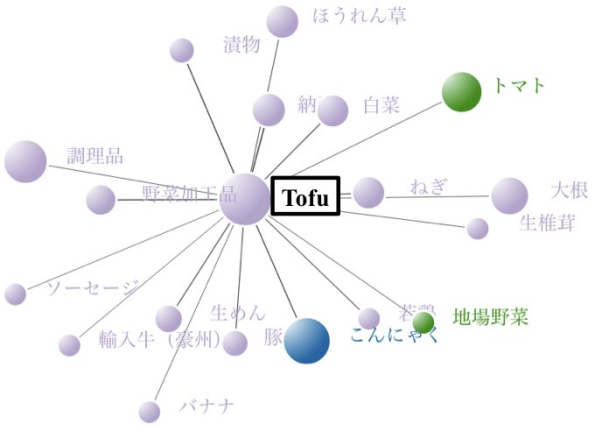


Figure 7: 2nd partition part

Figure 6 shows the first part of the partition in Figure 5. This figure shows that "Curry & Stew" is related to "potato," "onion," and "carrot." This indicates that it was possible to extract the typical ingredients from the network for curry and stew in Japan.

And Figure 7 shows the second part of the partition in Figure 5. This figure shows that "Tofu" is related to many goods and it is also involved in the other community goods. From this, it turns out that "Tofu" is a goods to be a hub that has relationships with various goods in supermarkets. When the particular goods extracted have relationships to many other goods and the threshold is high, they can indicate the core goods in the supermarket. Sales can be promoted by cross-selling with the core goods in the sales strategy.

Visualizing in this manner will help create the sales strategy and the analyst can easily understand the relationships between the normal goods and the core goods in an enormous network.

6 Experimental evaluation

In this section, we describe the evaluation of the system. We conducted an experimental evaluation based on a semi-structured interview. We set evaluation criteria beforehand to get the evaluation of the system. The evaluation criteria are shown below:

1. Quality of visualization (Is the relationship between the goods captured?)
2. Interactive processing (Is condition setting sufficiently interactive?)
3. Real-time processing (Is it used for a decision-making at a site?)

In the evaluation experiment, we identified the target analyst who had experience in management and working at a supermarket. However, she had limited knowledge of data structures and programming. The system was used by the analyst and we received the evaluation for each evaluation criteria, as shown in Table 2. In addition, we received the feedback on possible improvements in the system, as shown in Table 3.

Table 2: Evaluations

No.	Evaluation
1	Relationship between goods that tend to be bought at the same time is classified by colors. Therefore, the visualization is easy to understand and the appearance of the visualization is dynamic and interesting.
2	User can interactively perform various condition settings on the system and it is possible to visualize it using various patterns. Best of all, it can be expected to shorten analysis hours because there is no necessity to process the data by each condition setting.
3	It enables real-time analysis if the system is equipped with a server and the environment is adequately equipped. For this reason, the headquarters can cooperate and do quick strategic planning as well as on-site judgment.

Table 3: Areas of improvement

No.	Improvement points
1	Calculation speed and large scale database
2	Addition of the function for actual use
3	Documents for users are required

Based on the evaluation as per Table 2, we can conclude that the visualization display was found to be easy to understand. This is because of the appropriate grouping of goods, which makes it easier to understand the relationships between goods. The evaluation also showed that analysis hours could be reduced because there is no need to process the data by each condition setting. The proposed system enables real-time analysis if the system is adequately equipped. From the evaluation, it can be inferred that the proposed system is effective in terms of reducing the burden on the analyst for data processing.

We also studied the suggestions for improvement, as described in Table 3. Admittedly, the calculation speed of the system is inferior because of the usage of R script instead of compiler languages such as C and Java. The computational complexity is proportional to the amount of data and as the data becomes big, it takes more time to analyze it. For example, if the relationship between the goods is visualized for a 1-month time period, the operation becomes difficult and slow. Therefore, it is necessary to improve the interactive operations in such cases. The system is definitely just a prototype. Faster calculation time can be achieved by improving the environment and introducing compiler languages.

The lack of sufficient functions pointed out in Table 3 arises from the business experience at a supermarket. In particular, decisions such as elimination of the node (goods) that an analyst deemed unnecessary, multi-selection of stores, and condition setting

of the category of goods, etc. are planned for improving user convenience when the actual system is introduced. Therefore, the problem will be resolved by adding the function to the interface.

The third area of improvement in Table 3 points toward the necessity of a document explaining the analysis methods and usage. Although the proposed system can visualize the relationships between goods using various methods, it is necessary to take into account the kind of scene for each visualization indexes can be used. For example, it is necessary to demonstrate and recommend to the user that degree centrality should be used to extract goods that are purchased at a high frequency, such as bargain items.

7 Conclusion and future studies

In this study we proposed a system that allows users to visualize relationships between goods while manipulating interactively. Data extraction, processing, calculation, and visualization are automated even when an analyst does not have knowledge of programming, and an interface has been developed. We made it possible to freely set the threshold settings such as association rules and co-occurrence indexes. In addition, the system can support the proposal of real-time sales.

In this study, we used association rules and co-occurrence indexes to select the rule from a network. However, we could not demonstrate how the rule is useful by extracting with indexes. Therefore, we need to clarify index selection, which is a key topic for future studies. In the experimental evaluation of this study, some improvement areas were pointed out. In the future, it is necessary to improve and impart various functions in order to make the proposed system easier to use. Moreover, we could not obtain the quantitative evaluation about the system during the experiment, which is another necessary topic that needs to be explored.

Conflict of Interest The authors declare no conflict of interest.

Acknowledgment This work was supported by JSPS KAKENHI Grant Number 16K03944 and 17K13809.

References

[1] M. J. Shaw, C. Subramaniam, G. W. Tan and M. E.Welge : "Knowledge Management and Data Mining for Marketing," Decision Support Systems, Vol. 31, Issue. 1, pp. 127-137, 2001.

[2] M. S. Yadav, K. Valck, T. H. Thurau, D. L. Hoffman and M. Spann : "Social Commerce: A Contingency Framework for Assessing Marketing Potential," Journal of Interactive Marketing, Vol. 27, Issue 4, pp. 311-323, 2013.

[3] M. Shirota : Impact of Big Data, TOYO KEIZAI INC., pp22-26, 2012 (in Japanese).

[4] K. Hashimoto, K. Otake, and T. Namatame : "Proposal of the Visualization System of Simultaneous Purchasing Relation using POS Data with the ID of the Supermarket," Future Technologies Conference (FTC) pp.610-615, IEEE, 2016, https://doi.org/10.1109/FTC.2016.7821668.

[5] P. Holme, C. R. Edling and F. Liljeros : "Structure and Time-Evolution of an Internet Dating Community," Social Network, Vol. 26, pp. 155-174, 2004.

[6] K. Yuta, N. Ono and Y. Hujiwara : "Structural Analysis of Human Network in Social Networking Services," Transactions of Information Processing Society of Japan, Vol.43, No.3, pp.865-874, 2006.

[7] F.Toriumi, K.Ishida and K.Ishii : "Proposal for a Network Growth Model of Social Network Service," IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences of Japan, Vol.J93-D, No.7, pp.1135-1143, 2010.

[8] K.Tanaka and Y.Kadono : "A Study on Development Organizations for Information Systems through Network Analysis," Proceedings of Autumn 2015 National Conference Presentation Competition of JASMIN.

[9] A. Ito, T. Yoshikawa, T. Furuhashi, R. Ikeda, and T. Kato : "Search of Interesting Rules in Visualized Association Rule Mining," Proceedings of 26th Fuzzy System Symposium, pp.684-689, 2010 (in Japanese).

[10] Y. Matsuo, H. Tomobe, K. Hashida, H. Nakashima and M. Ishizuka : "Social Network Extraction from the Web information," Transactions of the Japanese Society for Artificial Intelligence, Vol.20, No.1, pp.46-56, 2005 (in Japanese).

[11] T. Pang-Ning, M. Steinbach, K. Vipin. : "Association Analysis (Chapter 6) : Basic Concepts and Algorithms," Introduction to Data Mining. Addison-Wesley, 2005.

[12] C.D. Manning and H. Schütze : Foundations of Statistical Natural Language Processing, The MIT Press, London, 2002.

[13] L. Freeman : "Centrality in Social Networks Conceptual Clarification," Social Networks, Vol.27, No.1, pp.215-239, 1979.

[14] M.E.J. Newman : "Detecting Community Structure in Networks," The European Physical Journal B, Vol.38, Issue 2, pp.321-330, 2004.

[15] K. Otake, Y. shinozawa, A. Sakurai, T. Uetake, M. Oka and M. Suzuki : "Examination of the Effect of Social Login Through Analysis of User's Purchasing Tendency —Case study of golf portal site—," 12th Annual IEEE Consumer Communications and Networking Conference (CCNC), pp.617-618, 2015.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 2, No. 3, 1298-1306 (2017)
www.astesj.com
Special Issue on Recent Advances in Engineering Systems

ASTES Journal
ISSN: 2415-6698

Public transportation network design: a geospatial data-driven method

Alexey Golubev*, Maxim Shcherbakov

Volgograd State Technical University, CAD, Lenin av., 28, 400131, Volgograd, Russia

ARTICLEINFO

Article history:
Received: 10 June, 2017
Accepted: 17 July, 2017
Online: 01 August, 2017

Keywords:
urban development
geospatial data analysis
PTN design
route planning
decision support
GIS

ABSTRACT

The paper explores an issue of efficient public transportation network design as a part of the urban developing process. Having data about everyday residents travelling inside of an urban area, we can consider this data as people's requirements for the public transport system. We propose a novel method for initial public transportation network design based on clustered geospatial data containing origin/destination point of travel patterns of residents. The core of a method is a set of four algorithms for selecting terminal clusters based on a number of centres (or focuses). The appropriate quality criteria are proposed: degree of transport demand satisfaction, the coefficient of non-straightness, and transport network density. Use cases allow evaluating a performance of proposed method and give sufficent conclusions about its application.

1 Introduction

Urban Computing is a scientific area which provides models and methods for designing convenient urban environments. Also, the set of models allows predicting how the urban area changes and supporting decision making for urban planning and reconstructing [1]. For instance, urban planning includes operations for transportation, communication and distribution networks design toward to improving residents quality of life [2].

The public transportation network (PTN) in general meaning (including buses, underground, taxi, etc) is sufficient part of the city. From the one side, the city is growing around transport infrastructure, from another side, the expansion of city leads to the implementation of infrastructure projects. Residents select their location for living based on many criteria, but the transport is the essential one. Generally, the ideal public transport networks provide a service with minimal travel time, also including time for waiting and changes. As PTN contains the different service provider, the concurrency leads to increasing quality of service. In real life, PTN is not so efficient, moreover for new developing areas. Commonly used methodologies for transport system design and development evaluates technical features, but they do not take into consideration the residents opinions and preferences into consideration. It leads to a fundamental contradiction: if people do not use public transportation network due to its inefficiency, they prefer personal carriers, but increasing the number of cars in a city leads to negative ecological and economical effect, like traffics. Developing new data-driven tools which consider residents' preferences of travel paths is challenging tasks for Urban Computing domain.

Actually, new data collecting technologies allow gathering data about different aspects of people's life. Cell operators can detect the position of a subscriber with acceptable quality level. A post in Twitter contains data about location rectangle event, where a post was made. Geospatial data is are a common part of most information gathered by software applications. In this case, if a person has his/her own common travel paths (the most used origin-destination points) they can be considered as personal requirements for transportation systems. Having this data, we are able to (i) evaluate the efficiency of the current transportation network and (ii) to suggest modifications based on citizen's preferences.

So, in the research, we consider the following question. Given anonymised travel data in the framework of an urban area, what the data-driven technologies which allow designing the optimal public transportation network? To find the answer this question, the following high-level steps need to be performed: (i) collecting geospatial travel data, (ii) reasonable

*Corresponding Author: Alexey Golubev, PhD Student, CAD, Faculty of Electronics and Computer Engineering, Lenin av., 28, Volgograd, Russia & ax.golubev@gmail.com

https://dx.doi.org/10.25046/aj0203164

data reduction for generalisation of key points in routes in a PTN, (iii) designing a set of (sub)optimal routes and (iv) picking up the appropriate PTN according to multicriteria decision analysis based on quality criteria [3]. If we have travel data of every single person in an urban area, we can evaluate the most optimal stops as nodes in public transportation routes. Suppose we get information about 5% of residents of a city with 1 million population. In this case, the correspondence matrix has 50000^2 elements. Reduction of original nodes count allows to detect the centres of clusters and to understand where stops could be located. Attempts to solving the problem of optimal PTN design lead to many subtopics including data processing algorithms over geospatial data [4]. As an example, a task of finding public transportation network with (sub)optimal routes based on defined clusters. This problem includes the well-known tasks like shortest path search problem as well as specific task need to be explored, e.g. how to create initial public transport network based geospatial data analysis. In contrast with shortest path search problem with its many efficient and fast program solutions e.g. OpenStreetMap (OSRM), the initial network design is the unstructured problem containing multi-criteria decision analysis.

The main contribution of the paper is a novel method for initial public transportation network design based on clustered geospatial data containing origin/destination point of residents. The core of a method is a set of four algorithms for selecting terminal clusters based on focuses.

A paper has the following structure. After the introduction, we observe literature referring to a problem of data-driven approaches for transportation network design. Next, we propose a novel method for initial PTN design with various approaches for terminal clusters choice. Use cases section describes how the method was implemented and estimated over data according to performance evaluation criteria. The conclusion contains the main findings and future work ideas.

2 Background

Classical planning and development of the public transport network is a well-studied field of science. Methodologies for planning the transport network of the city according to [5] includes the sequence of actions: (i) public transportation network design; (ii) setting timetables; (iii) scheduling vehicles to trips, (iv) and assignment of drivers and other maintenance work.

You can also observe the emergence of a number of theoretical studies. In work [6] author proposed a hypothesis about the routing of tables with a symmetric shortest trajectory. It describes the symmetric routing of the shortest path table and presented counterexamples for its hypothesis. In their article [7], Daniel Delling et al. investigate the problem of calculating all optimal Pareto travels in a dynamic public transport network and introducing an algorithm for a public transit route (called RAPTOR). Another of their work [8] offers a routing mechanism for calculating traffic directions in large-scale road networks. And in [9], we suggest public transit transportation as the solution of the journey in the future. Turning to other works, one can distinguish [10]. The authors study the problem of finding good alternative routes in road networks. They look for routes that (i) differ significantly from the shortest path, (ii) have a small extension, and (iii) are locally optimal. The authors of the paper [11] proposed the RICK algorithm for constructing popular routes from undetermined trajectories, using information on the sequence of locations and time intervals. In the article [12] Tim Duyer and Lev Nahmanson proposed fast cross-routing for large graphs. They focus on discussing two methods for achieving faster routing using approximate methods of finding shortest paths. In such sources as [1, 13–15] the issues of theoretical research, planning and optimization of the transport network are given very great attention. Special attention is paid to the best practices in [1, 15].

Also we analysed different well-known algorithm to solve short path problem such as $A^\star$, $IDA^\star$, Breadth-First-Search, Best-First-Search, Dijkstra's algorithm, Bi-directional $A^\star$, Jump Point Search, Orthogonal Jump Point Search, Two-stage algorithms (ALT, Reach) and genetic algorithms of random search. Such algorithms like $A^\star$, $IDA^\star$, Best-First-Search, ALT ($A^\star$ labeling and methods are based on the inequality of the triangles used to work without prior preparation data that leads to the intersection large graphs and create significant problems in the time cost. In contrast to these algorithms exist such as HLC [16], HL, TNR [17] and lookup tables, where the costs associated with the pre-processing and data storage, while the speed of the algorithms are very fast.

Besides theoretical results, software implementations for decision support in urban transportation planning have been observed. PTV Visum [18] is software for evaluation existed (defined by a responsible person) public transport network based on transportation matrix. Another software called Emme, proposed by "INRO Software" [19] and SaaS solution by Citilabs is named Cube Cloud [20]. The Transportation Analysis and Simulation System (TRANSIMS) developed by [21] is a set of tools for analysis of regional transportation systems. TRANSIMS is an open source project which is available for public usage under NASA Open Source Agreement Version 1.3. Software like Aimsun is used in urban development traffic modelling that allows to model fusing travel demand and simulate networks of difference complexity [22].

The task is closely related to the task of routing from point A to point B, but it has some differences. First, in the typical task of routing, the objective function is the journey time from the beginning to the end of the route, which must be minimized. In the

case of the construction of a network of public transport routes, the objective function is an integral function that takes into account the average time of the pedestrian travel to the stop, the length of the path, the number of transplants and other characteristics of [5, 23]. Secondly, in the typical task of routing for the movement, any intermediate points that reduce the route are selected, whereas in the problem under consideration the intermediate points are located in the same place as the cluster centers. All of these variables are averaged over the number of residents. Moreover, an important difference between the two algorithms is that the end point in the route can not be arbitrary since each route must contain nodes (stops) which represent the centers of clusters.

The main conclusion of the review can be drawn: in spite on different approaches for network design, the initial network structure design is not studied well. Especially, if these network obtained for data gathered from residents.

3 A geospatial data-driven method

Algorithm 1: The scheme of the method for route network design

Data: n_r, C_t, C_{nt}
Result: RN
for $i = 1 \dots n_r$ **do**
  Build R_i direct route, containing a pair of opposing points of C_t;
  Add R_i to RN;
end
while C_{nt} *is not empty* **do**
  for $i = 1 \dots n_r$ **do**
    Select R_i from the route RN,
    $R_i = [n_1, n_2, \dots n_R]$;
    Split R_i on a pair of nodes
    $PN^{(R_i)} = [[n_1, n_2], [n_2, n_3], \dots, [n_{R-1}, n_R]]$;
    Initialize RC list;
    for *pairs* $[n_x, n_y]$ *in* $PN^{(R_i)}$ **do**
      /* Find the node minimally increases the length of the direct route $[n_x, n_y]$ */
      Find c_j node from C_{nt} with j index, which $j = argmin(|len(n_x, n_y) - (len(n_x, c_j, n_y))|)$;
      Add $[n_x, c_j, n_y]$ to RC;
    end
    Compose $\|RC\|$ options for new routes: one variant of the RC, the rest of the $PN^{(R_i)}$ and compile a list of routes candidates to replace RCC;
    Evaluate the length of routes from the list RCC and choose the route R_i^* with minimal length;
    Replace R_i to R_i^* from RN;
    Remove c_j node entered to R_i^* from C_{nt};
  end
end

3.1 General description

The basic idea of the initial public transport network design is to find the network consistently adding new nodes to existed routes in respect with minimal length increasing of designed network.

Let's consider the task statement. Given n_r number of routes in the designed public transportation network, n_c number of nodes (clusters centers) which should be added into network, C_t a set of terminal nodes, C_{nt} a set of non-terminal nodes, correspondence matrix $\|C_{nt}+C_t\| \times \|C_{nt}+C_t\|$. Note, that $C_t+C_{nt} = n_c$. Output of proposed algorithm is a transportation network RN with routes containing non-overlapping set of nodes,

$r_i = [p_1^{(i)}, \dots, p_k^{(i)}]$, where $i = 1, \dots, n_r$.

We propose the algorithm which contains the following steps represented in the scheme 1.

3.2 Explanation

Consider the operation of the method using a simple example. We choose the number of routes in the network $n_r = 2$ and assume that the algorithm for the choice of terminal clusters has given us A, A', B, B' (C_t) and nonterminal clusters C, D, E, F, G, H (C_{nt}).

In the first step, we combine terminal clusters into pairs and put them in the list $R_i = [AA', BB']$. The image 1a represents this network configuration.

Next, select the first route from the list of R_i and divide it into parts (two clusters each). Because We have two clusters in the route, then we have only one route $PN = [AA']$.

Next, we compose a new set of routes by adding a new cluster to the middle of the existing routes from PN. In our case, we have $ACA', ADA', AEA', AFA', AGA', AHA'$.

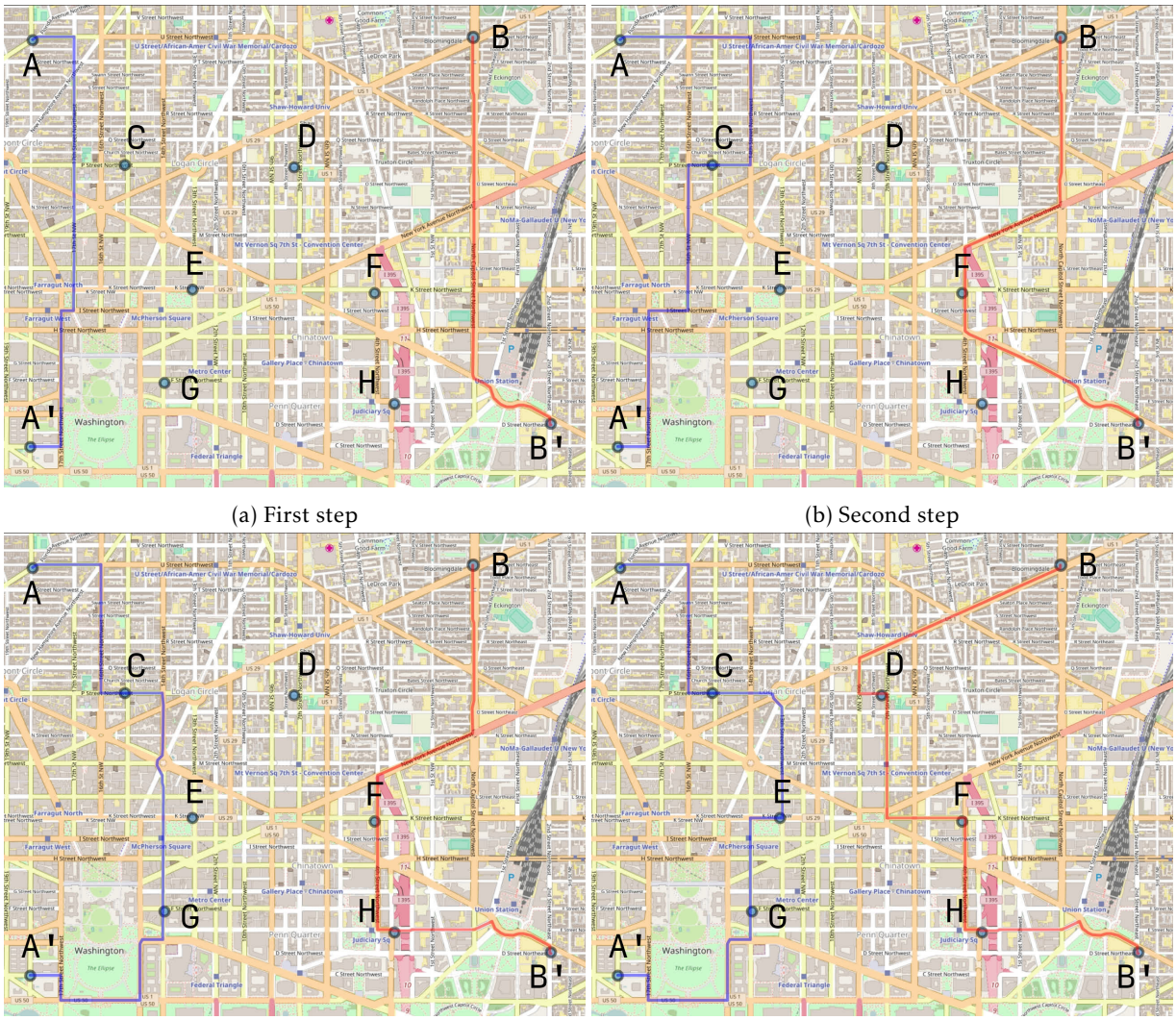
In the next step, we need to calculate the length for each new option and choose the one that has the shortest length – ACA'. Add this route to the list RC.

Next, we compose all variants of routes from the lists RC and PN. In our case, we have only one route ACA' and we add it to the list of $R*_i$. The used cluster ACA' is removed from C_{nt}. We perform similar operations for BB' and get a new route BFB'. The image 1b demonstrates this stage of construction.

On the next iteration of the algorithm we have $R_i = [ACA', BFB']$. Using ACA' we get two pairs $PN = [AC, CA']$. By successively adding nonterminal clusters, we get the following set: $ADC, AEC, AGC, AHC, CDA', CEA', CGA', CHA'$. Suppose that GA' is the shortest of all.

We make variants from CGA' and AC, CA' and get $ACGA'$. Add the resulting version to $R*_i$ and remove G from C_{nt}. Similarly, for BFB' we obtain $BFHB'$. The image 1c represents this stage of building a route network.

We continue to execute the algorithm while C_{nt} contains clusters. Finally, we get the network shown in the figure 1d.



(a) First step

(b) Second step

(c) Third step

(d) Fourth step

Figure 1: Illustration of the network design process.

3.3 Algorithms for choice of terminal clusters

In the last paragraph we mentioned some algorithm for selecting terminal clusters. We proposed 4 algorithms with which you can select the origin-destination clusters. Let us now examine them in more detail.

Algorithm 1: opposite clusters (opposite). The essence of the method is the connection of clusters lying on opposite sides of an imaginary circle.

The main goal of this method is to choose clusters that lie on opposite sides of the city. If the city has sleeping and industrial areas located in different parts of the city, then this method will allow transportation of a large number of people to and from work.

The algorithm contains the following steps.

1. Find the geometric center C of the convex hull
2. Find the furthest cluster from the center C on the convex hull and consider it to be the radius R
3. On the imaginary circle formed by R, marks uniformly $2 \cdot n_r$ points
   - the circle is given by the center of the shell C and the radius R
   - n_r – number of routes in PTN
4. Combine the diametric points into groups of two
   - for each point, we search for the nearest clusters in the Δ neighborhood
   - choose the points with the largest value of people

Algorithm 2: from the center to the suburbs (c2out). The essence of the method: pulling all the routes to the center of the congestion of people. The main idea of the method is transportation of people from the suburbs to the center. This method is designed for the type of city for which the center is the main place for the work of a large number of people, and the suburbs – sleeping areas. The ideal form of the city for this method is the circle.

The algorithm

1. Repeat steps 1-2 from method 1
2. On an imaginary circle, points n_r of points F are uniformly marked
3. Find the center G of the city (and accordingly the nearest cluster with the largest number of people)
4. Choose n_r clusters near G (it is permissible to repeat clusters)
5. Pair pairs of clusters from G and F

Algorithm 3: two focuses (2focus). The essence of the method: is similar to the second method, but with two focuses (cluster centers). This method is a logical continuation of method 2, but with the account that in a city there can be not one center, but two. The centers determined by this method need not be located near the geographical center of the city. Another important observation is that routes are also constructed between the focuses. The ideal shape of a city can be an ellipse or a circle.

The algorithm

1. Find F_1 and F_2 for the convex hull of a city
   - clusters with the largest concentration of people
   - and, accordingly, the closest clusters in their neighborhood along $n_r/2$ for each focus
2. On an imaginary circle, points B (K are uniformly $n_r - K$ – the number of routes connecting focal points)
3. The first K routes include two clusters (one from F_1 and the other from F_2)
4. Distribute the remaining points between B, F_1 and F_2 (one point from B, and the other from F_1 or F_2 (depending from distance))

Algorithms 4: N-focuses. The essence of the method: generalization of methods 2 and 3 for an arbitrary number of focuses (cluster centers).

This method is a generalization of the previous two methods: *c2out* and *2focus*. In this case, you can choose an arbitrary number of "focuses" for the best provision of passenger traffic between the focuses and the suburbs of the city. The form of the city for this algorithm can be arbitrary.

The methods implements the following steps

1. Choose N focuses F_N with the largest number of people
2. On the convex hull, $n_r - N \cdot K_N$ points B (K_N – the number of routes connecting N focal points)
3. The first $N \cdot K_N$ routes between the focal points
4. The remaining points are distributed between B and F_N (by analogy with method 3)

As an algorithm for searching for a convex hull, the Graham algorithm used in the methods is used.

4 Use cases

4.1 General information

To explain how the proposed methods are applied in urban development task, we showed the example for new public transportation network design for midsize city located near the large regional centre. For the subset of residents in the city, we obtain a pair of origin and destinations expressed as longitude and latitude. Origin might be where the certain resident lives and the destination where the same resident works. It means each pair shows the most popular path of the certain person in the city. This path contains a part when the public transportation is used, and walking zones. An ideal public network meets requirements according to the several criteria. We use such criteria as: the degree of satisfaction of demand for transport, the coefficient of tension and the density of the transport network.

We apply the straightforward procedure containing the following steps: (i) generation of origin-destination pairs for residents; (ii) clustering all the points into 82 clusters (the number of clusters has been chosen experimentally); (iii) creating the initial transport network using the methods and algorithms described in the article.

4.2 Data

We generated data about travel citizens preferences inside of mid-sized town with approximately 350,000 residents. We received 6,000 pairs of origin-destination points of 12,000 in total. In this case correspondence matrix has a large size 6,000 × 6,000 elements where every element with 0 value means no connection or 1 stands for connection existence.

4.3 Implementation

Proposed algorithms were implemented using Python and Open Source Routing Machine (OSRM) service for urban terrain distance calculation between the nodes. The program code is published online in the Git-like repository. The software used:

- Python v 3.5.2
  - numpy 1.12.1
  - scipy 0.19.0
  - geographiclib 1.48
  - polyline 1.3.2
  - requests 2.9.1
  - json 2.0.9
- OSRM v5.8.0

4.4 Performance evaluation

As an estimation method, an integral formula was used to evaluate the efficiency of the transport network with the following calculation criteria:

- Degree of transport demand satisfaction

$$U = \frac{1}{T}\sum_{i=1}^{N} F_i$$

where T is the total number of passengers in need of transportation, F_i is the number of passengers carried with i transplants, N is the number of traffic.

- Coefficient of non-straightness

$$P = \frac{\sum\limits_{i=1}^{N} D_i}{\sum\limits_{i=1}^{N} L_i}$$

where D_i is the direct distance from the initial to the final point of the route for the i route, L_i is the length of the i route, N os the umber of routes in the transport network.

- Transport network density

$$L = 1 - \frac{1}{4 \cdot S}\sum_{i=1}^{N} L_i$$

where S is the area of the city, L_i is the length of the i route, N is the number of routes in the transport network.

Integral formula for assessing the efficiency of the PTN is based on additive convolution function:

$$O(I, W) = \frac{\sum\limits_{i=1}^{n} I_i w_i}{\sum\limits_{i=1}^{n} I_i} \quad (1)$$

where O – integral assessment of the transport system; I – values of criteria; w – weights for the criteria that determine their importance; n – number of criteria. The decision maker's preferences are defined as weights w.

4.5 Experimental design

To evaluate the efficiency of the algorithm and to study its specifics, we carried out experimental design varying the number of nodes and routes. The most interesting case is big geospatial data processing where a lot of nodes we processed.

To reduce the size of correspondence matrix and to understand the most popular locations, we applied clustering algorithms for original/destination geospatial data. In contrast with typical and well-known cluster techniques such as k-means and MeanShift, our algorithms used OSRM for distance calculation according to urban terrain conditions. As centers of clusters are considered as stops in a transportation network with the most populated neighborhoods, we designed different use cases regarding a number of nodes as it is a variative parameter.

As the input data for the program, the data was used after the clustering stage: 82 clusters, as well as the correspondence matrix. The following parameters were chosen as the variable experimental parameters:

- n_r – the number of routes in the network (the number of routes varied from 10 to 60 in steps of 2)
- method used
  - *opposite* (method 1)
  - *c2out* (method 2)
  - *2focus* (method 3)
  - $*$*focus* (method 4, where the focus value varies from 3 to 9)

The experiment was carried out using a computer running the Linux Mint 18.1 Serena operating system (4.4.0-77-generic x86_64 core), Intel (R) Core (TM) i5-4690 processor.

5 Result and discussion

Based on method implementation, we consider results regarding the following specifics:

(i) analysis of the efficiency of the designed transport networks with respect to different methods for selecting terminal clusters, based on the proposed criteria;

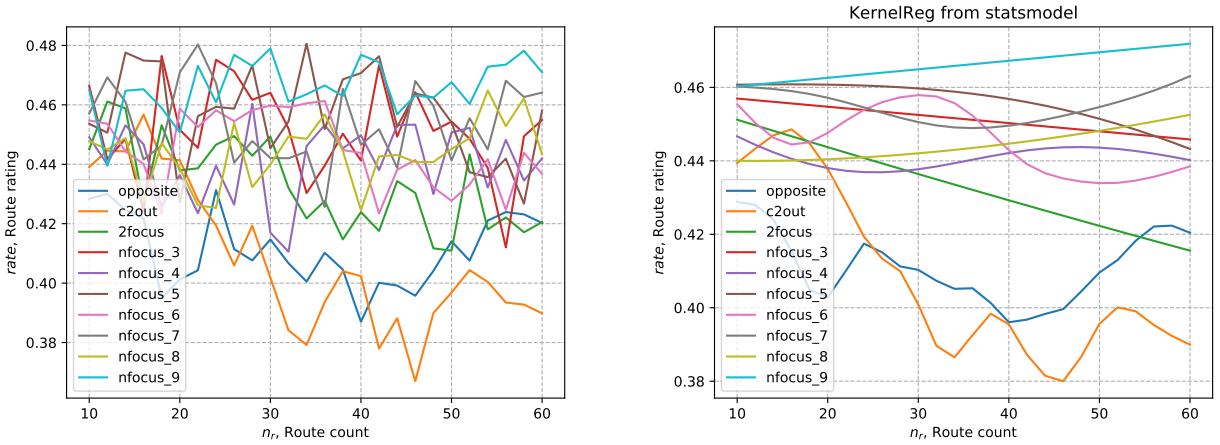
(ii) analysis the relationship between such quality criteria and parameters of methods;

(iii) analysis dependencies of performance (in terms of time complexity) and on a predefined number of routes in the network.

The graph of the network quality versus the number of routes is shown in the figure 2a. The trends of the methods are shown in the figure 2b.

The method ($*$*focus*) based on the idea of £allocation of clusters-focuses, where the largest number of people accumulate, proved to be better than all other implementations.

Depending on the number of focuses, the method behaved differently. For example, with 4 focuses, the quality of the network has the highest value in the current data set, for 8 and 9 focuses, with an increase in the number of marches, the quality of the transport network tends to grow, while for 4 and 6 – the periodical character is observed, and at 3 and 5 in general it starts to decrease (see figure 3b).



(a) Network assessment

(b) Results of smoothing of methods rate (e.g. performance)

Figure 2: Experimental results

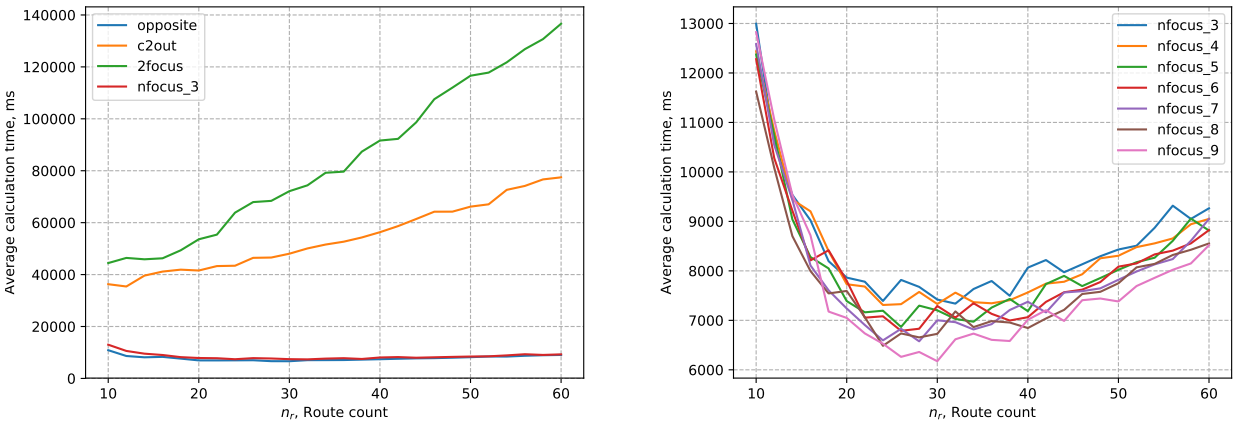
method \ n_r	10	20	30	40	50	60
opposite	10875.37 ± 192.12	6945.38 ± 1079.01	6627.15 ± 44.81	7400.85 ± 53.15	8185.97 ± 31.43	9045.62 ± 0.76
c2out	36297.57 ± 3596.22	41566.63 ± 21010.02	48040.82 ± 165417.80	56328.24 ± 14859.12	66171.95 ± 2052.92	77456.68 ± 28685.60
2focus	44437.48 ± 5473.17	53594.80 ± 9488.52	72121.6024± 4573.53	91627.76 ± 551.29	116600.52 ± 128.33	136617.8555± 53752.90
nfocus_3	12994.67 ± 720.38	7863.52 ± 1758.31	7420.7796 ± 577.53	8065.59 ± 2.63	8431.99 ± 44.88	9264.99 ± 72.74
nfocus_4	12437.60 ± 454.99	7729.52 ± 505.73	7330.43 ± 48.47	7563.89 ± 79.40	8305.97 ± 27.52	9050.09 ± 293.00
nfocus_5	12373.53 ± 230.70	7389.41 ± 805.00	7201.8921 ± 317.74	7185.38 ± 17.52	8021.30 ± 126.81	8814.65 ± 9.74
nfocus_6	12278.72 ± 128.41	7799.21 ± 761.17	7290.62 ± 937.70	7063.52 ± 146.78	8085.03 ± 264.92	8823.53 ± 113.01
nfocus_7	12582.99 ± 376.28	7239.58 ± 992.20	7001.10 ± 0.34	7375.48 ± 1.85	7820.72 ± 137.00	9050.35 ± 107.34
nfocus_8	11622.69 ± 1196.58	7593.90 ± 1458.24	6727.29 ± 65.67	6844.61 ± 446.14	7749.45 ± 33.92	8554.50 ± 51.91
nfocus_9	12823.84 ± 2203.70	7049.03 ± 369.11	6173.76 ± 99.90	7012.52 ± 8.99	7381.77 ± 5.23	8518.75 ± 44.39

Table 1: Results of calculation time (part of the results in increments of 10)

From the general trends, the result with 7 focus has a slight decrease in the quality of the network, It can be connected with an unbalanced network configuration when number of routes. For other methods of selecting terminal clusters, things are slightly worse. The methods opposite, *c2out* and *2focus* have the greatest value of network quality for small values of the number of routes, and then they only decrease. If for *2focus* the quality of the constructed network can be described by a linear decreasing function, then the opposite and *c2out* have non-periodic oscillations in their basis.

Turning to the analysis of time necessary for the calculation (Fig. 3a and 3b), we can say that the methods *∗focus* and opposite are the fastest, while *c2out* and *2focus* grow with the increase in the number routes. This can be explained by the fact that *c2out* and *2focus* methods use a costly function to search for clusters in a given radius, i.e. With the increase in the number of routes, the number of calls to this function also increases. The calculation time for the *∗focus* method for any n basically has the smallest parsing (Fig. 3b). Calculation of the algorithm running time using different methods for n_r with step 10 is presented in the table 1.

We can say that the method *∗focus* with a different number of foci (from 3 to 9) for the given set of geodata is the most optimal. It combines a good execution speed with the methods *c2out* and *2focus*, and the transport networks obtained with its use have the highest quality value for any n_r.



(a) Calculation time for different algorithms of terminal nodes choice

(b) Calculation time of *∗focus* algorithm depending on a number of route count

Figure 3: Experimental results

The main algorithm used to build a transport network by empirical evaluation of complexity (in terms of calculation time), the complexity of the algorithm 1 could be estimated as $O(n^3)$, where n is a number of nodes.

6 Conclusion

This article presents the method for initial a public transport network design based on the processing of geospatial data. The main idea of the proposed method is the successive addition of clusters in the transport route, taking into account its minimal increase in length. Also performance evaluation criteria is presented.

A method includes four different algorithms for terminal cluster choice: (i) algorithm based on opposite clusters, (ii) algorithm considering travelling experience from the center to the suburbs, (iii) algorithm for cities with two centers or focuses (*2focus*), and, finally, (iv) algorithm processing n focuses.

In accordance with the results obtained, it can be concluded that this algorithm can be used to construct the preliminary PTN. Based on the heuristic evaluation, it has cubic time complexity and the most choke point is in interaction with external services (e.g. OSRM). The proposed method *∗focus* with a different number of focuses (from 3 to 9) for the given set of geospatial data is the most suitable for terminal clusters choice, because of its speed and performance.

The evaluation is based on synthesized data and thus allows only an evaluation of the performance and scalability of the approach, rather than on its effectiveness in supporting planning transport managers decisions. A feedback from mobility expert on a real case study will increase the quality of evaluating the usefulness and effectiveness of the approach. So, these can be considered as future work based on theoretical findings.

7 Acknowledgments

The reported study was partially supported by RFBR, research project No. 16-37-60066 and research project MD-6964.2016.9.

References

[1] Gustav Nielsen and Truls Lange. Network design for public transport success–theory and examples. *Norwegian Ministry of Transport and Communications, Oslo*, 2008.

[2] Danila Parygin, Natalia Sadovnikova, Alla Kravets, and Elena Gnedkova. Cognitive and ontological modeling for decision support in the tasks of the urban transportation system development management. In *6th International Conference on Information, Intelligence, Systems and Applications, IISA 2015, Corfu, Greece, July 6-8, 2015*, pages 1–5, 2015.

[3] Alexey Golubev, Ilya Chechetkin, Konstantin S. Solnushkin, Natalia Sadovnikova, Danila Parygin, and Maxim Shcherbakov. Strategway: Web solutions for building public transportation routes using big geodata analysis. In *Proceedings of the 17th International Conference on Information Integration and Web-based Applications & Services*, iiWAS '15, pages 91:1–91:4, New York, NY, USA, 2015. ACM.

[4] Maxim Shcherbakov and Alexey Golubev. An algorithm for initial public transport network design over geospatial data. In *Smart Cities Conference (ISC2), 2016 IEEE International*, pages 1–7. IEEE, 2016.

[5] Avishai Ceder. *Public Transit Planning And Operation: Theory, Modelling And Practice*. Butterworth-Heinemann (Elsevier Ltd), 2007.

[6] Thomas L Rodeheffer. The symmetric shortest-path table routing conjecture. *Microsoft Research, Silicon Valley*, 2013.

[7] Daniel Delling, Thomas Pajor, and Renato F Werneck. Round-based public transit routing. *Transportation Science*, 49(3):591–604, 2014.

[8] Daniel Delling, Andrew V Goldberg, Thomas Pajor, and Renato F Werneck. Customizable route planning in road networks. *Transportation Science*, 2015.

[9] Daniel Delling, Julian Dibbelt, Thomas Pajor, and Renato F Werneck. Public transit labeling. In *Experimental Algorithms*, pages 273–285. Springer, 2015.

[10] Ittai Abraham, Daniel Delling, Andrew V Goldberg, and Renato F Werneck. Alternative routes in road networks. *Journal of Experimental Algorithmics (JEA)*, 18:1–3, 2013.

[11] Ling-Yin Wei, Yu Zheng, and Wen-Chih Peng. Constructing popular routes from uncertain trajectories. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 195–203. ACM, 2012.

[12] Tim Dwyer and Lev Nachmanson. Fast edge-routing for large graphs. In *Graph Drawing*, pages 147–158. Springer, 2009.

[13] Hannah Bast, Daniel Delling, Andrew Goldberg, Matthias Müller-Hannemann, Thomas Pajor, Peter Sanders, Dorothea Wagner, and Renato F Werneck. Route planning in transportation networks. In *Algorithm Engineering*, pages 19–80. Springer, 2016.

[14] Elena Georgievna Krushel, Ilya Victorovich Stepanchenko, Alexander Eduardovich Panfilov, and Elena Dmitrievna Berisheva. An experience of optimization approach application to improve the urban passenger transport structure. In *Joint Conference on Knowledge-Based Software Engineering*, pages 27–39. Springer, 2014.

[15] Paul Mees, John Stone, Muhammad Imran, and G Nielson. Public transport network planning: a guide to best practice in nz cities. Technical report, 2010.

[16] Daniel Delling, Andrew Goldberg, and Renato Werneck. Hub label compression. In *Proceedings of the 12th International Symposium on Experimental Algorithms (SEA'13)*. Springer Verlag, January 2013.

[17] Holger Bast, Stefan Funke, Peter Sanders, and Dominik Schultes. Fast routing in road networks with transit nodes. *Science*, 316(5824):566–566, 2007.

[18] PTV Group. PTV Visum. `http://vision-traffic.ptvgroup.com/en-uk/products/ptv-visum/`.

[19] INRO Software. Emme. `http://www.inrosoftware.com/en/products/emme/`.

[20] Citilabs. Cube. `http://www.citilabs.com/software/cube/`, 2015.

[21] NASA. Transims. `https://code.google.com/archive/p/transims/`.

[22] Transport Simulation Systems. Aimsun. `https://www.aimsun.com/wp/`.

[23] Natalia Sadovnikova, Danila Parygin, Maria Kalinkina, Bulat Sanzhapov, and Trieu Ni Ni. Models and methods for the urban transit system research. In *Creativity in Intelligent, Technologies and Data Science*, pages 488–499. Springer, 2015.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1307-1313 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

A Secure Communication Framework for ECUs

Ali Shuja Siddiqui*[1], Yutian Gui[1], Jim Plusquellic[2], Fareena Saqib[1]

[1]*Dept. of Electrical and Computer Engineering, University of North Carolina at Charlotte, U.S.A.*

[2]*Dept. of Electrical and Computer Engineering, University of New Mexico, U.S.A.*

ARTICLE INFO

Article history:
Received: 24 April, 2017
Accepted : 24 June, 2017
Online: 01 August, 2017

Keywords:
Automotive Security
Controller Area Network
Physical Unclonable Functions

ABSTRACT

Electronic Control Units (ECUs) generate diagnostic and telemetric data that is communicated over the internal vehicular network. ECUs are resource constraint devices and have limited resources to devote for data security. In recent times, threats against vehicular networks have emerged that require attention of the research community. In this paper, we demonstrate data security threats in automobile, present a hardware based security framework that provides real time secure communication using lightweight cryptographic primitives and propose hardware based authentication protocol. Implementation details, performance and security analysis of proposed framework are presented.

1. Introduction

Vehicles are no longer composed of just mechanical components with limited control operations of electrical components. Digital transformations in electronics have contributed in the advancement of automotive industry in terms of customer interactions and experience in improving the overall driving experience of a vehicle and critical operations. Some examples of on-board electronic components are air-fuel mixing, temperature-control, thermostats, vents, doors-locking, seat-adjustments, radio, wind-shield wipers, transmission, ignition, engine, steering, brakes, acceleration, air-bags, navigation, in-dash entertainment, cameras, braking system, etc. Electronic systems that are connected with on-board systems are called Electronic Control Units (ECU). ECUs work in cooperation with each other, and need to communicate their state with one another. Instead of using dedicated signal wires for information exchange, that may require a lot of wiring and effort on part of the car manufacturer, the automotive industry has moved towards using data communication networks. Similar to traditional data communication networks, data security of the ECU network is becoming a concern. Following section covers the vulnerabilities, and recent attacks on on-board car networks and the new security requirements. The security vulnerabilities are demonstrated and existing countermeasures and their limitations are discussed. This is an extension to the paper [1] which was presented in IEEE Vehicular Networking Conference 2016. Columbus, OH, U.S.A.

1.1. Recent Documented Attacks

In 2014, a group of researchers were able to remotely control a Jeep Cherokee[2]. The vehicle is equipped with an on-board infotainment system. To gather local information, it uses an internet connection to connect to its server. The researchers could remotely hack into the dashboard system, using software vulnerability. The dashboard system is also connected to the internal network of connected ECUs using a popular network standard known as Controller Area Network (CAN)[3]. This gives access to all the other actuation units such as the brakes, accelerator, steering control etc. A more recent attack is discussed in [4]. The authors extracted master global shared encryption keys from deciphering the software code using a wireless key fob. This gave the researchers access to shared keys used by all the vehicles using the wireless key fob system from the affected brand.

1.2. Current Security Needs

A robust and secure framework is needed to improve the resiliency to malicious attacks. Traditional cryptographic primitives are computation intensive and rely on secrecy of shared

*Corresponding Author: Ali Shuja Siddiqui, Dept. of Electrical and Computer Engineering, University of North Carolina at Charlotte, U.S.A.
Email: asiddiq6@uncc.edu

or session keys, and are applicable on large systems like servers and secure databases. These are unsuitable for embedded devices with fewer resources to allocate dedicated hardware for implementing security protocols with extra storage requirements. Alternatively, if external *Hardware Security Modules* (HSM) are considered to provide security, they are costly and require additional physical area and power.

This paper demonstrates a hardware based light authentication framework integrating strong physical unclonable function for secret key generation and light cryptographic primitives, such as Elliptic Curve Cryptography to improve authentication protocol for small and resource constrained embedded devices at lower levels of composition. We evaluate the security methods for CAN protocol and propose techniques to mitigate those attacks on ECUs. Furthermore, the framework addresses the issue of unauthorized physical access to a vehicle via electronic means, and to the connected digital components within a vehicle.

2. The CAN Bus Protocol

The CAN bus, known as ISO 11898 architecture is a multi-master bus connecting embedded components, I/O interfaces as well as gateways to communicate with the external world. It consists of three signals, CAN High, CAN Low and Ground, and the CAN standard specifies three data rates 125kbps, 500 kbps and 1Mbps. In the more recent standard for CAN Bus, named *CAN Flexible Datarate*, CAN allows transfers of additional frames at non-standard and faster speeds.

Figure 1 describes the standard data frame of CAN bus. The SOF bit denotes the start of frame, 11 bits store destination ID, Data field of 64 bits and CRC field is 15 bits to hold the checksum of the entire packet. The length of the standard CAN frame is 108 bits whereas 131 for an extended CAN frame. On the physical medium CAN uses Non-Return to Zero (NRZ) encoding, that is after five consecutive zeroes on the stream, an extra '1' bit will be added to that position before putting it on the bus.

SOF	Arbitration	Control	Data	CRC Field	ACK	End of Frame
1 bit	12 bits	6 bits	0 – 64 bits	16 bits	2 bits	7 bits

Figure 1 Standard CAN Bus Frame.

3. Threat Models

CAN is susceptible to replay, man in the middle and stealing identifiers attacks. We demonstrate these attacks on CANBus, with the testbed using Raspberry Pi 2, PiCAN Duo CANBus Shield, Arduino UNO and CAN-BUS Shield.

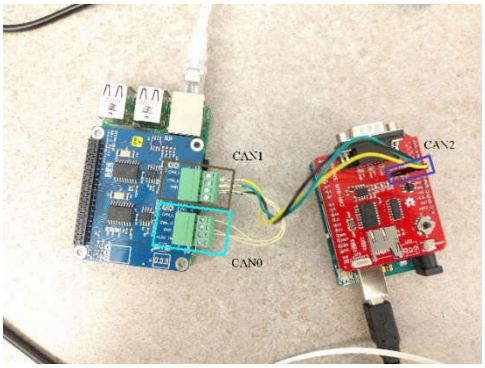


Figure 2 Experimental Setup.

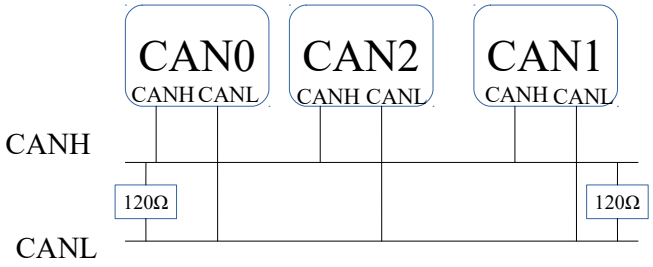


Figure 3 Block diagram of the Experimental Setup.

The PiCAN shield with two CAN nodes is connected with the Raspberry Pi. These nodes are considered to be legitimate nodes on the network. These nodes are referred as CAN0 and CAN1, as shown in Figure 2. The malicious node CAN2 is a CAN shield which is connected to an Arduino UNO. Figure 3 describes the block diagram of the complete setup.

3.1. Eavesdropping

CANBus is an insecure communication medium that transmits data in broadcasting the unencrypted packets to all connected nodes on the network. Malicious node getting physical or remote access to CANBus network has access to all the packet transfers and payload. Figure 4 shows traffic of packets on CANBus. The messages and destination identifiers are broadcasted and sent unencrypted on CAN. CAN0 messages for other nodes can be seen and recorded by CAN1. Similarly, any other device physically connected to the CAN bus can read all the communication over the network.

To mitigate eavesdropping, the information passing on the bus needs to be securely encrypted. In [5], the authors present an external security solution, and [6] proposes 3DES encryption and an external controller for the transfers and traffic monitoring. The added area, power requirements and slower performance of the software based encryption and decryption processes make the solution impractical.

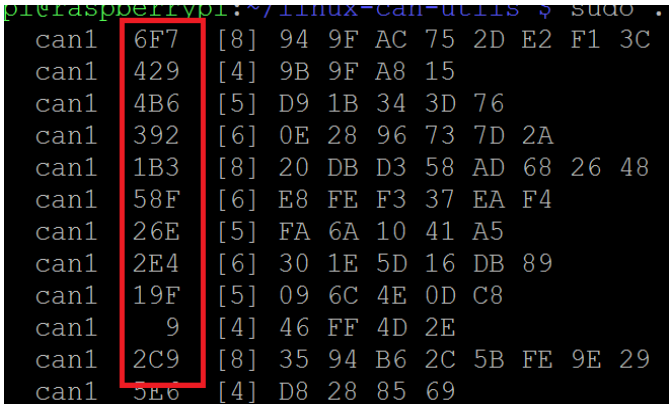


Figure 4 Screenshot of data transfer on the CAN.

3.2. Stealing Identifiers

As shown in the Figure 4, the destination address in the identifier field are visible to the hacker/ attacker on the CAN bus. All devices probe the ID field of each message received and based on the ID, deduce if the message was meant for them or not. Malicious device records these identifiers and can launch attacks against other nodes.

In [7], an authentication and encryption mechanism is described, but it requires about three times more CPU cycles than the cycles required for normal operation of a CAN node.

3.3. Non Repudiation

The frame consists of the destination identifier and not source identifier. The communication network cannot confine the attack or identify the originating node based on the network analysis. This results in incorporating non-repudiation requirements at each device (ECU).

3.4. Denial of Service

Denial of Service attacks makes the resources or services unreachable; the CAN communication channel is made inaccessible by the malicious node CAN2 by broadcasting random messages on the bus without pause as demonstrated in figure 5, while CAN0 and CAN1 are exchanging information. With the vendor provided code, the legitimate nodes stopped operation and notified about receiving a buffer overflow error as shown in Figure 6.

```
can1  7DF  [8] 02 01 05 00 00 00 00 00
can1  7DF  [8] 02 01 05 00 00 00 00 00
can1  7DF  [8] 02 01 05 00 00 00 00 00
can1  7DF  [8] 02 01 05 00 00 00 00 00
can1  7DF  [8] 02 01 05 00 00 00 00 00
can1  7DF  [8] 02 01 05 00 00 00 00 00
```

Figure 5 Screenshot of CAN2 flooding the network.

```
pi@raspberrypi:~/linux-can-utils $ sudo ./cangen can0
write: No buffer space available
pi@raspberrypi:~/linux-can-utils $
```

Figure 6 Screenshot demonstrating buffer-overflow occurring at CAN0.

In section 4 we propose a hardware based framework to introduce secure communication over CANBus and provide countermeasure against the discussed attacks.

4. Hardware Based Encryption and Digital Signature Solution

The limited processing power available on the hardware and real time constraints bound the traditional software based cryptographic solutions to provide secure communication. In embedded systems with a software solution, an attacker can extract the executable source code of the device using various methods to reverse engineer cryptographic algorithms and keys[8][9][4].

We propose hardware based secure enhancements to each connected electronic control unit (ECU). The secure framework can be embedded into the processor fabric itself as a co-processor. The hardware design is made to minimize software dependency on the resource scarce devices. Additionally, with incorporation of hardware based obfuscation techniques, reverse engineering attacks or attempts for extraction of keys using hardware based attacks can be further mitigated. The proposed framework requires no secret keys to be stored on non-volatile memory within the nodes/ECUs, thus eliminating the probing attacks.

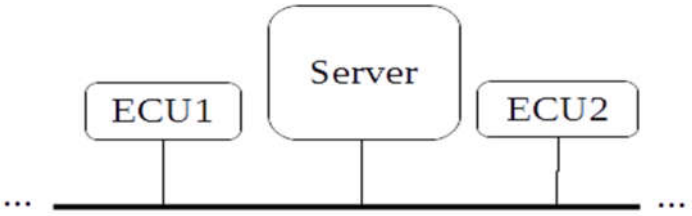


Figure 7 Server locally connected to the nodes.

The security framework is based on client server architecture. A local server node present on the network is responsible for handling all ECU nodes that are joined on the network. Each client must first be registered and authenticated at the server before they can communicate with the other nodes on a network. This process takes place in a secure and trusted environment, referred to as *registration*. Registration can be performed during manufacturing, assembly of vehicle components or at the trusted diagnostic centers. Figure 7 illustrates the local on-board server.

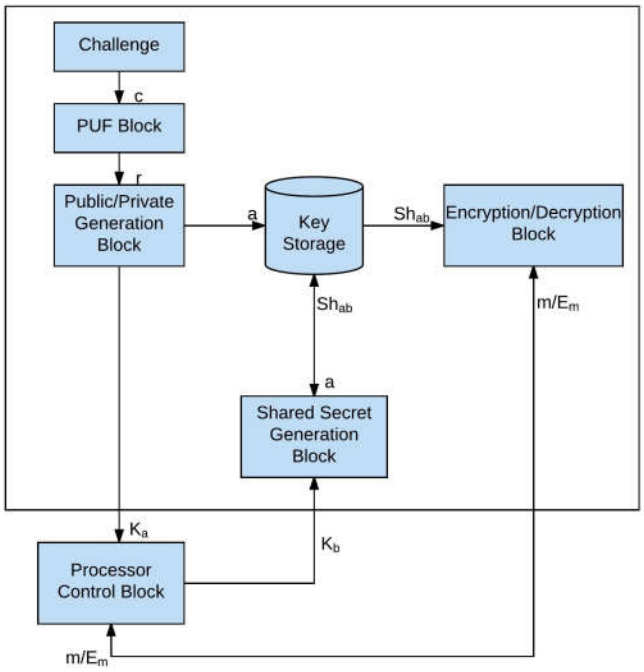


Figure 8 Secure block implemented at each client node.

Figure 8 shows the enhancements and design flow in form of block diagram of each secured ECU. Our framework assumes the server is secure and trusted. Following subsections discuss each of the components and their function in the framework.

4.1. PUF Block

Physical unclonable function (PUF) is an emerging physical layer cryptographic primitive used in hardware security and privacy protocols. These are embedded structures that utilize inherent manufacturing process variations to extract unique but reproducible secrets [10][11]. PUFs are based on a challenge-response pair (CRP) mechanism. The challenge for a PUF is defined as a digital input, usually in the form of a bitstring of '0's and '1's. The output of a PUF is also digital but for most PUFs, this requires an on-chip mechanism to convert the small analog variations leveraged by the PUF to be digitized.

There two different types of PUFs, namely *Weak* and *Strong* PUFs. A major difference between the two is that the weak PUFs have few challenges for which they can uniquely generate a key whereas strong PUFs have a large challenge space and therefore have a unique response for most of the challenges. The bit-

generation for cryptographic applications is a two-step process that is enrollment and registration. During the enrollment process, each PUF is given a set of challenges and the response pairs are recorded. Later when the PUF is in field, these responses are regenerated for use in identification and encryption applications.

During the registration, each ECU produces a unique reproducible response/key that serves as private key. This response is reproduced using the error correction code and is used to generate public private pair. Helper data for reproducing the secrets is stored on the NVM of ECU for the given challenge and public key is communicated to the secure server, where it is stored. PUF based authentication is implemented within vehicle network[11]. All client nodes (or devices) are enrolled at the server before they can be used. The registration process is summarized in Table I.

Table I. Registration of nodes in a trusted environment.

Algorithm: Registration process of client nodes in a trusted environment.
Input: Challenge c, configuration parameters for the ECC curve.
Output: Public key K_a of ECU.
1: At each client node, input a challenge *c* for the PUF Block.
2: PUF block computes a response *r*.
3: Generate a public private key pair (a, K_a) using PUF response r and configuration parameters.
4: K_a for each ECU is stored communicated to server, where it is stored in a public database accessible to every node on the bus.
5: Public key of the server, K_s is stored in non-volatile memory at each client node for later access.

4.2. Private/Public Key Generation Block

We implement Elliptic Curve Cryptography Diffie-Hellman (ECDH) based asymmetric key exchange and encryption engine in hardware. ECC algorithm is suitable for the resource constraint devices, providing same strength of security as RSA. ECC algorithms are implemented in Galois Field(GF) which limit the range of output values and thus reducing the resource requirements. ECC implementations have been presented over two types of Galois Fields, namely Prime Fields (GF(Fp)) and Binary Fields (GF(F2^m)). National Institute of Standards and Technology (NIST) provides a set of recommended elliptic curves (http://csrc.nist.gov/groups/ST/toolkit/documents/dss/NISTReCur.pdf).

This block is used to generate public and private keys for a node. It is used in the one-time registration process (as described in the previous subsection). Based on the reliable PUF response, public and private keys are generated at each ECU during the authentication process that is initiated every time the vehicle starts (turning on of the ignition).

4.3. Authentication and Shared Key Generation Process

During the startup, each connected device requires authentication with the server to enable trusted environment. ECU sends a packet with its own id as target identifier and its public key encrypted with the shared key formed using ECC in the payload. The shared key block is used to generate a shared key between two communicating nodes to have an encrypted communication. At the startup of vehicle each device/ECU, the private key of the current node and the public key of the server, are used in Elliptic curve Diffie-Hellman (ECDH) algorithm for the secret shared key generation. This shared key is used in symmetric key encryption algorithm. All generated keys are stored in the volatile memory for the whole time vehicle is running and requires real time communication.

Server decrypts the payload and compares the public key with the sender ID sent in the arbitration bits. Once the ECU is authenticated, server shares the public keys of the other ECUs the requesting node is allowed to communicate.

Table II. Algorithm for authentication performed at the beginning of each session.

Algorithm: Authentication process.
Input: None.
Output: None.
1: Each ECU generates PUF response r using stored challenge c, and helper data.
2: With *r* as the input to the Public / Private Key Generator Block, generate public and private keys, K_a and *a* respectively.
3: With public key of the server K_s and private key *a*, generate shared symmetric encryption key Sh_{ab} using the Shared Key Generation Block.
4: Encrypt the generated public key K_a with the shared key Sh_{ab}.to form message E_m.
5: Transmit E_m to the server while setting the node ID I_A of the current node as CAN message arbitration ID.
6: At the server, once an encrypted message E_m is received, Server decrypts it using the shared encryption key associated with I_A to get message *m*, which holds the public key of node A, K_a.
7: If K_a received is equal to the K_a stored in the database then consider node A authorized for the rest of the session.
8: Wait for other nodes to register themselves.
9: Server prepares a message for each ECU encrypting with the ECU's shared key containing the public keys of all the registered nodes.
10: Once this message is received at a node, decrypt it using the shared key.
11: Each node communicates with the other nodes and generates session keys using both its private key and other nodes public key.

The authentication process is a timed process. If any node fails to authenticate itself with the server in the time allocated, then it is blacklisted for communication for the entirety of that session. Private keys for all nodes are generated at run time and are not stored to ensure security. No private keys leave the node and are erased at the end of session. Figure 9 illustrates the sequence of communication during authentication of a node.

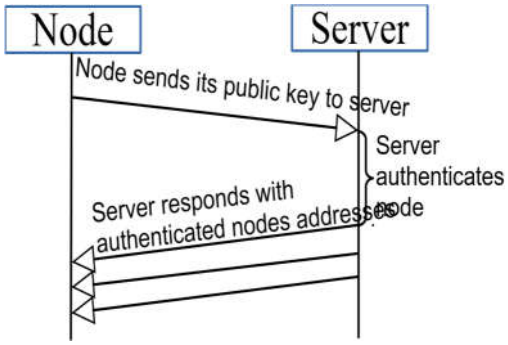


Figure 9 Sequence Diagram for Authentication.

4.4. Key Storage

The key storage block consists of a non-volatile memory (NVM) to store:

- Input challenges to capture unique responses from PUF on each ECU.
- Helper data consists of tuning stages configuration to re-produce PUF response for private key(discussed in the next section).
- Public key K_s of the server for server authentication.
- Parameters to generate public private key pairs.

The framework allows each ECU to generate the private key on fly thus does not require it to be stored in a NVM. Private keys can be stored temporarily in volatile storage along with the shared session keys between ECUs. For each pair of communicating nodes, the following elements are stored in the volatile key storage:

- Sh_{ab}: Generated Shared secret key for encryption for each pair of nodes (a being the source node and b the node being destination node).
- K_b: Public key of the communicating node.

No access is possible to the storage other than to the nodes connected to it.

4.5. Encryption / Decryption Block

Once shared keys are generated for each pair of nodes, these keys can be used for encryption and decryption for the given session. Every message that is sent must first be encrypted before it can be transmitted. Selection of a symmetric key algorithm is dependent on the choice of the system designer, for example an AES-128 engine. The message is sent to the Encryption / Decryption Block that encrypts the message at the sender and decrypts the message at the receiving node using the shared key.

5. Implementation and Performance Analysis

Following section discusses the implementation details of framework and its performance evaluation. Xilinx Kintex KC705 FPGA Evaluation Board is used to implement the design IP with the clock speed of 200 MHz.

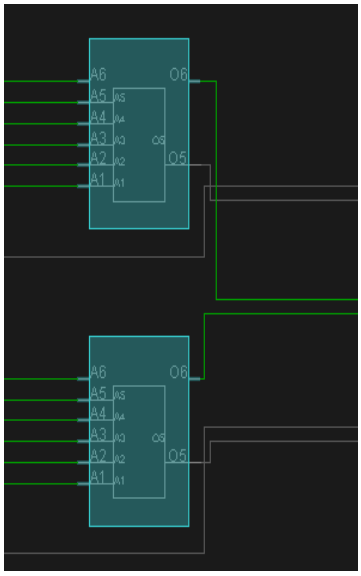


Figure 10 Implementaion of a single PUF stage.

5.1. PUF Implementation

The *Arbiter* PUF is a strong PUF with multiple CRPs that utilizes the process variations in the identical delay lines to generate response bits. The traditional Arbiter PUF implementation is hard to implement on FPGAs because of limited capability on place and route process. This causes systematic bias and the bitstream is not random. An alternative Arbiter PUF implementation is proposed in [12] that utilizes the internal delays of an LUT as a source of entropy for the process variation. The FPGA implementation design details with symmetric delay lines are in further explained in [12].

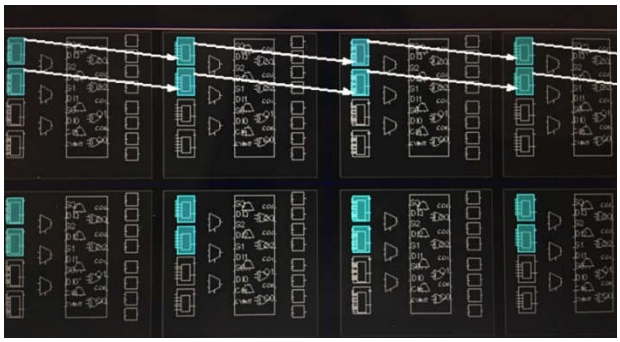


Figure 11 Section of a PUF delay line.

A single stage of the delay line is illustrated in figure 10. The output is a function of A1, where A2 to A6 inputs decides the route of input to output. LUT is configured as an invertor, when the input signal A1 is 0, the output changes to 1 and vice versa. A rising edge is propagated to the first LUT of both upper and lower path that propagated through 64 stages. In our implementation, we have instantiated 16 copies of lower and upper delay lines to produce 16-bit output response for each given challenge. The routing between the stages is shown in Figure 11.

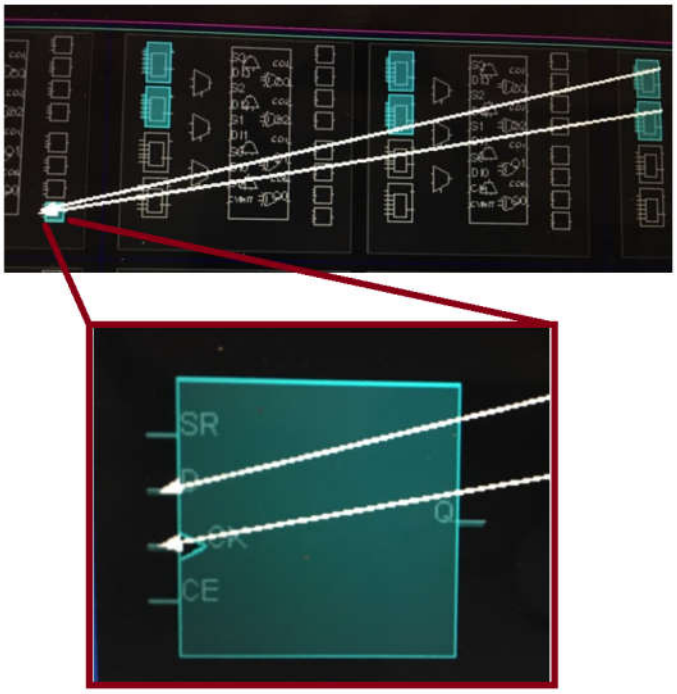


Figure 12 PUF Response stored in DFF.

In order to fine tune the delay difference due to routing, tuning stages are added to the end of the delay lines as described in [12]. Each delay line ends at an arbiter, in our implementation, we are using a D Flip Flop. The difference in the delay caused by the physical variation will result in either capture of a '0' or a '1' based on whether the lower or upper rising edge is reached on the flip-flop clk or D input respectively. The capture stage is illustrated in figure 12. The implementation uses a total of 2416 LUT slices.

5.2. Key Generation Subsystem

ECC engine operating over the binary field GF (2^163)is used [13]. The ECC engine is only used on boot up to generate the public and the shared keys. Figure 13 shows the internal configuration of the ECC core.

The selected implementation of the ECC engine is organized as a pseudo multi-core processor and has a total of three cores for finite field (FF) computation. They are connected using a RISC like instruction controller. For computation of a single point multiplication, the ECC core requires a total of 1428 cycles. Since our target platform is running on 200 MHz, the total time to compute a single multiplication point is 7.14 microseconds. In the process for shared key generation, there is a need for two single point multiplication operations at each node. On our target platform, the ECC core uses a total of 25,454 LUT slices.

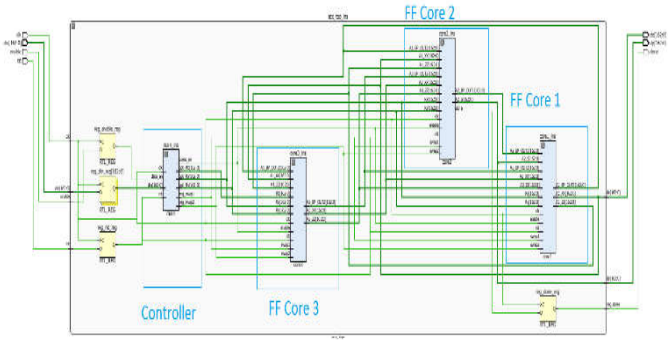


Figure 13 Block Diagram of the ECC Core.

5.3. Encryption/ Decryption

The encryption/decryption module consists of AES-128 for encryption. The AES engine uses the shared key generated by the ECC subsystem to encrypt any traffic leaving the node and decrypt any traffic coming in. AES-128 engine takes a total of 22 clock cycles for encryption and decryption and a total of 110 nanoseconds. The AES engine has a footprint of 2560 LUT slices.

Table III shows a comparison of different block computation speeds at different clock speeds.

Table III. Comparative analysis of block computation speeds at different system clock rates.

	AES-128	FF Multiplication
At 60 MHz	366.66 ns	23.8 µs
At 100 MHz	220 ns	14.28 µs
At 200 MHz	110 ns	7.14 µs

5.4. Storage Elements

For implementing the storage elements of the framework, we are using the FPGA's on-fabric block RAM resources. A block RAM of 128-bit width and 512 elements is instantiated. On our target board, a total of two 36K BRAM elements were used. This memory element has a read and write speed of one cycle.

5.5. Timing Analysis

The framework has system clock speed of 200MHz and CAN speed is 1Mbps. Time for transmitting one bit at the data rate of 1Mbps is 1µs. We are ignoring bit stuffing due to NRZ encoding in our computation. The time required to transmit one data frame is 1 x 108µ = 108 µsecs. Each frame has a total of 64 data bits, to send an AES-128 encrypted 128-bit message, two data frames are required. The total time for transmission for an encrypted message is 108µ x 2 = 216 µsecs.

During Authentication

At the time of boot up, at each node the first the public key is generated using the private key. This process requires one point multiplication operation to be performed in the ECC core. As stated in subsection 5.2, this operation requires a total of 7.14µsecs. Using the stored public key of the server, another point multiplication operation will be required to generate the shared encryption key. Per Table II, this key is encrypted before it is sent to the server. The encryption operation takes 110 ns. Therefore, the total minimum time required from boot-up to sending the server the shared key is:

$$(7.14\mu \times 2) + 110\text{ns} + (2 \times 108\mu) = 230.39\mu\text{s}.$$

For each node that has been authenticated by the server, the server will send a 3-part message (as described in figure 12) to all the connected nodes. The first part contains the node ID of the node authenticated, and the two parts (total of 128 bits) will consist of the encrypted public key of an authenticated node.

Time taken to send these three messages is:

$$108\mu \times 3 = 324\ \mu\text{s}.$$

This process will be repeated for all successfully authenticated nodes.

Since the public key of an authorized node is encrypted in the message, it needs to be decrypted. This process will take another 110 nanoseconds. To generate the shared key for that public key requires another 7.14 µs by the ECC unit.

Receiver ID	ID Packet 1	ID Packet 2

Figure 14 Server's reply to a node after authentication.

During normal operation

During normal operation, CAN message is encrypted, it will take a total of two frames to transmit a message. Table IV gives a message overhead incurred for authentication on CANBus communication operation.

Table IV. Overhead overview at standard can connection speeds.

	125 kbps	500 kbps	1 Mbps
Time for sending one frame	864µs	216µs	108µs
During Authentication			
Node sending encrypted message to server (2 frames)	1.728ms	432µs	216µs
Server's reply (3 frames)	2.592ms	648µs	324µs
Normal Operation			
For a message (2 frames)	1.728ms	432µs	216µs

6. Security Analysis

The proposed framework assumes a secure server, and the legitimate/authorized ECUs are not compromised and does not allow unauthorized code execution. In the proposed framework, the payload of CAN Bus packets are encrypted. There are no private key storage or exchanges in the untrusted field. An ECU node stores only the public key of the server. This makes it difficult for an attacker to retrieve keys through probing or man in the middle attack.

No new device can be added in the field, since the device needs to register with the server in a trusted environment before the server can send the shared key of this node to the other connected ECUs, it must be authenticated. Since the malicious node's public key is not stored in the server, it will not authenticate it and therefore will not communicate its ID with the other nodes. Thus, eliminating backdoors for accessing the network physically. All the nodes have different keys and as such on compromise of a single node, even if an attacker can retrieve a private key from one node, the private key to the other node will still be unknown to the attacker.

In the scenario where an attacker has acquired the public key of a legitimate node on the network or in case of a server database compromise, an adversary can only retrieve the public keys of participating ECUs. Once a communication from the adversary is initiated, the legitimate node will reject it since public key was not initially sent by the server.

7. Conclusion

In this paper, we investigate the threat models of internal vehicular network communication using CANBus. We present a secure framework for authentication and point to point encrypted communication for ECUs. Paper presents the architecture and implementation details and area overhead of FPGA implementation. The hardware based security solution for the resource constraints and time critical applications are viable for the automotive industry.

Acknowledgement

This work is supported by National Science Foundation under grant no. 1566530 and 1623299.

References

[1] A. S. Siddiqui, Y. Gui, J. Plusquellic, and F. Saqib, "Poster: Hardware based security enhanced framework for automotives," in *2016 IEEE Vehicular Networking Conference (VNC)*, 2016, pp. 1–2.

[2] C. Miller and C. Valasek, "Remote Exploitation of an Unaltered Passenger Vehicle," *Black Hat USA*, 2015.

[3] BOSCH, "BOSCH CAN Specification," 1991.

[4] F. Garcia, D. Oswald, and T. Kasper, "Lock It and Still Lose It–On the (In) Security of Automotive Remote Keyless Entry Systems," in *USENIX Security 2016*, 2016.

[5] A. Hanacek and M. Sysel, "Design and Implementation of an Integrated System with Secure Encrypted Data Transmission," Springer International Publishing, 2016, pp. 217–224.

[6] G. Singh, "A study of encryption algorithms (RSA, DES, 3DES and AES) for information security," *Int. J. Comput. Appl.*, 2013.

[7] Q. Wang and S. Sawhney, "VeCure: A practical security framework to protect the CAN bus of vehicles," in *2014 International Conference on the Internet of Things (IOT)*, 2014, pp. 13–18.

[8] S. Ravi, P. Kocher, R. Lee, G. McGraw, and A. Raghunathan, "Security as a new dimension in embedded system design," in *Proceedings of the 41st annual conference on Design automation - DAC '04*, 2004, p. 753.

[9] D. Papp, Z. Ma, and L. Buttyan, "Embedded systems security: Threats, vulnerabilities, and attack taxonomy," in *2015 13th Annual Conference on Privacy, Security and Trust (PST)*, 2015, pp. 145–152.

[10] C. Herder, M.-D. Yu, F. Koushanfar, and S. Devadas, "Physical Unclonable Functions and Applications: A Tutorial," *Proc. IEEE*, vol. 102, no. 8, pp. 1126–1141, Aug. 2014.

[11] W. Che, F. Saqib, and J. Plusquellic, "PUF-based authentication," in *2015 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, 2015, pp. 337–344.

[12] M. Majzoobi, F. Koushanfar, and S. Devadas, "FPGA PUF using programmable delay lines," in *2010 IEEE International Workshop on Information Forensics and Security, WIFS 2010*, 2010.

[13] Y. Zhang, D. Chen, Y. Choi, L. Chen, and S.-B. Ko, "A high performance pseudo-multi-core ECC processor over GF(2 ^ 163)," in *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*, 2010, pp. 701–704.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1314-1322 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Content Reordering Based on Mouse-tracking for Web Applications

Md.Tanjim-Al-Akib*, Lutfullahil Kabir Ashik, Krishanu Chowdhury, Md. Zahid Hasan, Hosne Al Walid

Computer Science & Engineering, Ahsanullah University of Science & Technology, Dhaka-1208, Bangladesh

A R T I C L E I N F O

Article history:
Received: 30 April, 2017
Accepted : 15 July, 2017
Online: 01 August, 2017

Keywords:
HCI
Web Application
User experience
User interest
User engagement (UE)
Mouse-tracking
User Interest Capturing Algorithm (UICA)
Content Reordering Algorithm (CRA)

A B S T R A C T

Rising usage of web applications in every aspect of life has created intense need of enhancement of user experience and formation of intelligent web applications. This research work was conducted with a view to do so. This paper focuses on the reorganization of web application content based on user's specific interests while browsing any web application by providing new algorithms and a new approach.

1. Introduction

This paper is an extension of work originally presented in 19th International Conference on Computer and Information Technology (ICCIT). IEEE, 2016 [1]. The number of users using the internet is rapidly rising with the advancement of the internet as it helps users to find information of their interest with in very short possible time. Compared to the number of internet users in 1995 which was only 1% of world's total population, it has increased up to 40% in recent years. This mass communication media has gained so much popularity that people have been using this on daily basis.

The internet serves its users through web applications and websites. The reasons why web applications have gained so much popularity are given below-

- Zero installation cost
- Reduce business costs
- Reach anybody, anywhere in the world
- 24/7 accessibility
- Access to the latest information

Web applications are being used in medical, banking transactions, online buying selling, online marketing, communication etc. Web application can help doctors to prescribe elder people [2]. With the escalation in number of web applications, the amount of information is increasing. Any kind of web applications have numerous number of contents within it. With large number of contents it is often difficult to satisfy the users. Most of the time users have to go through a large number of information to find out their required information which is leading towards usability problems. To evaluate and improve web applications, usability issue has become an important field of study. Nowadays Facebook, YouTube etc. have become so much popular because of their advanced recommendation systems. In order to solve the usability problem researchers are consistently working on how user experience can be improved. How the user's experience can be improved is described in [3]. This paper focuses on how web application contents can be made more user friendly

*Corresponding Author: Md. Tanjim-Al-Akib, 575/2/B, Noyatola, Moghbazar, Dhaka-1217, +8801673582219 & akibtanjim@gmail.com.

through mouse tracking. We propose a generalized approach that will solve any kind of web application's usability problems at the same time improving the user experience.

2. Background

A large number of contents often make it hard for the users to find their items of interest. Solution to this problem is to make user-centered web applications. Developing a user centered web application is not an easy task as it's often very difficult to take user's feedback manually. There have already been many works for knowing user's preferences and expectations automatically without bothering the user [4]. An earlier study [5] shows adaptive navigation mechanisms and illustrates how these can help web applications. The paper [6] focuses on architecture, database design, local concepts, and user interface design for developing content management system. User interest can be predicted by eye tracking [7]. An alternative way for predicting user's engagement can be mouse activity and cursor movement. Mouse activity can be used for web designing to make the user interface more precise for user usage [8]. In our previous work [1] it's described how user-modeling and recommendation by mouse-tracking can be used for the construction of an intelligent e-commerce website. In the previous version of work, we only worked with the registered users of ecommerce websites. We tracked their mouse movement and clicked information for modeling their interest in database. Based on modeled data, we provided them recommendation so that they can easily find their product of interest. In this version of work, we suggested generalized algorithms which will support any kind of web applications. User interest data will be stored in the browser cookie rather than database which will improve the loading time of application and overall performance. More detailed brief of data collection process and content reordering is provided. We have compared some exiting approaches defining their problems and solving those issues with our system. This paper mainly focuses on how web application content can be organized in a manner that responds to user interest. For achieving this goal, this paper will focus on providing algorithms, observe user engagement, find user interest, store data in browser cookie, re-organize the web application contents based on interest data.

3. Overview of Previous Work

Previous approach that was taken earlier was to design a recommendation model based on the interests of users of ecommerce web applications. The overall process was to learn about the interests of users through mouse-tracking, analyze the interest data to create user model and recommend user according to the best match of their interest using. To complete the process, we proposed two algorithms: Modeling Algorithm (MA) & Recommendation Algorithm (RA). Modeling algorithm is used to collect user interests data through some specific button click (like search button, compare button & detail button) and mouse movement over various content. These information are stored in database. Recommendation algorithm uses these data from database and try to find out any matched product from the intersection of clicked data and mouse movement data and provide recommendation based on obtained product/products category/categories. On finding no matched product at first, it checks for maximum clicked data as it reflect more user interest specifically and recommends product according to the category of that/those product/products. If this happens that user have not yet clicked anything, then it checks for maximum mouse movement data and recommends product/products according to this. If both mouse clicked and movement data are null, then it recommends nothing but the default items.

4. Proposed Algorithm

4.1. User Interest Capturing Algorithm (UICA)

UICA or User Interest Capturing Algorithm is used to observe user behavior in a web application to capture user interest and store interest data in browser cookie. A web application has several content section. We mark each of them with id. At first, we find out the type of activity in which user is involved. In our case, there were two possible scenarios. One is button or hyperlink click and the other one is mouse movement over content. Each of the activity is stored in different way (different array for both click information and moue move information) in browser cookie. Whenever this section is clicked or mouse moved over the content for threshold time, we get that id. Figure 1 shows the code snippet how we identified each of the content.

```
<div class="row">
    <div class="col-md-12">
        <div class="content1">
            <div class="main-content" id="1"
                onmouseover="cookied_mousemove_entry(1)"
                onmouseout="cookied_mousemove_exit(1)"
            >
                <p>Some Information</p>
            </div>
            <div class="link content">
                <a href="some_link" id="1"
                    onClick="cookied_click_entry(1)"
                >
                    Some Link
                </a>
            </div>
        </div>
    </div>
</div>
```

Content Type ID for Data Collection

Figure 1 Code snippet of identification of a content

After getting the content id depending on what type of activity user is involved in, we find out whether that item existed in the browser cookie or not. Upon earlier existence we increased the counter of that product. If the item is not already stored in the then we store it in the browser cookie on relevant place (for mouse movement we store the information in cookied_mousemove_data array and for click activity we store the information in cookied_click_data array). Details discussion about the whole process is presented in the implementation section of this paper.

Figure 2 shows the visualization of how mouse data is stored

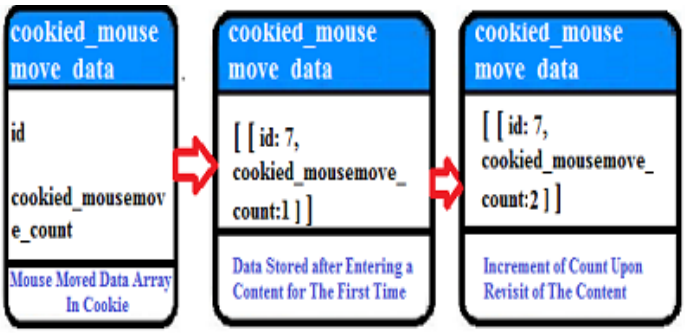


Figure 2 Visualization of how mouse data is stored.

Figure 3 shows User Interest Capturing Algorithm.

```
1.  cookied_data_click_count[ ] ← null;
2.  cookied_mousemove_count [ ] ← null;
3.  value =threshold;
4.  if (btn_or_hyperlink_click ( ) ==true) then
5.     id = get_item_id ( );
6.     existence_confirm = find_item_in_cookie_click_data(id);
7.     if (existence_confirm == false) then
8.        cookied_data_click_count[ count ] = cookied_data_click_count[ count ] +1;
9.     else
10.       cookied_data_click_entry (cookied_data_click_count[ ],    ['id':id,'count':1]);
11. else if (mouse_movement (value) then
12.    id = get_item_id ( );
13.    existence_confirm = find_item_in_cookie_mouse_data(id);
14.    if (existence_confirm == false) then
15.       cookied_mousemove_count [count]=cookied_mousemove_count [ count ] +1;
16.    else
17.       cookied_mousemove_entry (cookied_mousemove_count [ ], ['id':id,'count':1]);
18. else
19.    no_data_enrty ( );
```

Figure 3 User Interest Capturing Algorithm

4.2. Content Reordering Algorithm (CRA)

CRA or Content Reordering Algorithm as the name suggest is used for content reorganization based on browser cookie data which was collected through UICA. The process of reordering is discussed below:

First we get all the mouse moved and content clicked data from browser cookie. We find out the max clicked and max mouse moved item/items id/ids. Assuming that both of the max clicked and mouse moved items are not empty, the system will find out intersection of both information and try to find out related items. Finding upon any related items the content will be reorganized or else default content will be served to the user. If any of the mouse moved and clicked data is empty then the system will find out related items based on the kind of data that was available. When any related item/items is found, then the system will automatically reorganize its content accordingly. If there is no information stored in the cookie both for mouse moved and clicked data than the default content will be displayed. More brief discussion is covered in the Implementation section.

Figure 4 shows Content Reordering Algorithm.

Content Reordering Algorithm (CRA)

```
1.  clicked_data [ ] =  get_cookied_data_click_count();
2.  mouse_movemenet_data [ ] = get_cookied_mousemove_count();
3.  max_clicked_items [ ] = find_max_clicked_items(clicked_data );
4.  max_mouse_moved_items[ ]=find_max_mousemove_items(mouse_movemenet_data);
5.  if(max_clicked_items  !=null && max_mouse_moved_items !=null) then
6.       result = intersection(max_clicked_items ,max_mouse_moved_items );
7.       realated_items_id=get_item_id(result);
8.       if(realated_items_id !=null) then
9.              show_related_items(realated_items_id);
10.      else
11.             no_reordaring();
12.  else
13.      if(max_mouse_moved_items ==null && max_clicked_items != null) then
14.        realated_items_id=get_item_id(max_clicked_items);
15.        if(realated_items_id !=null) then
16 .            show_related_items(realated_items_id);
17.        else
18.             no_reordaring();
19.      else if(max_clicked_items ==null && max_mouse_moved_items != null) then
20.        realated_items_id=get_item_id(max_mouse_moved_items);
21.        if(realated_items_id !=null) then
22 .            show_related_items(realated_items_id);
23.        else
24.             no_reordaring();
25.      else
26.           no_reordaring();
```

Figure 4 Content Reordering Algorithm

5. Proposed Approach for Web Applications

Web applications engage users. Web interfaces that has confusing navigation and huge content bores the user. People often avoid these applications and go elsewhere to perform their tasks. The most challenging task for any web application is how to design so that it can satisfy all of its users. Successful web applications serves appropriate content which stimulate and maintain the users interest, motivation, and attention through relevant and exquisite information related to individual. In short, they ease users enjoyment of their tasks. User engagement (UE) and user experience can help user in judging product quality during interaction [9]. Our approach will solve this problem as discussed in previous section.

For any kind of user-centered application, it is mandatory to know about what each user interested in. User interest is very difficult to find out as it varies with age and culture. Click data can identify user interest [10]. Interest model can be built upon user reading and organizing of documents [11]. Interest of users can be known explicitly or implicitly. Explicit learning includes feedback from users or the rating system. As there is no benefit of giving feedback or rating has no influence on their user experience they may not give any rating [12]. In implicit learning procedure users remain completely unaware of the fact that they are giving information about their liked contents. Mouse scroll movement and time spent any content can be hint of user's interests [13]-[14].

Eye tracking is the best possible way to learn about user interest for this reason human eye has been studied in HCI [15]. The advertising industry, entertainment industry, packaging industry

and web design [16] can be significantly benefited from studying the visual behavior of the user. The automotive, medical and defense industries have also been using eye tracking technology to make us safer. Understanding of how eye moves has become an extremely important consideration in research and design. There are some problems in eye-tracking. They are:

- Eye-tracking equipment's are costly.
- For getting correct data these devices need proper alignment with user's eyes. As user's eye position cannot be fixed to a place as body moves so the task becomes more difficult.
- If one or both eyes do not move smoothly, accurately and quickly across a line or from one object to another it creates problems. This problem includes the inability to fixate (lock one's eyes) onto a single target (like a word on a printed page).
- If a person's word to word eye movement is not tracked precisely then it creates major problems like reading, writing or copying.

The remedy of this problem lies in mouse tracking. As users have to use mouse to interact with any web application, it is the best and most cost effective way to find out user interest from their mouse movement. Earlier studies shows that how mouse movement is related to eye movement [17]-[20]. We would like to mention here that this paper is beyond the debate of the relationship between eye movement and mouse movement. The paper [21] represents an estimation of visual attention on Search Engine Result Pages (SERP) by log-based mouse movement model. This includes snippets, aggregated results and advertisements. The mouse movement and click data can reflect user's behavior while interaction [8], [22], [23]. Previous research shows that user interest can be traced by clicking, reading, hesitating, mouse hovering and scrolling [24]-[27]. Biometric identification can be done by mouse movement [28]. It is essential for any web application to learn whether existing user interface provides support to different kinds of users [29]. The effectiveness, efficiency and usability of the website is improved by the involvement of the users in design and evaluation [30]. In the field of web marketing mouse tracking plays a vital role as it improves user experience in web applications by effective mouse movement and user hesitation levels [31].

Any software usually uses database to store information. Large scale application contains huge number of information offering to its users. In such situation it would not be feasible to use database to store user interest information as it can lead to crash or the application being slow. Other possible way is browser cookies.

User preferences, shopping carts information and anonymous data of third party applications like Google Analytic etc. are retained in cookies. User authentication is made dynamic and validated with the help of browser cookies [32]. Cookies also make user browsing experience better. The paper [33] defines how structured data should be stored in browser.

Continuous growth of web application has led to overload problem both for information and its usage. User often finds it difficult to find their necessary information when it is needed the most [34]-[35]. Recommendation system identifies the problem and comes up with a solution by finding user interest data from large number of contents [35]-[36]. Recommendation systems use details of user's views, opinions, liking, disliking, habits etc. to find out the best match information that is nearly or in some case exactly the same as their interests..

6. Implementation

Implementation steps are followed as below:

6.1. Demo Website

First and the most obvious necessary thing is a website. We developed a e-commerce web application for our testing with an attractive outlook.

6.2. Data Collection

The data were collected from user implicitly. We asked users to visit our system. System will then observed user behavior while interacting with our system and reorganize its contents according to user specific behavior. For knowing user interest we used our User Interest Capturing Algorithm (UICA). The whole process of data collection are described below:.

6.2.1 Click Data Collection: In any kind of interaction with the web application user click is a very important feature. As user is intended to click the things that attracts them most or the things that fulfill their needs, user clicked data were taken into consideration in our content reordering. Our system will check if user have clicked on button or hyperlink or not, if so then it will search for that content in the browser cookie where all clicked data are stored. If any match is found then system will increase the number of click of that item otherwise it will add the content in the clicked data as a new entry. The below figure shows how button click data will be stored in the cookie. Figure 5 how button click data will be stored in the cookie.

6.2.2 Mouse Data Collection: What if user only browses a system without clicking any content? User can only read the summery in this situation. If user mouse enters any specific content over and over again, this indicates that user is interested in that specific content. This also should be taken into consideration while collecting user interest information. Mouse movement data are collected on user's each movement. We defined a threshold time in our case it is 1 sec. If at any time user mouse remains still over a content for the threshold time then we will take that action in our consideration and check if the content is already in the mouse movement data in the cookie. If this happens then we will increase the number of mouse movement data of that item otherwise it will add the content in the mouse movement data as a new entry. Figure 6 shows how user mouse movement data would be stored in the browser cookie.

6.3. Content Reordering

For content reordering our developed system will use Content Reordering Algorithm (CRA). As user mouse may remain still at some content even if user has no interest in that content. This will not provide us with an actual solution. That is why we will mainly focus on user clicked data.

Figure 5 Process of collecting user clicked data.

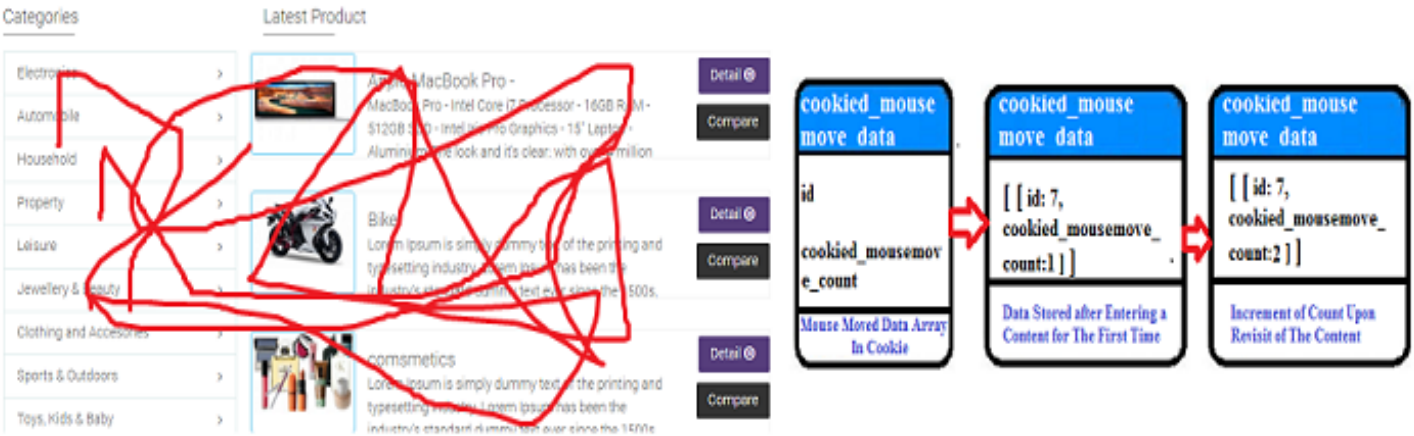


Figure 6 Process of collecting user mouse movement data.

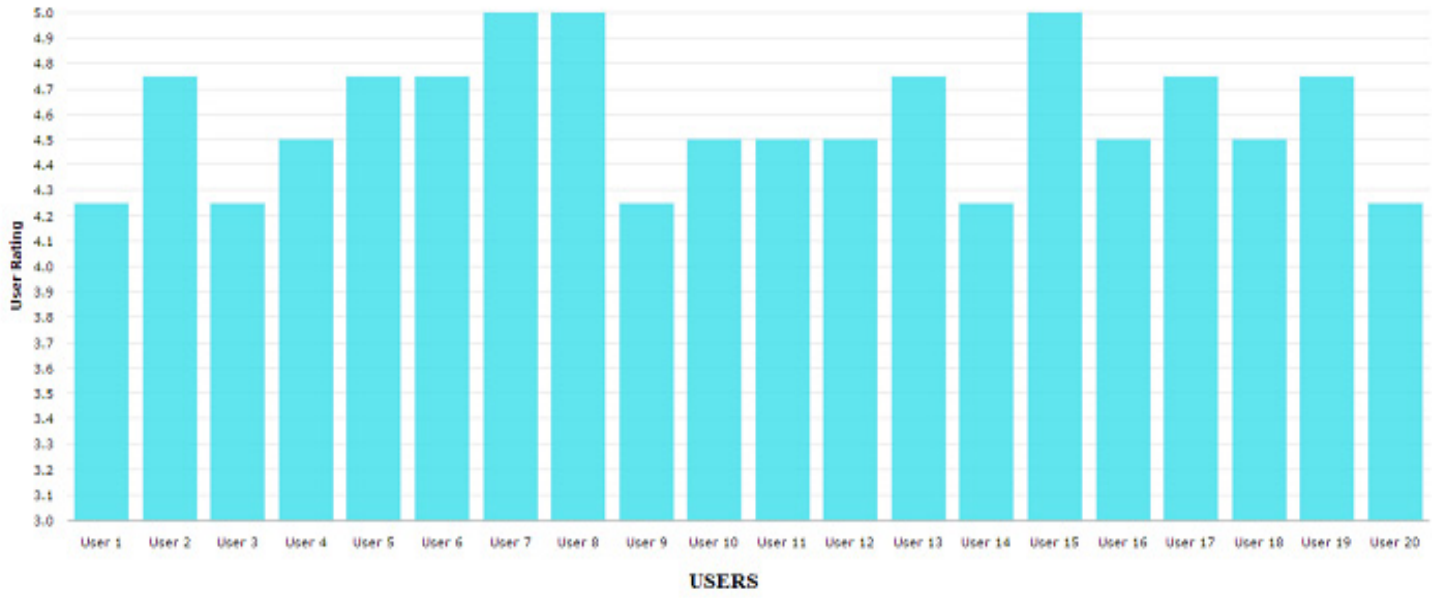
The total process of content reordering is given below in brief:

- First of all the system will get all the clicked and mouse movement data from cookie.
- Finding out the max clicked and max mouse moved items from the cookie is the next step.
- The next step of our procedure is to check whether both of the max clicked and max mouse moved items are empty or not. When this happens the system will then find the intersection of max clicked and max mouse moved items. On finding any item or items then it organizes its contents in manner that responses to the related items to the item or items from the intersection. If it fails to find any item, the system will not change or reorder its contents at all.
- On the failing of above situation system will check out for max clicked items. If it finds one or more item/items then content of the web application will adapt to the situation by reordering its contents.
- Whenever step 3 and step 4 fails then the system will prioritize on the max moused moved items. In the state where the system finds one or more items, then it will reorder its contents related to those item/items. Otherwise, there will be no reordering.
- Lastly if a situation arises that user visits the application for the first time or above step 3, 4, 5 don't provide with the related item/items required then the content will remain the same as it was earlier.

7. Result Analysis

For evaluating our algorithms and data collection process our demo website was put into a web server. That was there for about a month. We asked the visitors to browse our application for 5-10 minutes. By this time using our UICA, we collected user interest data and system adapted to its contents using CRA. We also asked users explicitly about their liking's (contents) in our system for better understanding of the efficiency of our proposed algorithm UICA and CRA.

Over the period of a month data was collected from 120 users of different age and profession. Figure 7 shows rating from anonymous users. Over all findings are quite satisfactory. Rating from user averaged 4.43 out of 5.00.



The system incredibly got accuracy of 91.67% (110 out of 120 users) of serving user according to their interests. Figure 8 shows the accuracy of our evaluation.

For better understanding of our approach we will briefly describe content reorganization for a random specific user among the 120 users. Figure 9 shows the content before user browse our system. The user is a businessman aged around 42 years. He visited our developed e-commerce web application. Figure 10 shows the visualization of his interaction with the system.

While he was interacting with the system, system closely monitor his mouse movement and click on web content. We stored the type id/ids in our case the category of the product both for mouse movement and click in browser. Figure 11 shows how data is stored in user's browser and findings of user interests.

Reorganization of content is done with the help of CRA. Using the algorithm, the system found out that he has mostly visited the electronic goods or clothing type items (type/category id 7 and 15) which was the intersection of mouse moved data and clicked. The system than reorganized its products which were of electronics and clothing (serve products of electronics category first as it has the most count value which was acquired from the sum of click and mouse moved count value). Figure 12 shows the content after reordering based on his interest data stored in the browser cookie.

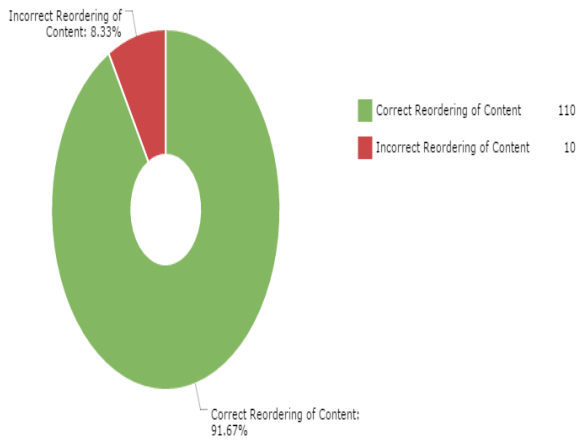


Figure 8 Content reordering accuracy

Figure 9 Content before user browsing

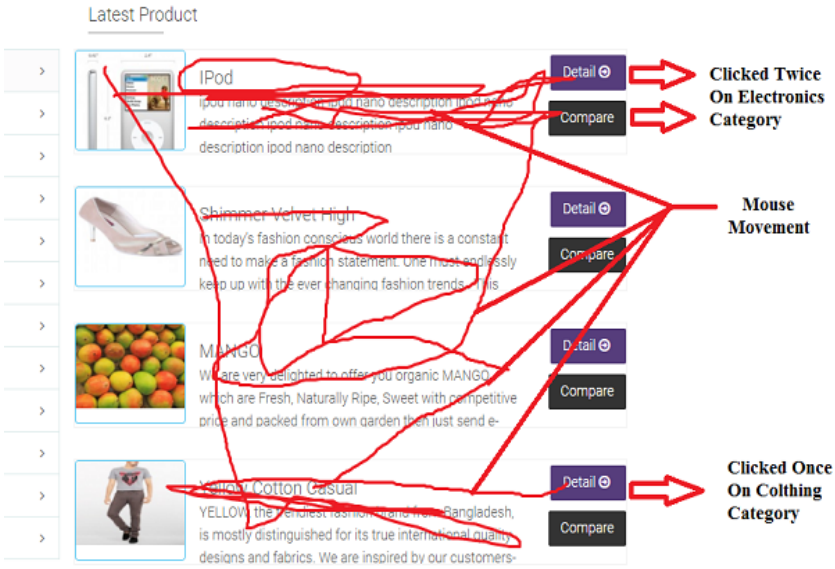


Figure 10 Visulaization of user interaction with the system

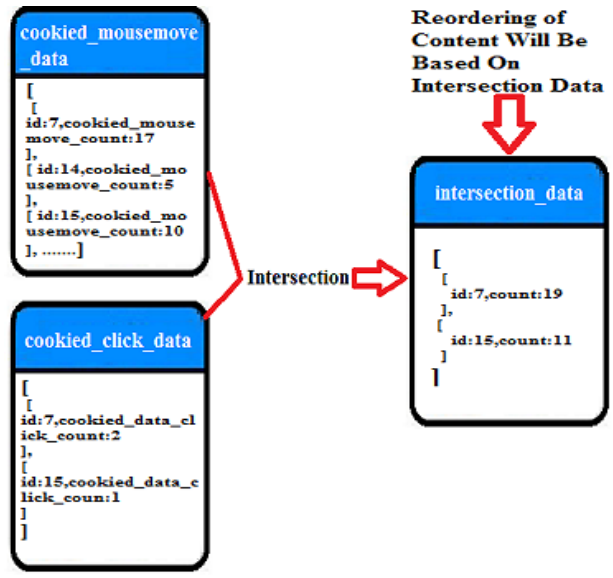


Figure 11 Data storage and user interest finding process

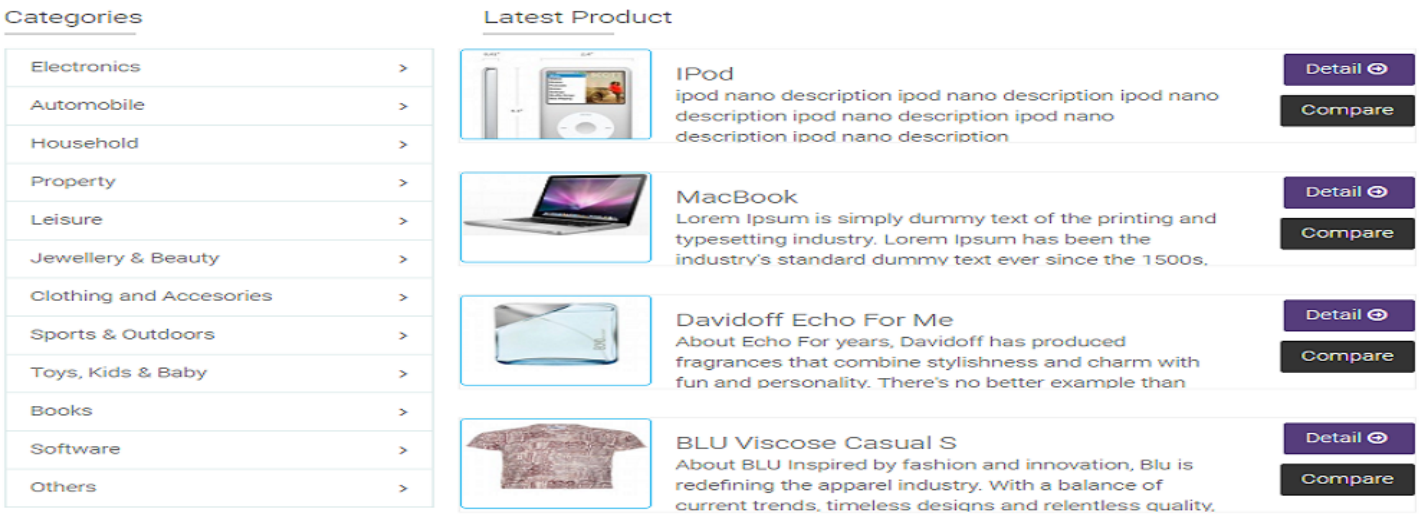


Figure 12 Content after user browsing

At the completion of his browsing we asked him for his experience. He told that he was overwhelmed to see contents being changed almost immediately after his each activity. He gave us rating of 4.75 out of 5.00.

8. Comparative Analysis

The paper [8] introduced mouse movements as an additional layer of information that interprets user interest. In our previous work [1] we implemented the idea. We gave a new approach and algorithms which gives the accuracy of 87.50 %.

Our proposed system [1] solved the problem of e-commerce website by storing registered user interest in database and providing recommendation based on the interest data but can face problem in large scale application if we approach with same approach. With huge number of users the system loading time may increase. This will be quite frustrating for the users. The approach discussed in this paper can solve all the existing problems using our newly given two algorithms UICA, CRA and storing data in browser cookie. This will not affect the performance of the application and there is no need to distinguish between registered or unregistered user anymore. New approach gained accuracy of 91.67% compared to earlier accuracy which was 87.50%. So the accuracy has increased 4.17%. Figure 13 shows the accuracy of different systems.

There was 1.2% improvement in the rating system. Our previous system loading time was 2.8749 sec where new system loads in 2.6531 sec leading to 22.14% improvement of performance. Figure 14 shows the improvement of performance using our new approach.

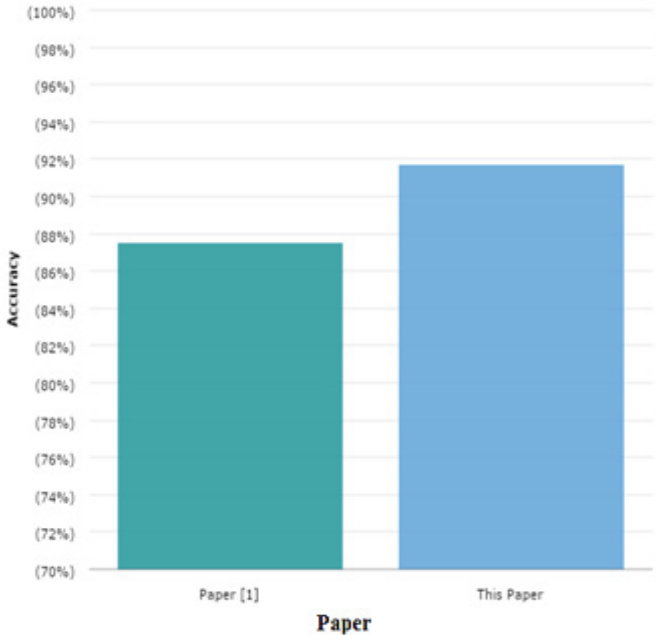


Figure 13 Accuracy comparison graph

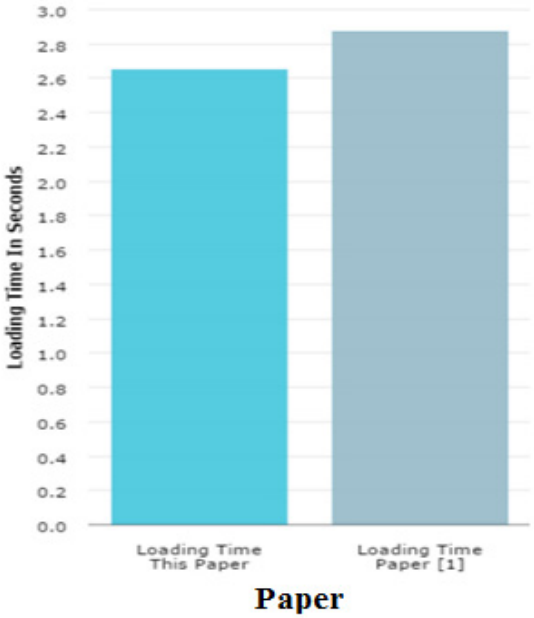


Figure 14 Loading time comparison graph

9. Conclusion

Purpose of this research is to make web applications more user oriented for better interaction and better user experience. Our proposed approach is hassle free as our system will observe user's common actions (mouse movement and click) without bothering them and reorganize its contents according to their needs. This approach can be helpful for both large and small scale web applications. Our system also provided great accuracy in identifying user's interests and reordering its content according to each user interest. We would like to introduce machine learning to raise the accuracy level in future.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] Md. Tanjim-Al-Akib, Lutfullahil Kabir Ashik, and Krishanu Chowdhury, "User-modeling and recommendation based on mouse-tracking for e-commerce websites," Computer and Information Technology (ICCIT), 2016 19th International Conference on. IEEE, 2016. https://doi.org/10.1109/ICCITECHN.2016.7860252.

[2] Polona Selic, Eva Cedilnik Gorup, Savin Gorup, Marija Petek Ster, Janez Rifel and Zalika Klemenc Ketis, "The Effects of a Web Application and Medical Monitoring on the Quality of Medication, Adverse Drug Events and Adherence in the Elderly Living at Home: a Protocol of the Study" Materia socio-medica 28, no. 6 (2016): 432-436. https://doi.org/msm.2016.28.432-436.

[3] Song, Hengjie, Ruoxue Liao, Xiangliang Zhang, Chunyan Miao, and Qiang Yang, "A Mouse-Trajectory Based Model for Predicting Query-URL Relevance" in Twenty-Sixth AAAI Conference on Artificial Intelligence. 2012.https://www.aaai.org/ocs/index.php/AAAI/AAAI12/paper/viewPaper/4968

[4] Olivier Chapelle, and Ya Zhang, "A dynamic bayesian network click model for web search ranking" in Proceedings of 18th international conference on World Wide Web. ACM, 2009. https://doi.org/10.1145/1526709.1526711.

[5] Peter Brusilovsky, "Adaptive navigation support", The adaptive web (2007): 263-290.

[6] Bc Josef Kunhart, "Web content management system: architecture, concepts and application," PhD dissertation, Master Thesis, Czech Technical University, Prague: Department of Computer Science and Engineering, 2013.

[7] Hans-Joachim Bieg, Lewis L. Chuang, Roland W. Fleming, Harald Reiterer, and Heinrich H. Bülthoff, "Eye and pointer coordination in search and selection tasks," in Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications, ACM, 2010. https://doi.org/10.1145/1743666.1743688.

[8] Florian Mueller, and Andrea Lockerd, "Cheese: tracking mouse movement activity on websites, a tool for user modeling," CHI'01 extended abstracts on Human factors in computing systems. ACM, 2001, 279-280. https://doi.org/10.1145/634067.634233.

[9] Jennefer Hart, Alistair Sutcliffe, and Antonella di Angeli, "Evaluating User Engagement Theory," in Proceedings of CHI Conference on Human Factors in Computing Systems, 5th to 10th May 2012. http://openlab.ncl.ac.uk/uxtheory/files/2011/11/5_Hart.pdf.

[10] Vikas Chahar, "An Analytical Study on HCI" International Journal of Computer Science and Management Studies, 2012, 251-254. http://www.ijcsms.com/journals/Volume%2012,%20Issue%2001,%20January%202012_Volume%2012,%20Issue%2001,%20January%202012_Vikas%20Paper.pdf.

[11] Rajiv Badi , Soonil Bae , J. Michael Moore , Konstantinos Meintanis , Anna Zacchi , Haowei Hsieh , Frank Shipman , Catherine C. Marshall, "Recognizing user interest and document value from reading and organizing activities in document triage," in Proceedings of the 11th international conference on Intelligent user interfaces. ACM, 2006. https://doi.org/10.1145/1111449.1111496.

[12] Jonathan Grudin, "Groupware and social dynamics: Eight challenges for developers" Communications of the ACM, 37(1), 1994, 92-105. https://doi.org/10.1145/175222.175230.

[13] Philip K. Chan, "A non-invasive learning approach to building web user profiles" in Proceedings of WEBKDD '99 Revised Papers from the International Workshop on Web Usage Analysis and User Profiling 1999. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.35.2866.

[14] Mark Claypool, Phong Le, Makoto Wased, and David Brown, "Implicit interest indicators" in Proceedings of the 6th international conference on Intelligent user interfaces, ACM, 2001. https://doi.org/10.1145/359784.359836.

[15] R. J. Jacob and Keith S. Karn, The Mind's Eye Cognitive and Applied Aspects of Eye Movement Research Vol. 12, Issue 01, 2003.

[16] Robert J.K. Jacob, Eye tracking in advanced interface design, Virtual environments and advanced interface design, 1995.

[17] Atsuo Murata, Masakazu Higashijima, and Takehito Hayami, "Relationship between trajectory of mouse pointer and eye movements characteristics during search activity," in Proceedings of SICE Annual Conference (SICE), IEEE, 2012.

[18] Daniel J. Liebling, and Susan T. Dumais, "Gaze and mouse coordination in everyday work" in Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication, 2014. https://doi.org/10.1145/2638728.2641692.

[19] Kerry Rodden, and Xin Fu, "Exploring how mouse movements relate to eye movements on web search results pages" in SIGIR Workshop on Web Information Seeking and Interaction, 29-32, 2007. www.rodden.org/kerry/wisi2007.pdf.

[20] Kerry Rodden, Xin Fu, Anne Aula, and Ian Spiro, "Eye-mouse coordination patterns on web search results pages," in CHI'08 extended abstracts on Human factors in computing systems, 2997-3002. ACM, 2008. https://doi.org/10.1145/1358628.1358797.

[21] Fernando Diaz, Ryen White, Georg Buscher, and Dan Liebling, "Robust models of mouse movement on dynamic web search results pages," in Proceedings of the 22nd ACM international conference on Conference on information & knowledge management, ACM, 2013. https://doi.org/10.1145/2505515.2505717.

[22] Hideo Joho, Robert Villa, and J. Jose, "Interaction pool: Towards a user-centered test collection," in SIGIR'07 Web Information-Seeking and Interaction Workshop, 2007. https://pdfs.semanticscholar.org/44ed/e9d34dda8676f118ee76e34d47e66cd2bd01.pdf.

[23] Jeff Huang, Ryen W. White, Georg Buscher, and Kuansan Wang, "Improving searcher models using mouse cursor activity," in Proceedings of SIGIR '12 Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, 2012. https://doi.org/10.1145/2348283.2348313

[24] Luis A. Leiva Torres, and Roberto Vivo Hernando, "A gesture inference methodology for user evaluation based on mouse activity tracking," in IHCI 2008, Proceedings of the IADIS International Conference on Interfaces and Human Computer Interaction, Amsterdam, 2008. https://smt.speedzinemedia.com/docs/pubs/smt-ihci08.pdf

[25] Nazirah Abd Hamid, Suhailan Safei, Siti Dhalila Mohd Satar, Suriayati Chuprat, and Rabiah Ahmad, "Randomized Mouse Movement for Behavioral Biometric Identification," International Journal of Interactive Digital Media, 1(2), 52-57, 2013. https://ijidm.org/wp-content/uploads/2013-1-2-10-nazirah.pdf

[26] Mark D. Smucker, Xiaoyu Sunny Guo, and Andrew Toulis, "Mouse movement during relevance judging: implications for determining user attention," in Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval, ACM, 2014. https://doi.org/10.1145/2600428.2609489

[27] Anuradha Gupta, and Vartika Srivastava, "PREDICTING USER INTEREST FROM MOUSE DATA". http://cerc.iiitd.ac.in/spsymp15/papers/18.pdf.

[28] Alfred Kobsa, "Generic user modeling systems" User modeling and user-adapted interaction, 11(1), 49-63, 2001. www.umuai.org/anniversary/2001-UMUAI-kobsa.pdf.

[29] Cristina Gena, "Evaluation methodologies and user involvement in user modeling and adaptive systems" User Modeling and User-Adapted Interaction, 74-78, 2003.

[30] Cristina Gena, and Stephan Weibelzahl, Usability engineering for the adaptive web, The adaptive web: methods and strategies of web personalization, 2007.

[31] Tzafilkou Katerina, Protogeros Nicolaos, and Yakinthos Charalampos, "Mouse tracking for web marketing: enhancing user experience in web application software by measuring self-efficacy and hesitation levels" International Journal of Strategic Innovative Marketing, 1(4), 233-247, 2014. http://dx.doi.org/10.15556/IJSIM.01.04.005.

[32] Theodore Jack London Shrader, Garry L. Child, and William H. Gengler, "Dynamic use and validation of HTTP cookies for authentication" U.S. Patent 6,374,359, issued April 16, 2002.

[33] Eric A Benson, "Use of browser cookies to store structured data," U.S. Patent No. 6,714,926. 30 Mar. 2004.

[34] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl, "GroupLens: an open architecture for collaborative filtering of netnews" in Proceedings of the 1994 ACM conference on Computer supported cooperative work, ACM, 1994. https://doi.org/10.1145/192844.192905.

[35] Magnus Mortensen, "Design and evaluation of a recommender system" Master's thesis, Universitetet i Tromsø, 2007.

[36] Resnick, Paul, and Hal R. Varian, "Recommender systems" Communications of the ACM, 40(3), 56-58, 1997.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1323-1329 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Multi-attribute Reverse Auction Design Based on Fuzzy Data Envelopment Analysis Approach

Deyan Chen*, Dingwei Wang

Institute of Systems Engineering, College of Information Science and Engineering, Northeastern University, 110000, Shenyang, China

ARTICLE INFO

Article history:
Received: 30 April, 2017
Accepted : 29 June, 2017
Online: 01 August, 2017

Keywords:
Multi-attribute reverse auction
Multiple winner
Fuzzy data envelopment analysis
Winner determination
Bidding strategy

ABSTRACT

Multi-attribute reverse auction is widely used for the procurements of enterprises or governments. To overcome the difficulty of identifying bidding attribute weight and score function of the buyer, the multi-round auction and bidding models with multiple winners are established based on fuzzy data envelopment analysis. The winner determination model of the buyer considers the integrated input-output efficiency of k winners. The bidding strategy of seller is divided into two parts: the first one estimates the weight of the ideal supplier that is thought to be the buyer's preference; the second one is to calculate the weight of the test supplier which reflects the change trend of current weights and the seller's weakness. The final predicted weight is the weighted sum of both. On the basis of known weight, the test supplier can improve his efficiency to increase the winning chance in the next round auction. Our models comprise crisp numbers and fuzzy numbers. Finally, a numerical example verifies the validity of the proposed models.

1. Introduction

This paper is an extension of work originally presented in 2016 31st Youth Academic Annual Conference of Chinese Association of Automation [1].

With the rapid development of network technology and electronic commerce, multi attribute reverse auction has been widely applied to the enterprise or government procurement of goods or services due to saving cost, increasing income and conducting comprehensive assessment of multiple indexes. Especially, online procurement auction plays an important role in E-commerce [2].

In the multi-attribute reverse auction, winner determination problem of the auctioneer and bidding strategies of bidders are two key problems. Many scholars have done a lot of work. At present, there are two main research methods: one is from the perspective of game theory. An optimal auction model was designed with two attributes of price and quality and the best bidding strategy for a bidder was provided where a linear score function was used to evaluate the bids in [3]. This model was extended to an extensive application from different aspects in [4, 5]. A decision support model was used to solve the winner determination problem of the multi-attribute reverse auction in [6]. Another is from the perspective of decision theory. The multi-attribute auction problem based on the technique of order preference was studied in [7]. A fuzzy multi-attribute matching and negotiation mechanism of single round reverse auction with fuzzy numbers was studied in [8]. A PT–BOCR decision method of risk-averse bidders was developed where prospect theory was incorporated into the "benefits, opportunities, costs and risks" framework in [9]. Mathematical programming model was applied for the selection of suppliers in [10]. In [11], the winner determination problem was solved through an evolutionary algorithm of goal programming model.

In all these studies, the score function of supplier evaluation and the weight of bidding attribute need to be predefined ahead of time. But in the real auction process, it is hard to achieve because of complex environment. Furthermore, accurate description of uncertain or vague attributes is another critical issue in practice, especially in a multi-round multi-attribute reverse auction.

Since data envelopment analysis (DEA) was firstly proposed by Charnes in [12], it has been broadly used in various fields. DEA is a nonparametric linear programming method for evaluating the input-output efficiency of decision making unit (DMU) [13]. Thus,

*Corresponding Author: Deyan Chen, Institute of Systems Engineering, College of Information Science and Engineering, Northeastern University, 110000, Shenyang, China | Email: chendy0413@163.com

to avoid the difficult of the determination of score function and weight of bidding attribute, DEA is proved to be an advantageous tool in multi-attribute reverse auction. Bogetoft and Nielsen developed a DEA based auction mechanism with asymmetric information and correlated costs [14]. DEA was combined with an integer programming model to elicit preference information of the buyer in [15]. Data envelopment analysis was applied to the selection problem of multi-attribute reverse auction from the perspective of the suppliers in [16]. But all these DEA models are solved under the assumption of the crisp input–output data [17]. However, the input–output data in most auction models are often vague or imprecise in a real world when the auction relates to multiple different attributes. Thus fuzzy DEA provide a powerful tool to solve the vague or imprecise information. Meanwhile, a great of works has already been done in many fields [18–21].

In this paper, by the fuzzy DEA and possibility theory, we develop a winner determination decision model with multiple winners, in which sealed-bid multi-attribute reverse auction can be conducted regularly, that means a multi-attribute multi-round simultaneous auction. From the perspective of unselected supplier, we consider weight information of ideal supplier and winning supplier simultaneously, and predict the possible weight in the next round auction. Under the premise of increasing input-output efficiency, the test supplier will stand a chance to win bids in the next round auction by improving the value of bidding attribute. The developed methods comprise two new characteristics. To begin, Fuzzy numbers are allowed for existence in the multi-attribute reverse auction model. The score function of buyer is not essential to auctioneer and bidders. Moreover, strategies of buyer and sellers are given based on the same principle. The decision of bidding attribute weight is no longer a complicate problem. All the models can be transformed into linear programming problems and it is easy to compute for the practical application.

The rest of this paper is organized as follows: The multi-attribute multi-round reverse auction design is depicted in section 2. In section 3, the winner determination model for the buyer with multiple winners is suggested in view of fuzzy DEA. The bidding model of unselected seller is also presented for the same scene. A numerical example is given to demonstrate the proposed models in section 4. In section 5, the conclusions are presented.

2. Auction Mechanism Design

In this section, the proposed multi-attribute reverse auction mechanism is described and the related application background and conditions are presented.

There exists a buyer and multiple suppliers in the proposed multi-round sealed-bid reverse auction. The single buyer needs to procure the same k indivisible commodities provided by the different k suppliers at regular intervals. All the suppliers want to make a bid to become the winners in each procurement auction. It is similar to a multi-unit multi-round multi-attribute simultaneous auction. When k=1, it is a multi-attribute single–item reverse auction for an indivisible commodity.

In every auction cycle, the buyer first sends the request for quotes which provides the basic demand and requirement for the bidders. For example, the buyer has a minimum or maximum threshold value for bidding attributes. At the same time, the last round winners' quotes information is shared by all the bidders. All the bidders know that the buyer evaluates suppliers mainly through the input-output performance. However, they don't know the true attribute weight. Second, the qualified suppliers, who fulfill the demand and requirement of buyer, submit bids on multiple attributes of procured commodities. And some of attribute value are fuzzy numbers which describe the vague and imprecise information of procured commodity. Here the previous winners' bidding information is a common reference for all the suppliers. Each supplier only has one chance to bid and the submission cannot be revised. Finally, the buyer decides the final k winners according to their integrated input-output efficiencies and declares the corresponding information of the k winning suppliers.

This auction mechanism is common among enterprise and government procurement. Therefore, the decision of strategies of buyer and sellers are significant. In the next section, winner decision making model of buyer and the bidding model of sellers are established respectively.

3. Winner Decision Models of the Buyer

In this section, basic knowledge of fuzzy number and fuzzy DEA is introduced. Then the winner determination model of buyer with multiple winners is proposed based on fuzzy DEA model. At last, the bidding model of sellers is constructed, where the preference weight of buyer and possible weight reflecting the weakness of the seller are reckoned simultaneously.

3.1. Fuzzy Arithmetic and Fuzzy DEA

A fuzzy number is a fuzzy set characterized by a given real number interval, whose value of membership is between 0 and 1. It is often used to describe a vague, imprecise variable or a linguistic variable in a more direct form. The triangular fuzzy number is most commonly used in practice. It is defined by the following membership function $\tilde{\mu}_X(x)$.

$$\tilde{\mu}_X(x)=\begin{cases}(x-x^1)/(x^2-x^1), & x^1<x\le x^2,\\ (x^3-x)/(x^3-x^2), & x^2<x\le x^3,\\ 0, & otherwise,\end{cases} \tag{1}$$

where x, x^1, x^2, x^3 are non-negative and $x^1 \le x^2 \le x^3$ holds. A crisp real number is regarded as a special case of triangular fuzzy number. We usually denote the triangular fuzzy number $\tilde{X}$ as $\tilde{X}=(x^1,x^2,x^3)$.

We denote two triangular fuzzy numbers as $\tilde{X}=(x^1,x^2,x^3)$ and $\tilde{Y}=(y^1,y^2,y^3)$ respectively. Then the fuzzy arithmetic operations on $\tilde{X}$ and $\tilde{Y}$ are defined as follows [19]:

Addition: $\tilde{X}+\tilde{Y}=(x^1+y^1,x^2+y^2,x^3+y^3)$;

Subtraction: $\tilde{X}-\tilde{Y}=(x^1-y^3,x^2-y^2,x^3-y^1)$;

Multiplication: $\tilde{X} \times \tilde{Y} \approx (x^1 y^1, x^2 y^2, x^3 y^3)$;

Division: $\tilde{X} / \tilde{Y} = (\frac{x^1}{y^3}, \frac{x^2}{y^2}, \frac{x^3}{y^1})$.

Fuzzy DEA model has the following primal linear programming statement for the CCR model [18].

$$Max \quad \sum_{r=1}^{s} u_r \tilde{y}_{r0}$$

$$s.t. \quad \sum_{r=1}^{s} u_r \tilde{y}_{rj} - \sum_{i=1}^{m} v_i \tilde{x}_{ij} \le 0, \; j = 1,2,\cdots,n$$

$$\sum_{i=1}^{m} v_i \tilde{x}_{i0} = 1 \tag{2}$$

$$v_i, u_r \ge 0, i = 1,2,\cdots,m; r = 1,2,\cdots,s.$$

where the subscript 0 is the decision making unit(DMU) under the evaluation. When input and output data of (2) are be characterized by positive triangular fuzzy numbers $\tilde{x}_{ij} = (x_{ij}^1, x_{ij}^2, x_{ij}^3)$ and $\tilde{y}_{ij} = (y_{rj}^1, y_{rj}^2, y_{rj}^3)$ including crisp input and output data as a special case of triangular numbers, we can obtain the following fuzzy DEA model.

$$Max \quad \tilde{E}_0 \approx (e_0^1, e_0^2, e_0^3) = (\frac{\sum_{r=1}^{s} u_r y_{r0}^1}{\sum_{i=1}^{m} v_i x_{i0}^3}, \frac{\sum_{r=1}^{s} u_r y_{r0}^2}{\sum_{i=1}^{m} v_i x_{i0}^2}, \frac{\sum_{r=1}^{s} u_r y_{r0}^3}{\sum_{i=1}^{m} v_i x_{i0}^1})$$

$$s.t. \quad (\frac{\sum_{r=1}^{s} u_r y_{rj}^1}{\sum_{i=1}^{m} v_i x_{ij}^3}, \frac{\sum_{r=1}^{s} u_r y_{rj}^2}{\sum_{i=1}^{m} v_i x_{ij}^2}, \frac{\sum_{r=1}^{s} u_r y_{rj}^3}{\sum_{i=1}^{m} v_i x_{ij}^1}) \le 1 \tag{3}$$

$$v_i, u_r \ge 0, i = 1,2,\cdots,m; r = 1,2,\cdots,s; j = 1,2,\cdots,n$$

where $\tilde{E}_0$ represents the fuzzy efficiency of DMU0 under evaluation and it is also a triangular fuzzy number. In accordance with the existing method, (3) can be transformed into the following three linear programming models [19].

$$Max \quad e_0^1 = \sum_{r=1}^{s} u_r y_{r0}^1$$

$$s.t. \quad \sum_{i=1}^{m} v_i x_{i0}^3 = 1 \tag{4}$$

$$\sum_{r=1}^{s} u_r y_{rj}^3 - \sum_{i=1}^{m} v_i x_{ij}^1 \le 0, j = 1,2,\cdots,n,$$

$$v_i, u_r \ge 0, i = 1,2,\cdots,m; r = 1,2,\cdots,s.$$

$$Max \quad e_0^2 = \sum_{r=1}^{s} u_r y_{r0}^2$$

$$s.t. \quad \sum_{i=1}^{m} v_i x_{i0}^2 = 1,$$

$$\sum_{r=1}^{s} u_r y_{rj}^3 - \sum_{i=1}^{m} v_i x_{ij}^1 \le 0, j = 1,2,\cdots,n \tag{5}$$

$$v_i, u_r \ge 0, i = 1,2,\cdots,m; r = 1,2,\cdots,s$$

$$Max \quad e_0^3 = \sum_{r=1}^{s} u_r y_{r0}^3$$

$$s.t. \quad \sum_{i=1}^{m} v_i x_{i0}^1 = 1,$$

$$\sum_{r=1}^{s} u_r y_{rj}^3 - \sum_{i=1}^{m} v_i x_{ij}^1 \le 0, j = 1,2,\cdots,n \tag{6}$$

$$v_i, u_r \ge 0, i = 1,2,\cdots,m; r = 1,2,\cdots,s.$$

The fuzzy efficiency of each DMU is derived from the above models. All the DMUs are ranked according to efficiencies of the above objective value.

A variety of methods are proposed to compare fuzzy number, we use the preference degree approach to rank the fuzzy numbers [19]. Let $\tilde{X} = (x^1, x^2, x^3)$ and $\tilde{Y} = (y^1, y^2, y^3)$ be two triangular fuzzy numbers, then the degree of preference of $\tilde{X} > \tilde{Y}$ is defined as follows.

$$r_{\tilde{X}>\tilde{Y}} = \frac{S_{\tilde{X}>\tilde{Y}}}{S_{\tilde{X}>\tilde{Y}} + S_{\tilde{Y}>\tilde{X}}}$$

$$= \begin{cases} 1, & x^1 \ge y^3 \\ 0, & x^3 \le y^1 \\ \dfrac{(x^3 - y^1)^2}{(x^3 - y^1 + y^2 - x^2)(x^3 - x^1 + y^3 - y^1)}, & if \; x^2 \le y^2, x^3 > y^1 \\ 1 - \dfrac{(y^3 - x^1)^2}{(y^3 - x^1 + x^2 - y^2)(x^3 - x^1 + y^3 - y^1)}, & if \; x^2 > y^2, x^1 < y^3 \end{cases}$$

where $S_{\tilde{X}>\tilde{Y}}$ represents the area of $\tilde{X} > \tilde{Y}$ in the imagine of membership function of $\mu_{\tilde{X}-\tilde{Y}}(x)$ and $S_{\tilde{X}<\tilde{Y}}$ represents the area of $\tilde{X} < \tilde{Y}$ in the imagine of membership function of $\mu_{\tilde{X}-\tilde{Y}}(x)$. From the definition of the degree of preference $r_{\tilde{X}>\tilde{Y}}$, we know that $r_{\tilde{X}>\tilde{Y}} + r_{\tilde{X}<\tilde{Y}} = 1$. In general, if $r_{\tilde{X}>\tilde{Y}} > r_{\tilde{Y}>\tilde{X}}$, we think that $\tilde{X}$ is bigger than $\tilde{Y}$ and if $r_{\tilde{X}>\tilde{Y}} < r_{\tilde{Y}>\tilde{X}}$, we think that $\tilde{X}$ is smaller than $\tilde{Y}$. When $r_{\tilde{X}>\tilde{Y}} = r_{\tilde{Y}>\tilde{X}} = 0.5$, $\tilde{X}$ is indifferent to $\tilde{Y}$.

Based on the degree of preference, we can compare any two fuzzy numbers, and then a complete priority order of fuzzy

efficiencies of suppliers can be obtained in the following multi-attribute reverse auction.

3.2. Auction Models of the buyer with Multiple Winners

In the multi-attribute reverse auction, multiple attributes of procured commodity may be vague or imprecise which cannot be impressed by the crisp real numbers. For example, the design level of a commodity, the reputation of a seller and so forth. It is necessary to consider the evaluation problem of suppliers with fuzzy attribute data. Thus, on the base of the fuzzy DEA theory, we construct a winner determination model of multi-attribute reverse auction with multiple winners from the perspective of buyer.

As we need to determine k winners once in an auction, total input-output efficiency of selected k suppliers is taken for the objective function. After adding some appropriate constraints, (3) can be transformed into the following three linear programming models to determine the final k winners of multi-attribute reverse auction.

$$Max \quad \tilde{E} = (e^1, e^2, e^3) = \left(\frac{\sum_{t=1}^{k}\sum_{r=1}^{s} u_r y_{rj_t}^1}{\sum_{t=1}^{k}\sum_{i=1}^{m} v_i x_{ij_t}^3}, \frac{\sum_{t=1}^{k}\sum_{r=1}^{s} u_r y_{rj_t}^2}{\sum_{t=1}^{k}\sum_{i=1}^{m} v_i x_{ij_t}^2}, \frac{\sum_{t=1}^{k}\sum_{r=1}^{s} u_r y_{rj_t}^3}{\sum_{t=1}^{k}\sum_{i=1}^{m} v_i x_{ij_t}^1}\right)$$

$$s.t. \quad \left(\frac{\sum_{r=1}^{s} u_r y_{rj}^1}{\sum_{i=1}^{m} v_i x_{ij}^3}, \frac{\sum_{r=1}^{s} u_r y_{rj}^2}{\sum_{i=1}^{m} v_i x_{ij}^2}, \frac{\sum_{r=1}^{s} u_r y_{rj}^3}{\sum_{i=1}^{m} v_i x_{ij}^1}\right) \le 1 \qquad (7)$$

$$\left(\frac{\sum_{t=1}^{k}\sum_{r=1}^{s} u_r y_{rj_t}^1}{\sum_{t=1}^{k}\sum_{i=1}^{m} v_i x_{ij_t}^3}, \frac{\sum_{t=1}^{k}\sum_{r=1}^{s} u_r y_{rj_t}^2}{\sum_{t=1}^{k}\sum_{i=1}^{m} v_i x_{ij_t}^2}, \frac{\sum_{t=1}^{k}\sum_{r=1}^{s} u_r y_{rj_t}^3}{\sum_{t=1}^{k}\sum_{i=1}^{m} v_i x_{ij_t}^1}\right) \le 1$$

$$U_1 u_p \le u_i \le U_2 u_p, \quad 0 < U_1 < U_2 \in R^+$$

$$V_1 v_q \le v_r \le V_2 v_q, \quad 0 < V_1 < V_2 \in R^+$$

$$v_i, u_r \ge \varepsilon, i = 1,2,\cdots,m; r = 1,2,\cdots,s; j = 1,2,\cdots,n,$$

where $\tilde{E} = (e^1, e^2, e^3)$ is the optimal fuzzy efficiency of the combination $(j_1, j_2, \cdots, j_k)$ of k suppliers, compared with the previous fuzzy DEA model. Here the input and output are the total sum of k suppliers respectively. U_1, U_2, V_1, V_2 are the proportional coefficients of preference of the buyer corresponding to the inputs and outputs. They indicate the relative preference of the buyer for the different attributes, and $\varepsilon > 0$, which ensures nonzero solution of weight in (7).

Similarly, (7) can also be transformed into the following three linear programming models.

$$Max \quad e^1 = \sum_{r=1}^{s} u_r \left(\sum_{t=1}^{k} y_{rj_t}^1\right)$$

$$s.t. \quad \sum_{i=1}^{m} v_i \left(\sum_{t=1}^{k} x_{ij_t}^3\right) = 1,$$

$$\sum_{r=1}^{s} u_r \left(\sum_{t=1}^{k} y_{rj_t}^3\right) - \sum_{i=1}^{m} v_i \left(\sum_{t=1}^{k} x_{ij_t}^1\right) \le 1 \qquad (8)$$

$$\sum_{r=1}^{s} u_r y_{rj}^3 - \sum_{i=1}^{m} v_i x_{ij}^1 \le 0, \quad j = 1,2,\cdots,n$$

$$U_1 u_p \le u_i \le U_2 u_p, \quad 0 < U_1 < U_2 \in R^+$$

$$V_1 v_q \le v_r \le V_2 v_q, \quad 0 < V_1 < V_2 \in R^+$$

$$v_i, u_r \ge \varepsilon, i = 1,2,\cdots,m; r = 1,2,\cdots,s.$$

$$Max \quad e^2 = \sum_{r=1}^{s} u_r \left(\sum_{t=1}^{k} y_{rj_t}^2\right),$$

$$s.t. \quad \sum_{i=1}^{m} v_i \left(\sum_{t=1}^{k} x_{ij_t}^2\right) = 1$$

$$\sum_{r=1}^{s} u_r \left(\sum_{t=1}^{k} y_{rj_t}^3\right) - \sum_{i=1}^{m} v_i \left(\sum_{t=1}^{k} x_{ij_t}^1\right) \le 1 \qquad (9)$$

$$\sum_{r=1}^{s} u_r y_{rj}^3 - \sum_{i=1}^{m} v_i x_{ij}^1 \le 0, j = 1,2,\cdots,n,$$

$$U_1 u_p \le u_i \le U_2 u_p, \quad 0 < U_1 < U_2 \in R^+$$

$$V_1 v_q \le v_r \le V_2 v_q, \quad 0 < V_1 < V_2 \in R^+$$

$$v_i, u_r \ge \varepsilon, i = 1,2,\cdots,m; r = 1,2,\cdots,s.$$

$$Max \quad e^3 = \sum_{r=1}^{s} u_r \left(\sum_{t=1}^{k} y_{rj_t}^3\right)$$

$$s.t. \quad \sum_{i=1}^{m} v_i \left(\sum_{t=1}^{k} x_{ij_t}^1\right) = 1,$$

$$\sum_{r=1}^{s} u_r \left(\sum_{t=1}^{k} y_{rj_t}^3\right) - \sum_{i=1}^{m} v_i \left(\sum_{t=1}^{k} x_{ij_t}^1\right) \le 1 \qquad (10)$$

$$\sum_{r=1}^{s} u_r y_{rj}^3 - \sum_{i=1}^{m} v_i x_{ij}^1 \le 0, \quad j = 1,2,\cdots,n$$

$$U_1 u_p \le u_i \le U_2 u_p, \quad 0 < U_1 < U_2 \in R^+$$

$$V_1 v_q \le v_r \le V_2 v_q, \quad 0 < V_1 < V_2 \in R^+$$

$$v_i, u_r \ge \varepsilon, i = 1,2,\cdots,m; r = 1,2,\cdots,s.$$

Calculate the above three linear programming models, the optimal fuzzy efficiency of multiple winners can be obtained and the corresponding k winners are determined.

3.3. Bidding Models of Sellers

In the proposed multi-attribute procurement auction, the attribute value of winning suppliers is the common knowledge in each period. A seller knows that the input-output efficiency of suppliers is the principal index that the buyer evaluates a supplier. In this specific context, the only available information is the bidding attribute values of k winners and the seller's own quote. The seller thinks the accurate weight value of bidding attributes of procured commodity is affected by the following two aspects: one is the preference of the buyer which can be characterized by the weight of the ideal supplier; the other is the current impact which can be described by the weight of the unselected supplier reflecting his weakness. The predicted weight in the next round auction can be seen as a weighted sum of above two parts after being introduced a weighting factor. In view of the above two points, (11) and (12) are established to compute the preference weight of buyer and the current possible weights respectively.

$$Max \quad \tilde{E}_* = (e_*^1, e_*^2, e_*^3) = \left(\frac{\sum_{r=1}^{s} u_r y_{r*}^1}{\sum_{i=1}^{m} v_i x_{i*}^3}, \frac{\sum_{r=1}^{s} u_r y_{r*}^2}{\sum_{i=1}^{m} v_i x_{i*}^2}, \frac{\sum_{r=1}^{s} u_r y_{r*}^3}{\sum_{i=1}^{m} v_i x_{it*}^1}\right)$$

$$s.t. \quad \left(\frac{\sum_{t=1}^{k}\sum_{r=1}^{s} u_r y_{rj_t}^1}{\sum_{t=1}^{k}\sum_{i=1}^{m} v_i x_{ij_t}^3}, \frac{\sum_{t=1}^{k}\sum_{r=1}^{s} u_r y_{rj_t}^2}{\sum_{t=1}^{k}\sum_{i=1}^{m} v_i x_{ij_t}^2}, \frac{\sum_{t=1}^{k}\sum_{r=1}^{s} u_r y_{rj_t}^3}{\sum_{t=1}^{k}\sum_{i=1}^{m} v_i x_{ij_t}^1}\right) \le 1 \qquad (11)$$

$$\left(\frac{\sum_{r=1}^{s} u_r y_{r*}^1}{\sum_{i=1}^{m} v_i x_{i*}^3}, \frac{\sum_{r=1}^{s} u_r y_{r*}^2}{\sum_{i=1}^{m} v_i x_{i*}^2}, \frac{\sum_{r=1}^{s} u_r y_{r*}^3}{\sum_{i=1}^{m} v_i x_{it*}^1}\right) \le 1$$

$$\left(\frac{\sum_{r=1}^{s} u_r y_{rd}^1}{\sum_{i=1}^{m} v_i x_{id}^3}, \frac{\sum_{r=1}^{s} u_r y_{rd}^2}{\sum_{i=1}^{m} v_i x_{id}^2}, \frac{\sum_{r=1}^{s} u_r y_{rd}^3}{\sum_{i=1}^{m} v_i x_{id}^1}\right) \le 1$$

$$U_1^d u_p \le u_i \le U_2^d u_p, \quad 0 < U_1^d < U_2^d \in R^+$$

$$V_1^d v_q \le v_r \le V_2^d v_q, \quad 0 < V_1^d < V_2^d \in R^+$$

$$v_i, u_r \ge \varepsilon, i = 1,2,\cdots,m; r = 1,2,\cdots,s.$$

where supplier * is the ideal supplier, that is the best supplier among all the bidders. It is obtained in the ways listed below. First, let $x_{ij}^* = \min_j \{x_{ij}\}$, then we get the m input data of the ideal supplier; Second, suppose $y_{ij}^* = \max_j \{y_{ij}\}$, the s output data are also decided.

When the bidding attribute values are triangular fuzzy numbers, the degree of preference is used to compare them. The ideal supplier * is an optimal result for the selected suppliers which includes minimum input variables and maximum output variables. Supplier d is the test supplier who is not selected in the auction. Furthermore, $U_1^d, U_2^d, V_1^d, V_2^d$ are the proportional coefficients of preference of the buyer that are predicted by the supplier d, in view of calculation results of previous weights and his own experience.

The first constrain means the integrated efficiency of k winning suppliers is less and equal to1. The second and third constrains make the efficiencies of the ideal supplier and test supplier fall in the interval [0, 1].

$$Min \quad \tilde{E}_d = (e_d^1, e_d^2, e_d^3) = \left(\frac{\sum_{r=1}^{s} u_r y_{rd}^1}{\sum_{i=1}^{m} v_i x_{id}^3}, \frac{\sum_{r=1}^{s} u_r y_{rd}^2}{\sum_{i=1}^{m} v_i x_{id}^2}, \frac{\sum_{r=1}^{s} u_r y_{rd}^3}{\sum_{i=1}^{m} v_i x_{id}^1}\right)$$

$$s.t. \quad \frac{\sum_{t=1}^{k}\sum_{r=1}^{s} u_r y_{rj_t}^3}{\sum_{t=1}^{k}\sum_{i=1}^{m} v_i x_{ij_t}^1} \le 1 \qquad (12)$$

$$\frac{\sum_{t=1}^{k}\sum_{r=1}^{s} u_r y_{rj_t}^1}{\sum_{t=1}^{k}\sum_{i=1}^{m} v_i x_{ij_t}^3} \ge \hat{e}$$

$$\left(\frac{\sum_{r=1}^{s} u_r y_{rd}^1}{\sum_{i=1}^{m} v_i x_{id}^3}, \frac{\sum_{r=1}^{s} u_r y_{rd}^2}{\sum_{i=1}^{m} v_i x_{id}^2}, \frac{\sum_{r=1}^{s} u_r y_{rd}^3}{\sum_{i=1}^{m} v_i x_{id}^1}\right) \le 1$$

$$U_1^d u_p \le u_i \le U_2^d u_p, \quad 0 < U_1^d < U_2^d \in R^+$$

$$V_1^d v_q \le v_r \le V_2^d v_q, \quad 0 < V_1^d < V_2^d \in R^+$$

$$v_i, u_r \ge \varepsilon, i = 1,2,\cdots,m; r = 1,2,\cdots,s.$$

where $\hat{e}$ is given by the test supplier. The first two inequalities ensure the efficiency of k winners falls in the interval [$\hat{e}$, 1], where it is considered to be optimal objective value that makes k suppliers become the final winners. In (11), the final weight solution vector is written as $\vec{S}^* = (u_1^*, u_2^*, \cdots, u_s^*, v_1^*, v_2^*, \cdots, v_m^*)$. Solve the linear programming model (12), we can obtain another weight vector $\vec{S}^d = (u_1^d, u_2^d, \cdots, u_s^d, v_1^d, v_2^d, \cdots, v_m^d)$ which reflects the current change trend of weight and the weakness of the test supplier d.

In the auction process, inexact information and random factors are allowed. When including fuzzy data, (11) and (12) can be transforms into three linear programming models to compute the final fuzzy efficiency, three weight vectors will be obtained corresponding to the $\tilde{E}_* = (e_*^1, e_*^2, e_*^3)$ or $\tilde{E}_d = (e_d^1, e_d^2, e_d^3)$ respectively. Here we take $\vec{S}^*$ corresponding to e_2^* , and $\vec{S}^d$ corresponding to e_2^d since the middle bound value of a fuzzy number means the maximum possibility. Given a weighting factor δ , the predicted weight is written as follows

$$\vec{S} = \delta \vec{S}^* + (1-\delta)\vec{S}^d, \quad \delta \in [0,1] \qquad (13)$$

where a seller can select a appropriate weighting factor δ to describe the change trend of attribute weight in the next round auction.

Once the next period weight of bidding attributes is given, we can calculate the efficiencies of the k winning suppliers in the next auction when they do not change their bidding strategies. Because the efficiency of the supplier is a triangular fuzzy number, our goal is to make the lower bound of fuzzy efficiency of the test supplier is greater than or equal to the middle bound of fuzzy efficiency of one or more winning suppliers through adjusting bidding attribute values properly. Then the test supplier will have the opportunity to become the winner in the next round auction.

4. Numerical Example

In this section, an application of the established models is be demonstrated by the following datasets of bidding attribute value of suppliers. An enterprise need to purchase two new machines for production from two different suppliers regularly. There is a trade-off among price, delivery performance, quality and design in the multiple period auctions. The price of procured machine is expressed in thousand dollars and quality is expressed in percentage of qualified products manufactured by this type of machine. Delivery performance is expressed by days of service and it is a triangular fuzzy number. Design is a linguistic variable and it is expressed by a symmetric triangular fuzzy number. Specific bidding attribute data of five suppliers in the third round auction are shown in Table 1.

Table 1 Bidding attribute value of the suppliers

Supplier (DMUs)	Inputs		Outputs	
	Price ($)	Delivery	Quality (%)	Design
1	48	(32,36,40)	96	(90,95,100)
2	46.5	(36,38,41)	99	(85,90,95)
3	47.5	(34,37,41)	97	(80,85,90)
4	46	(34,36,38)	97	(85,90,95)
5	47	(35,38,41)	98	(90,95,100)

First, the buyer need determine the winners using (8), (9) and (10). Conforming to specification of models, we assumed that the weight of quality is somewhere between two and three times than the weight of design. This constrain condition reflects the inexact preference of the buyer for quality and design. It is easier to describe this kind of imprecise information than to give an exact result in practice. At the same time, the buyer thinks $\varepsilon \geq 0.001$.

Because only two suppliers of five will become the final winners, efficiencies of 10 groups of suppliers need to be compared, and it is shown in Table 2. We use previously introduced degree of preference to rank them, supplier 2 and supplier 5 will be the final winners.

Table 2 Fuzzy efficiencies of suppliers with two winners

Suppliers	Efficiency	Suppliers	Efficiency
{1,2}	(0.9465,0.9657,0.9938)	{2,4}	(0.9649,0.9825,0.9978)
{1,3}	(0.9238,0.9440,0.9835)	{2,5}	**(0.9615,0.9809, 1)**
{1,4}	(0.9465,0.9639, 1)	{3,4}	(0.9414,0.9725,0.9835)
{1,5}	(0.9434,0.9644, 1)	{3,5}	(0.9384,0.9585,0.9818)
{2,3}	(0.9415,0.9607,0.9793)	{4,5}	(0.9616,0.9802, 1)

Then we use (11) and (12) to analysis the data of the test supplier and the winning suppliers for three auction cycles in Table 3 and Table 4. From the perspective of an unselected supplier, we need to determine his bidding strategy in the next phase.

Table 3 Bidding attribute value of the winning supplier 1

Period	Attributes of winner 1			
	Price	Delivery	Quality	Design
1	46	(34,36,37)	98	(85,90,95)
2	47	(33,35,38)	97	(75,80,85)
3	**46.5**	**(36,38,41)**	**99**	**(85,90,95)**

Table 4 Bidding attribute value of the winning supplier 2

Period	Attributes of winner 2			
	Price	Delivery	Quality	Design
1	47	(36,39,41)	98	(85,90,95)
2	46.5	(36,38,41)	99	(85,90,95)
3	**47**	**(35,38,42)**	**98**	**(90,95,100)**

Supplier 4 isn't selected in the third round auction, thus supplier 4 is taken as the test supplier. In Table 1, we know that his bidding attribute value is 46, (34, 36, 38), 97 and (85, 90, 95) in the third round auction. The ideal supplier's bidding attribute value is 46, (34, 36, 37), 99 and (90, 95, 100) respectively (as shown in Table 3). The comparisons of fuzzy numbers are based on the degree of preference when calculating the bidding attribute value of the ideal supplier. Furthermore, in view of supplier 4's experience, the weight of quality is somewhere between two and four times than the weight of design.

Computing the linear programming models (11) and (12), we obtain the efficiency of the ideal supplier is $\tilde{E}_* = (0.9769, 0.9879, 1)$ and the efficiency of test supplier is $\tilde{E}_d = (0.9342, 0.9663, 0.9898)$ which also indicate how ineffective the test supplier 4 is. We take the solution vectors of the middle value of above two triangular fuzzy numbers as the preference weight of the buyer and the current possible weight. They are weight vectors $\vec{S}^* = (0.0080, 0.0020, 0.021, 0.0010)$ and $\vec{S}^d = (0.0081, 0.0020, 0.0125, 0.0118)$. Let $\delta = 0.3$, we obtain the predicted weight is $\vec{S} = (0.0081, 0.0020, 0.0150, 0.0086)$. It is noted that δ is decided by the test supplier which reflects his understanding to the future change trend of weights. The derived weight $\vec{S}$ is regarded as attribute weight for the next round auction. Because the winning suppliers don't change the bidding strategies, upon the new attribute weight, the middle bound value of supplier 2's efficiency is 0.9586 and the middle bound value of supplier 5's efficiency is 0.9535. At present, supplier 4 need improve his bidding attribute value to make the lower bound value of his efficiency is greater or equal to 0.9535 at least. For example, suppose quality and design is hard to revise, we might cut down the bidding value of input attributes. Let price is 45 and the others are unchanged, the lower bound value of efficiency of supplier 4 raises to 0.9590 that is greater than the middle bound value of supplier 5's efficiency. Moreover, let delivery performance is (34, 36, 36), and the lower bound value of supplier 4's efficiency is 0.9561 which also meets the requirement. Here the test supplier

has great flexibility to change bid attribute value to be competitive in the fourth round auction.

5. Conclusions

Multi-attribute reverse auction is widely used in the procurement of enterprise and government. In this paper, a winner determination model of the buyer and a bidding model of the sellers are proposed with multiple winners in the multi-attribute multi-round reverse simultaneous auction. The bidding attribute value can be a real number or a triangular fuzzy number which describes the vague or inaccurate information. Basing on the fuzzy DEA theory, we construct a winner determination model with multiple winners for the buyer to compare the overall efficiency of k winners. The k suppliers with the highest efficiency are the final winners. From the unselected seller's view, the weight of each attribute in the next round of auction is a crucial problem. We first compute the weight of the ideal supplier reflecting the buyer's preference. Second, the weight of the test supplier is calculated to indicate the current change trend and the weakness of the failed supplier. In the end, a weighted value of above two parts is regarded as the next round weight. As the k winning suppliers do not change the original bidding strategies, a test supplier might compute their efficiencies and adjust his own bidding attribute value to make his efficiency greater than the known winners so as to have an opportunity to become the final winner in the next round auction. Overall, the proposed models can handle the vague and imprecise data in view of the actual circumstance of the auction and synthesis the previous information. At last, a numerical illustration verifies the feasibility of the proposed method.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This study was supported by the National Natural Science Foundation of China under Grant No. 70931001 and 61273203.

References

[1] D. Chen, D. Wang, "Solving a multi-attribute reverse auction problem by fuzzy data envelopment analysis method" in 2016 31st Youth Academic Annual Conference of Chinese Association of Automation (YAC), China, 2016. http:// doi.org/10.1109/YAC.2016.7804866

[2] L. Pham, J. Teich, H. Wallenius, J. Wallenius, "Multi-attribute online reverse auctions: Recent research trends" Eur. J. Oper. Res., 242, 1–9, 2015. http://dx.doi.org/10.1016/j.ejor.2014.08.043

[3] Y. K. Che, "Design competition through multidimensional auctions" Rand J. Econ., 24, 668–680, 1993. http:// doi.org/102307/2555752.Source:RePEc

[4] F. Branco, "The design of multidimensional auctions" Rand J. Econ., 28, 63–81, 1997. http:// doi.org/102307/2555940.Source:RePEc

[5] E. David, R. A. Schwartz, and S. Kraus, "Bidding in sealed-bid and English multi-attribute auctions" Decis. Support Syst., 42, 527–556, 2006. http:// doi.org/10.1016/j.dss.2005.02.007

[6] M. Marzouk, O. Moselhi, "A decision support tool for construction bidding" Constr. Innov., 3, 111–124, 2003. http:// doi.org/10.1108/ 14714170310814882

[7] C. B. Cheng, "Slove a sealed-bid reverse auction problem by multiple-criterion decision-making methods" Comput. Math. Appl., 56, 3261–3274, 2008. http:// doi.org/10.1016/j.camwa.2008.09.011.Source:DBLP

[8] K. Tsai and F. Chou, "Developing a fuzzy multi-attribute matching and negotiation mechanism for sealed-bid online reverse auctions" J. Theor. Appl. Electron. Commer. Res., 6, 685–963, 2011. http:// doi.org/ 10.4067/s0718-18762011000300007.Source:DBLP

[9] M. Huang, X. H. Qian, S. C. Fang and X. W. Wang, "Winner determination for risk aversion buyers in multi-attribute reverse auction" Omega, 59, 184–200, 2016. http:// doi.org/10.1016/j.omega.2015.06.007

[10] J. E. Teich, H. Wallenius, J. Wallenius, and A. Zaitsev, "A multi-attribute e-auction mechanism for procurement: theoretical foundations" Eur. J. Oper. Res., 175, 90–100, 2006. http:// doi.org/10.1016/j.ejor.2005.04.023

[11] F. T. S. Chan, M. Shukla, M. K. Tiwari, R. Shankar, K. L. Choy, "B2B multi-attribute e-procurement: A artificial immune system based goal programming approach" Int. J. of Prod. Res., 49(2), 321–341, 2011. http:// doi.org/10.1080/00207540902922802

[12] A. Charnes, W. W. Cooper, E. Rhodes, "Measure the efficiency of decision making unit" Eur. J. Oper. Res., 2(6), 429–444, 1978. http:// doi.org/10.1016/0377-2217(78)90138-8

[13] M. Azadi, M. Jafarian, R. F. Saen, S. M. Mirhedayation, "A new fuzzy DEA model for evaluation of efficiency and effectiveness of suppliers in sustainable supply chain management context" Comp. Oper. Res., 54, 274–285, 2015. http:// doi.org/10.1016/j.cor.2014.03.002

[14] P. Bogetoft and K. Nielsen, "DEA based auctions," Eur. J. Oper. Res., 184, 685–700, 2008. http:// doi.org/10.1016/j.ejor.2006.11.010.Source:RePEc

[15] S. Talluri, R. Narasimhan, S. Viswanathan, "Information technologies for procurement decisions: A decision support system for multi-attribute e-reverse auctions" Int. J. of Prod. Res, 45(11), 2615–2628, 2007. http:// doi.org/10.1080/00207540601020585

[16] R. Narasimhan, S. Talluri, and S. Mahapatra, "Effective response to RFQs and supplier development: A supplier's perspective," Int. J. Prod. Econ., 115, 461–470, 2008. https://doi.org/:10.1016/j.ijpe.2008.07.001

[17] K. D. Kaveh, T. Madjid, H. S. Elham, "A data envelopment analysis model with interval data and undesirable output for combined cycle power plant performance assessment" Expert Syst. Appl., 42, 760—777, 2015. http:// doi.org/10.1016/j.eswa.2014.08.028

[18] S. Saati, A.Memariani, G. R. Jahanshahloo, "Efficiency analysis and ranking of DMUs with fuzzy data" Fuzzy Optim. Decis. Ma., 1, 255–267, 2002. http:// doi.org/10.1023/A:1019648512614

[19] Y. Wang, Y. Luo, L. Liang, "Fuzzy data envelopment analysis based upon fuzzy arithmetic with an application to performance assessment of manufacturing enterprises" Expert Syst. Appl., 36, 5205–5211, 2009. https://doi.org/:10.1016/j.eswa.2008.06.102

[20] F. Cebi, I. Otay, "A two-stage fuzzy approach for supplier evaluation and order allocation problem with quantity discounts and lead time" Inform. Sciences, 339, 143–157, 2016. http:// doi.org/10.1016/j.ins.2015.12.032

[21] K. D. Kaveh, T. Madjid, J. S. Francisco, "A comprehensive fuzzy DEA model for emerging market assessment and selection decision" Appl. Soft Comput., 38, 676–702, 2016. http:// doi.org/10.1016/j.asoc.2015.09.048

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1330-1341 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

A Category Based Threat Evaluation Model Using Platform Kinematics Data

Mustafa Çöçelli*, Ethem Arkın

Aselsan A.Ş., Defense Systems Technologies, 06172, TURKEY

ARTICLE INFO

Article history:
Received: 01 May, 2017
Accepted : 24 June, 2017
Online: 01 August, 2017

Keywords:
Threat Evaluation
Decision-Making
Situational Awareness
Information Fusion
Command and Control (C2)

ABSTRACT

Command and control (C2) systems direct operators to make accurate decisions in the stressful atmosphere of the battlefield at the earliest. There are powerful tools that fuse various instant piece of information and brings summary of those in front of operators. Threat evaluation is one of the important fusion method that provides these assistance to military people. However, C2 systems could be deprived of valuable data source due to the absence of capable equipment. This situation has a bad unfavorable influence on the quality of tactical picture in front of C2 operators. In this paper, we study on the threat evaluation model that take into account these deficiencies. Our method extracts threat level of various targets mostly from their kinematics in two dimensional space. In the meantime, classification of entities around battlefield is unavailable. Only, category of targets are determined as a result of sensors process, which is the information of whether entities belong to air or surface environment. Hereby, threat evaluation model is consist of three fundamental steps that runs on entities belongs to different environment separately: the extraction of threat assessment cues, threat selection based on Bayesian Inference and the calculation of threat assessment rating. We have evaluated performance of proposed model by simulating a set of synthetic scenarios.

1. Introduction

This paper is an extension of the study that originally presented in 2016 IEEE Symposium Series on Computational Intelligence (SSCI) [1]. The original work focuses on a threat evaluation model that prioritizes entities without any characteristics at all in tactical picture. It accepts all targets without any differentiation and applies same threat evaluation model even if they have great difference because their dynamics. In this work, we separate targets from each other according to their category and apply appropriate methods to each categories independently.

Military operations take place in a stressful environment with a large amount of changing data. This atmosphere could drag C2 operators to erroneous decisions easily and leave them with their biases alone. Therefore, the aid of computer power is indispensable along OODA (Observe-Orient-Decide-Act) in battlefield [2]. The level of assistance changes according to capabilities of equipment deployed on the naval platform. Decision support systems serve operators the compiled tactical picture the combat field by processing raw sensor detections. Operators try to make their best while evaluating identity and kinematics of entities together before reaching final decisions. However, it is very difficult to establish relation between various data and extract value from instant data at a limited time without any automatic assistance tool [3]. Due to uncertain risks of battlefield, operators need to recognize, identify and prioritize surrounding objects continuously without any interruption [4]. The decision cycle of C2 operators are modeled and simulated in lots of studies in order to build up decision support systems [5, 6]. It shows that these systems are one of the core elements in C2 operations.

Situational awareness phase completes the observation step of the C2 processes by working on sensor data. This phase generates necessary tactical information used by orientation step of the C2 processes [7]. At this point, threat evaluation is defined as the analysis of surrounding objects continuously in order to detect their intents against defended assets and rank these objects according to their threat level. After the reveal of intents of these suspected objects, action sequence is determined to neutralize danger on friendly forces [8]. In the meanwhile, threat evaluation methods gain importance to accelerate decision cycle of C2

*Corresponding Author: Mustafa Çöçelli, Aselsan A.Ş., Defense Systems Technologies, 06172, TURKEY | Email: mcocel@aselsan.com.tr

operators by decreasing time needed between operator perception and reaction time in the tactical area.

Various methods, such as, rule based techniques, fuzzy logic based techniques, neural networks and graphical models, are used in the threat evaluation algorithms to measure the threat level of assets surrounding defended forces [9]. Threat assessment cues are fundamental input data of these threat evaluation methods. As a result of various long-term studies that benefiting from experiences of qualified military personal, the most effective threat assessment cues are determined. Hereby, the environment of the target has a major influence on threat assessment cues. Kinematics of entities changes depending on where they operates. While air objects are able to make fast and sharp moves, surface objects on sea does not have the rapidity belonging to air objects. Similarly, different characteristics of target arising from environment lead to different threat assessment cues. Therefore, the list of threat assessment cues differs according to target's environmental category. The complete list of cues used for air forces is airlane, altitude, coordinated activity, course heading, closest point of approach (CPA), ESM/Radar Electronic Support, feet wet/dry, IFF mode, maneuvers, origin/location, own support, range/distance, speed, visibility, weapon envelope, wings clean/dirty in the study of Liebhaber and Feher [10]. Same researchers defines the list of threat assessment cues for surface forces as follows: cargo, coordinated activity, destination, ESM, heading, number of vessels, own support, range/distance, recent history, regional intelligence, sea lane, speed, voice communication with track and weapon envelope [11]. While important portion of lists are common, there are some significant difference between lists.

Johansson's study approaches threat assessment cues with a different perspective and categorize them as capability, proximity and intent parameters [9]. Mostly, classification of entities forms the list of capability parameters. Those parameters measures the possible deadliness of suspected object against friendly object. Meanwhile, the concern of proximity parameters is the closeness of objects in regards of position and orientation with respect to each other. Finally, intent parameters are interested with the actual aim of the suspected object on friendly asset [9].

Threat evaluation cues are originated from the characteristic of the target, the kinematic data of the target itself and the kinematic relationship of the target and the defended asset. These cues are classified in this regard. The kinematic relationship of objects find place in proximity and intent parameters itself. Classifying and identifying suspected object feeds capability parameters among cues. Small-scale naval platform has a limited foresight about the classification of target due to the lack of technologies, such as IFF and ESM, providing more modal information about the suspected target. In this paper, we examine the naval system that supply whether the target is an air track or a surface track. Moreover, the system provides kinematics of surrounding objects in two dimensional space. We present a threat evaluation model that initially generating scores for each threat assessment cues from kinematics of objects, then performing threat selection by using scores generated from initial phase and finally calculating threat assessment rating by using again cue scores. Data fusion is progressed at different paths for each environmental category. While air objects are evaluated in air container, surface objects are interpreted in surface container. Each category follows same threat evaluation model with different threat assessment parameters that convenient to use in that category.

The remaining part of this paper is organized as follows: Firstly, section II gives information about challenges while designing such a fusion based system. After that, section III presents the reason of why this threat evaluation model needed and the definition of the threat evaluation problem in literature. Then, section IV introduces whole steps of the threat evaluation model in detail. Section V presents the measurement of the performance of the mentioned model by evaluating couple of synthetic scenarios. Section VI gives the brief information about other researches for the threat evaluation problem. Lastly, section VII states the summary of this paper. Significant comparison is performed with this work and original work in discussions at the section VII. Further potential work after this study to complete research will be discussed in the end.

2. Challenges

Fusion problem is a challenging issue in itself apart from threat evaluation problem. Real world is measured by sensors and turns into signals. Then, these information is translated to discrete values to make mathematical calculations possible. There are many questions rising about the registration of data before beginning of the fusion process. In low-level information fusion processing, sensor's detections are associated and concluded as estimated tracks and target identities. After this step, high-level information fusion processing is initiated in order to detect repetitive behaviors, association of entities, prediction of future behavior and classification of situation [12].

Threat evaluation problem is generally related with high-level fusion processing since the problems tries to define tactical situation more clearly. Therefore, challenges for HLIF processing is addressed difficulties for threat evaluation problem, also. From the Fusion08 Panel discussion, these challenges are listed as proper fusion process models, estimation-based capabilities for HLIF relations, formal models for HLIF and use of other domains to contribute fusion solutions [13].

Threat evaluation problem can be assessed within situation assessment process widely discussed by fusion community. Explanation of the process, graphical displays to reveal internal process and interactive control to correct resulting situation are listed as common challenges in situation assessment [13]. C2 operator could be aware of the internal process and interact with system by the aid of graphical interfaces. However, the information displayed and the way of interaction is not fully described and resolved in the threat evaluation problem. There is a guide and studies for the design of graphical interfaces [3, 10]. However, it would be assertive to evaluate this design as a remedy to all needs.

There are various challenges that the community faced with while studying on situation and threat assessment. In general, following challenges are highlighted as a result of heavy discussions performed in the fusion community: situation and environment modeling, representation of knowledge, system design approaches, decision support process and performance and effectiveness evaluation methods for implemented systems. These challenges describes the problems while designing systems that needs fusion study in it.

3. Motivation

There is no complete definition and solution for threat evaluation problem since C2 operator behaviors and evaluations

could not be guessed easily during ongoing military operation [14]. Since the fusion of large amount of instant information on combat field is still needed excessively, the problem is still debatable topic and new approaches are introduced to handle situation continuously [15]. The problem is getting more compulsive in the situation of naval platforms deprived of obtaining adequate information to process. The lack of important detection systems, such as 3D radar, ESM, IFF, lead to less reliable estimation results in threat evaluation problem. In this study, we study on similar case and propose an approach to meet the need of C2 systems that have limited capability.

Any content word related with threat evaluation process is named parameter, factor, characteristic or specifically cue [14, 16]. Fundamental input of threat evaluation process is threat assessment cues. These parameters are determined in different researches. US Navy uses Origin, IFF Mode, Intelligence Report, Altitude, Proximity to an Airlane and ESM parameters as most effective threat assessment cues in their air defense system. Moreover, U.S Navy officers within a Combat Information Center rank the importance threat assessment cues for the surface warfare in another Liebhaber's study. While platform, weapon envelope, ESM, origin-flag parameters dominates the top of the list, range, heading, CPA/Speed cues follows these capability parameters [11]. As a result, the identity and the classification of the platform has a major influence on determining threat level of surrounding objects according to experienced users. Unfortunately, naval vessels with limited detection systems has no chance to use those parameters directly in their defense systems if there will be no operator intervention to the system. Therefore, these platforms need effective threat evaluation process that benefits from kinematics of objects. In this study, we focus on the use of parameters related with kinematics. Only available information in our case is whether the entity is an air object or surface object. Since dynamics of air and surface entities are different from each other, they are evaluated separately. Method section presents how this evaluation differentiate two distinct environment from each other.

The formal definition of the threat evaluation problem is described as follows: While $\mathbf{T} = \{T_1, \dots, T_n\}$ represents targets and $\mathbf{A} = \{A_1, \dots, A_m\}$ is the symbolic definition of friendly assets that need protection. V_{ij} is the relation function between target and defended asset. Moreover, V_{ij} represents the threat assessment rating of target-defended asset pair (T_i, A_j), where $T_i \in$ T, $A_j \in$ A. The normalization is applied to V_{ij} and it takes values between 0 and 1 in the end. Predictably, while 0 stands for safe force, 1 defines dangerous one. The functions describing the relation between these assets is defined as follows [17]:

$$f: \mathbf{T} \text{ x } \mathbf{A} \rightarrow [0,1] \quad (1)$$

In threat evaluation function, each friendly assets establish relationship with each suspected assets separately. When the identification of surrounding object is not possible, own platform is left as only friendly asset. If the function take into account only the platform itself and ignore all other friendly asset, then (1) changes as follows:

$$f: \mathbf{T} \text{ x } A_1 = \mathbf{V} \rightarrow [0,1] \quad (2)$$

In (2), $\mathbf{V} = \{V_1, \dots, V_n\}$ symbolizes threat assessment rating of each target with respect to A_1 which stands for our own platform. Similarly, the main purpose of this paper is to evaluate threat level of targets as function does. Our method will be introduces in following section of this paper in detail.

4. Method

It is compelling duty to separate friendly assets from hazardous ones when there is no identification information about surrounding asset. Kinematics of objects indicates limited foresight about the intent of objects. By fusing information extracted from objects' movement and orientation, it is possible to strength this foresight. Data fusion operation need attention while evaluating kinematics objects, because entities belongs to different environment has dissimilar features and it is impossible to evaluate air and surface objects with same perspective.

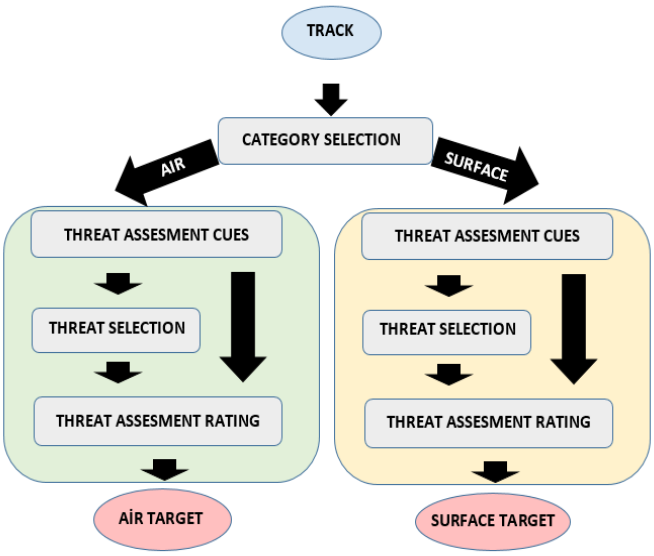


Figure 1 Threat Evaluation Model

The flow of each categories for threat evaluation process differentiate from the beginning of the model. The information about the environment of track is available. If track is a surface, then it is directed to flow handling surface tracks. Similarly air tracks are taken care of by air flow. Inside of each flow, steps are very similar to each other. Mainly, there is a three phase method inside each flow. Firstly, threat assessment cues phase generates various scores by benefiting from kinematics of objects. Then, these scores are transferred to threat selection phase that determined whether the entity show friendly or unfriendly attitude. This step marks the unfriendly objects as threat. Then, scores generated from first phase also transferred to final phase that calculates threat assessment rating of each object. This final phase provides to ranking of surrounding object depending on their threat level.

The Mentality of flows for each category is very similar as mentioned. Difference between each flow is to use convenient threat assessment cue in the right category.

Figure1 summarizes the general situation of threat evaluation model in this paper. Potential objects move to the related container depending on their environmental information. In each container, firstly, flow related threat assessment cues scores are generated from kinematics of objects. Then, threat selection phase determines dangerous objects by benefiting from these scores according to Bayesian Inference based algorithm. Dangerous objects are marked as threat at this phase for the use of the operator.

Finally, threat assessment rating phase calculates the threat rating of each track by using category related cue scores.

In brief, the model accepts potential objects as input and produces two different target list for each category. Surface objects are ranked among themselves. Similarly, air objects are listed on their own list. Objects carries threat rating and target mark with themselves.

4.1. Category Selection

At the beginning of the model, tracks are directed to different models according to their categories. If track is air track, then the remaining model specializes to air object; otherwise, surface dynamics are taking into account in the remaining model. Steps performed after this selection is differentiated from each other. This step is source of main difference with original work. Category selection could be seemed a very simple step. In fact, its effect spread out the remaining model in deep.

4.2. Threat Assessment Cues

It is important to use accurate threat assessment cues that generate convenient relationship between friendly assets and dangerous ones in right place while estimating threat level of an object [18]. In lots of studies, many threat assessment cues have been suggested for the use of threat evaluation models. However, it is impossible to benefit from all parameters since sensors of the naval platform could not provide all necessary information. Therefore, threat evaluation model is restricted to present parameters. Kinematic of objects serves position, speed and course values as input to the model in order to generate threat assessment scores.

Table I: Threat assessment cues with respect to environment

Air Tracks	Surface Tracks
• Speed • Distance • Heading • Maneuver • Closest Point of Approach (CPA) • Time Before Hit (TBH)	• Speed • Deceleration • Distance • Heading • Closest Point of Approach (CPA) • Time Before Hit (TBH)

Table 1 gives the full list of threat assessment cues that is generated from only kinematics objects which is basically position, speed and course of objects. Parameter lists are very similar to each other with a few difference. The way of generating scores does not change in each environment flow. However, weights of these parameters in algorithm changes according to category of asset. While determining parameters itself and weights of them, we benefit from mostly air warfare and surface warfare studies separately [9-11]. After the performance evaluation of the algorithms, weights of the parameters adjusted according to observation made during trials. Details of each parameters, namely speed, deceleration, distance, heading, maneuver, CPA and TBH are given as follows:

1) Speed: Speed parameter has a potential to be used to reveal classification of targets. An ordinary single engine aircraft could not reach the speed of fighter which is much more dangerous. However, a fighter could travel at lower speed. Similarly, for a surface track, fish boat can reach limited speed while boats are able to travel at higher speed. However, assault boat can stop in the middle of ocean and hide its characteristics easily in manner of speed. Therefore, there is no certain judge that classification could be made by only using speed parameter. Moreover, it is not close to the top of threat assessment cues with respect to importance. Still, speed is a good indication of target capability when it exists [10, 11, 14, 16].. Threat evaluation model uses this parameter considering seriously weight of it to final result. Speed parameter contributes the model as follows:

$$s_i = \frac{s_c}{s_{max}} \tag{3}$$

The score generated by speed parameter is found by dividing current speed value of target to maximum value of platforms. Here, maximum value of platforms differentiate according to category of target given in the beginning of model.

2) Deceleration: Change of speed does not imply important indication for air tracks. There is no pattern in speed parameter that indicates danger when observing air platforms. Therefore, the deceleration parameter does not used in air flow part of the threat evaluation model. When it comes to surface platforms, deceleration of naval platforms indicates abnormal intent for friendly assets [11]. If a naval vessel follows its route with steady speed, this vessel is supposed to be safe object for our own ship. Moreover, acceleration at speed of target indicated a danger for our own ship, but it is not much as deceleration case. The formal definition of the deceleration/acceleration score is indicated as follows:

$$r_i = \frac{s_i - s_{i-1}}{s_{max}}\ ,\ \begin{cases} r_i > 0\ , then\ w_a r_i \\ r_i < 0\ , then\ |r_i| \end{cases} \text{ where } w_a < 1 \tag{4}$$

As (4) shows, deceleration of objects has more influence on the score than acceleration does. This influence is adjusted with weight coefficient that decrease the score of acceleration with a specific ratio.

3) Distance: Location information of both own ship and target on earth sphere is available. Therefore, the haversine formula is able to compute the distance from our platform to target object [19]. The calculated distance is found in two dimensional space since there is no altitude data of surrounding air objects. Since the third axis in space is missing on calculation for air objects, the distance does not reflect the real situation for tactical picture of air entities. However, the amount of distance in two dimensional space is still serve valuable opinion about the intent of objects. At least, the objects are not closer than the amount calculated in two dimensional space. As expected, farther target is thought as safer than closer one. Therefore, it has lower score than closer one [10, 11, 14, 16].

$$d_i = 1 - \frac{d_c}{d_{max}} \tag{5}$$

The score of this cue is calculated in similar way for both air and surface objects. However, the contribution of this cue to total score is different with respect to importance level. While range is the most important kinematic cue for surface objects[11], some of other cues pass the importance level of this cue in air threat assessment studies [16]. Therefore, the contribution of the threat assessment cues is not same in air and surface tactical picture. It is adjusted with weight parameters according to their importance in algorithms while calculating total score.

4) Heading: The direction of target's heading is another important cue to understand target intent against defended asset [10, 14, 16]. As expected, if the target's heading is pointed out to defended asset, then it is a signature of danger for the defended one. If target's heading is not related with the positon of defended asset, then it means that target does not show any interest to defended object in manner of heading cue.

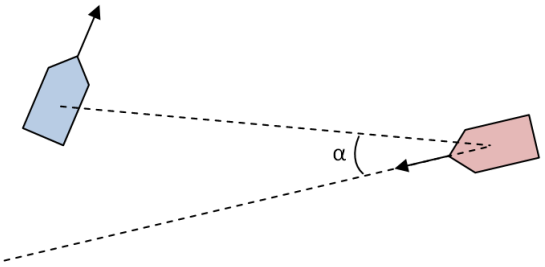


Figure 2 Heading Angle

The measurement of the danger level in this cue is computed by determining the orientation of target with respect to own ship position. In Figure 2, the red object is target and its heading is pointed out to blue object which is the defended entity with a degree of α angle. If α angle is getting bigger, then it means that target is not interested with the position of defended asset. Therefore, this situation decrease the danger on blue object. When α angle is getting smaller, then it means that red object point out its heading to blue object and it will come closer in time to defended asset. As a result, this positioning cause danger for the defended asset in manner of heading cue. As a result, the score generated by heading cue is calculated as follows:

$$h_i = 1 - \frac{\alpha}{\pi} \tag{6}$$

Since air objects has a hidden altitude value in two dimensional space, they would fly over defended asset even if their heading directly point out the defended asset. However, in surface tactical picture, this behavior is the danger of impact in the end. Therefore, the heading of target reveal more danger in surface tactical picture when compared to air tactical picture.

5) Maneuver: Maneuvers around the defended target is another unusual behavior that implies unfriendly act for it. Therefore, it is useful to take into account this kinematic data to threat evaluation process [10, 14, 16]. There are various options to measure the value of maneuver cue. The method could count maneuvers made at a specified time interval. However, this approach needs more frequent in order to determine the specific number of maneuvers. What behavior is thought as a maneuver is another discussion that needs to be solved. Instead of this complexity, a simple method come forward to handle situation. Firstly, the difference between consecutive course values is calculated and this value is divided to angle of maximum maneuver.

$$m_i = \frac{c_i - c_{i-1}}{c_{max}}, \text{ where } c_{max} = \frac{\pi}{2} \tag{7}$$

Air objects are capable of making sharp maneuvers. However, a plane that follows steady path does not requires to make this move frequently. Therefore, if a maneuver is observed from kinematics of air objects, then it make contribution to threat level of them. However, maneuver is not considered as a serious indication of danger for surface objects. Therefore, it is not used in the calculation of threat level for surface objects indeed.

6) Closest Point of Approach (CPA): Another key parameter in threat evaluation models is CPA of the target according to defended asset [20]. In many threat evaluation techniques, threatening tracks are prioritized by measuring the distance of threatening track to its CPA by own platform.

$$p_i = 1 - \frac{p_c}{p_{max}} \tag{8}$$

The score is generated in same manner of other cues. There is an acceptable maximum value for CPA distance, and the method divide the current value to maximum value.

CPA cue is beneficial to understand behaviors of air tracks and surface tracks. Air tracks are able to change their heading rapidly. This affects the CPA of target asset very quickly, also. According to air warfare studies, it has respectable impact on CPA. Actually, it is the one the most effective cue that measuring threat level of surrounding asset [11]. However, CPA for surface tracks does not produce same effect as air track does. Still, it is not an ignorable parameter and has specific effect on threat level of assets.

7) Time Before Hit: CPA is very useful indication to estimate behavior of threating asset. It is possible to strengthen this cue by taking account other kinematic information. In the study of Johnson and Falkman, TBH term is proposed in that manner. At that case, the speed of target and both CPAs are decisive factors. The time that suspected object follow the route to reach defended asset by passing through CPAs is measured. While calculating this time, defended asset is presumed to be stationary. If this time is smaller, then it means that there is undeniable threat against defended asset. Otherwise, defended asset is in safe territory...

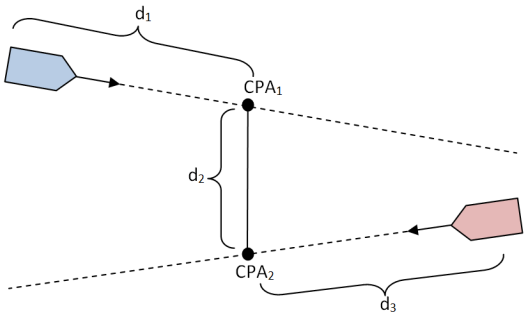


Figure 3 The Route Passing Through CPAs

Figure 3 shows the route of the suspected object needs to follow to get defended asset by passing through CPA points. Explanation of each leg of the route is described as follows:

- d_1: The distance between friendly asset and its CPA to the threating object.
- d_2: The distance between two CPA points.
- d_3: The distance between threating object and its CPA to friendly asset.

The time that takes the catch friendly asset in its current position by passing CPA points determines TBH: The score of this is generated by dividing TBH to a reasonable maximum value expected from TBH.

$$t_i = 1 - \frac{TBH_i}{TBH_{max}}, \text{where } TBH = \frac{d_1 + d_2 + d_3}{s_o} \tag{9}$$

TBH is taken into account for both air and surface tracks with a specific weight. This weight is adjusted by running scenarios consecutively and interpreting output of them.

4.3. Threat Selection Process Based on Bayesian Inference

It is difficult to observer dozens of surrounding objects even if a threat assessment rating is assigned to each of them by a threat evaluation algorithm. Classically, operator begin to examine targets from top of the list. After enough observation and investigation, C2 operator classify unknown target as friend and go on examining next target in the list; or C2 operator classify unknown target as hostile and take the necessary action needed. There is no border in the middle of list, which shows remaining part of the list is not worth to further investigation.

It is very critical to save time for C2 operator while making these observation and spend valuable time on real suspected object. We brings a mechanism based on Bayesian Inference that highlight the targets that need more attention than others. Formally, the method draws the border line on ranked list, which separates targets needs to be seriously assessed.

In Bayesian Inference, a hypothesis is put forward and the reality of this hypothesis is tried to be proved by evaluating evidences related with hypothesis [21]. After the evaluation, the hypothesis is interpreted as true or false. In our case, the hypothesis is whether a track is a target or not with a given scores of threat assessment cues. Here, threat parameters take the role of evidences in Bayesian Inference. The answer of the hypothesis gives the information whether track is target. If the result is positive, then related track is marked in the list. Therefore, C2 operator can concentrate on more these tracks that interpreted as target.

Bayesian Inference is an useful method showing the formula learning unknown status of the situation from available data [21]. Simply, Bayes' rule let us reach the posterior probability (the posterior) of a hypothesis with given prior probability (the prior) and compatibility of the observed evidence with the hypothesis (the likelihood) [22]. Bayes' formula points out the method to change probability statements by the use of evidences [15]. Formally, Bayes' theorem is formulized as follows [21]:

$$P(H \mid E) = \frac{P(E \mid H).P(H)}{P(E)} \tag{10}$$

Definition of each term in Bayes' theorem and the corresponding instances of these expressions inside the threat selection problem can be described as follows:

- H stands for the hypothesis that is tested according to Bayes' rule. The tested hypothesis in this paper is whether track is a target or not.
- E stands for the evidence that is available information to calculate posterior probability from prior probability in Bayes' formula. Threat assessment cues take role of evidences in this paper.
- P(H) stands for the prior probability of the hypothesis before the evidence is observed. In this paper, previous result of threat selection algorithm hold the place of the prior probability.
- $P(H \mid E)$ stands for the posterior probability of a hypothesis (H) after observing the evidence (E). In this paper, initial situation is the previous result of the threat selection algorithm, which is formally P(H). The fusion of the threat assessment cues bring to pass the observation of evidences. Then, posterior probability is obtained after application of this observation on prior probability.
- The remaining term in the formula, $\frac{P(E \mid H)}{P(E)}$, is the impact of evidence on the prior probability. Similarly, the data fusion of threat assessment cues' scores is determinant factor for the threat selection process.

Bayes' formula takes the following form after reinterpretation of terms according to threat selection process:

$$P(T_i \mid \mathbf{C}) = \frac{P(\mathbf{C} \mid T_{i-1}).P(T_{i-1})}{P(\mathbf{C})} \tag{11}$$

Similarly, explanation of each term in (11) as follows:

- T symbolizes the threat selection hypothesis that if a track is target.
- $\mathbf{C}$ denotes threat assessment cues which are evidences of hypothesis T.
- $P(T_{i-1})$ is the prior probability of the hypothesis before evidences are observed, namely, the probability found at the previous iteration of threat selection method.
- $P(T_i \mid \mathbf{C})$ stands for the posterior probability of the threat selection hypothesis (T) after observing threat assessment cues ($\mathbf{C}$).
- The remaining term in (11) is $\frac{P(\mathbf{C} \mid T_{i-1})}{P(\mathbf{C})}$ and it is the impact of the fusion of threat assessment cues on previous result of the threat selection method.

Threat selection method uses threat assessment cues according to their rank of importance is listed in Table 2 below. The importance ranking of the threat assessment cues mostly determined by taking reference of air warfare studies and surface warfare studies [11, 16].

Table II: Threat assessment cues according to the importance in threat selection

Air Tracks	Surface Tracks
• Closest Point of Approach (CPA) • Speed • Maneuver • Heading • Distance	• Distance • Heading • Deceleration • Speed • Closest Point of Approach (CPA)

TBH cue is removed from the list because of the assumption of conditional independence. This assumption does not allow to fuse evidences that are dependent to each other [22, 23]. Since TBH is dependent to speed cue, there is a need to remove one of them from the list. CPA holds the part of the information that TBH has. However, there is no way to reach speed effect in TBH. Therefore, threat selection algorithm does not accept TBH as evidence.

Scores of threat assessment cues reference multiple evidences in Bayes' rule. These evidences are fused to locate in Bayes' formula together. Therefore, there is a need to redefine (11) according to combination of these evidences. Equation (11) is turned to following form after processing multiple evidences rule:

$$P(T_i|\ \wedge_{j=1}^{n} C_j) = \frac{P(T_{i-1}) \prod_{j=1}^{n} P(C_j \mid T_{i-1})}{P(T_{i-1}) \prod_{j=1}^{n} P(C_j \mid T_{i-1}) + P(N_{i-1}) \prod_{j=1}^{n} P(C_j \mid N_{i-1})}$$

where $P(N) = 1 - P(T)$ and $1 \leq n \leq 5$
n : number of threat assessment cue
N : opposite hypothesis of T (4)

In (12), $P(T)$ denotes the probability of track to be a target Moreover, $P(N)$ stands for the probability of a track to not being a target. After the computation of probability of T hypothesis, decision making process of this process is simply performed as follows:

$$f(t) = \begin{cases} P(T_i \mid \mathbf{C}) \geq \beta, & t \text{ is threat} \\ otherwise, & t \text{ is not a threat} \end{cases} \quad (5)$$

Threshold value is chosen somewhere middle of the scale. Because, Bayesian Inference results tends to get maximum or minimum value when it becomes stable after couple of iteration. This result could be observed in the evaluation section of this paper.

4.4. The Calculation of Threat Assessment Rating

Threat selection algorithm separates the list of suspected objects as target that needs more attention and other tracks that does not have priority to be observed. There is still need of ranking on this list if a scenario includes dozens of objects.

Threat assessment rating of each object is calculated by taking account into scores generated from threat assessment cues. Each threat assessment cue does not affect the final result equally. They influence the result with respect to the importance on threat level against defended asset. While their weight on threat assessment rating is determined, studies done for air and surface warfare are considered as reference. The list of threat assessment cues and weight of them changes depending on category of surrounding objects. Table 3 gives this lists and importance order of cues as follows:

Table III: Threat assessment cues according to the importance in threat assessment

Air Tracks	Surface Tracks
• Time Before Hit (TBH) • Speed • Maneuver • Heading • Distance	• Distance • Heading • Deceleration • Speed • Time Before Hit (TBH)

As explained in section 4.2, each threat assessment cue does not carry same weight in calculations. Therefore, formal definition of the calculation of threat assessment rating can be described as follows:

$$v_1 + v_2 + v_3 + v_4 + v_5 = 1 \quad (6)$$

$$w_1 + w_2 + w_3 + w_4 + w_5 = 1 \quad (7)$$

$$V_a = \sum_{k=1}^{5} w_k C_k,\ \mathbf{C} \rightarrow [0, 1],\ V_a \in \mathbf{V} \quad (8)$$

$$V_i = \begin{cases} air\ track, & vt_i + v_2 s_i + v_3 m_i + v_4 h + w_5 d_i \\ surf.track, & w_1 d_i + w_2 h_i + w_3 a_i + w_4 s_i + w_5 t_i \end{cases} \quad (9)$$

First phase of the threat evaluation model generates the score for each threat assessment cue after the selection of the category. These scores are generated by following methods introduces in section 4.2. In (16), $\mathbf{C} = \{C_1, C_2, C_3, C_4, C_5\}$ denotes scores generated in that phase. Then, threat assessment rating is calculated by adding up these scores with a specified weight. (14) defines weights of air tracks and (15) shows the relation of weights for surface tracks.

Equation (16) defines the calculation of threat assessment rating formally. (17) is more clear version of (16), which gives details of formulas for each category. Threat assessment rating is calculated at each iteration. Whenever kinematics of track or own ship is changed, then threat assessment rating of each track is updated.

5. Evaluation

Threat evaluation model is evaluated with synthetic scenarios. Similar scenarios are formed for both air and surface objects. While playing these scenarios, scores of threat assessment cues and resulting threat selection value and threat assessment rating are recorded in combat management system. These records are visualized by using ordinary charts that show probability of threat assessment cue as y-axis and iteration of each calculations as x-axis.

Synthetic scenarios are generated with the tool that capable of produce track data and own ship data. Objects in simulation tool are able to follow route defined by the user. Moreover, category and speed of objects can be entered to system by the user. While playing scenario, simulation tool directs related data to combat management system.

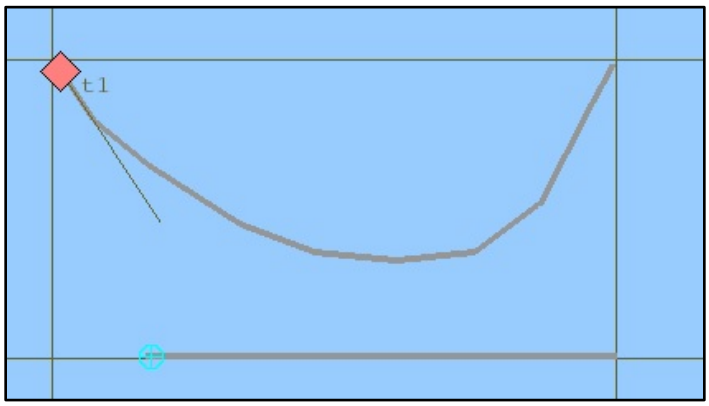


Figure 4 Syntetic Scenario 1

Figure 4 shows the first scenario generated in simulation tool. This scenario is played for both an air object and a surface object. The mentality behind this type of strategy is to see whether category affect algorithms results or not. We expects the difference results since threat assessment cues and their weight are different in categories.

According to first scenario, red object closes to the route of own ship that symbolized with blue mark, then point out its heading away and become distant to own ship as shown in Figure 4. In surface scenario, speed of the red object is convenient to a naval vessel platform.

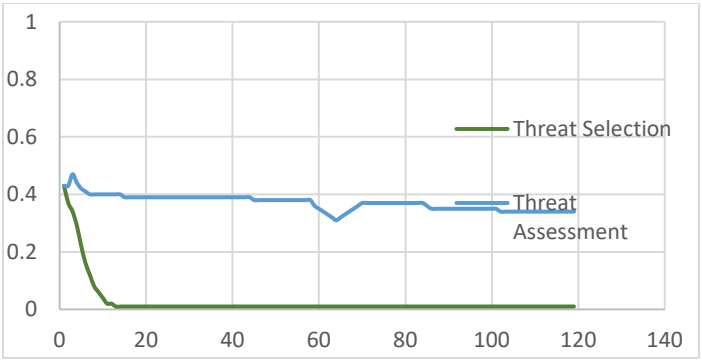


Figure 5 Results of Previous Study for Synthetic Scenario 1 Used for Surface Object

Figure 5 indicates that previous implementation [1] results seems to be very steady. There is no critical change on the threat level even if surface object changes its route significantly. Since common parameters and weights used for both categories in this implementation, kinematics of surface object could not exceed barriers of those parameters. Therefore, result is not a good indication for operator.

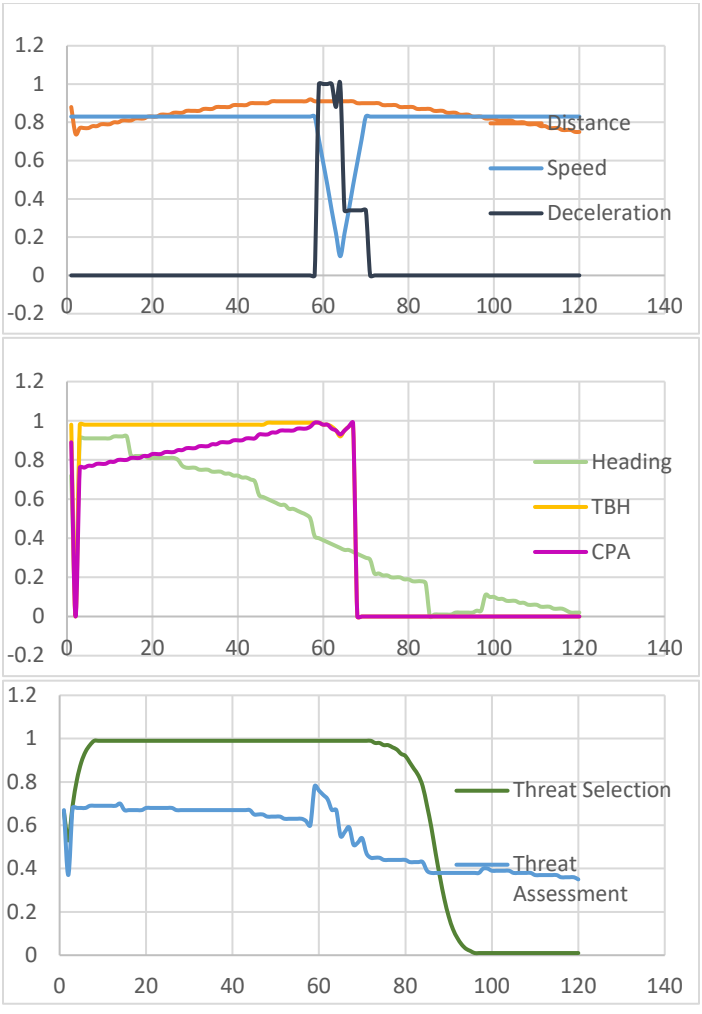


Figure 6 Threat Assessment Parameters Scores and Resulting Threat Selection Probability and Threat Assesment Rating for Synthetic Scenario 1 Used for Surface Object

Figure 6 indicates scores of threat assessment cues and results of algorithms. At the beginning of the scenario, TBH and CPA are very dominant over results. Since orientation of red object is very threatening for own ship, it classified as target by threat selection algorithm and takes around 0.7 threat assessment rating. In the middle of scenario, red object slower down its speed. This behavior causes an increase at deceleration score. Then, threat assessment rating increase as expected. I nth second part of the scenario, red object changes its orientation and give up being threat for own ship. Its threat assessment rating drops with this movement. However, threat selection algorithm does not classify the res object as friendly despite of last movement. This is a reasonable results since there is no guarantee that danger has passed away. Algorithm warns the operator for a while to follow actions of red object. After an enough number of iterations, red object seem to be not dangerous. Threat selection algorithm decrease its total score while it goes away.

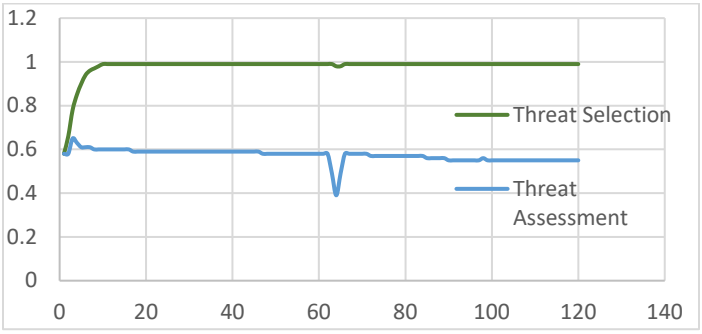


Figure 7 Results of Previous Study for Synthetic Scenario 1 Used for Air Object

Similar to the surface scenario, previous implementation [1] results seems to be very steady again in Figure 7. Air object is selected as a threat all the time and threat assessment rating does not change sharply.

Figure 8 reveals results of similar scenario with an air object. Red object follows again typical scenario shown in Figure 4 with the speed making sense for an air object.

In the beginning of the scenario, red object is categorized as target again by the threat selection algorithm. It takes around 0.8 threat assessment rating. When it turns out its heading way from defended asset, its threat assessment rating begins to drop down. After a while, TBH and CPA becomes low and it drop down the threat assessment rating sharply. Then, threat selection algorithm begin to decrease its threat level. In the middle of the scenario, speed of red object is slower for a while. This situation lead to decrease in threat assessment rating. Since deceleration does not mean anything for an air object, results of algorithm behaves different then what they do in surface scenario.

Through the end of the scenario, red object move away and become distant to friendly asset. As a result, threat assessment rating becomes close to 0.

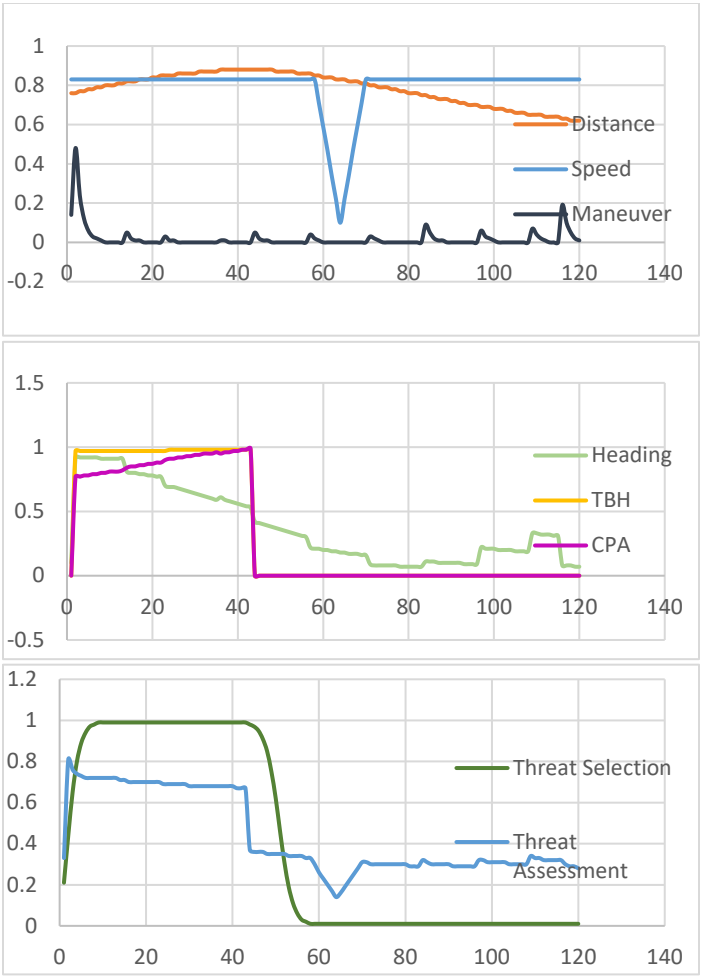


Figure 8 Threat Assessment Parameters Scores and Resulting Threat Selection Probability and Threat Assesment Rating for Synthetic Scenario 1 Used for Air Object

Figure 9 indicates second general scenario by the user in simulation tool. According to scenario, red object moves around defended asset by not changing its distance too much.

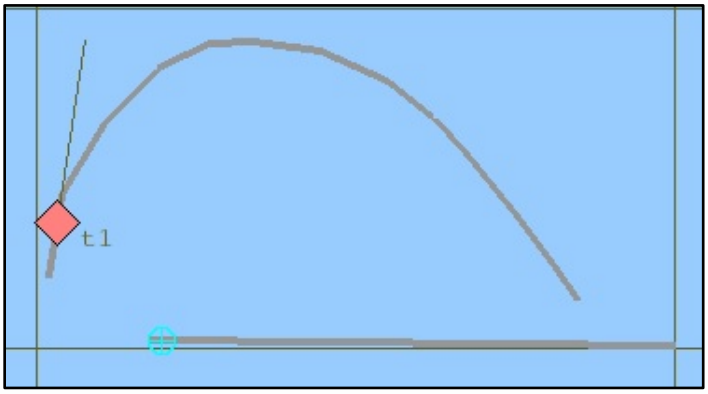


Figure 9 Syntetic Scenario 2

In Figure 10, it is not difficult to observe that previous implementation [1] could not give a valuable feedback about the foreigner entity as shown in figure. The object is classified as safe through the scenario.

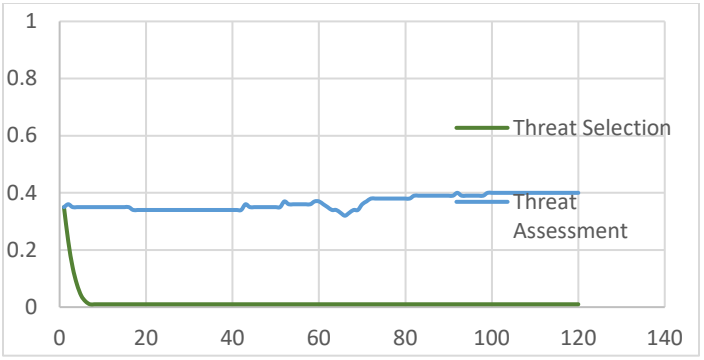


Figure 10 Results of Previous Study for Synthetic Scenario 2 Used for Surface Object

Figure 11 indicated results of second scenario for surface object. Since there is a peek for CPA and TBH in the beginning of the scenario, red object is sensed as threat. When red object sits on its road, threat selection algorithm give up threat classification for red object. In the middle of scenario, CPA and TBH becomes higher again. Threat selection algorithm notice the danger and classifies red object again as threat. However, threat assessment rating of red object does not affected from this change too much.

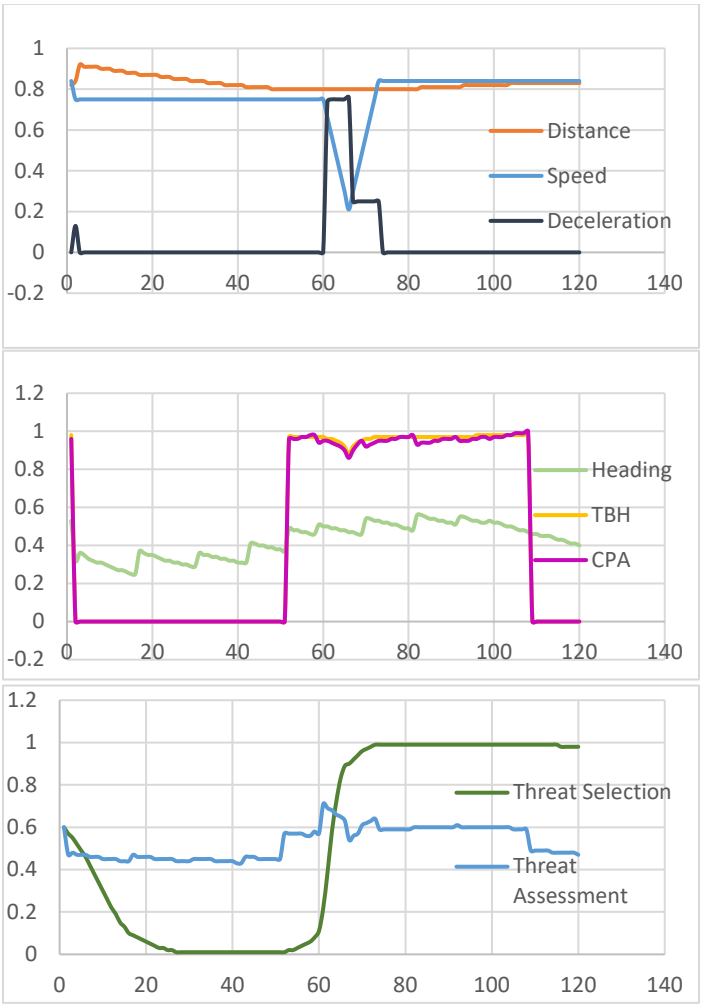


Figure 11 Threat Assessment Parameters Scores and Resulting Threat Selection Probability and Threat Assesment Rating for Synthetic Scenario 2 Used for Surface Object

Because, TBH is bottom of the list according to importance level of threat assessment cue. Threat assessment rating gets its highest value when surface object slows down. This behavior activates deceleration cue and increase threat assessment rating to highest value.

Similarly, previous implementation [1] does not provide beneficial picture to the operator as shown in Figure 12. Results does not affected from kinematics of object very much.

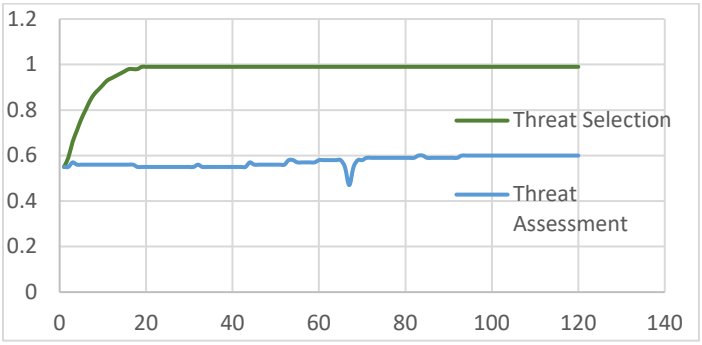


Figure 12 Results of Previous Study for Synthetic Scenario 2 Used for Surface Object

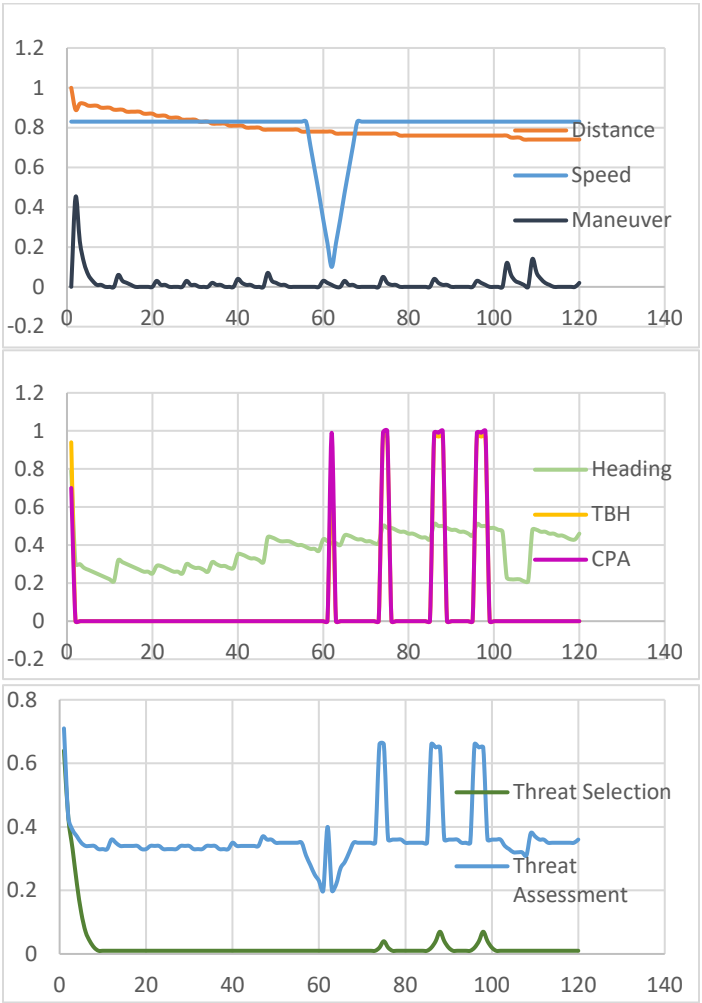


Figure 13 Threat Assessment Parameters Scores and Resulting Threat Selection Probability and Threat Assesment Rating for Synthetic Scenario 2 Used for Air Object

The route in Figure 9 is repeated for air object that has higher speed than surface object normally. As can be seen in Figure 13, red object is classified as threat at the beginning of the scenario. Because, TBH and CPA is effective at that time. Then, threat assessment rating begin to decrease. CPA and TBH shows themselves time to time. While the response of the threat assessment rating is a sharp increase, threat selection algorithm does no affected from this short period movement. Red object end scenario as safe object.

6. Related Work

There are many researches from different perspectives for threat evaluation problem. Mainly, rule based approaches, fuzzy logic methods and Bayes network studies are popular ones in this field as follows.

U.S navy officers have made serious contributions to a rule based study by answering question of survey related with threat evaluation in air defense domain [16]. Moreover, a different survey is conducted with U.S navy officers to contribute study done for threat evaluation in surface warfare [11]. These studies reveals most important threat assessment cues in both air and surface domain. After the extraction of these cues, rule based techniques applied on them to find out threat assessment rating of surrounding entities.

Fuzzy sets theory is used in another study for air threat evaluation process [15]. Range, altitude, CPA, speed, maneuver, weapon envelope, visibility, own support, fire, target support and IFF are used as threat assessment cues in proposed fuzzy inference system. It is claimed that simulation results indicate that proposed method is correct, accurate and reliable and it has minimum error in evaluating threats. By using fuzzy logic, they handle with imperfect situations, also.

Another study use altitude, speed, CPA and range as parameters to compute threat assessment rating of entities in tactical picture. This method is a fuzzy knowledge based reasoning model [24].

Bayesian belief networks takes important place in many threat assessment systems while performing fusion operations. Since Bayesian networks provides flexibility in terms of taxonomy of threats, it is chosen as a fusion method in Beaver's study [25]. Moreover, a security threat assessment model based on Bayesian Network is proposed in order to prevent subjective judgement of the information and bring objective perspective to OWA operators [26].

Another study uses Bayesian network to measure the overall probability of a threat that shows five anomalous ship behaviors. These behaviors are listed as deviation from ordinary routes, abnormal AIS activity, unexpected port arrival, abnormal close approach and zone entry [27]. For each behavior, the method shows whether the specific behavior occurs with a probability value.

In study based on Bayesian Network, target type, weapon range, speed, TBH and distance are taken into account while calculating threat basement rating [9]. Nodes of Bayesian Network is formed from these cues. There is hierarchical relationship between them, which could be examined more in detail at mentioned article.

A Bayesian Network based method is introduced by claiming advantages over other methods such as logical and fuzzy [18]. This study works for air defense scenarios. Moreover, linear Gaussian approximation seems to be necessary for their implementation.

Another threat evaluation system based on Bayesian Network is developed in Johansson's study [9]. Mainly, target type, weapon range, speed, TBH and distance are used as threat assessment cues. These parameters become the nodes of Bayesian Network while constructing the relationship among them.

Dynamics of threat evaluation problem changes depending on capabilities of platform and target's category. Systems that deployed on platforms has effective influence the problem. This situation is main reason behind the selection of threat assessment cues in the combat management system of the platform.

7. Conclusion

In this paper, we extent the work that originally presented in 2016 IEEE Symposium Series on Computational Intelligence (SSCI) [1]. Original paper presents threat evaluation model depending on kinematics of target without any indication about target's classification. Proposed model accepts all targets and produces results for them without checking any capabilities of objects. This paper updates threat evaluation model by adding category information to the model. In this time, model has category information about surrounding objects and there is a chance to benefit from this critical information.

Updated threat evaluation model begins with the category selection step. If an object is an air force, then the model apply air procedure on it; otherwise, surface parameters are taken into account to evaluate the threat level of the object. Basically, threat evaluation model for air and surface objects follows same path within themselves. Firstly, threat assessment cues are extracted from target's kinematics. Then, threat selection algorithm are applied on target by benefiting from scores generated at first phase in both air and surface flow. Finally, threat assessment rating are calculated by fusing threat assessment cues in both flows. However, there are significant differences that affect the resulting values between air and surface flow of the model. Firstly, air objects and surface objects are separated from each other by indication of danger. While maneuver cue is a indication of dangerous target for air objects, it does not have this kind of effect on surface object. Moreover, the deceleration cue has an impact on threat level of surface object, there is no influence of deceleration cue on air objects. Therefore, set of threat assessment cues change according to the category information of the target. Other threat assessment cues are common for both categories. However, their importance ranking are very different. While TBH is very important for air objects, distance dominates cores of surface object. This situation generates critical difference between flows of the model and it could be observed from results of synthetic scenarios. As a result, this study brings more confident results since it takes into account of accurate dynamics of related category while evaluating threat level of objects.

Original work apply same model to all targets without considering target's nature and capabilities. Current model eliminate these deficiencies at certain level. Similar synthetic scenarios are played for different categories. It is observed that behavior of models changes accordingly.

Evaluation section reveals that original work does not make an impact as current one while running on defined synthetic scenarios in this paper. Mostly, results of previous work seems to be steady. Once, a target is classified as threat, it remains at the same classification through the rest of the scenario. However, target moves and changes its direction in all scenarios. It seems to be Common parameters and cues for both air and surface objects does not perform well for these scenario.

Original work results with one common list that includes air and surface objects together by applying same methods on them. In this study, air and surface objects follows different flows with parameters special for them. In the end, threat evaluation model generated two list, one for air objects and one for surface objects. There is no fusion of these two lists. As a future work, threat evaluation model could be updated to produce one common list by combining results coming from different flows. Ranking surface and air objects will require serious investigation to reach confident results.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

We wish to thank our company, ASELSAN A.Ş, and in particular our division, Defense Systems Technologies, for their encouragement and vision that allow us to prepare this work.

References

[1] Çöçelli, M. and E. Arkın. "A threat evaluation model for small-scale naval platforms with limited capability". in *Computational Intelligence (SSCI), 2016 IEEE Symposium Series on*. 2016. IEEE.

[2] Boyd, J.R., "The essence of winning and losing". Unpublished lecture notes, 1996.

[3] Irandoust, H., A. Benaskeur, F. Kabanza, and P. Bellefeuille. "A mixed-initiative advisory system for threat evaluation". in *Proceedings of the 15th International Command and Control Research and Technology Symposium: The Evolution of C*. 2010.

[4] Riveiro, M., T. Helldin, M. Lebram, and G. Falkman. "Towards future threat evaluation systems: user study, proposal and precepts for design". in *Information Fusion (FUSION), 2013 16th International Conference on*. 2013. IEEE.

[5] Park, S.C., Y. Kwon, K. Seong, and J. Pyun, "Simulation framework for small scale engagement". Computers & Industrial Engineering, 2010. **59**(3): p. 463-472.

[6] Mury, B.N., Bao, "A Recursive Engagement Simulation Tree (REST) For Use in Maritime Defence". 2007, Defence R&D Canada – Atlantic.

[7] Bolderheij, F. and P. Van Genderen. "Mission driven sensor management". in *Proceedings of the 7th International Conference on Information Fusion*. 2004.

[8] Paradis, S., A. Benaskeur, M. Oxenham, and P. Cutler. "Threat evaluation and weapons allocation in network-centric warfare". in *Information Fusion, 2005 8th International Conference on*. 2005. IEEE.

[9] Johansson, F. and G. Falkman. "A Bayesian network approach to threat evaluation with application to an air defense scenario". in *Information Fusion, 2008 11th International Conference on*. 2008. IEEE.

[10] Liebhaber, M.J. and B. Feher, "Air threat assessment: Research, model, and display guidelines". 2002, DTIC Document.

[11] Liebhaber, M.J. and B. Feher, "Surface warfare threat assessment: Requirements definition". 2002, DTIC Document.

[12] Waltz, E. and J. Llinas, "Multisensor data fusion". Vol. 685. 1990: Artech house Boston.

[13] Blasch, E., et al. "High level information fusion developments, issues, and grand challenges: Fusion 2010 panel discussion". in *Information Fusion (FUSION), 2010 13th Conference on*. 2010. IEEE.

[14] Liebhaber, M.J. and C. Smith, "Naval air defense threat assessment: Cognitive factors and model". 2000, DTIC Document.

[15] Azimirad, E. and J. Haddadnia, "A New Data Fusion Instrument for Threat Evaluation Using of Fuzzy Sets Theory". International Journal of Computer Science and Information Security, 2015. **13**(4): p. 19.
[16] Liebhaber, M.J., D. Kobus, and B. Feher, "Studies of US Navy air defense threat assessment: Cues, information order, and impact of conflicting data". Studies, 2002.
[17] Johansson, F., "Evaluating the performance of TEWA systems". 2010.
[18] Okello, N. and G. Thorns, "Threat assessment using Bayesian networks". Information Fusion, 2003: p. 1102-1109.
[19] Robusto, C., "The cosine-haversine formula". The American Mathematical Monthly, 1957. **64**(1): p. 38-40.
[20] Roy, J., S. Paradis, and M. Allouche. "Threat evaluation for impact assessment in situation analysis systems". in *AeroSense 2002*. 2002. International Society for Optics and Photonics.
[21] Box, G.E. and G.C. Tiao, "Bayesian inference in statistical analysis". Vol. 40. 2011: John Wiley & Sons.
[22] de Vos, A.F., "A primer in Bayesian Inference". preprint, 2004.
[23] Russell, S. and P. Norvig, "AI a modern approach". Learning, 2005. **2**(3): p. 4.
[24] Liang, Y. "An approximate reasoning model for situation and threat assessment". in *Fuzzy Systems and Knowledge Discovery, 2007. FSKD 2007. Fourth International Conference on*. 2007. IEEE.
[25] Beaver, J.M., R.A. Kerekes, and J.N. Treadwell. "An information fusion framework for threat assessment". in *Information Fusion, 2009. FUSION'09. 12th International Conference on*. 2009. IEEE.
[26] Wu, K. and S. Ye, "An information security threat assessment model based on Bayesian network and OWA operator". Applied Mathematics & Information Sciences, 2014. **8**(2): p. 833.
[27] Lane, R.O., D.A. Nevell, S.D. Hayward, and T.W. Beaney. "Maritime anomaly detection and threat assessment". in *Information Fusion (FUSION), 2010 13th Conference on*. 2010. IEEE.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1342-1349 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

The effect of employing knowledge workers from technologically advanced countries: The knowledge spillover caused by the mobility of knowledge workers in electronic industries in Asia

Ayano Fujiwara*

Senior Research Fellow, Ministry of Education, Culture, Sports, Science and Technology (MEXT), National Institute of Science and Technology Policy (NISTEP), Chiyoda-ku, Tokyo 100-0013 JAPAN

ARTICLE INFO

Article history:
Received: 24 May, 2017
Accepted: 21 July, 2017
Online: 01 August, 2017

Keywords:
Mobility of knowledge workers
Knowledge spillover, Social network, Tacit knowledge, Star inventor

ABSTRACT

This paper is an extension of work originally presented in 2017 6th International Conference on Industrial Technology and Management (ICITM). This research focuses on the mobilization of human resources in the Asian electrical industry and analyzes the impact of engineers' international mobility on the innovation of enterprises in the host countries. I examine the characteristics of engineers who moved from Japan to China and South Korea to explore which types of engineers successful firms in China and South Korea select from rival companies in Japan. In addition, this study focuses on the role of the inventors' networks to investigate the significance of researcher mobility and knowledge spillover as mechanisms that facilitate the flow of tacit knowledge. This study reveals that during the process of innovation, informal networks play an important role in the mobility of engineers. The findings suggest that successful firms require people with analytical skills and problem-solving capacity, not "star inventors."

1. Introduction

To spur innovation and economic growth, the endowments of knowledge factors, such as research and development and human capital, are a necessary condition to generate new technologies. The greatest source for the generation of new economic knowledge is generally considered to be R&D (Research and Development) [1], [2]. R&D not only generates innovation but also develops firms' ability to identify, cultivate, and exploit knowledge and ideas developed by other firms. In other words, firms are exogenous, whereas their performance in generating technology changes is endogenous [2]. Endogenous growth models suggest that R&D activities undertaken by profit-maximizing firms represent an important investment in obtaining new knowledge, whereas knowledge is an input in the process of generating endogenous growth [3]-[5].

In this sense, knowledge spillover is an important mechanism for endogenous growth. Small and young enterprises can supplement their lack of R&D through knowledge spillover [6]. The mechanisms by which knowledge is transmitted across firms and individuals include, for example, publications; the mobility of employees, scientists and engineers; and reverse engineering. The mobility of knowledge workers in general and new firm start-ups in particular has become an important mechanism of knowledge spillover [7]. Michelacci (2003) implies that a low rate of return on R&D may be attributable to an inability to transform new knowledge into economic opportunities through skills, aptitudes, and insights [8].

Knowledge that can be easily codified can be transmitted and understood over long distances at very low cost [9]. However, tacit knowledge is difficult to communicate, and it therefore relies on face-to-face interaction to be transmitted effectively [10]. Thus, the knowledge spillover resulting from the mobility of knowledge workers is especially important because it facilitates the transfer of tacit knowledge, which is not readily transferable via codified sources. The mobility of knowledge workers, such as scientists, inventors, and engineers, has a significant and positive impact on new firms' output and productivity growth [11], [12]. Knowledge workers from large, established firms may have knowledge and ideas about inventions and ultimately about innovations. The mobility of knowledge workers therefore plays a central role in the spillover of knowledge and ultimately in economic growth. Additionally, informal networks and labor mobility are closely linked with respect to their emergence and re-configuration [13]. Cross et al. (2001) investigate methods of improving employees'

*Corresponding Author: Ayano Fujiwara, Science and Technology (MEXT), National Institute of Science and Technology Policy (NISTEP), Chiyoda-ku, Tokyo 100-0013 JAPAN | Email: ayanokaayano@gmail.com

ability to create and share knowledge in social networks [14]. However, the literature identifying the mechanisms underlying the transmission of knowledge spillover by social networks remains underdeveloped.

Engineers' mobility plays an important role in the transfer of knowledge from stable and mature enterprises to high growth and young enterprises, especially in high-technology sectors. Many studies have examined the relationship between the mobility of knowledge workers and innovation; however, the relationship between the characteristics of engineers who move abroad and innovation is a less-researched field. Only a few studies have analyzed the impact of R&D investment and human capital on innovation. Therefore, this study empirically analyzes the impact on innovation of the quality of the talent that migrated from developed countries.

This paper focuses on the link between the mobility of individual engineers and knowledge spillover. The purpose of this study is to identify the influence of engineers' mobility on firm development and innovation. I investigate the role of engineers who migrated from large, established companies to young companies in transmitting knowledge. Engineers from large, established corporations often have new knowledge that they acquired through research and development work in their firms' labs.

I propose a hypothesis to investigate the link between researchers' mobility and R&D performance. The hypothesis is that young, successful firms in China and South Korea are selective in their hiring, overwhelmingly recruiting talent from large, established firms in countries such as Japan with the expectation that the knowledge spillover embedded in their new hires will allow the young firms to catch up technologically. I explore this hypothesis by investigating patent data to estimate the effect of knowledge transfers from Japanese firms to Chinese and South Korean firms. I analyze the mobility of engineers from Japanese firms to Chinese and South Korean firms by drawing on patent data from 1976 to 2013. Young, successful firms in China and South Korea employ many engineers who were formerly employed by Japanese firms. I therefore examine the characteristics of engineers who moved from Japan to China and South Korea to explore what types of engineers Chinese and South Korean firms select from rival companies in Japan. In addition, this study focuses on the role of engineers' networks to investigate the significance of engineer mobility and knowledge spillovers as mechanisms facilitating the flow of tacit knowledge. Thus, I use social network indices to analyze co-invention networks. Additionally, I examine whether these engineers play an essential role in innovation within their new companies.

Next, I examine what types of researchers absorbed technology and knowledge from developed companies and contributed to innovation in young companies such as Chinese and South Korean firms. To this end, I use a knowledge production function model to obtain insights into the process of innovation using the knowledge embedded in new hires. For the specific methods of analysis, the knowledge production function developed by Pakes and Griliches (1984) is used; for the explained variables, patent citation counts are used as a representative index to measure innovation through the number and quality of patents [15]. I use panel data from 5 young, successful firms in China and South Korea over the past two decades.

A few studies have focused on the effect of talent migration on knowledge spillover via informal networks. One of these papers focuses on scientific labor as a measure of spillover [16] and studies the personal characteristics of scientists employed at a biotech company. Here, I extend this framework by examining engineers who moved from developed firms in Japan as the source of knowledge transmission. This study examines the relationship between the characteristics of individual knowledge workers and knowledge spillovers. It also investigates the relations between the mobility of knowledge workers and their informal networks during the process of innovation.

2. Theory and Hypotheses

2.1. Mobility of knowledge workers and innovation

This study conducted an empirical analysis of the effect of knowledge workers' mobility on innovation. In the previous literature, the influence of investment in R&D and human capital has been analyzed [15], [17]-[22]. These studies examine the influence of R&D input on R&D output by using the knowledge production function and considering the patent productivity of R&D using company-level data from North America and Europe. In addition, several models related to innovation have been presented, including some that use a knowledge-production function to analyze the effect on innovation activities produced by human capital and investment in R&D. Those studies focus on the role of human capital and R&D expenditures in obtaining knowledge capital in return.

Unlike those previous analyses of innovation and technological change that focus on the firm, in this paper, I focus on the individual. Some researchers have examined scientist mobility in the context of individual careers. Many of these studies focus on Ph.D. graduates [23], [24]. Zucker et al. (1997) examine the mobility of "star scientists" as a knowledge transfer mechanism [25].

The most prevalent model in the literature on innovation is that of the knowledge production function, which was formalized by Zvi Griliches (1979) [19]. The inputs in the model of the knowledge production function are R&D cost, human capital, skilled labor, and educational levels [19], [26]. The model of the knowledge production function from the literature on innovation and technological change can be represented as

$$IA_i = \alpha RD_i^{\beta} X_i^{\gamma} \varepsilon_i$$

In the equation, IA represents the degree of innovative activity, RD represents R&D input, and X represents human capital inputs.

Conventionally, R&D investment costs have been widely used as surrogate variables for innovation. However, the size of R&D costs is an input index, which cannot be used to measure the output of innovation. In this study, I use the number of patents as an output index of innovation. However, both the quality of patents and the number of patents are important issues. In this regard, patent quality has been studied, and the number of times a patent has been cited has been proposed as a proxy for patent quality. Trajtenberg (1990), Jaffe et al. (2000), Reitzig (2003), Hall et al. (2005) and Carpenter et al. (1981) have found a strong correlation between the number of citations and the quality of patents [27]-[30]. For example, Harhoff et al. analyze American and German patents and find that the number of citations in the patents that were updated before expiration was greater than that for the patents that expired

earlier [31]. For these reasons, I use the number of citations as an index to measure innovation in this study.

2.2. The characteristics of engineers hired by successful firms in China and South Korea

A combination of various types of technologies can be interpreted as an indicator of the increasingly complex nature of innovation [32]. Small enterprises can make up for a lack of innovation through knowledge spillovers from engineers who have experience with the complex nature of an innovation. This study analyzes the impact of the transfer of technology and knowledge on innovation in a corporation caused by the mobility of engineers who previously held positions with mature enterprises and then moved to young, successful firms. For this analysis, the characteristics of the knowledge workers hired by young, successful firms in emerging nations must first be analyzed. In particular, I shed light on the link between engineers' personal characteristics and inventors' attraction to a rapidly growing company. To measure the characteristics of talent, various indices have been considered. From the recruitment perspective, superiority as a researcher and whether the person held a position where in-house technology and knowledge could be accumulated are important factors.

Young, successful firms in China and South Korea must be selective in their hiring when recruiting engineers from Japanese firms. Warren (1997) emphasizes that a person who has produced results in his or her former work will have a positive influence on the new firm's productivity after migration [33]. Thus, successful firms will select engineers who have had excellent results in their former job. A high value for the number of citations received (forward citation) implies that a technology is used in diverse fields of future invention. I hypothesize that successful firms will hire inventors whose patents have been extensively cited. A high score on the number of patents cited (backward citation) means that the patents are based on past inventions drawn from diverse categories. This leads to my first hypothesis:

Hypothesis 1-1. Successful firms employ new hires who have had excellent results in their former jobs and have experience with the complex nature of innovation.

Because recruited talent is expected to lead local inventors in developing technology and knowledge in emerging nations' corporations, their rich experience and knowledge should have a strong impact. Therefore, successful firms will hire engineers who have long-term experience in stable and mature enterprises. Additionally, successful firms will hire engineers who have rich experience in a variety of technical domains. Thus, I propose as follows:

Hypothesis 1-2. Young, successful firms select engineers as new hires who have long-term experience.

Hypothesis 1-3. Young, successful firms select engineers as new hires who have rich experience in a variety of technical domains.

Similar statements can be made about informal networks. Successful firms employ people who occupy central positions in inventor networks. In other words, "star engineers" are selected by Chinese and South Korean companies. Zucker et al. (1996) study the mobility of "star scientists". They emphasize that the very best "star scientists" play central roles in both the development of the science and its successful commercialization [34]. Hence, I propose as follows:

Hypothesis 1-4. Successful firms select "star inventors" as new hires.

Zellner (2003) studies the various elements of knowledge transfer via scientists' migration [35]. He investigates the migration of scientists formerly employed at the Max Planck Society in Germany by using survey data obtained from a sample of scientists who moved from the Max Planck Society into the commercial sector. The main result of this analysis is that knowledge that is specifically related to a certain research question is not the most important issue for the scientists employed by the Max Planck Society who transferred into the private sector. The results suggest that non-specific knowledge was valued more highly than specific knowledge. In other words, broad knowledge and analytical skills are more important than propositional knowledge. It is said that those who are strongly tied to the star people in the network can acquire a wide range of knowledge [36]. Hence, I propose as follows:

Hypothesis 1-5. Successful firms select engineers as new hires who are strongly connected to "star engineers."

2.3. The type of talent contributes to innovation in emerging nations' enterprises

Knowledge workers who have rich experience with inventions will contribute to innovation. Amalya (2004) indicates that scientists who have rich experience with inventions increase the number and citation rate for firm patents and contribute to firms' success [37]. Another study [25] finds that "star" bio-scientists play a central role in their firms' success. Thus, knowledge workers from developed countries are expected to have vast scientific knowledge and communicative competence. This discussion leads to the following hypothesis:

Hypothesis 2-1. "Star engineers" who were formerly employed in a developed country's enterprises contribute to innovations by emerging nations' enterprises.

The ability to exploit external knowledge is a critical component of innovative capabilities, and the ability to evaluate and utilize outside knowledge is related to the level of prior knowledge. The ability to recognize the value of new information, assimilate it, and apply it to innovation is defined as "absorptive capacity" [1]. An inventor must have a strong absorptive capacity or the ability to successfully replicate new knowledge to access the new knowledge developed by other units [38]. A similar statement can be made about firms' absorptive capacity. While encompassing a firm's ability to imitate new process or product innovations, absorptive capacity can also include a firm's ability to exploit outside knowledge. Cohen (1990) suggests that the development of a firm's absorptive capacity will build on its prior investment in R&D [1]. Thus, I hypothesize that engineers who moved from Japanese firms will contribute more to Chinese and South Korean firms' innovation in the growth stage. Hence, I propose as follows:

Hypothesis 2-2. To promote innovation through knowledge spillover, enterprises in emerging nations need an adequate "absorptive capacity."

3. Data and Methodology

3.1. Data

In this study, I examine the characteristics of engineers who moved from Japan to China and South Korea and the characteristics of engineers who contribute to Chinese and South Korean firms' innovation. I used US patent data for filings by Japanese, Chinese and South Korean companies between 1976 and 2013. I calculated the number of years of experience from the year that the inventor first filed for a patent, so it was necessary to use data from as far back as 1976. It is said that a particularly high number of engineers who moved from Japanese companies to Chinese and South Korean companies are in the electrical equipment industries. Therefore, I used patents from the H section of the IPC classes. IPC stands for International Patent Classification and is a universal patent classification that was created based on the "Strasbourg Agreement on International Patent Classification" (effective in 1975). The H section indicates a patent in the electric field. I targeted approximately 270,000 patents from Japanese electrical manufacturers, approximately 50,000 Chinese patents, approximately 70,000 South Korean patents, and approximately 50,000 Taiwanese patents. I extracted all of the inventor names from those patents and confirmed people with the same name as Japanese inventors among the Chinese and South Korean inventors. Because inventors with identical names as inventors in Section H are highly likely to be the same people, the fact that their names appear on patents filed by both Japanese companies and Chinese and South Korean companies means that these people have moved between these companies. However, it cannot be denied that it is possible that the same name might refer to different people. Therefore, I decided to establish a criterion for the degree of similarity between IPC numbers at the IPC main group level, and only in cases in which that criterion was met did I judge the people to be equivalent. In this way, I judged the names of inventors who appeared in both Japanese and Chinese or South Korean patents to be equivalent when they met the requirements for a connection between IPC numbers, and I judged them to have moved between Japanese companies and foreign companies, thus raising the probability that these were actual movements. After identifying the movements of all of the inventors, I judged whether they moved from or to a Japanese company. Moreover, using the last patent application by their former company and the first patent application by their destination company, I calculated the probability of the timing of the movement for all transferees. I checked the patent data for the names of the companies where the inventors previously worked in Japan and the names of the companies to which they moved in China and Korea. I deleted data that conflicted with the period in which they were assigned to one of the countries.

Then, I calculated a network index for all of the inventors that were employed by Japanese companies to check the characteristics of the inventors who were also employed by Chinese or South Korean companies.

I focused on 5 successful companies in China and South Korea: Samsung, LG, Hyundai, Huawei and Hon Hai Precision. Samsung, LG, and Hyundai are South Korean companies. Huawei is a Chinese company. Hon Hai Precision is Taiwanese company.

For Hypothesis 1, the analysis is based on the notion that rapidly growing companies are selectively recruiting engineers from developed countries. The engineers who migrated from corporations in Japan to the 5 successful companies and to other companies in China and South Korea were identified from the patent data filed by the Japanese companies. All of the engineers' names listed as inventors on the patents mentioned above are extracted. Excluding those with identical names, there are approximately 170,000 inventors from Japanese companies and approximately 50,000 inventors from Chinese and South Korean companies. Then, all of the names of the inventors mentioned above who appear in both patents filed by Japanese companies and those filed by non-Japanese companies are extracted. Because engineers with identical names who are mentioned as inventors in Section H are highly likely to be the same people, the fact that their names appear on patents filed by both Japanese companies and Chinese and South Korean companies means that these people have worked in both countries. However, even given identical names, some of the data may refer to different engineers. Therefore, certain conditions were set to eliminate these cases by checking the International Patent Classifications. Next, with regard to patents with inventor names that match those extracted as above, the company that employed them was identified by examining the applications. Then, by considering the order of the years in which the applications were filed, the researchers' career changes could be identified regarding the company they left and the company they then joined.

In Hypothesis 2, to analyze how talent from developed countries contributes to innovation by emerging nations' enterprises, patent data and financial data from a 22-year period (1990–2011) are used to perform an analysis using panel data for a Chinese corporation (Huawei), a Taiwanese corporation (Hon Hai), and three South Korean companies (Samsung, LG, and Hyundai).

3.2. *Explained variable*

Regarding Hypothesis 1, to indicate which type of knowledge workers migrate to emerging nations' enterprises, the explained variable can have two values: 1 for migrating to emerging nations' enterprises and 0 for not migrating.

For Hypothesis 2, the number of patents is used in this study to determine innovations by emerging nations' enterprises when measuring the results for knowledge workers from developed countries' corporations. Other well-known studies have used the knowledge-production function approach [15], [17]-[22]. Thus, the knowledge production function employing the number of patents is used in this study.

Furthermore, from the perspective of measuring the quality of patents, the citation count is also used as an explained variable because the analysis shows that in many cases, the citation count of American patents (forward citation) is strongly correlated with the value of the patent [27]-[30]. For example, Hall et al. (2005) demonstrate that the aggregation of the citation count of a patent owned by a corporation is correlated with the company's share price [30]. In this way, the citation count is an excellent index to verify the impact of the quality of patents.

3.3. *Explanatory variable*

In this study, to analyze the impact caused by the mobility of human resources and inventor networks on innovation by Chinese and South Korean firms, numerous indices representing the quality of talent are established. To analyze the role played by the mobility of knowledge workers and their influences on innovation and growth in Chinese and South Korean companies, I examine the

characteristics of knowledge workers based on the aspects of skills and inventor networks. Specifically, a social network index and some indices to measure the features of the talent are used as explanatory variables based on patent data.

I use social network indices of co-invention by inventor networks using patent data. I calculate the social network indices for each inventor in the company that employs him or her, and I examine the structural positions in the network. Social networks are composed of different nodes of individuals and the social relationships that individually connect them. By observing the centrality index of each node, it is possible to analyze the characteristics and roles of each node in the overall network. I investigate co-inventor networks by using patent data and develop social network indices based on degree centrality, closeness centrality, betweenness centrality (Freeman, 1979), eigenvector centrality (Bonacich, 1972), clustering coefficients, and page ranks. Degree, betweenness, closeness, and eigenvectors are the four main measures of centrality. These centrality measures capture the relative position of an individual within the network. Degree centrality is defined as the number of links incident upon a node—in other words, the number of ties that a node has [40]. Degree centrality is based on a measured focus on the level of communication activity. In contrast, betweenness centrality reflects the intermediary location of a node along the indirect relationships linking other nodes [40]. Betweenness centrality captures stress control and the capacity to interrupt communication. The maximum theoretical value of betweenness centrality is given by (n-1)(n-2)/2 [41]. Closeness centrality measures depend on the geodesic distance, which is the minimal length of an indirect path. Freeman (1979) suggests that measures of centrality based on closeness reflect a node's freedom from the controlling actions of others and their capacity for independent action within the network [40], [42]. Eigenvector centrality reflects the fact that individuals might profit from well-connected friends [43]. I focus on degree centrality and eigenvector centrality.

A node with high degree centrality has many direct connections with other nodes, and people with high degree centricity are central people in communication within the network. [44]. High degree centrality means that the inventor has experience cooperating with many other inventors from his company. As an individual becomes exposed to increased degree centrality, his or her position in the network will attain a higher status [45], providing more opportunity to learn, acquire and utilize knowledge. Thus, an inventor who has high degree centrality is considered a "star inventor."

Eigenvector centrality is a centrality index proposed by Phillip Bonacich; it evaluates not only high-scoring nodes but also all nodes that are connected to high-scoring nodes. In other words, a high score for eigenvector centrality means that the node has connections with high-scoring nodes. This means that eigenvector centrality is a measure of the influence of a node in a network. Being related to a node that has many relationships with other nodes implies an important position within the network. In other words, an inventor who has a high score on eigenvector centrality is a person with whom "star inventors" always want to work. This is because individuals in such positions can obtain a great deal of important information and have both analytical skills and problem-solving capacity. Recently, the strategic importance of patent analysis has been highlighted by high-technology management as the process of innovation has become more complex [46]. Furthermore, patents facilitate analytical work through their relative advantages [48]. In addition, to measure the features of migrating talent, variables such as the number of migrated inventors, the size of the company of origin, the number of years in the inventor's career, the citation count, the forward citation count, the Herfindahl-Hirschman Index of the IPC number, and eigenvector centrality are used.

The patent citation count is used as a representative index to indicate patent quality. This approach has been validated by various studies [49], [31], indicating that the forward citation count is effective for measuring the quality of a patent. In addition, an inventor who has been involved with many high-quality patents can be regarded as excellent. The total citation count of a patent in which the inventor in question has been involved is used as a surrogate index for his/her excellence. Thus, the forward citation number is an important performance indicator to measure the performance of individual inventors. I therefore count all forward citations received by each patent for each inventor.

Furthermore, as an index to measure patent quality, in addition to the citation count, it has been noted that the number of a patent's backward citations correlates with the quality of the patent [31]. The number of backward citations is effective for measuring a patient's originality [50]. A high score on the number of backward citations means that the patent is complex, and a low score means that the patent is highly original. Therefore, the number of times a patent has been cited has been added as a variable to measure the quality of talent. I use the total number of patent citations (backward citations) as a representative index of originality and complexity.

The number of years in the inventor's career is calculated as the number of years elapsed since the year in which the inventor in question filed his/her first patent application.

In addition, the Herfindahl-Hirschman Index of the IPC number can measure whether an inventor has been involved with patents in a specific technical domain or in a variety of technical domains. In other words, a high HHI value indicates that the inventor specializes in a specific technical field, whereas a low HHI value indicates that the inventor has wide knowledge covering a variety of technical fields.

Another factor affecting the quality and quantity of patents is an increase in R&D costs. There have been numerous analyses [15], [17], [18], [21] of the cost of R&D and patent applications. In general, the number of patents increases and their quality improves as R&D cost increases, regardless of inventor quality. Therefore, in this study, R&D cost in the current period is also used as a variable.

3.4. *Control Variables*

Most empirical studies of firm performance include firm size as either a control variable or an independent variable. Following this tradition, I use firm size as a control variable. In terms of the scale of the company of origin, corporations with $100 trillion or greater in sales are classified as large, whereas corporations with less than $100 trillion in sales are classified as small.

3.5. *Estimation technique*

To test the first hypothesis and determine the characteristics of knowledge workers migrating to Chinese and South Korean enterprises from Japanese corporations, a logistic model is built for

the analysis using two-value objective variables. The model formula is as follows:

$$Pi(i=1)=F(C+\sum \alpha ixi)$$

$$Pi=\exp(C+\sum \alpha ixi)/(1+\exp(C+\sum \alpha ixi))$$

Pi represents the probability of mobility of engineer i, taking the value of 1 if the inventor migrated to an emerging nation and the value of 0 if he/she did not. F is the cumulative distribution function, αi is a coefficient, Xi is an explanatory variable, and C is a constant.

To test the second hypothesis and determine the influence of engineer mobility on innovation by Chinese and South Korean firms, the knowledge production function developed by Griliches and Regev (1995) is applied [26]. The knowledge production function measures the effectiveness of R&D investment for increasing knowledge. Griliches and Regev (1995) establish the following formula to compare labor productivity between companies [26].

$$Y = Xb + Zc + lm + \mu$$

Here, Y represents the logarithm of productivity, X is the logarithm of the input variable (intermediate inputs and capital services per person), and Z represents the control variables (e.g., dummy variables for size, age, and location). The authors establish eight variables: 1) intermediate inputs and fixed capital services, 2) R&D capital and labor quality (as a proxy for human capital), 3) size, 4) sector and type of ownership, 5) industry grouping, 6) establishment year, 7) mobility status (entry and withdrawal status), and 8) year dummies.

In this study, the impact of the migration of engineers from corporations in developed countries on innovation in emerging nations' enterprises is examined by entering the variable related to the knowledge worker's characteristics. Here, I extend this framework to examine the link between knowledge workers' characteristics and innovation. The model formula is as follows.

$$Y = Xb + Zc + Wd + lm + \mu$$

Y stands for the degree of innovative activity, X represents R&D inputs, Z represents local human capital inputs, and W represents the characteristics of knowledge workers who migrated from Japan.

In this study, I use the number of patents and the number of citations as a surrogate variable for innovative activity. The number of patents used as an explained variable represents non-negative numeric data. In this study, a fixed-effect model and a random-effect model are used to perform the calculations, following the example of Hausman et al. (1984) [17]. Hall and Mairesse (1995), Mairesse and Hall (1996), and Harhoff (1998) analyze French, American, and German manufacturers using a fixed-effect model to control for the deviation of a company's fixed effect [51]-[53].

4. Results and discussion

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

4.1. Identify the Headings

The main regression results are reported in Table 1. I analyze the characteristics of knowledge workers recruited as new hires by 5 successful firms in China and South Korea. A logistic regression model is used to verify the influence of the basic explanatory variables, such as inventors' individual social network index, the number of years in their career, and the evaluation of their invention. Model 1-6 reports the results for South Korean companies and Model 7-10 reports the results for Chinese companies.

Consistent with Hypothesis 1-1, the results indicate that the coefficient of forward citations is significantly high for South Korean companies; however, it is negative for Chinese companies. Conversely, the results for the number of backward citations are negative for South Korean companies and positive for Chinese companies. These results mean that South Korean firms employ knowledge workers who have had excellent results in their former jobs and have the ability to be original. In contrast, Chinese firms employ knowledge workers who have not had excellent results in their former jobs but who have experience with the complex nature of innovation. These results suggest that there is a high probability that South Korean companies select new talent by considering the number of times that a given patent has been cited, which is the most commonly used index to indicate a patent's importance.

In addition, consistent with Hypothesis 1-2, the variable for years of experience is negative in Model 5 (Hyundai) and Model 7 (Huawei) and positive in Model 1 (Samsung), Model 3 (LG), and Model 9 (Hon Hai).

The coefficient of HHI is negative for all of the companies. This result means that successful firms in China and South Korea select engineers who have experience not in a specified technical domain but in a variety of technical domains with Japanese firms. This result supports Hypothesis 1-3.

Consistent with Hypothesis 1-4, the coefficient for degree centrality is statistically positive in Model 3 and statistically negative in Model 7. Inventors who have high degree centrality tend to have experience cooperating with many other inventors in their company. I define them as "star inventors." This result suggests that LG acquires "star inventors" and Huawei does not. In other words, the specific knowledge that an inventor absorbed while employed at a mature enterprise is not necessarily the most important element transferred into emerging nations' companies.

However, consistent with Hypothesis 1-5, the results show that the coefficient for eigenvector centrality is significantly positive in Model 2, Model 4, and Model 10. This result means that Chinese and South Korean companies prefer talent with high eigenvector centrality. Inventors who have a high score on eigenvector centrality are inventors with whom the "star inventors" always want to work. In other words, young, successful firms acquire knowledge workers who connect strongly with the "star inventors." This result supports Hypothesis 1-5.

Table 2 shows the results analyzing the link between the quality of human capital inputs and innovation using a fixed- effect model and a random-effect model. Model 1-4 reports the results for 1990-2011. The null hypothesis can be dismissed based on the Hausman test regarding the selection of the two calculation methods. Therefore, the fixed-effect model is preferable.

Consistent with Hypothesis 2-1, regarding knowledge workers employed by Japanese firms, the coefficient for eigenvector centrality is positive in Model 1 and Model 3 in Table 3. Additionally, the coefficient for degree centrality is positive in Model 1 and Model 3. However, only the coefficient for eigenvector centrality in Model 1 is positive. The results showing a positive coefficient for eigenvector centrality mean that inventors who strongly connect with "star inventors" contribute to innovation in emerging nations' companies. In other words, the inventors who connect strongly with "star inventors" contribute more to young companies' innovation as measured by the number of patents than do "star inventors" who were formerly employed by mature companies. Thus, Hypothesis 2-1 is not supported.

Table 3 shows the results of analyzing the link between the quality of human capital inputs and innovation using a fixed effects model and a random effect model. Model 5-8 reports the results for 2001-2011. The null hypothesis can be dismissed based on the Hausman test regarding the selection of the two calculation methods. As a result, the fixed-effect model is preferable.

Consistent with Hypothesis 2-2, the coefficient for the number of Japanese inventors is statistically positive in Model 5. This result shows the impact on innovation by emerging nations' enterprises as measured by the number of patents won by Japanese researchers employed at these companies from 2001-2011. Similarly, the coefficient for the number of Japanese inventors is statistically positive in Model 1. However, the level of statistical significance is different. The coefficient for the number of Japanese inventors in Model 5 is significant at the 5% level; however, the coefficient for the number of Japanese inventors in Model 1 is significant at the 10% level. These results mean that inventors with experience in a developed country contributed more to quantity of innovation after 2001 than in the period from 1990-2001. This tendency is not similar in quality of innovation (see Model 3 and Model 7). The results suggest that after emerging nations' enterprises strengthened their "absorptive capacity," there was a high possibility that new hires who had formerly been employed in developed countries would contribute to innovation, as measured by the quality of patents. Thus, Hypothesis 2-1 is partially supported.

5. Conclusion

In this study, I identify the influence of engineers' mobility on firms' development and innovation. In particular, this study focuses on knowledge spillover and transnational networks between companies in developing and developed countries from the perspective of inter-organizational labor mobility. I investigate the role of engineers who migrated from Japan to China and South Korea in transmitting knowledge from large, established companies to young companies.

This analysis has yielded important insights into how knowledge workers' characteristics condition their transnational mobility and engagement in international knowledge development. First, this analysis of the characteristics of inventors who migrate from Japanese corporations to China and South Korea indicates that rapid growth companies recruit knowledge workers from Japanese companies based on differences in priority characteristics such as skills, the length of their careers, and their experience in technical fields. Interestingly, top firms, such as Samsung and LG, are more selective in their hiring than other firms in that they overwhelmingly recruit talent from Japan. Successful companies in China and South Korea select engineers who do not have experience cooperating with many other inventors but who do have analytical skills and problem-solving capacity. In other words, successful firms require people with analytical skills and problem-solving capacity, not "star inventors." This result means that for successful companies, non-specific knowledge is more important than specific knowledge. These findings dovetail reasonably well with the findings from Zellner's study (2003) [35].

Second, the knowledge production function is applied to analyze what types of knowledge workers contribute to innovation in emerging nations using panel data. With respect to innovation, inventors who connect strongly with "star inventors" from developed companies play an important role, as do the "star inventors" themselves.

Based on the above results, it can be argued that at the time of hiring, it is most important to select people with the applicable skills, which suggests that corporate growth requires not only "star inventors" but also inventors who connect strongly with the "star inventors."

My study has notable limitations. At this time, only analyses of Chinese and South Korean enterprises were conducted. However, in terms of the relationship between innovation and talent migration, it would be worthwhile to expand the scope of analysis to all emerging nations. Additionally, this research focused on the electronics industry. However, the effect of inventor mobility may differ considerably from industry to industry. Therefore, further research is needed on other industrial sectors.

Nevertheless, the theoretical framework developed and applied herein draws on a broad range of academic literature. The empirical findings are well aligned with the framework regarding the link between the mobility of knowledge workers and knowledge spillover. The approach developed herein will hopefully inspire future empirical studies to analyze the role of migrating knowledge workers.

Conflict of Interest

The author declare no conflict of interest.

Acknowledgment

This research was supported by JSPS KAKENHI, Grant Number JP16K13383. This work was supported by the National Institutes of Science and Technology Policy. In particular, I am grateful to Prof. Toshiya Watanabe (The University of Tokyo) who provided many suggestions and comments on a previous version of this paper.

References

[1] Cohen, Wesley, Levinthal, Daniel, Absorptive Capacity: A New Perspective on Learning and Innovation. Administrative Science Quarterly, Vol. 35 Issue 1, 128-152. 1990.

[2] Cohen, W. M., & Levinthal, D. A.. Innovation and learning: the two faces of R & D. *The economic journal*, *99*(397), 569-596. 1989.

[3] Romer, Paul M., Endogenous Technological Change: The Journal of Political Economy, Vol. 98, No. 5, 71-102. 1990.

[4] Segerstrom, Paul S, Anant, T C A, Dinopoulos, EliasA, 1990. Schumpeterian Model of the Product Life Cycle: The American Economic Review,80, 1077-1091.

[5] Zoltan J. Acs, Pontus Braunerhjelm, David B. Audretsch, Bo Carlsson, 2009. The knowledge spillover theory of entrepreneurship: Small Business Economics, 15-30.

[6] Audretsch, David B., and Erik E. Lehmann. "Does the knowledge spillover theory of entrepreneurship hold for regions?." Research Policy 34.8: 1191-1202.2005.

[7] Audretsch, David, Dirk Dohse, and Annekatrin Niebuhr. "Cultural diversity and entrepreneurship: a regional analysis for Germany." The Annals of Regional Science 45.1: 55-85. 2010.

[8] Michelacci, Claudio. "Low returns in R&D due to the lack of entrepreneurial skills." *The Economic Journal* 113.484: 207-225.2003.

[9] Donald Patton, Martin Kenney, The spatial configuration of the entrepreneurial support network for the semiconductor industry: R&D Management, Vol. 35 Issue 1, 1-16. 2005.

[10] Feldman, M.P. Location and innovation: the new economic geography of innovation, spillovers, and agglomeration: The Oxford Handbook of Economic Geography, 373–394. 2000.

[11] Romer, Paul M., 1990. Endogenous Technological Change: The Journal of Political Economy, Vol. 98, No. 5, 71-102.

[12] Grossman, Gene M., and Elhanan Helpman. Protection for sale. No. w4149. National Bureau of Economic Research, 1992.

[13] Zellner, Christian, and Dirk Fornahl. "Scientific knowledge and implications for its diffusion." *Journal of Knowledge Management* 6.2: 190-198.2002.

[14] Cross, Rob, Parker, Andrew, Prusak, Laurence, and Borgatti, Stephen, Knowing What We Know: Supporting Knowledge Creation and Sharing in Social Networks, Organizational Dynamics, Vol. 30 Issue 2, 100-120. 2001.

[15] Pakes, Ariel and Zvi Griliches, Estimating distributed lags in short panels with an application to the specification of depreciation patterns and capital stock constructs: Review of Economic Studies 51(2), 243-262. 1984.

[16] Audretsch, David B., and Paula E. Stephan. "Company-scientist locational links: The case of biotechnology." The American Economic Review 86.3: 641-652. 1996.

[17] Hausman, Jerry, Bronwyn Hall, and Zvi Griliches, Econometric models for count data with an application to the patents R&D relationship: Econometrica 52, 909-938. 1984.

[18] Hall, Bronwyn, Avi Griliches, and Jerry Hausman. Patents and R&D: Is there a lag?: International Economic Review 27, 265-83. 1986.

[19] Griliches, Zvi,, 1990. Patent statistics as economic indicators: A Survey: Journal of Economic Literature, 28, 1661-1797.

[20] Kortum, Samuel and Josh Lerner, Stronger protection or technological revolution: what is behind the recent surge in patenting?: Carnegie-Rochester Conference Series on Public Policy, 48, 247-304. 1998.

[21] Crépon, B, E. Duguet and Jacques Mairesse, Research and development, innovation and productivity: an econometric analysis at the firm level: Economics of Innovation and New Technology. 1998.

[22] Hall, Bronwyn and Rose Marie Ham, 1999. The patent paradox revisited: determinants of patenting in the U.S. semiconductor industry 1980-94, UC Berkeley and NBER.

[23] Martinelli, Daniel. "Labour market entry and mobility of young French PhDs." *Innovative People–Mobility of skilled personnel in national innovation systems. OECD Proceedings*. 2001.

[24] Mangematin, Vincent. "Individual careers and collective research: is there a paradox?." International Journal of Technology Management 22.7-8: 670-675.2001.

[25] Zucker, L.G., Darby, M., 1997. Individual action and the demand for institution: American Behavioral Scientist 4, 502–513.

[26] Griliches, Zvi and Haim Regev, Productivity and firm turnover in Israeli industry: 1979-1988: Journal of Econometrics 65, 175-203. 1995.

[27] Trajtenberg, Manuel, 1990. A penny for your quotes: patent citations and the value of innovations: The Rand Journal of Economics, 21(1), 172-187.

[28] Jaffe, Adam, Manuel Trajtenberg, and Michael Fogarty. The meaning of patent citations: reports on the NBER/Case-western reserve survey of patentee: NBER Working Paper No. 7631. 2000.

[29] Reitzig, Markus, 2003. What determines patent value? Insights from the semiconductor industry: Research Policy 32, 12-26.

[30] Hall, Bronwyn, Adam Jaffe and Manuel Trajtenberg., Market value and patent citations, RAND Journal of Economics 36, 16-38. 2005.

[31] Harhoff, Dietmar, et al. "Citation frequency and the value of patented inventions." *Review of Economics and statistics* 81.3: 511-515.1999.

[32] Hall, Bronwyn H., Adam B. Jaffe, and Manuel Trajtenberg. The NBER patent citation data file: Lessons, insights and methodological tools. No. w8498. National Bureau of Economic Research, 2001.

[33] Warren Boeker, Executive Migration and Strategic Change- The Effect of Top Manager Movement on Product: Administrative Science Quarterly: Johnson Graduate School of Management, Cornell University, Vol. 42, No. 2, 213-236. 1997.

[34] Zucker, Lynne G., and Michael R. Darby. "Star scientists and institutional transformation: Patterns of invention and innovation in the formation of the biotechnology industry." Proceedings of the National Academy of Sciences 93.23: 12709-12716.1996.

[35] Zellner, C., The Economic Effects of Basic Research: Evidence for Embodied Knowledge Transfer Via Scientists' Migration: Research Policy, 32,1881-1895. 2003.

[36] Borgatti, S. P. Centrality and network flow. Social networks, 27(1), 55-71. 2005.

[37] Amalya L Oliver, Scientific and Technical Human Capital: Research Policy, Volume 33, Issue 4, 583–597. 2004.

[38] Tsai, Wenpin. "Knowledge transfer in intraorganizational networks: Effects of network position and absorptive capacity on business unit innovation and performance." *Academy of management journal* 44.5: 996-1004.2001.

[39] Jaffe, Adam B., and Manuel Trajtenberg. Patents, citations, and innovations: A window on the knowledge economy. MIT press, 2002.

[40] Freeman, Linton C. "Centrality in social networks conceptual clarification." Social networks 1.3: 215-239.1978.

[41] Stefano Breschi, Christian Catalinia, 2010. Tracing the links between science and technology: An exploratory analysis of scientists' and inventors' networks: Research Policy, Volume 39, Issue 1, 14–26.

[42] Marsden, Peter V. "Egocentric and sociocentric measures of network centrality." Social networks 24.4: 407-422.2002

[43] Wehrli, Stefan. "Personality on social network sites: An application of the five factor model." Zurich: ETH Sociology (Working Paper No. 7) .2008.

[44] Scott, J. 2000. Social network analysis, SAGE express.

[45] Ibarra, H. 1992, Homophily and differential returns: Sex differences in network structure and access in an advertising firm: Administrative Science Quarterly, Vol. 37, No. 3,422-447.

[46] Yoon, Byungun, and Yongtae Park. "A text-mining-based patent network: Analytical tool for high-technology trend." The Journal of High Technology Management Research 15.1: 37-50.2004.

[47] Kuznets, Simon. "Inventive activity: problems of definition and measurement." The rate and direction of inventive activity: Economic and social factors. Princeton University Press, 19-52. 1962.

[48] Carpenter, Mark, Francis Narin and Patricia Woolf, Citation rates to technologically important patents: World Patent Information 3, 160-163.1981.

[49] Akira Goto, Kazuyuki Motohashi, Construction of a Japanese Patent Database and a first look at Japanese patenting activities: Research Policy Volume 36 Issue 9, 1431–1442. 2007.

[50] Hall, Bronwyn and Jacques Mairesse, Exploring the relationship between R&D and productivity in French manufacturing firms: Journal of Econometrics 65(1), 263-293. 1995.

[51] Mairesse, Jacques and Bronwyn Hall. Estimating the productivity of research and development in French and United States manufacturing firms: NBER, 285-315.1996.

[52] Harhoff, Dietmar. "R&D and productivity in German manufacturing firms." Economics of Innovation and New Technology 6.1: 29-50.1998.

[53] Ayano Fujiwara, "The Knowledge Spillover Resulting from the Mobility of Knowledge Workers" in 2017 6th International Conference on Industrial Technology and Management (ICITM), United Kingdom, 2017.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1350-1357 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Compact wireless control network protocol with fast path switching

Yasutaka Kawamoto*[1], Makoto Kubomi[2], Yuichi Kado[3]

[1]*Smart Network Technologies R&D Department, Corporate R&D Center, OKI Electric Industry Co., Ltd., 541-0051, Japan*

[2]*Industrial Network Architecture Group Embedded Network Technology Department, Information Technology R&D Center, Mitsubishi Electric Corporation, 247-8501, Japan*

[3]*Department of Electronics, Kyoto Institute of Technology, 606-8585, Japan*

A R T I C L E I N F O

Article history:
Received: 24 May, 2017
Accepted: 20 July, 2017
Online: 01 August, 2017

Keywords:
Sensor network protocol
Communication failure
Wireless control networks

A B S T R A C T

Sensor network protocol stacks require the addition or adjustment of functions based on customer requirements. Sensor network protocols that require low delay and low packet error rate (PER), such as wireless control networks, often adopt time division multiple access (TDMA). However, it is difficult to add or adjust functions in protocol stacks that use TDMA methods. Therefore, to add or adjust functions easily, we propose NES-SOURCE, a compact wireless control network protocol with a fast path-switching function. NES-SOURCE is implemented using carrier sense multiple access/collision avoidance (CSMA/CA) rather than TDMA. Wireless control networks that use TDMA prevent communication failure by duplicating the communication path. If CSMA/CA networks use duplicate paths, collisions occur frequently, and communication will fail. NES-SOURCE switches paths quickly when communication fails, which reduces the effect of communication failures. Since NES-SOURCE is implemented using CSMA/CA rather than TDMA, the implementation scale is less than one-half that of existing network stacks. Furthermore, since NES-SOURCE's code complexity is low, functions can be added or adjusted easily and quickly. Communication failures occur owing to changes in the communication environment and collisions. Experimental results demonstrate that the proposed NES-SOURCE's path-switching function reduces the amount of communication failures when the communication environment changes owing to human movement and others. Furthermore, we clarify the relationships among the probability of a changing communication environment, the collision occurrence rate, and the PER of NES-SOURCE.

1. Introduction

This paper is an extension of a work originally presented at the 2016 TRON Symposium [1].

Recently, sensor network technology has been applied to smart networks and Internet of Things networks, such as wireless control networks (WCN) for factory and plant automation. Such practical applications require a low delay, a low packet error rate (PER), and delay guarantees [2, 3].

In general sensor network protocols, e.g., the ZigBee PRO specification [4], nodes in the network determine the communication path autonomously. Therefore, it is difficult to consider communication delay when designing a system because the communication route is unknown. Therefore, many WCN protocols that require delay guarantees adopt a fixed routing system, i.e., external routing information is the input.

Some WCN protocols employ multiple routes simultaneously during data transmission to eliminate the delay caused by packet loss. For example, WirelessHART [5] increases communication reachability by duplicating all communication paths from each node to the base station. However, duplicating communication paths can cause collisions.

*Corresponding Author: Yasutaka Kawamoto, Smart Network Technologies R&D Department, Oki Electric Industry Co., Ltd. , +81662600700, Japan
Email: kawamoto728@oki.com

Thus, most existing WCN protocols adopt TDMA [5–9] because it is contention-free.

However, TDMA has the following problems.

(1) Handling difficulty

TDMA requires high synchronization accuracy. Generally, it is difficult to implement programs that can maintain synchronization accuracy in less than 10 ms because, in embedded software environments, unexpected delays of several milliseconds occur randomly, such as OS task switches and interrupts.

For example, the flooding time synchronization protocol [10], which is a general time synchronization technology, handles such random delays by including time information in the send frame under the RFLSI driver sequence. Software that crosses network layers is difficult to develop without errors. The WCN stack must operate with various types of hardware wherein processing capacity and how connected sensors are handled vary. In addition, customer requirements must be considered for each application. When using a stack with TDMA, we must adjust the stack each time the hardware changes. In particular, when using sensors that generate various interrupts or require time-consuming operations, it is difficult to maintain accurate TDMA synchronization.

As described above, TDMA systems are difficult to implement, i.e., it is difficult to add or modify functions to such systems.

(2) High procurement cost

Note that rigid hardware specifications must be satisfied to maintain high TDMA synchronization accuracy.

For example, Linear Technology's WirelessHART module maintains high synchronization accuracy by measuring temperature changes and clock error changes during the manufacturing process and writing this information to memory.

However, fabricating such hardware is difficult, and procurement costs are high.

Some WCN studies have employed CSMA/CA protocols [11, 12]. However, CSMA/CA protocols are designed for large networks and are ineffective in typical WCNs (approximately 25 nodes) [3, 5].

In this paper, we propose the NES-SOURCE protocol stack for a WCN of practical size.

NES-SOURCE has compact implementation. Its code size is approximately one-quarter that of a general sensor network protocol stack and approximately one-half that of the TDMA-based WCN protocol stack. The cyclomatic complexity, which is a measure of source code complexity, of NES-SOURCE's functions is at most 10, which is less than other WSN [19] and WCN [20] protocol stacks. This indicates that NES-SOURCE stacks can be customized and adjusted quickly.

NES-SOURCE has a high-speed path-switching function, i.e., paths are switched quickly when communication failures are detected. With this function, NES-SOURCE achieves a low delay and a low PER without duplicating the communication path, which is the case with TDMA-based technology. Compared to a simple fixed-path protocol, NES-SOURCE has the following features that allow routes to be changed quickly.

(1) Reduced back-off time

Sensor data and commands used in a WCN are at most several dozen bytes. By limiting the frame size to approximately 100 bytes, the random back-off time with CSMA/CA can be reduced from a maximum of approximately 15 ms to a maximum of 2.8 ms, as specified in IEEE 802.15.4g.

(2) ACK omission

If the communication path comprises a single hop, when NES-SOURCE detects packet loss in the MAC layer, it resends the frame using a backup route without waiting for the network layer ACK timeout.

With the above features, NES-SOURCE's communication delay can be reduced to approximately 60 ms for a communication path using a normal single-hop route and a backup two-hop route, which has been confirmed experimentally.

Packet loss occurs owing to packet collision or changes in the communication environment. We demonstrate experimentally that the PER and communication delay can be improved by implementing NES-SOURCE's route change function in a changing communication environment, e.g., in a factory environment where people and equipment interfere with wireless communications. Furthermore, we clarify the relationships between the probability of interference from obstacles in a communication environment and the collision occurrence rate and the PER of NES-SOURCE. The NES-SOURCE path-switching method can be considered an optimal method during system design.

2. Problems with Standardized Technologies and Related Work

2.1. Standardized Technologies

Several technologies have been standardized for WCNs.

IEEE 802.15.4 compliant technologies [8] are most frequently used in short-range wireless sensor networks (WSN), e.g., IEEE 802.15.4 is used as the MAC layer protocol in ZigBee IP [13].

IEEE 802.15.4 has two operational modes, i.e., non-beacon and beacon modes. The non-beacon mode is a CSMA/CA-based communication mode, and the beacon mode is a hybrid of TDMA and CSMA/CA. If nodes are in beacon mode, they send packets during a contention-free period (CFP). During the CFP, nodes communicate using TDMA. If the delay request for packets is short, nodes send packets during a contention access period (CAP). Note that nodes communicate using CSMA/CA during the CAP. IEEE 802.15.4's beacon mode is a well-thought-out protocol; however, it has not been implemented broadly because the specification is complicated and difficult to implement.

Time Slotted Channel Hopping (TSCH) is a TDMA-based protocol standardized in IEEE 802.15.4e [15]. TSCH nodes change communication channels each time the nodes change their TDMA timeslot. TSCH network systems are highly noise resistant. Owing to these characteristics, TSCH is used in WirelessHART [5]. TSCH is a suitable MAC layer protocol for WCNs. However, it requires improved clock accuracy, which is expensive. For example, in WirelessHART, clock correction information for

nodes is measured and written during the manufacturing process; however, such processes cannot be performed in all factories.

WirelessHART [5] uses TSCH to enhance communication reliability by incorporating a redundant communication path. However, this protocol control network information resides on the base station side; therefore, the system will fail if nodes cannot communicate with the base station. In addition, these protocols use TDMA, and it is difficult to guarantee delay on the order of 50–100 ms [3].

Sub-GHz bands are suitable for WCNs because they allow communication over greater distances compared to the 2.4-GHz band, which is used in IEEE 802.15.4 [8]. For example, IEEE 802.15.4g [14] is a wireless communication standard that uses a sub-GHz band. The theoretical communication distance of IEEE 802.15.4g is greater than 1 km at 20 mW. This is greater than the IEEE 802.15.4 communication distance (approximately 250 m at best). In addition, compared to 2.4-GHz bands, which are very noisy ISM bands, sub-GHz bands are not noisy.

The maximum frame size of IEEE 802.15.4g is 2048 bytes. Therefore, back-off time during CSMA/CA communication is approximately 15 ms at maximum. This back-off time is too large for a WCN, which requires a communication delay of less than approximately 50–100 ms.

2.2. Related Work

Researchers have improved the reliability of and delays in WCNs. WCN protocols can be classified as TDMA contention-free and CSMA/CA contention-based methods.

WirelessHART [5], PEDAMACS [7], and GinMAC [6] are representative contention-free schemes.

In WirelessHART [5], the base station regularly updates the routing and channel information for the network nodes. Therefore, the system does not function if nodes cannot communicate with the base station [15].

PEDAMACS [7] solves this problem by increasing the base station's transmission power to implement single-hop transmission from the base station. However, single-hop transmission often cannot be implemented in a WCN application environment.

GinMAC [6] avoids this problem by pre-installing node routing information. However, GinMAC uses TDMA and is difficult to implement.

MMSPEED [11] and Dwarf [12] are representative contention-based schemes.

MMSPEED [11] guarantees delay by controlling data stream QoS attributes, and this method is effective for large-scale data streams and networks. The network scale in [11] is approximately 150 nodes, which is greater than a typical WCN (approximately 25 nodes)

Dwarf [12] improves reliability by utilizing unicast-based flooding. This is effective when node density is high in a large-scale network. Therefore, Dwarf is unsuitable for WCNs.

Table 1 Practical WCN requirements

Network size	Up to 25 nodes
Packet occurrence frequency	0.1-1.0Hz (for monitoring) 3×10^{-3}Hz (for FA or PA)
Packet loss rate	10^{-4}order
Delay guarantees	1s (for monitoring) 50-100msec(for FA or PA)
Network topology	Most nodes will be located 1 hop from the station node and at most a distance of 3 hops will be observed.

3. NES-SOURCE Design

3.1. WCN Requirements

According to Zandra et al. [3], control applications can be categorized into three levels depending on the importance of message timing. Wired network communication is reasonable for applications with high-level requirements (e.g., control of life-threatening emergency stop devices). On the other hand, applications with low-level requirements (e.g., applications that safely allow manual intervention and/or monitoring) can use general WSNs.

WCNs can be used for applications with medium-level requirements. Table 1 lists the requirements generally assumed for WCNs. Note that the information given in Table 1 is based on the requirements provided by Samarasinghe et al. [6] and Kumar et al. [2]. The main application of a WCN that satisfies the requirements in Table 1 is assumed to be "a network for automatic control requiring responsiveness on the order of one second from measurement to response." [2, 15]. Note that "automated valve control at the factory" is often given as an example [15].

NES-SOURCE is assumed for use in environments wherein people and equipment move frequently, such as manufacturing plants. When a person or piece of equipment enters the communication path, the communication environment changes, and packet loss can occur. A WCN node cannot use an obstructed path until the obstruction is removed. Therefore, a WCN protocol requires a route-changing function.

As mentioned previously, in the general sensor network protocol, each node changes the communication path autonomously. In contrast, in the general WCN protocol, nodes do not construct routes autonomously. For example, in WirelessHART, nodes use duplicate communication paths simultaneously to transmit packets rather than changing routes.

In WirelessHART and comparable technologies that adopt TDMA, collisions do not occur even if the communication path is duplicated because TDMA is collision-free. However, when using CSMA/CA, collisions tend to occur if communication paths are duplicated because packets are transmitted to the same destination at essentially the same time. In this case, transmission and ACK packet collisions cause contention.

Therefore, when considering a WCN stack using CSMA, we must also consider a method to detect communication failure and switch communication paths quickly.

3.2. NES-SOURCE Features

We designed and developed NES-SOURCE to satisfy the conditions listed in Table 1 [16]. NES-SOURCE has the following features.

- PHY and MAC layers

  The PHY/MAC layer of NES-SOURCE is based on IEEE 802.15.4g. The usable frequency band is the 920-MHz band, the modulation method is GFSK, and the output power is 20 mW.

  Generally, if the single-hop communication distance is long, the number of hops in a multi-hop network decreases, and as a result, communication delay decreases. The communication distance of IEEE 802.15.4g is four times that of IEEE 802.15.4 at 2.4 GHz. Therefore, with IEEE 802.15.4g, it is highly probable that the effective communication delay will be less than that of IEEE 802.15.4. Furthermore, the diffraction characteristic of the 920-MHz band is better than that of the 2.4-GHz band. Therefore, IEEE 802.15.4g is suitable for the application environment assumed by NES-SOURCE (e.g., a factory with complicated piping).

  The CSMA random back-off time of IEEE 802.15.4g is up to approximately 60 ms, which is greater than the maximum random back-off time of IEEE 802.15.4d [17] (11.2 ms) because the maximum frame size of IEEE 802.15.4g is 2048 bytes. This maximum frame size is larger than the maximum frame size of IEEE 802.15.4d (128 bytes). Note that a WCN does not need to communicate large volumes of data; thus, 128 bytes is a suitable maximum frame size.

  Therefore, NES-SOURCE adopts the random back-off time of IEEE 802.15.4d. Our experimental results demonstrate that communication delay is improved by 6.5 ms (an average improvement of 27%) when transmitting 20 bytes by reducing the random back-off time [18].

- Network layer

  Similar to GinMAC, NES-SOURCE routing is a source routing method that uses an input fixed route. With NES-SOURCE routing, it is possible to set multiple paths to a single destination node, and these paths can have predetermined priorities.

  When NES-SOURCE receives transmission requests from the upper layer, it first transmits the packet using the highest priority path. When a packet transmission failure is detected at the network layer, NES-SOURCE retransmits the packet using the next priority path.

  NES-SOURCE uses ACK communication in the MAC and network layers to improve communication reliability. ACK communication in the network layer is required to implement path switching in the network layer. However, when an ACK from the network layer is used in a single-hop route, two transmissions (i.e., network layer data and ACK communications) are performed to confirm the transmission. However, such communication is useless, and reducing the frequency of such useless communication benefits the system.

  NES-SOURCE only uses ACK communications from the MAC layer, i.e., it does not use ACK communications from the network layer when the communication path to the destination node is a single hop. In this case, NES-SOURCE retransmits using the next priority path when the ACK communication of the MAC layer fails. As a result, an average communication delay improvement of approximately 9 ms is confirmed for a single-hop communication path [16].

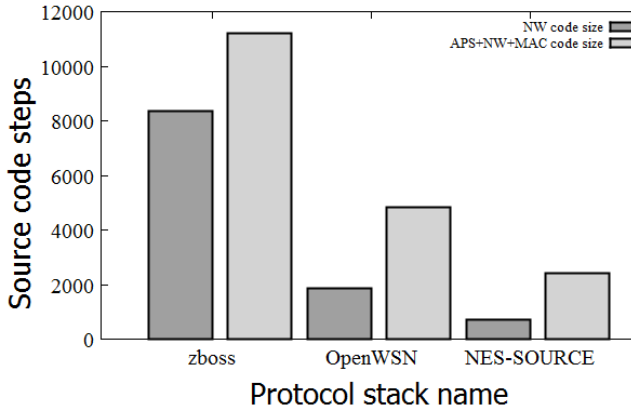


Figure 1 Protocol stack code size comparison

- Software

  The NES-SOURCE source code is more compact compared to the protocol stacks for WSNs and WCNs. A source code size comparison is shown in Figure 1.

  As an implementation of a general sensor network protocol, we refer to zboss [19], which is an open-source implementation of ZigBee PRO. The zboss code size (excluding the PHY driver) is 11217 steps in total (the NW layer code is 8366 steps).

  As an implementation of the control radio protocol stack, we refer to OpenWSN [20], which is an open-source implementation of TSCH. The OpenWSN code size (excluding the PHY driver) is 4833 steps (the NW layer code is 1859 steps).

  For NES-SOURCE, the total code size (excluding the PHY driver) is 2403 steps (the NW layer code is 727 steps). Even if only the number of code steps is compared, NES-SOURCE is a very compact implementation compared to other implementations.

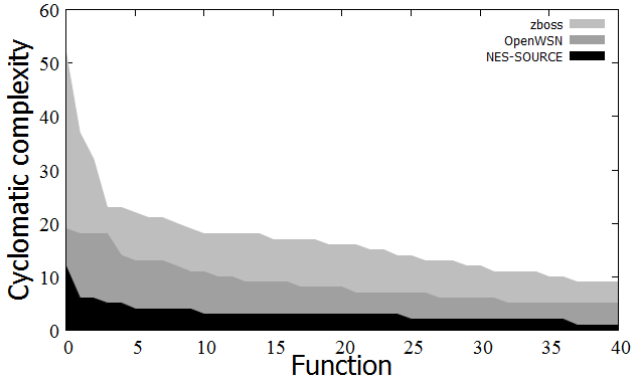


Figure 2 Program complexity

Figure 2 compares the function complexity of each protocol stack. The complexity of a function can be measured by its cyclomatic complexity [21]. Note that adding new code to a function is difficult if the cyclomatic complexity is large. Generally, there is no problem if the cyclomatic complexity is 10 or less. Figure 2 compares the cyclomatic complexity of the top 40 functions in the source code of each protocol stack. For example, zboss has 35 complex functions (the cyclomatic complexity is greater than or equal to 10), and OpenWSN has 10 complex functions. In contrast, NES-SOURCE has only one complex function. Thus, we consider that NES-SOURCE stacks can be customized and adjusted faster than general WSN and WCN protocol stacks.

A size comparison of the highly complex functions in each protocol stack is shown in Figure 3. Generally, fewer steps per function result in better code readability, which makes the code easier to maintain. As can be seen, NES-SOURCE has fewer steps per function than the other stacks; thus, NES-SOURCE is easy to maintain.

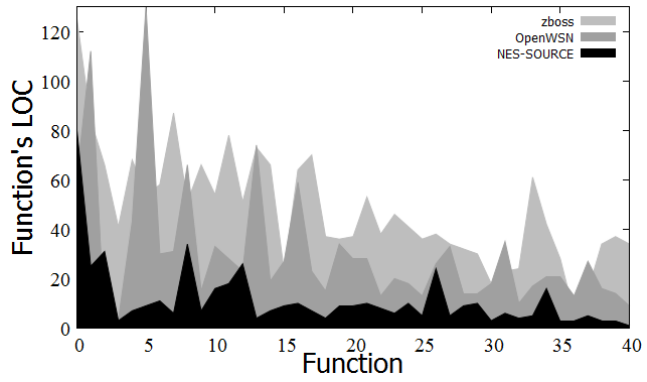


Figure 3 Size comparisons of complex functions

4. Evaluation on Actual Machine

To evaluate and verify NES-SOURCE, we implemented it on an OKI MH 920-Node-232. The MH 920-Node-232 is a sensor node that supports IEEE 802.15.4g and is equipped with a Cortex-M3 processor. Here we adopted FreeRTOS [22] as the embedded OS, and we developed the stack using the C programming language. The other experimental variables are shown in Table 3.

In this experiment, receiving nodes were arranged to form a hexagon, and the transmitting node was installed at the center of the hexagon. The transmission packet was 53 bytes (including overhead), the distance between nodes was 2 m on average, the communication system's transmission frequency was 1 Hz, the number of people in the room was always approximately 10 people, and the number of trials was 10000 for each node (approximately 16 hours).

The experiment evaluated the following packet loss management methods.

- Method 1: Retransmission by the same route

  The transmission node retransmits using the same single-hop route if communication fails. Here, the maximum number of retransmissions is four (Figure 4).

- Method 2: Retransmission by another route
  The transmission node retransmits using a two-hop route if single-hop communication fails. Here, the maximum number of retransmissions is one (Figure 5).

- Method 3: Retransmission by the same and another route

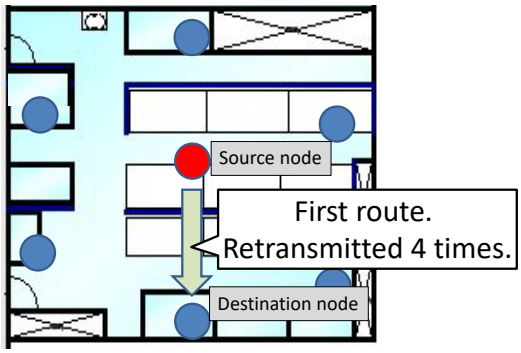


Figure 4 Method 1: first route only

The transmission node retransmits using a two-hop route if single-hop communication fails five times. Here, the maximum number of retransmissions for two-hop communication is one (Figure 6).

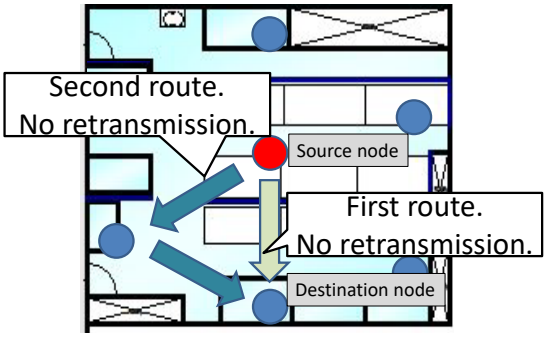


Figure 5 Method 2: two routes without retransition

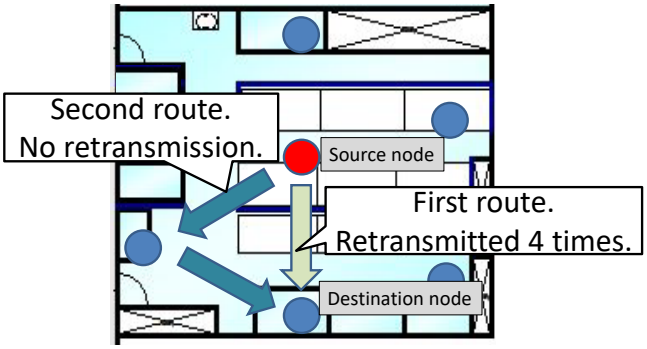


Figure 6 Method 3: two routes with retransition

Figures 7 and 8 show the experimental results. Figure 7 shows the maximum communication delay for each method (i.e., 55.37, 51.84, and 95.93 ms for Methods 1, 2, and 3, respectively). Figure 8 shows the PER for each method (8.71×10^{-3}, 5.83×10^{-4}, and 5.33×10^{-4} for Methods 1, 2, and 3, respectively).

The number of single- and two-hop communication routes for Method 2 was 58138 and 1827, respectively. The number of single- and two-hop communication routes for Method 3 was 59422 and 546, respectively.

5. Discussion

The experimental results indicate that the route change function in the proposed NES-SOURCE stack works effectively.

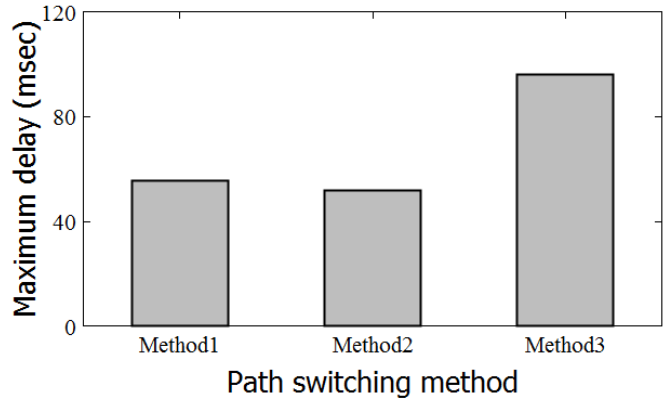


Figure 7 Maximum communication delay

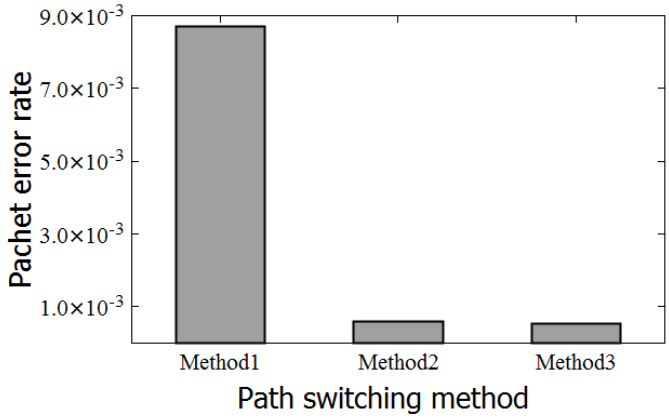


Figure 8 Packet error rate

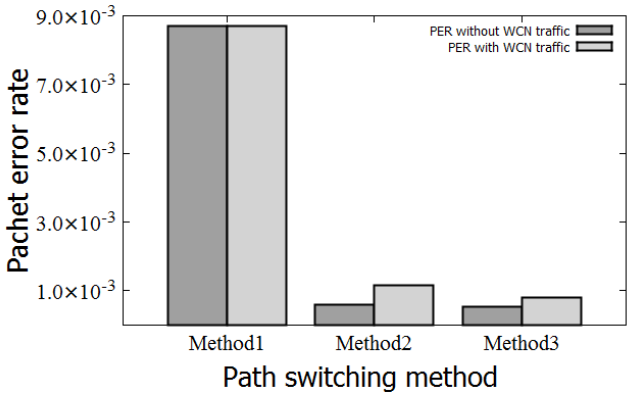


Figure 9 Packet error rate with/without WCN traffic

Table 2 Simulation parameter values

aUnitBackoffPeriod [8]	200 usec
aCCATime [8]	100 usec
macMinBE [8]	3
macMaxBE [8]	4
macMaxCSMABackoffs [8]	5
The number of peripheral nodes	25
Frame send period	3.5 msec
Frame send interval	1 sec
Measurement time	30000 sec × 50

Table 3 Variable for calculating the PER

Number of successful communications method 1 by 1 hop	N_{m1_1hop}
Number of failed communications method 1	N_{m1_failed}
Number of successful communications method 2 by 1 hop	N_{m2_1hop}
Number of successful communications method 2 by 2 hops	N_{m2_2hop}
Number of failed communications method 2	N_{m2_failed}
Number of successful communications method 3 by 1 hop	N_{m3_1hop}
Number of successful communications method 3 by 2 hops	N_{m3_2hop}
Number of successful communications method 3 by 2 hops	N_{m3_failed}
Number of trials	N_{trial}
PER of 1 hop without retransmission under the WCN traffic	$PER_{1hop\_noretry\_c}$
PER of 2 hops without retransmission under the WCN traffic	PER_{2hop_c}
PER of 1 hop with 4 retransmissions under the WCN traffic	$PER_{1hop\_4retry\_nc}$

In the experimental environment, Method 2 was more advantageous relative to PER (Figure 5) than Method 1 as a measure against packet loss.

Furthermore, Method 2 was clearly superior to Method 3 relative to maximum delay (Figure 4). However, when compared relative to PER (Figure 5), the difference between Methods 2 and 3 is small.

These results demonstrate that the effect of link retransmission on packet loss was low in our experimental environment.

Thus, in a changing communication environment (e.g., a person enters the communication path), changing the path is more

effective relative to PER and maximum delay compared to link retransmission.

As stated previously, changes in a communication environment cause packet loss and collision. Note that our experiment did not consider packet collisions. Thus, to calculate a more realistic PER, the collision occurrence rate in WCN traffic was calculated in a simulation.

The simulation was created based on the communication frequency assumed by a WCN (Table 1) and the communication device settings given in Section 4. The simulation parameters are shown in Table 2.

The simulation results show that, for 25 nodes in the vicinity and a packet transmission frequency of 1 Hz for each node (i.e., a WCN general data transmission environment; Table 1), the PER of one-to-one communication without retransmission ($PER_{1\text{hop\_noretry\_c}}$) was 0.8%. In addition, if four retransmissions were allowed ($PER_{1\text{hop\_4retry\_c}}$), the PER was 0%, and the PER of two-hop communication without retransmission ($PER_{2\text{hop\_noretry\_c}}$) was 1.59%.

On the basis of the simulation and experimental results given in Section 4 and Table 3, respectively, we can calculate the PER of NES-SOURCE in WCN traffic. The relationships among the values shown in Table 3 are expressed as follows.

$$N_{trial} = N_{m1\_1hop} + N_{m1\_failed} \quad (1)$$

$$N_{trial} = N_{m2\_1hop} + N_{m2\_2hop} + N_{m2\_failed} \quad (2)$$

$$N_{trial} = N_{m3\_1hop} + N_{m3\_2hop} + N_{m1\_failed}\,. \quad (3)$$

The PER of each method in a collision-free environment ($\text{PER}_{\text{m1\_nc}}$, $\text{PER}_{\text{m2\_nc}}$, $\text{PER}_{\text{m3\_nc}}$) can be expressed as follows.

$$PER_{m1\_nc} = N_{m1\_failed} / N_{trial} \quad (4)$$

$$PER_{m2\_nc} = N_{m2\_failed} / N_{trial} \quad (5)$$

$$PER_{m3\_nc} = N_{m3\_failed} / N_{trial}\,. \quad (6)$$

On the basis of Equations 4–6, the PER of NES-SOURCE under WCN traffic conditions can be expressed as follows.

$$PER_{m1\_nc} = \frac{N_{m1\_failed} + N_{m1\_1hop} \times PER_{1hop\_4retry\_c}}{N_{trial}} \quad (7)$$

$$PER_{m2\_nc} = \frac{N_{m2\_failed} + (N_{m2\_1hop} \times PER_{1hop\_4retry\_c} + N_{m2\_2hop} \times PER_{2hop\_noretry\_c})}{N_{trial}} \quad (8)$$

$$PER_{m3\_nc} = \frac{N_{m3\_failed} + (N_{m3\_1hop} \times PER_{1hop\_4retry\_c} + N_{m3\_2hop} \times PER_{2hop\_noretry\_c})}{N_{trial}} \quad (9)$$

On the basis of the measurement results (Section 4), simulation results, and Equations 7–9, the PER values of NES-SOURCE under WCN traffic conditions for Methods 1 to 3 were 8.71×10^{-3}, 1.16×10^{-3}, and 8.05×10^{-4}, respectively (Figure 9).

According to Zandra et al. [3], the PER request of a WCN is on the order of 10^{-4}, and the delay request is on the order of 50–100 ms. Therefore, NES-SOURCE can be used in an experimental environment if the given WCN application has a slightly lower PER requirement.

When we construct WCN systems, we cannot predetermine traffic conditions or the frequency of communication problems. To mitigate this uncertainty, NES-SOURCE can flexibly change the packet loss countermeasure depending on the application's requests and the environment. For example, if the network traffic is assumed to be high, Method 1 is reasonable. On the other hand, when the network is constructed in an environment where people frequently enter the communication path, Method 2 is reasonable. Method 3 is appropriate if the goal is to increase the PER, even at the expense of the maximum delay.

"When a person enters the communication path, packet loss occurs."

"Packet loss due to a person can always be avoided by changing the communication path."

Assuming the above points are correct, the relationship between the probability that a person is within the communication path (P_{ps}) and the PER of each method can be expressed as follows.

$$PER_{Method1} = (1 - P_{ps}) \times PER_{1hop\_4retry\_c} \quad (10)$$

$$PER_{Method2} = (1 - P_{ps}) \times PER_{1hop\_noretry\_c} + P_{ps} \times PER_{2hop\_noretry\_c} \quad (11)$$

$$PER_{Method3} = (1 - P_{ps}) \times PER_{1hop\_4retry\_c} + P_{ps} \times PER_{2hop\_noretry\_c}\,. \quad (12)$$

For example, Figure 10 shows the relationship between P_{ps} and PER when $PER_{Method1} = 0.8 \times 10^{-3}$, $PER_{Method2} = 1.59 \times 10^{-3}$, and $PER_{Method3} = 0$ (this is the same as the simulation results). In this case, when a person's appearance rate in the communication path exceeds 0.8%, it is better to adopt Method 2 relative to PER.

6. Conclusions

In this paper, we have proposed, implemented, and clarified the effectiveness of NES-SOURCE as a WCN protocol that adopts CSMA/CA, which is easy to handle.

WSN protocol stacks must be customized and adjusted for each application, and the same is true for a WCN protocol stack, which is a WSN protocol stack with strict delay and PER requirements.

Many WCN stacks are implemented using TDMA. However, TDMA is difficult to handle; thus, customization and adjustment are difficult.

Therefore, we have developed the compact NES-SOURCE WCN protocol stack. The proposed NES-SOURCE employs the CSMA/CA method rather than the TDMA method. As a result, NES-SOURCE's code size is approximately 2700 steps, which is less than one-half the size of the TDMA-based WCN stack. In addition, NES-SOURCE has a lower code complexity than conventional WCN stacks and can be customized and adjusted quickly.

The WCN protocol with TDMA ensures communication reliability by duplicating the communication path. In contrast, the proposed NES-SOURCE ensures communication reliability by fast path switching when communication fails.

We conducted experiments in an environment where people entered and exited the communication path, and we confirmed that NES-SOURCE's fast path-switching function works effectively. The experimental results demonstrate that NES-SOURCE's communication failure detection and path-switching mechanism are fast and that communication delay was within general WCN requirements.

We have also clarified the relationships between the probability of a person entering the communication path and the PER due to the occurrence of collisions relative to each of the PER of the path-switching methods that the proposed NES-SOURCE employs. Thus, when designing systems that use NES-SOURCE, it will be possible to predetermine the optimum path-switching method relative to communication frequency and the varying degrees of obstructions in the physical environment.

Conventionally, TDMA must be adopted in WCN systems; however, this is difficult to customize and incurs high costs. In contrast, according to our experimental and simulation results, we can construct less expensive CSMA/CA-based WCN systems that can be customized quickly.

Acknowledgment

The names of products or services are trademarks or registered trademarks of their respective companies.

References

[1] Y. Kawamoto, Y. Kado, "NES-SOURCE: Indoor small-scale wireless control network protocol that has a communication failure point avoidance function" in TRON Symposium (TRONSHOW), Tokyo Japan, 2016.

[2] Ovsthus. K, Kristensen. L. M, "An industrial perspective on wireless sensor networks—a survey of requirements, protocols, and challenges" IEEE Communications Surveys & Tutorials, 16(3), 1391-1412, 2014

[3] Zand, P., Chatterjea, S., Das, K., & Havinga, P., "Wireless industrial monitoring and control networks: The journey so far and the road ahead." Journal of sensor and actuator networks 1.2, 123-152, 2012.

[4] ZigBee Alliance, ZigBee PRO specification, Standard, 2007.

[5] J. Song, S. Han, A. Mok, D. Chen, M. Lucas, M. Nixon, W. Pratt, "WirelessHART: Applying wireless technology in real-time industrial process control" in Real-Time and Embedded Technology and Applications Symposium, IEEE RTAS'08, 377-386, 2008

[6] K. Samarasinghe, T. Voigt, L. Mottola, U. Roedig, "Network coding with limited overhearing" in 8th European Conference on Wireless Sensor Networks, Bonn Germany, 2011

[7] Ergen, S. C., Varaiya, P., "PEDAMACS: Power efficient and delay aware medium access protocol for sensor networks", IEEE Transactions on Mobile Computing, 5(7), 920-930, 2006

[8] IEEE Computer Society, "IEEE Standard for Local and metropolitan area networks—Part 15.4: Low-Rate Wireless Personal AreaNetworks (LR-WPANs) ", IEEE Std, 2011

[9] IEEE Computer Society, " IEEE Standard for Local and metropolitan area networks—Part 15.4: Low-Rate Wireless Personal Area Networks (LR-WPANs) - Amendment 1: MAC sub-layer" IEEE Std, 2012

[10] M. Maróti, B. Kusy, G. Simon, A. Lédeczi, "The flooding time synchronization protocol" in Proceedings of the 2nd international conference on Embedded networked sensor systems, Baltimore Maryland, 2004

[11] E. Felemban, C. Chang-Gun, Eylem. EKICI, "MMSPEED: multipath Multi-SPEED protocol for QoS guarantee of reliability and Timeliness in wireless sensor networks", IEEE transactions on mobile computing, 5(6), 738-754,2006

[12] M. Strasser, A. Meier, K. Langendoen, P. Blum, "Dwarf: Delay-aware robust forwarding for energy-constrained wireless sensor networks" in International Conference on Distributed Computing in Sensor Systems, Springer Berlin Heidelberg, 2007

[13] ZigBee Alliance, ZigBee IP specification. ZigBee Standard,2011

[14] IEEE Computer Society, "IEEE Standard for Local and metropolitan area networks—Part 15.4: Low-Rate Wireless Personal Area Networks (LR-WPANs) Amendment 3: Physical Layer (PHY) Specifications for Low-Data-Rate, Wireless, Smart Metering Utility Networks", IEEE Std, 2012

[15] T. O'Donovan, J. Brown, U. Roedig, C.J. Sreenan, A. Dunkels, A. Klein, L. Wolf, "Ginseng: Performance control in wireless sensor networks", in Sensor Mesh and Ad Hoc Communications and Networks (SECON), Boston UnitedStates,2010

[16] M. Kubomi, Y. Kado, Y. Kawamoto, "Improving reliability of 920 MHz-band wireless networks using advanced source routing function to guarantee real-time property.", in Internet of Things and Applications (IOTA), Pune India, 2016.

[17] IEEE Computer Society, "IEEE Standard for Local and metropolitan area networks—Part 15.4: Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for Low-Rate Wireless Personal Area Networks (WPANs) Amendment 3: Alternative Physical Layer Extension to support the Japanese 950 MHz bands", IEEE Std, 2009

[18] Y. Kado, A. Kobayashi, M. Kubomi, Y. Nogami, Y. Kawamoto, "Application of Economical 920-MHz Band Wireless Communication to Power Routing in HVDC Networks", in Proceedings of 2014 IEEE International Conference on Smart Grid Communications, Venice Italy, 2014

[19] A. Cunha, A. Koubaa, R. Severino, M. Alves, "Open-ZB: an open-source implementation of the IEEE 802.15. 4/ZigBee protocol stack on TinyOS", in Mobile Adhoc and Sensor Systems, Pisa Italy, 2007

[20] T. Watteyne, X. Vilajosana, B. Kerkez, F. Chraim, K. Weekly, Q. Wang, K. Pister, "OpenWSN: a standards-based low-power wireless development environment.", Transactions on Emerging Telecommunications Technologies, 23(5), 480-493, 2012

[21] T.J. McCabe, C. W. Butler, "Design complexity measurement and testing.", Communications of the ACM, 32(12), 1415-1425, 1989

[22] R. Barry, FreeRTOS. Internet, 2008.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1358-1362 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

A Meta-analysis of Content and Language Integrated Learning in English

Jong-Keol Kim, Jeong-ryeol Kim*

Korea National University of Education, Chung-Buk, Korea

ARTICLE INFO

Article history:
Received: 30 May, 2017
Accepted: 20 July, 2017
Online: 05 August, 2017

Keywords:
Meta-analysis
Content-based
Integrated learning

ABSTRACT

The purpose of this study is to investigate the effects of integrated content and English education. To fulfill this, 30 studies conducted in elementary school setting matching the meta-analysis criteria, such as being quantitative, experimental, or having affective statistical results after searching for the key words of "content-based", "theme-based" and "integrated learning" on the accessible databases such as RISS and Google scholar were selected. The results of meta-analysis are as follows: 1) Content-based integrated English education is more effective in improving English skills than improving affective factors. 2) There is no discernable difference in effect sizes among different grades. 3) Integrating with knowledge-based subjects were more effective than integrating with all the subjects.

1. Introduction

This paper is an extension of work [1] originally presented at the 2017 International Conference on Platform Technology and Service (PlatCon).

The main goal of learning a target language is to facilitate students' ability to communicate well. In countries that use English as a second language, the goal is to find effective English language learning in a limited period of time. While Korea is known as a country with a high level of education, many students in Korea who get high scores in reading and grammar lack the ability for authentic communication. Therefore, what could be a solution to solve this problem? Integrated English learning may be a solution.

Integrated English learning is an effort to learn English in an English class mixed with other subjects. Learning English at the same time as learning a subject saves time and brings meaningful language learning to learners.

Studies collected and analyzed had some weaknesses. For example, it was hard to identify whether the studies were integrated by the contents or the themes, which is a collection of common points of contents; as such, it can be said that theme-based integration is content-based integration. Furthermore, some of the studies analyzed were difficult to classify as theme-based integration. Therefore, in order to complement weaknesses, the title was changed to "A meta-analysis of content and language integrated learning in English."

Content-based integrated English education (CBIEE) refers to a method of teaching a foreign language whose main theme is extracted from a regular curriculum or curriculum area. Studies about CBIEE were conducted a few times [2, 3, 4] but analyzation of what part was effective for learning language was not conducted. Thus, it is needed to collect and analyze studies about CBIEE and propose directions to investigate via research on CBIEE that have been combined and sorted, and where the effects of CBIEE are found through meta-analysis.

2. Background

2.1. Content-based Integrated English education

The biggest positive effect of content learning through language instruction or language is that they can learn concepts and language in a natural environment; thus, learners can acquire language at the same time as general knowledge. For example, if they learn about other country's environment in class, they will be able to learn about the environment and its related language naturally while communicating through it. The learners who have this meaningful communication are instinctively stimulated to acquire a high level of language. It is, therefore, highly effective for learners' language learning to learn language and learn content knowledge naturally while participating in meaningful, purposeful life activities. The language learned through content has merits not only in terms of words, but also in integrating with higher dimensional thinking because learners' thoughts are reflected in communication [5].

2.2. Meta-analysis

Meta-analysis is an analytical method that synthesizes variouss

*Corresponding Author: Jeong-ryeol Kim, Korea National University of Education Chung-Buk, Korea | Email: jrkim@knue.ac.kr

www.astesj.com
https://dx.doi.org/10.25046/aj0203171

articles and derives them as a single result. It draws a standard on papers of similar themes and derives the overall result. In short, it is a way to reanalyze the results so far to derive general results. In order to achieve the main goal of this paper, CBIEE related papers needed to be effectively analyzed and integrated through meta-analysis.

There are five steps for meta-study before analyzing statistics: 1) Problem specification, 2) Study identification, retrieval and eligibility criteria, 3) Study features coding book, 4) Coding, and 5) Analysis and interpretation [6].

3. Method

3.1. Subject of study

1) Primary target classification for meta-analysis

In this research studies related to CBIEE domestic and overseas papers about foreign language education were collected. Key words or phrases for searching studies were content-based, integrated learning, or theme-based. Among the research, the ones that matched the condition were studies that published between the years 2000 to 2015, with 15 were domestic papers, 7 were from overseas papers, and 14 were domestic master's theses and doctoral dissertations. These 36 research papers were where meta-analysis was gleaned.

The number of studies for finding effect sizes in meta-analysis was 51 and the number of effect size for detail factors was 110 each, excluding cases that overlapped. Among these studies, studies that were directly related to searching the effect size for improving English skills and affective factors were chosen. The number of studies that suits the condition was 35 and the number of effect size is 110, excluding the cases which overlapped.

Table 1. Number of cases and number of effect sizes on detail case

Domain	Factor (Sub-domains)	Number of case	Detail case	Number of effect size
Cognitive	English skills	26	Listening	18
			Speaking	15
			Reading	14
			Writing	10
			Vocabulary	2
Affective	affective factors	25	Confidence	16
			Interest	22
			Attitude	13
Total (excluding overlapped)		35		110

2) Secondary target for meta-analysis

After classifying primary targets for meta-analysis, a Q-value test for homogeneity was taken and a random-effects model was chosen for meta-analysis. To raise the reliability and conduct a more valid meta-analysis, studies that were conducted only for elementary school students were only included, with studies that had different variables or a small number of cases were removed. Also, the studies that showed relatively large effect sizes without persuasive reasoning were dismissed. In the end, 13 research formats from domestic papers, 5 from overseas papers, 12 from master's theses for a total of 30 studies with 107 effect sizes were chosen as secondary targets for meta-analysis. Most of the research tried to find the effect of improving language skills and affective factors through CBIEE and the results show that it is both effective in improving language skills and affective factors.

3.2. Research design

1) Setting up variables of analysis

To conduct a meta-analysis related to CBIEE results, variables of analysis were explored and divided into two categories of variables: moderator variable and dependent variable. Moderator variables are integrated subjects, method of learning, and grade.

Table 2. Variables and coding value Conflict of Interest

Variable	Coding value
1. Integrated subject	(1) Knowledge-based subject (2) Skill-based subject (3) Both
2. Learning method	(1) Lecture (2) Group work (3) Storytelling (4) Game (5) Singing
3. Grade	(1) 4th (2) 5th (3) 6th
4. Dependent variable	(1) Language skills (2) Affective factors

Dependent variables are divided into language skills and affective factors. A meta-analysis to investigate the effects of the dependent variables on the detailed variables was conducted. The effects of language skills and affective factors were measured from the results of post-tests of experimental groups and control groups.

The term confidence, which is a detail factor in affective factors, included anxiety in learning foreign language and confidence in learning foreign language as the same factor, and the term attitude included attitude in foreign language lessons and attitude for learning a foreign language as the same factor. Similarly, interest included interest in foreign language and Interest in subject.

Table 3. Detail variable in dependent variable

Dependent variable	Detail variable
1. Language skills	(1) Listening (2) Speaking (3) Reading (4) Writing
2. Affective factors	(1) Confidence (2) Attitude (3) Interest

2) Instrument for meta-analysis

With the criteria of variables and detail variables mentioned earlier, coding is produced for meta-analysis in Excel. The coding of each quantitative data and variable is inserted and based on this CMA program is used to obtain the result from meta-analysis.

3) Data analysis and interpretation

Research for the meta-analysis of CBIEE was searched and

collected for this study. From 2000 to 2015 a total of 30 research papers form annals and graduate-level theses were chosen. By the criteria, moderator and dependent variables are organized, coded, and then analysis performed.

The number of students participating in each study and the experiment method were different. In order to compensate for the errors in each of the different methods weight (*w*) was utilized. For each effect size, an average effect size for moderator was calculated. *Thus, Hedge's g* was used to fix the standard difference in means (*Cohen's d*). After the analysis, the results with some implications were interpreted.

4. Analysis

4.1. Coding for variables and adjusting weights

Based on the coding criteria, the collected studies were divided into certain categories, such as grade, learning method, and the classification of integrated subject. Thirty studies were represented as experimental and control groups, and individual effect sizes were presented. Each effect size was measured for detailed factors of language skill and affective factors.

Table 4. Coding for dependent variables and effect sizes

Study	*k*	*Experimental group*			*Controlled Group*			*ES*
		N	*m*	*sd*	*N*	*m*	*sd*	
K** (2012)	Listening	20	89.90	9.59	20	88.80	9.12	0.12
	Speaking	20	76.80	7.92	20	75.65	7.93	0.14
	Reading	20	49.70	6.81	20	48.40	6.31	0.19
	Writing	20	40.30	9.63	20	38.50	8.08	0.20
	Interest	20	78.70	9.14	20	77.65	8.37	0.12
	Confidence	20	59.60	5.50	20	57.80	5.39	0.32
L** (2015	Writing	11	29.55	7.54	11	19.18	10.30	1.11
K** (2015)	Listening	32	80.63	11.90	32	76.56	14.28	0.31
	Speaking	32	7.84	2.33	32	7.21	2.23	0.27
	Interest	32	3.41	0.24	32	3.25	0.11	0.84
	Confidence	32	3.28	0.09	32	3.17	0.05	1.55
K** (2014)	Reading	27	67.41	11.04	25	56.20	15.91	0.81
	Writing	27	20.56	5.60	25	17.00	6.92	0.56
	Interest	27	3.85	0.74	25	3.80	0.71	0.07
	Attitude	27	4.19	0.88	25	3.70	0.80	0.57
J** (2003)	Listening	39	3.46	0.68	39	3.43	0.86	0.04
	Speaking	39	3.77	0.96	39	3.38	1.21	0.35
	Interest	39	3.740	1.07	39	3.34	0.58	0.41
	Attitude	39	3.72	0.62	39	3.22	0.47	0.88
C** (2013)	Speaking	27	37.00	3.49	27	35.81	3.45	0.34
	Interest	27	4.49	0.24	27	4.38	0.18	0.47

(N=number of students, ES = effect size)

To compensate any error that was made from the different number of students and quality of study, weight (*w*) was added to offset error. As such, each effect size was measured again with weight. The effect sizes of each study with adjusted weight are shown in Table 5.

Table 5. Effect sizes with weight (*w*)

Study	*k*	Experimental Group	Controlled Group	*ES*	*w*
		N	*N*		
K** (2012)	Listening	20	20	0.12	0.80
	Speaking	20	20	0.14	0.80
	Reading	20	20	0.19	0.80
	Writing	20	20	0.20	0.80
	Interest	20	20	0.12	1.12
	Confidence	20	20	0.32	1.10
L** (2015)	Writing	11	11	1.11	0.39
L** (2015)	Listening	32	32	0.31	1.25
	Speaking	32	32	0.27	1.62
	Interest	32	32	0.84	1.35
	Confidence	32	32	1.55	1.44
K** (2014)	Reading	27	25	0.81	0.95
	Writing	27	25	0.56	1.00
	Interest	27	25	0.07	1.38
	Attitude	27	25	0.57	2.09
C** (2003)	Listening	39	39	0.04	1.54
	Speaking	39	39	0.35	1.51
	Interest	39	39	0.41	1.95
	Attitude	39	39	0.88	1.45
C** (2013)	Speaking	27	27	0.34	1.06
	Interest	27	27	0.47	1.53

4.2. Comparing effect of factors in CBIEE

1) Overall CBIEE effect on English skills versus affective factors

Overall average CBIEE effect sizes on language skills and affective factors are 0.68 and 0.48, respectively where the difference is statistically significant. CBIEE has larger effect size on language skills than affective factors which means CBIEE has more effects on improving listening, speaking, reading, and writing than improving students' interests, confidence, and attitude in learning a foreign language.

Table 6. Overall CBIEE effects on language skills versus affective factors

	n	k	*Q*	*p*	*ES*	*Z*	95 percent CIs	
							Lower	Upper
Cog	24	57	187.67	0.05	0.68	24.33	0.62	0.73
Aff	24	50	227.67	0.05	0.48	14.56	0.41	0.54

(Cog = Cognitive skills, Aff = Affective factors, *n* = number of studies, *k* = number of detail case, *Q* = *Q-value*, *p* = *p-value*, *Z* = *Z-value*)

2) CBIEE effects on different grades

Table 7. CBIEE effects on different grades

G	Var	*n*	*k*	*Q*	*p*	*ES*	*Z*	95 percent CIs	
								Lower	Upper
4th	Cog	8	26	124.43	0.05	0.80	22.59	0.73	0.80
	Aff	9	20	87.40	0.05	0.40	8.10	0.31	0.50
5th	Cog	6	16	11.90	0.05	0.38	22.59	0.26	0.51
	Aff	5	10	9.22	0.05	0.53	6.57	0.37	0.68
6th	Cog	6	12	7.96	0.05	0.67	9.27	0.53	0.81
	Aff	7	15	121.46	0.05	0.52	8.90	0.40	0.63

(G = Grades)

Students in 4th grade had an average effect size on language skills (0.80) which were larger than that on affective factors (0.40), but the difference was not statistically significant. Conversely, 5th and 6th grade students (0.53, 0.52) had larger effect size on affective factors compared to 4th grade learners (0.40). As the grade progressed, the gap between minimum and maximum effect sizes grew, which means the effect of CBIEE matters on the choice of content and a teacher's method of teaching.

3) CBIEE effects on language skills

Table 8. CBIEE effects on language skills

Skills	*n*	*k*	*Q*	*p*	*ES*	*Z*	95 percent CIs	
							Lower	Upper
Listening	18	18	63.73	0.05	0.66	12.86	0.56	0.75
Speaking	15	15	55.75	0.05	0.65	11.82	0.54	0.75
Reading	14	14	43.50	0.05	0.70	12.23	0.59	0.81
Writing	10	10	20.03	0.05	0.76	11.91	0.64	0.89

There was no difference in effect sizes among language skills. CBIEE's effect on students' reading and writing skills was minutely larger than that of listening and speaking skills. Since most activities take place within similar content or themes, they focus on output activities such as reading and writing.

This results show that effect sizes in improving written language is slightly larger than spoken language, although the difference was not statistically significant.

4) CBIEE effects on affective factors

Table 9. CBIEE effects on affective factors

	n	*k*	*Q*	*p*	*ES*	*Z*	95 percent CIs	
							Lower	Upper
Confidence	16	16	91.33	0.05	0.44	7.84	0.33	0.55
Attitude	12	12	60.84	0.05	0.54	8.06	0.41	0.67
Interest	22	22	74.21	0.05	0.47	9.32	0.37	0.57

There was no difference in effect sizes among affective factors. Improving attitude was the largest effect (0.54) where CBIEE helped to give a positive influence on attitude by removing cognitive burden while learning a foreign language.

By analyzing the statistics, CBIEE proved to have positive effects on affective factor development; especially, CBIEE helped students to raise their concentration better by eliminating cognitive pressure. When integrating different subjects, teachers should know that integrating can raise students' affective factors when students face difficult tasks.

5) Knowledge-based subjects vs skill-based subjects

Table 10. Knowledge-based versus skill-based integration

Subjects	*n*	*k*	*Q*	*p*	*ES*	*Z*	95 percent CIs	
							Lower	Upper
Skill-based	13	45	184.79	0.000	0.67	10.44	0.55	0.80
Knowledge-based	9	30	112.31	0.000	0.52	6.01	0.35	0.68
All subjects	8	38	99.94	0.000	0.41	6.32	0.29	0.54

Integrating with knowledge-based subjects (0.67) shows larger effect than that of skill-based subjects (0.52), but the difference was not statistically significant. However, integrating with knowledge-based subjects (0.67) is more effective than integrating all subjects (0.41), which the difference is statistically significant. Integrating with only knowledge-based subjects (i.e., mathematics, science, and social studies) is the most effective way for content integration.

6) Teaching method in CBIEE

Table 11. Effect sizes by different teaching methods in CBIEE

Method	*n*	*k*	*Q*	*p*	*ES*	*Z*	95 percent CIs	
							Lower	Upper
Lecture	11	45	292.96	0.000	0.48	8.22	0.38	0.78
Group work	5	15	3.37	0.001	0.49	3.37	0.20	0.78
Game	4	16	47.42	0.000	0.69	5.86	0.46	0.92
Singing	2	4	0.99	0.002	0.34	3.10	0.13	0.56
Storytelling	8	30	39.76	0.000	0.42	7.91	0.31	0.52

Using games (0.69) shows the largest effect compared to the rest of the learning methods, although the difference was not statistically significant. Students can enjoy their learning while playing games and they are able to release the stress of learning foreign language. It is, however, challenging to see the effect of using singing in CBIEE, because there are only few cases.

7) Overall results of meta-analysis

CBIEE shows larger effects on improving language skills than improving affective factors. We think that most of the experiments are focused on English courses, not regular courses, and these results seem to have resulted. There are no effect size differences among different methods. Integrating specifically with knowledge-based subjects is more effective than integrating with all subjects in improving language skills, which means in order to develop the English language skills, it is important to emphasize the importance of the mixture of subjects rather than the way of teaching.

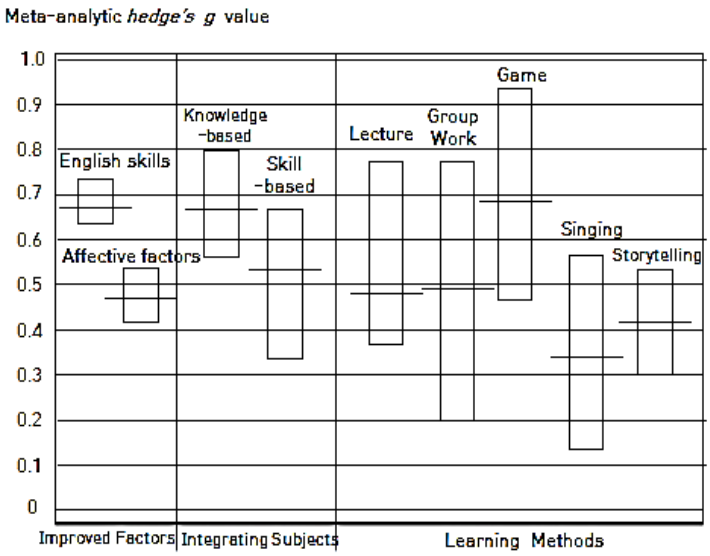


Figure 1 Overall/Meta-analytic effects of CBIEE across sub-domains.

5. Conclusions

It was explored in this study the effect of CBIEE by meta-analysis. Through the meta-analysis, it was found that CBIEE is more effective in improving skills than affective factors and it depicts different effect size across sub-domains.

Compared to a previous meta-study of integrated language learning [7], different results were found in this study. After integrated learning, two studies have different results on the improved factors, such as language skills versus affective factors, and spoken language versus written language. Two studies also have opposite effects on integrating knowledge-based subjects versus skill-based subjects.

Due to this it means that integrating with other courses can have different effects depending on whether the teacher focuses on content or language when the language is taught. This meta-analysis displays not only the findings of the effects of CBIEE itself, but also it may be gleaned that different kind of integration with language, for example, task-based or experience-based integration could give different effects to language learners.

In future studies, it will be necessary for the researchers and teachers to pay much attention and effort to examine the effects of applying various types of integrated foreign language classes and to find appropriate education methods for each country's educational environment.

References

[1] Jong-Keol Kim & Jeong-ryeol Kim, "A Meta-analysis of Theme-based Integrated English Education Effects on English Skills," in proceedings of the 2017 International Conference on Platform Technology and Service (PlatCon), Busan, Korea, Feb. 2017.

[2] Jeong-ryeol Kim & Yeo-rim Ko, "Effects of content-based language teaching in elementary English classes," Journal of the Korea English Education Society, vol. 7, no. 2, pp.131-152, 2008.

[3] Tae-Hyeok Moon, A study of syllabus design for integrated English education in elementary school (Unpublished doctoral dissertation), Korea National University of Education, Chungbuk., 2010.

[4] Jun-Eon Park, "Applicability of immersion program in Korean English education," English Teaching, vol. 53, no. 4, pp. 247-26, 1998.

[5] Mimi Met, "Learning language through content: Learning through language," Foreign Language Annals, vol. 24 no.4, pp.281-295, 1991.

[6] Seong-Sam Oh: Theory and practice of meta-analysis. Seoul: Keonguk University Press, 2002.

[7] Tae-Hwan Kwon & Jeong-ryeol Kim, "Meta-analysis of the effects of content-integrated English education in knowledge-based subjects and skill-based subjects in elementary school," Journal of Learner-Centered Curriculum and Instruction, vol. 12, no. 2, pp. 29-49, 2012.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1363-1371 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Hybrid Ontology-based knowledge with multi-objective optimization model framework for Decision Support System in intercropping

Kornkanok Phoksawat[*1, 2], Massudi Mahmuddin [2]

[1]*Department of Information System, Rajanangala University of Technology Srivijaya, 80110, Thailand*

[2]*School of Computing, Universiti Utara Malaysia, Sintok, 06010, Malaysia*

A R T I C L E I N F O

Article history:
Received: 30 May, 2017
Accepted: 18 July, 2017
Online: 05 August, 2017

Keywords:
Ontology-based knowledge
Multi-Objective Optimization
Multi-Criteria Selection problem
Intercropping in rubber plantation
Decision Support System

A B S T R A C T

The problem of multi-criteria or multi-objective intercropping crop planning makes it vital to consider all related factors under the constraints that will produce the highest revenue and minimum cost. Principles of intercropping elements including soil type, plant area, plant diseases, planting and harvesting time and economics factors (e.g. price, cost) are some of the factors in making decisions. Intercropping is important in the situation such as during inability to harvest main crops, co-cultivation to increase productivity, or even to increase extra revenue. Therefore, the decision-making system requires a wise decision support system, which can advise farmers on economics matters. In this article, we present the decision support system (DSS) model framework for planting rubber with intercropping by a hybrid approach using ontology-based knowledge consuming rule concepts and relationships for intercropping with integrated multi-objectives optimization to recommend the crop to be planted and the suitable proportion of planting areas or planting co-suited to the rubber plantation of farmers. This approach could be applied as a guideline for another field, such as production problems or other resource allocation issues.

1. Introduction

This paper is an extension of the original work which was presented in 2016 at the International Computer Science and Engineering Conference (ICSEC) [1]. This article explains in more detail about the Research Framework and Research Methodology.

The agriculture in Thailand comprises of 45.84 million acres or 35.7% of area of the country where 25.9% of households are farmers with an average 7.76 acres per person. These farmers are divided into 79.1% monoculture farmers and another 19.9% integrated farming. Other main agricultural activities in Thailand are livestock production, fish or other important aquatic animal raising or salt ponds [2].

Recently, the weather uncertainty and low price of the agricultural products especially hits the monoculture farmer severely. This can be seen in rubber planting (rubber is a major Thailand economic crops), where the rubber price affected the farmers. In 2015, the rubber prices dropped 20.1% as compared to the previous year which made export of rubber drop by THB 52,000 million (21.4% decrease) [3].

Southern Thailand is the main producer of rubber in Thailand, and 70% of the labor in this region are rubber farmers in monoculture. The farmers have been impacted from this terrible situation where their income in 2015 has declined THB 33,000 million (or 17.7 %) as compared to on the previous year. As a result, this decline has raised the household's average debt from THB 134,893 per household in 2013 to THB 153,144 in the first half of 2015. Findings from the Bank for Agriculture and Cooperative shows that the farmers were giving some signal that their income is unsustainable for them to pay the bills normally due to the fluctuation of rubber's price since 2014 and many of them decided to uproot their rubber trees and preferred to plant other more profitable crops. In this condition, monoculture farming might not be the possible answer although the economic return was high, at the same time, the risk is still there when the price fluctuates by the market [4].

[*]Corresponding Author: Kornkanok Phoksawat, Department of Information System, Rajanangala University of Technology Srivijaya, 80110, Thailand
Email: kornkanok.p@rmutsv.ac.th

www.astesj.com
https://dx.doi.org/10.25046/aj0203172

Due to that, it is unwise to stay in the type of farming approach which only concentrates on one single type of crop. One solution is to introduce or convert single monoculture into poly or intercropping farming system. In a general sense, intercropping is a planting system that uses a multiple cropping system with at least two or more crops planted in the same field/plot during a growing season [5]. In this regard, intercropping is a potential alternative for a sustainable way for the farmers to overcome the above-mentioned problem by reducing risks of product price fluctuation, pathogens or insect attacks [6]. At the same time, the Thailand government have made a lot of effort to support this idea of intercropping system.

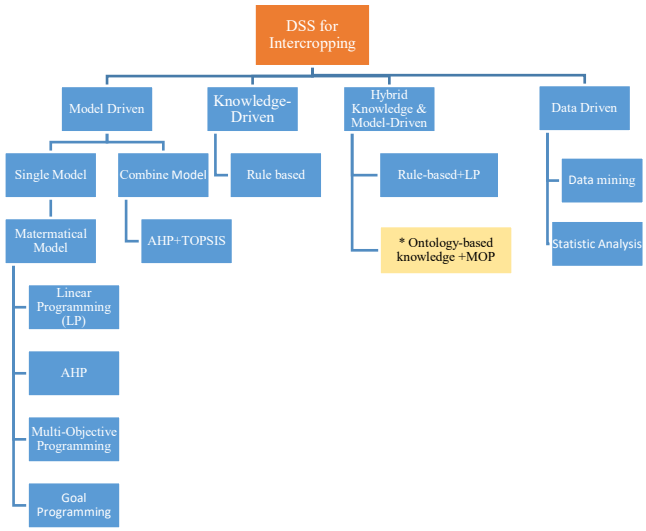
To clarify this idea, we have organized the information in figure 1 to make understandable our proposed concept to assist the farmer in this specific region or it might be applied or adopted in many other area. The interconnectedness of this model is based on the previous works that are discussed fully in Section 2. The proposed concept of integrating the overall concept framework and evaluation is discussed precisely in Sections 3 and 4 and we conclude our idea in Section 5.

2. Related Work

2.1. DSS for Intercropping

Many DSS for intercropping planning used mathematical models or are model-driven to solve this problem such as Linear Programming (LP) [7], Analytic Hierarchy process (AHP), Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) or Simple additive weighting (SAW) [8] , or hybrid technique e.g. AHP and TOPSIS[9]. Nevertheless, that approach will provide just an answer to a priority of alternatives and does not answer about living space to grow. For example, the proportion of main crop and secondary crops and the calculations in this area, considering the distance between the plants and the second plants as well as the principles of the intercrop. After the system recommends a suitable priority for particular co-planting of plants, the farmers need to determine the specific area of use in order to figure out a maximum overall return. Such types of calculation may not result in an answer with maximum income under their restricted resources, and hence their opportunity of profit making is lost with not including the extraction of knowledge from experts to optimize knowledge management and decision-making.

[10] presented an integrated DSS for intercropping by using a rule-based expert system as a model. This method can be extended to cover more decision criteria to cost the cropping. The objective of the work is to manage cropping activities for improved crop production. However, approaches ignore important features due to the paper not offering price statistics for the decision. This factor is necessary for planning crops. [11] showed the model-based support by using evolutionary algorithms to solve multi-objective optimization and design of mixed farming systems aimed to maximize the operating profit and organic matter balance, and to minimize the labor requirement and soil nitrogen losses. [12] presented the goal programming for rubber plantation intercropping with tea trees. The purpose of this work is management of rubber plantation yield before getting a return for seven years.



Note: * New guideline in DSS for intercropping

Figure 1 Summary of approach for DSS-Intercropping

However, literature reviews to analyze approaches in the DSS cannot ignore the knowledge-driven decisions, sometimes generically called expert systems or recommender systems because knowledge-driven DSSs might use artificial intelligence and/or some of the statistical inference technologies to suggest or recommend actions to managers. These systems have knowledge codified to provide specialized problem-solving assistance [13]. Especially in agriculture, the farmers who succeed in farming collect long-standing experiences. The knowledge has been tested through experimentation and theory and then becomes tacit knowledge which is necessary to be maintained and adopted in a knowledge-base for a decision [14, 15].

Therefore, the challenge is to design an ontology in the field of mixed intercropping which refers to the practice handbook of planting and a rule base from domain experts. The system can suggest the plants that grow well together in the suitable areas in a format that can be processed by a computer program. It focuses on knowledge sharing and reuse and sends related parameters about plants that suggest related plants from the recommender engine to calculate multi-objective optimization programming in order to give the answers which can support the intercropping decision supporting system for an individual agriculture.

2.2 Ontology-knowledge base for DSS

Ontology is a structured establishment and knowledge representation in knowledge management system or acts as a tool similar to that system whose indices were used to manage and access the subject of the question using controlled vocabularies comprising vocabularies, terms, syntax and certain rules based on making use of appropriate word patterns [16]. The results of using an ontology-based illustration of the important data structure of a DSS are the distinguishable state from what happened in the past with DSS as closed systems, the primary ontologies' objective are to determine proportionated domain models and knowledge from a variety of sources. Also, some of the inference stages of the DSS can be carried out using state-of-the-art logical reasoning performance, as rule logic providers or justification providers [17].

Advantages of ontologies include seeming well suited to an evolutionary approach to domain knowledge and requirements specification [18] in contrast to traditional knowledge-based approaches, e.g. formal specification languages. Requirements management and traceability can be supported by ontologies. Semi-formal or informal approaches provide no logical formalism or model theory compared to automated validation and consistency checking which are considered as a potential benefit. Finally, the formal specification may be a prerequisite to realizing model-driven design and implementation approaches [19].

Deciding to intercrop does not focus on finding answers to the income or economics alone but it must be based on the principles of integrated crop plants to reduce disease risk and the suitability of the area under the constraints of farmers. Therefore, this research uses ontology-based knowledge for driving DSS and providing the type of plant and design optimization model for calculating the answer for maximum income. This phase uses knowledge acquisition and ontology modeling for ontology design and experiments for economic design modeling.

Ontology-based knowledge representation that allows information sharing, reuse and collaboration on design and obsolescence issues across different organizations provides the basis for decision-making in an intercropping selection and presents concepts, relationships, and instances. To establish a comprehensive knowledge representation scheme, an ontology that can serve as the backbone of the needed information model is required [20].

Ontologies technology can be applied in the work of various systems by considering the purpose of the deployment: (1) information retrieval; (2) systems that require integration of knowledge for a basic understanding of the knowledge domain; (3) systems that have the need for knowledge reuse and such systems may be defined ontologies technology to use allowing another system to be used again by the system later. This may define semantic information that does not have contrary information of semantic ontologies taken back to another system and there is no need to explain that there is no change, such as static domain knowledge or knowledge-based system. In particular; ontology has recently been adopted in DSS in various application domains [21].

Currently, ontologies are semantic knowledge management which have been successfully implemented in DSS in various application domains, and to determine the tasks and parameters of some system modules [17], applied in knowledge management [22]. However, an ontology applies in knowledge management through the web service system, question answering [23] or DSS [24] in planting trees. Nevertheless, that answer or knowledge is about planting monoculture crops only and lacks a decision to plant mixed or intercropping to maximize income while minimizing cost in a limited farmer's resources.

2.3 Multi-Objective Optimization Methods

This section and figure 2 presents a summary approach for Multi-Objective Optimization. There are classical methods and Evolutionary Multi-Objective Optimization (EMO) Method [25].

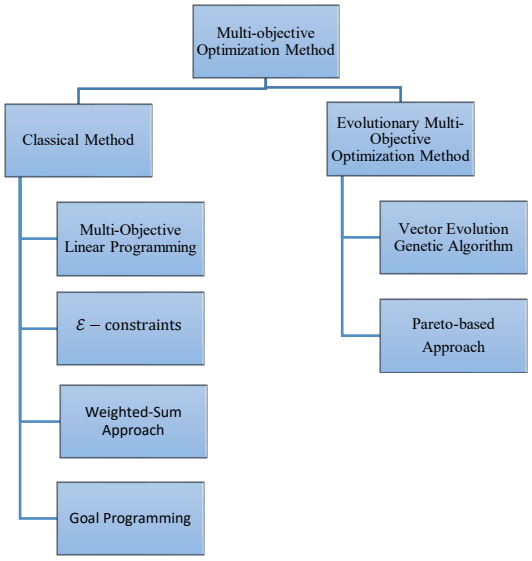


Figure 2 Some approaches for Multi-Objective Optimization

- Multi-objective is a vector of constraints, objective functions, and variables and is considered by Multi-objective Programming Model. A reduced subset of the feasible set exceeds all of the aspiration levels of each objective contains a satisfactory solution a set of which has acceptable alternatives that do not need to be non-dominated. A final choice of a non-dominated solution selected through decision makers' involvement in the information processing stage is a preferred solution [26].
- Ɛ-constraint methods are those of a posteriori using scalarization. By these methods, a multi-objective problem will be converted into a single-objective method and can solve the problem and obtain its solution as a Pareto-optimal single solution.
- Weighted Sum Approach is an easy method to apply to weighted specific functionalities and then combine the weight and objective functions to build a single objective but the problem of using this method is if there is not enough information, it will result in difficulty in defining the objective weight as can be seen from the research of [27].
- Goal Programming Approach is a method that rank evaluates a set of results which uses the stage of Multi-Objective Goal Programming (MOGP), as can be seen from the research of [28] which presents that this method is well known in methodology and development of the proposal which has efficiency in defining goals (Multiple Goal Linear Programming-MGLP) that are trying to deviate between various objectives. The lowest formula is used in calculating outputs from various goals and compares those values until it gets the best results suitable for an allocation problem such as limited resources management [29].
- Vector Evaluation Approach is a method that defines Fitness Assignment which is the first method that expands the steps of Genetic Algorithms, which results

in finding the most appropriate results for various objectives. [30] explained the process consists of 1) Initialization Population 2) Evaluation Objective Function 3) Define Fitness Assignment 4) Selection of the results from the Fitness Assignment of strings: each result during this process will crossover. This is a process where a right string answer which comes paired to exchange to get the new string result and mutation to evaluate new strength until the calculation reaches the maximum number of answers looked for [31].

- Pareto-based Approach: This method is used to rank the Pareto to build relationships between the function, purpose, and strength by using the Pareto principle to calculate the strength. For how to configure the strength to make this the best result, it will have more than one possible result that could be the best answer and it will be in the format of set, or group of the best answers [32].

3. Conceptual Framework

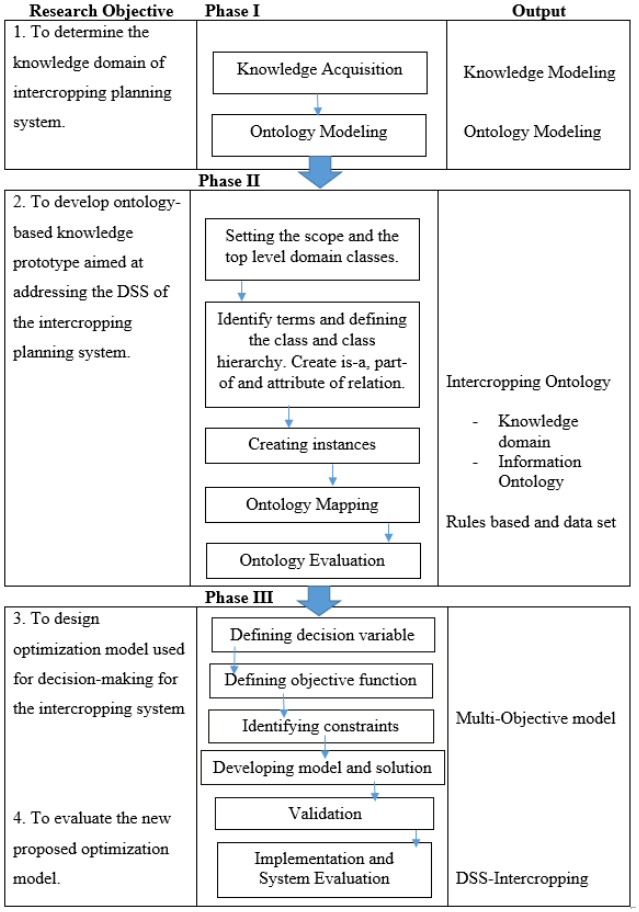


Figure 3 Research Conceptual Framework of DSS-Intercropping

The DSS for growing rubber plants is a hybrid method which works together in two parts. The knowledge-based part uses the process of knowledge management with ontology to obtain a rubber plantation model and to develop an intercropping ontology and rule base that will lead to a more intelligent recommender system which is mixed with a multi-objective optimization model part. It will helps to calculate and allocate each crop area to yield maximum yield and the lowest cost that the system will help farmers or users make decisions as a guideline or primary recommendations before making a decision on plantation operation as indicated by soil, time, season, area, cost, and market price or other critical parameters.

4. Research Methodology

The method of this work is divided into three main phases namely, Phase I knowledge acquisition and ontology modeling, Phase II ontology development, and Phase III optimization modeling and system implementation. This is shown in Figure 3 and details in each phase are presented in Sections 4.1, 4.2 and 4.3, respectively.

4.1. Phase I: Knowledge Acquisition and Ontology Modeling

This research is information to extract knowledge and expert knowledge for decision making and decision criteria and a variety of organizations. The Ministry of Agriculture and Cooperatives of Thailand is the main organization for promoting agriculture. It was split into several entities with different responsibilities. In addition, certain information is required from other ministries. These are summarized as the sources of data for research purposes to support farmers for intercropping shown in Table 1.

Table 1 Summarized the sources of data to support decision making for intercropping

Organization	Type of information
Department of Agricultural Extention	providing: - Knowledge and media for cropping - Disease and pets
Office of Agricultural Economic	- Sale Price Statistic per kg (THB) - Cost per square meters (THB) - Productivity per square meters (kg.)
Royal Irrigation Department	The water situation and rainfall statistics.
Land Development Department	Soil information each province
Provincial Agriculture Office	Advising agriculture in each province for intercropping.
Farmer who are success in intercropping	The information about intercropping from experience.

Knowledge Acquisition

In this section, the knowledge acquisition approach was a combination of text analysis and expert interview approach [33]. The study included three steps:

- Concept and domain knowledge. Identified using secondary resource document analysis e.g. electronic resources, journal, or book. Based on intercropping system principles, in this research merely three factors: spacing, cultivation period, plant disease will be employed as important factors. Selection of factors are also consistent with research that had been performed by a summary of factors from previous research and from the selection of sub rubber plantations booking to pursue an

alternative career at the Department of Agricultural Extension, Ministry of Agriculture and Cooperatives, Thailand which recommended considering the use of tree age criteria. In addition, due to time constraints of performing this research data some important related factors are already sufficiently provided and the high cost of extension of the research to other less important or less prioritized topics and the selection is based on recommendations provided by plant production experts. The selected factors are soil, spacing, time, and diseases. Spacing is important as individual plants require a different area for growth of root, and size of branching. These factors are expected to produce a significant impact on light, water and mineral resource use of plants. Time is a critical factor as well, and a proper arrangement or organization of short-term and long-term co-planted plants will cause an increased use and rate of light, and mineral resource for plants compared to single cropping that is usually suitable for some particular seasons and period of harvesting so as to generate constant revenue throughout the year. Even though mixing or intercropping may not produce complete immunity to related plants, as some diseases can destroy more than one type of plant in one single invasion, a selection of proper but different plants that usually can't be destroyed by the same diseases will much reduce a destruction rate of disease destruction. However, economic factors including price and cost need to be involved because the decisions affect production rates[8].

- Interview with experts who are in the areas of intercropping. Two groups classify the population and sample: (1) government agency officials obliged to provide agriculture information, and (2) farmers successful in intercropping. Purposive sampling from the population is used because they have to choose from the expertise of experienced farmers in the agricultural field and the knowledge and experience of rubber farmers in the province. So the experts came from two groups, the first group were five state employees from five agencies of the Ministry of Agriculture and Cooperatives. As depicted in Table 1, this is an agency that provides information related to agricultural extension which has a direct duty to provide advice on cropping for local farmers, and Group 2 are three farmers who have specialized and succeeded in intercropping in rubber plantations. After that drafting knowledge classification was categorized by evaluating the data types and characteristics of the data using a hierarchical clustering of intercropping system terms. Questions to ask are open-ended questions to extract knowledge of intercropping cultivation. The criterion for selecting a plant to plant with rubber are the age range of tree and plants suitable for growing any plant with the same plant disease, the risk of damage to the main crops or rubber trees. Then the draft knowledge domain is summarized to be verified in the next step.
- Knowledge domain and concept is summarized and results reconfirmed with the experts. Consider the appropriateness of the criteria and concepts that are included in the manual and from expert interviews. Then reconfirm from experts by the focus group method.

Ontology Modeling

Identifying terms of related information is required for the design of ontologies for intercropping which starts from defining the scope of knowledge about intercropping. A combination method is then used for the system to classify the terms into classes and hierarchy of class levels of ontology. Important concepts are sequenced first e.g. the selecting crops principle. For instance, requirements of co-planting plants in terms of disease and insect, soil, and these co-planting plants harvesting time should be overlapped to help farmers have a continuous or relatively constant revenue throughout a year. Then each class qualification is identified including heritage of classes consideration and defining conditions for qualification of a class. For example, classes types of qualifications are identified. The tool to build knowledge modeling is CmapTools. CmapTools collaboration and publishing features are used widely for sharing and representing all knowledge domains including organizing information, brainstorming and knowledge management.

4.2. Phase II: Ontology Development and the Evaluation

Ontology Development

The ontology life cycle has seven steps [34] but: "there is no one way or methodology for developing ontologies and ontology development is necessarily an iterative process." Ontology development should be based on application objectives, data characteristic, information system rules and context and future reuse feasibility [35]. The intercropping ontology will be developed using a top-down approach. This is a way to start with the concept of a generic domain which starts with the most common definitions based on concepts in the domain and take that expertise to be guidelines and define a specific concept. For example, starting by creating a class for the general idea of cropping. Then, create a subclass in that class and go to the bottom of the ontology structure which are instances or data, followed by the modeling:

- Setting the scope and the top level domain classes. To define scope and the top level domain classes it is defined in the field of intercropping and this ontology is used to make decisions in intercropping. It helps answer questions such as "What are characteristic of plants that can grow with rubber plantation?", "Which season is an appropriate season to plant?", "What is the cost and output per square meter?", and "What is the total revenue and cost?" Knowledge-based ontology development uses the ontology editor. Several software platforms and application interfaces (APIs) have been developed to support the Semantic Web vision of making machine-readable content available on the web to permit automatic creation and use of OWL and RDF. Currently there are many tools for ontologies editors such as Protégé, OntoEdit, OILDd, Hozo. However, Protégé is an open-source platform developed by Stanford University and is a kind of popular tool for ontology editor. Protégé is plug-in architecture that can

be adapted to build both simple and complex ontology-based applications. Developers can integrate to output of Protégé with rule systems or other problem solvers to construct a wide range of intelligent systems [36].

- Identify terms or concept and defining the class and class hierarchy. Create *Is-a*, *Part-of* and *attribute-of* relation. Examples of important terms or concepts in this work are such as economic crops, soil, water, plant diseases, seasons, etc. Define relationship between concept types *Is-A* such as rubber is an economic plant. For *Part-of* relationship, for example, the address of the farmer consists of district, etc. Then define the properties of Concept (Term) such as rubber is a species, type of soil suitable for planting, plant diseases, yield per rai (3000 m^2), price etc. There are two types of intrinsic property such as variation, appropriate soil type, planting period, plant disease, and 2 extrinsic properties such as cost and price.
- Creating data instances and ontology mapping. Recommender engine requires a rule engine to construct and additional criterion [37] applied JENA and SPARQL for developing the backend of this component [38].

Ontology evaluation

Assessment of ontology in this work is through clarity, completeness, consistency, and correctness criteria by ontology development experts. Based on an ontology evaluation by an expert also known as assessment by humans it can evaluate all 6 levels including lexical, taxonomy, semantic relations, application, syntactic, and design. When compared to Gold Standard, application-based, and data-driven evaluation [39] and the number of experts to use for validity, [40] advised a minimum of three experts, on the other hand indicated that more than ten was probably unnecessary [41]. Therefore, in evaluating the ontology in this research, it employs three developers with three to five years ontology experience. This step provided evaluation in terms of the concepts, and relationships related to the intercropping system and includes an accuracy evaluation of rules, a concept of result calculation, screening of cultivated plants suitable for a particular area as farmer requests and screening of co-planting plants are assessed by three experts by a comparison between system results and results obtained by experts' presented recommendation. Then it is improved before using an optimization model to calculate economic factor in the next step.

4.3. Phase III: Optimization Modeling and Implementation

Optimization Modeling

After creating the ontology-based knowledge to introduce crops that can be planted with major crops suited to an area by intercropping we will get parameters for design models and import data into the model including data from the database in the recommender system such as the type of crop, prices, and cost and input data from users consisting of: name, address, size of area, cost, and start of planting.

Because this problem is one of resource allocation which are land and cost the mathematical model used to solve this limited resource allocation problem, decision variables are objective functions or the limitations of the model. Each decision variable is often multiplied by a constant and then added as a function, objective or equation or inequality of constraints. These variables have a linear relationship. The models in this characteristic are called Linear models such as x_1 and x_2 instead of the 1st and 2nd crop where z represents the profit or revenue from the sale of both crops. There are three main components: Decision variable, Objectives, and Constraints. A detailed explanation of the objective function and the constraints of the proposed model are given in the following sections.

- **Decision Variables**

From the problem, it is found that the farmers have to choose intercrop plants that fit their own space and resources. Therefore, the decision variable is what economic crops are suitable to grow with rubber trees and how much of the planted area will produce the highest yield and lowest cost. The calculation model contains two parts: Part 1 Data is added to the user's planting area of intercropping, identifying the district to determine the type of soil from the Land Development Department, the size of the area measured in square meters. Secondly one must determine which information is stored in the data set. This includes production per square meter, price (THB / kg.), cost per square meter, and the number of days it takes to grow. The result of this calculation, a multi-objective optimization model, is shown in the background of each type of crop, illustrating the potential highest income and lowest total cost.

- **Create the Objectives and Constraints**

Intercropping has several objectives to prevent pests, adding nutrients to increase productivity, mulch, help increase revenue before harvesting main crops, etc. [42]. However, [43] has studied factors of decision making for multiple crops or new theory farming practices of farmers in north-eastern, Thailand. It was found that the main factors that affected the changing behavior of decision making or the selection of crop system was the price or income and that most farmers' problems are capital, plant diseases and lacking irrigation. This is in line with the study by [44] which presented the factors affecting the decision to change the use of land or change the crops of farmers Thailand were revenue and costs. The problem is caused by the risk of price of output and the problems caused by the monoculture as mentioned above. Helping farmers get the highest annual income with lowest cost was the objective of this research through focusing on the allocation of space to grow various crops to increase revenue. Maximizing plant cultivation income and minimizing cost is the purpose of the model.

Many variables are used to calculate incomes for listed types of plantings as follows: particular productivity per square meter (m^2), sale price / kilogram (calculated as local currency as Thai currency), planting period starting from initial cultivation until harvested and cultivation period (month), area (m^2), related expenditure or production cost is calculated as sum of cost the individual productivity types (kilogram) multiplied by cultivation area (m^2). Constraints include size of the total area of the farm, where the number of plants will be not be equal to one. The

prototype consists of Decision Variable, Objective Function, and Constraints.

- **Developing Model and Validation Solution**

Multi-Objective Programming (MOP) is formulated with the objective of maximizing economic values obtained from an intercropping system as well as minimizing its cost. The optimal levels for various decisions regarding seeding and harvesting time periods and cultivation areas are determined by solving the MOP. The tool used in the test model is MATLAB.

MATLAB is a proprietary fourth-generation programming language developed by MathWorks and MATLAB provides a multi-paradigm numerical computing environment allowing C, C++, Java, Fortran and Python interfacing, user interface creation, algorithms implementation, data and function plotting and manipulations of matrices and offers tools for application development, visualization capabilities, data analysis and high-performance numerical computation [45].

The solution allows one to find an answer within the set area of the possible answers, to find a minimum or maximum of each function simultaneously. The result or best answer consists of m objectives and decision variables as you can see in Equation 1 written as general equations as:

$$Minimize\ (or\ maximize): \{f_1(x), f_2(x), \dots, f_m(x)\} \quad (1)$$

where x is a vector of decision variables

f_i (x) is objective function, where i=1,2, …, m

Therefore, patterns of solving multi-objective optimization function is finding x vector solutions within m constraints as shown in Inequality 2 or within m constraints as shown in Equation 3 that determines the possible solution area. This defines a set of possible solutions.

$$g_i(x) \leq 0, i = 1,2, \dots, m \quad (2)$$

$$h_i(x) = 0, i = 1,2, \dots, m \quad (3)$$

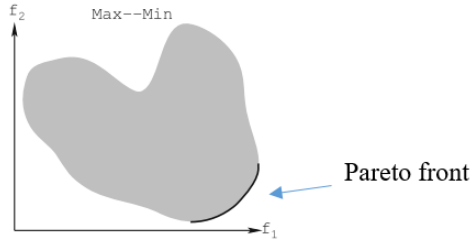


Figure 4 Pareto front of multi-objective functions f_1 (Max) and $f2$ (Min)

[46] define the term Pareto Dominance and appropriate answer Pareto Optimality as seen below:

Definition 1: Pareto dominance: A given vector $u=(u_1, \dots, u_m)$ if and only if u is partially less than v. As can be seen in Equation 4:

$$\forall_i \in \{1, \dots, m\}, u_i \leq v_i \wedge \exists_i \in \{1, \dots, m\}: u_i < v_i \quad (4)$$

Where in the above equation u and v are defined.

Subsequently, there will also be some instances where u is greater than v.

"Definition 2: Pareto optimality: A solution $x_u \in [a, b]$ is said to be Pareto-optimal if and only if there is no $x_v \in u$ for which $v=f(x_v)=(v_1, \dots, v_m)$ dominates $u=f(x_u)=(u_1, \dots, u_m)$.

Pareto-optimal solutions are also called efficient, non-dominate, and non-inferior solutions. The corresponding objective vectors are simply called non-dominated. The set of all non-dominate vectors is known as the non-dominated set, or the tradeoff surface of the problem."

Implementation and System Evaluation

This stage includes coding and testing, installation and training and the final step is system evaluation. Knowledge-driven DSS uses a knowledge base extracted from the tacit knowledge of expert domain which makes the knowledge base available for interaction with decision makers by the structure of the Multi-Objective Optimization model and the overall evaluation system. It is rule based to integrate with the data found in the system database and the decision maker's input to get the answer in planning rubber plantations.

- Checking to confirm the accuracy of recommender engine the introduction of intercropping planted with rubber. Using Ontology-based knowledge includes validation of agricultural information, agricultural economics, and the accuracy of the rule base. According to the principle of rubber plantation, using ontology-based knowledge in intercropping the ontology of the rubber plantation created consists of two parts: part 1 , the agricultural and agricultural economics data is the data of each economic crop such as harvesting season, harvesting Planting Period Planting Cost from data obtained from the public sector with reliable data storage and statistics and the experts in cropping are evaluating the validity of the criteria and parameters used in the model and rule based trust and knowledge validation mappings [47] by ontology development experts and intercropping experts. The validation solution by Mean Absolute Error (MAE) is used to calculate the rating difference between the actual and the predicted ratings.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| = \frac{1}{n}\sum_{i=1}^{n}|e_i| \quad (5)$$

where:

$$AE = |e_i| = |y_i - \hat{y}_i|$$
$$Actual = y_i$$
$$Predicted = \hat{y}_i$$

- The data collection using a triangulation method was conducted by literature review, expert validation and

review, and user evaluation. The user evaluation focused on decision makers or users focused on decision makers' satisfaction. In order to evaluate overall satisfaction of DSS, the users were officers at various levels of the agricultural extension office such as provincial or district level from 23 districts. Thus, sampling of choice was multi-stage mixed random sampling and selection and the sampling was conducted by division of the sample into district level and purposive selection was made by officers who are responsible for giving related advice and recommendation. Statistics used in this case are mean $\bar{x}$ and standard deviation ($s.d.$).

5. Conclusion and Future Work

Ontology-based knowledge and multi-objective optimization programming to drive a decision support system for intercropping is based on constraint factors from an individual farmer for illustrating maximum income. The research methodology is divided into three phases as follows: (1) knowledge acquisition and ontology modeling. This methodology is based on knowledge's extraction from experts or specialists due to the fact of long experience based knowledge; (2) ontology development and recommender system is a creation of ontology as it is designed as well as a preparation of instance and ontology mapping as rules governing intercropping and evaluation and assessment of ontology; and the last stage is (3) design of multi-objective programming in order to obtain the answers of what types of plants should be planted, the proportion of each co-planting plants, overall generated revenue and production cost. The future work is knowledge acquisition and ontology modeling which is an important step to collect knowledge from experts and published papers issued by Department of Agriculture Extension, Ministry of Agriculture and Agricultural Co-operatives before being used for developing ontology at Phase II.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

Many thanks to Rajamangala University of Technology Srivijaya who sponsored the research grants and, also thanks to Nakhon Si Thammarat Agricultural Extension Office for the information and their time to explain about the problems of farmers.

References

[1] K. Phoksawat and M. Mahmuddin, "Ontology-based knowledge and optimization model for Decision Support System to intercropping," in *2016 International Computer Science and Engineering Conference (ICSEC)*, 2016, pp. 1-6. https://doi.org/10.1109/ICSEC.2016.7859927

[2] Office of Agricultural Economics, "Agricultural Statistics of Thailand 2015," 2015.

[3] Bank of Thailand, "Report on major agricultural price trends in southern Thailand," 2016.

[4] M. Stamp, "Risks of monoculture," *Communication of the ACM,* vol. 47, p. 120, 2004.

[5] S. R. Mousavi and H. Eskandari, "A general overview on intercropping and its advantages in sustainable agriculture," *Journal of Applied Environmental and Biological Sciences,* vol. 1, pp. 482-486, 2011.

[6] S. Arreeras, T. Angskun, and J. Angskun, "A Design of an Expert System for Intercropping Planning," in *Proceedings of the 16th Tri-University International Joint Se miner and Symposium*, 2009, pp. 19-22.

[7] A. Gadge, A. Kedar, and K. Aglawe, "Optimization of cost for farmers of cotton based farming system by Linear Programming Model," *International Journal of Mechanical Engineering and Robotics Research,* vol. 3, p. 505, 2014.

[8] C. Srisawat and J. Payakpate, "Comparison of MCDM Methods for Intercrop Selection in Rubber Plantatios," *Journal of ICT,* vol. 15, pp. 165-182, 2016.

[9] P. Worawimolwanich and K. Kesorn, "Decision Support System for Economic Crops using Hybrid Approaches," *The eleventh National Conference on computing and Information Technology,* pp. 425-431, 2015.

[10] A. S. Sodiya, A. T. Akinwale, K. A. Okeleye, and J. A. Emmanuel, "An Integrated Decision Support System for Intercropping," *International Journal of Decision Support System Technology,* vol. 2, pp. 51-66, 2010.

[11] J. C. Groot, G. J. Oomen, and W. A. Rossing, "Multi-objective optimization and design of farming systems," *Agricultural Systems,* vol. 110, pp. 63-77, 2012.

[12] Y.-C. Chang, P.-P. Lee, and C.-E. Kan, "Farm-level land diversity of smallholders with limited input resources during urbanization in Taiwan," *Paddy and water environment,* vol. 12, pp. 155-162, 2014.

[13] D. J. Power, R. Sharda, and F. Burstein, *Decision support systems*: Wiley Online Library, 2015.

[14] K. Asanee, "Ontology Engineering and Knowledge Services for Agriculture Domain," *Journal of Integrative Agriculture,* vol. 11(5), pp. 741-751, 2012.

[15] S. Barthel and C. Isendahl, "Urban gardens, agriculture, and water management: Sources of resilience for long-term food security in cities," *Ecological Economics,* vol. 86, pp. 224-234, 2013.

[16] A. Dawson, D. Brown, and V. Broughton, "The need for a faceted classification as the basis of all methods of information retrieval," in *Aslib proceedings*, 2006, pp. 49-72.

[17] M. R. a. L. Serafini, "An Ontological Framework for Decision Support," *Springer-Verlag Berlin Heidelberg,* pp. 239-254, 2013.

[18] B. Wouters, D. Deridder, and E. Van Paesschen, "The use of ontologies as a backbone for use case management," in *European Conference on Object-Oriented Programming (ECOOP 2000), Workshop: Objects and Classifications, a natural convergence*, 2000.

[19] H.-J. Happel and S. Seedorf, "Applications of ontologies in software engineering," in *Proc. of Workshop on Sematic Web Enabled Software Engineering"(SWESE) on the ISWC*, 2006, pp. 5-9. https://doi.org/10.1.1.89.5733

[20] X. Chang, L. Zheng, and J. Terpenny, "Ontology Development and Optimization for Data Integration and Decision-Making in Product Design and Obsolescence Management," in *Ontology Modeling in Physical Asset Integrity Management*, ed: Springer, 2015, pp. 87-132.

[21] R. Saa, A. Garcia, C. Gomez, J. Carretero, and F. Garcia-Carballeira, "An ontology-driven decision support system for high-performance and cost-optimized design of complex railway portal frames," *Expert Systems with Applications,* vol. 39, pp. 8784-8792, 2012. https://doi.org/10.1016/j.eswa.2012.02.002

[22] D. Fensel, "Ontology-based knowledge management," *Computer,* vol. 35, pp. 56-59, 2002.

[23] A. Thunkijjanukij, A. Kawtrakul, S. Panichsakpatana, and U. Veesommai, "Ontology development: a case study for Thai rice," *Kasetsart Journal (Natural Science),* vol. 43, pp. 594-604, 2009.

[24] S. Y. Mawardi, M. H. Abdel-Aziz, A. M. Omran, and T. A. Mahmoud, "An Ontology-Driven Decision Support System for Wheat Production," *International Journal of Computer Science and Telecommunications,* vol. 4, pp. 11-17, 2013.

[25] K. Deb, "Multi-objective optimization," in *Search methodologies*, ed: Springer, 2014, pp. 403-449. https://doi,org/10.1007/978-1-4614-6940-7_15

[26] G. Zhang, J. Lu, and Y. Gao, *Multi-Level Decision Making*: Springer, 2015.

[27] P. V. Radtke, T. Wong, and R. Sabourin, "A multi-objective memetic algorithm for intelligent feature extraction," in *International Conference on Evolutionary Multi-Criterion Optimization*, 2005, pp. 767-781.

[28] F. Blancas, R. Caballero, M. González, M. Lozano-Oyola, and F. Pérez, "Goal programming synthetic indicators: An application for sustainable tourism in Andalusian coastal counties," *Ecological Economics,* vol. 69, pp. 2158-2172, 2010.

[29] B. Mishra and S. Singh, "Optimal Land Allocation in Agricultural Production Planning Using Fuzzy Goal Programming," in *Proceedings of Fifth International Conference on Soft Computing for Problem Solving*, 2016, pp. 287-298. https://doi.org/ 10.1007/978-981-10-0448-3_23

[30] J. D. Schaffer, "Some experiments in machine learning using vector evaluated genetic algorithms," Vanderbilt Univ., Nashville, TN (USA)1985.

[31] P. Ammaruekarat and P. Meesad, "Multi-Objective Optimization using Evolutionary Algorithms," *Information Technology Journal,* vol. 8, pp. 73-80, July-December 2012 2012.

[32] Z. He, G. G. Yen, and J. Zhang, "Fuzzy-based Pareto optimality for many-objective evolutionary algorithms," *IEEE Transactions on Evolutionary Computation,* vol. 18, pp. 269-285, 2014. https://doi.org/ 10.1109/TEVC.2013.2258025

[33] Y. I. Liou, "Expert system technology: knowledge acquisition," *The handbook of applied expert systems,* pp. 2.1-2.11, 1999.

[34] N. F. Noy and D. L. McGuinness, "Ontology development 101: A guide to creating your first ontology," ed: Stanford knowledge systems laboratory technical report KSL-01-05 and Stanford medical informatics technical report SMI-2001-0880, Stanford, CA, 2001.

[35] N. Kaewboonma, K. Tuamsuk, and W. Kanarkard, "Ontology development for drought management information," *International Journal of Metadata, Semantics and Ontologies,* vol. 9, pp. 324-332, 2014. https://doi.org/10.1504/IJMSO.2014.065443

[36] H. A. Salah, "Ontology development (OWL&UML) methodology of web-based Decision Support System for water management," in *Electronics, Computers and Artificial Intelligence (ECAI), 2014 6th International Conference on*, 2014, pp. 11-22. https://doi.org/ 10.1109/ECAI.2014.7090217

[37] T. Supnithi, P. Paiboonrat, M. Buranarach, A. Kawtrakul, S. Lekawatana, A. Intalak*, et al.*, "Ontology Based Orchid Knowledge Platform for Knowledge Services in Orchid Cluster," in *2011 Annual SRII Global Conference*, 2011, pp. 540-549. https://doi.org/ 10.1109/SRII.2011.62

[38] Y.-F. Zhang, L. Gou, Y. Tian, T.-C. Li, M. Zhang, and J.-S. Li, "Design and Development of a Sharable Clinical Decision Support System Based on a Semantic Web Service Framework," *Journal of medical systems,* vol. 40, pp. 1-14, 2016. https://doi.org/10.1016/ j.cmpb.2015.90.020

[39] H. Hlomani and D. Stacey, "Approaches, methods, metrics, measures, and subjectivity in ontology evaluation: A survey," *Semantic Web Journal,* pp. 1-5, 2014.

[40] M. R. Lynn, "Determination and quantification of content validity," *Nursing research,* vol. 35, pp. 382-386, 1986.

[41] D. F. Polit and C. T. Beck, "The content validity index: are you sure you know what's being reported? Critique and recommendations," *Research in nursing & health,* vol. 29, pp. 489-497, 2006. https://doi.org/10.1002/nur.20147

[42] Department of Agriculture Extension, "Alternatives way for small para-rubber farmers as supplemented revenue," 2014.

[43] W. Phromthong, "Factors of Decision Making on New Theory Farming Practices by Farmers in North-Eastern Region of Thailand," 2008.

[44] P. Muangmoon, "Land use changes from expansion of rubber plantation in Chattrakan, Phitsanulok," 2015.

[45] H. Aithal, O. Kumar, and M. R. Narasimhan, "Introducing Complex Software Design in MATLAB via Enterprise Architect," in *Engineering and Technology (ICETECH), 2016 IEEE International Conference on*, 2016, pp. 981-983. https://doi.org/ 10.1109/ICETECH.2016.7569396

[46] C. M. Fonseca and P. J. Fleming, "Multiobjective optimization and multiple constraint handling with evolutionary algorithms. I. A unified formulation," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans,* vol. 28, pp. 26-37, 1998.

[47] A. Vermesan and F. Coenen, *Validation and Verification of Knowledge Based Systems: Theory, Tools and Practice*: Springer Science & Business Media, 2013.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1372-1380 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Evaluating the Effect of Mozart Music and White Noise on Electroencephalography Pattern toward Visual Memory

Syarifah Noor Syakiylla Sayed Daud, Rubita Sudirman*

Faculty of Electrical Engineering, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia

ARTICLE INFO

Article history:
Received: 30 May, 2017
Accepted: 18 July, 2017
Online: 05 August, 2017

Keywords:
Mozart
White noise
Visual memory
Electroencephalography

ABSTRACT

Listening to auditory stimuli during study can give positive and negative influence on human cognitive processing. Thus, it has attracted researchers to conduct studies using various types of auditory stimuli. Some researchers believe that Mozart music and white noise are able to give positive influence on cognitive performance. However, most of the past studies gave more attention towards spatial task. Very little studies have been made on the effect of Mozart music and white noise towards memorizing task. Besides, the effect of these auditory stimuli on task difficulty has also not been studied deeply. Hence, the aims of this study were to investigate the effect of Mozart music and white noise on memory performance with different task difficulty levels; and to propose an effective background stimuli condition for memorization. Experiments have been conducted involving 60 healthy adults that required them to memorize the visual memory task with two difficulty levels; i.e. easy and difficult. Brain signal was recorded during memorization duration using 10-20 electrode placement system of electroencephalography (EEG) machine. EEG is a neurological test for measuring and recording the electrical activity of the brain. The effect of sound stimuli on memory performance was evaluated based on memorization test score and brain activity. The wavelet approach was used in processing the EEG data. Based on the memorizing test score result, the subjects are able to memorize better when listening to white noise compared to Mozart music at different difficulty levels. Listening to auditory stimuli can influence the electroencephalography pattern and brain activity. The level of sensory processing and attention increases when listening to white noise which cause the increase of relative gamma (easy level: p-value = 0.005; difficult level: p-value = 0.007) and beta power (easy level: p-value = 0.001; difficult level: p-value = 0.003). Thus, in this study, it is found that listening to white noise is far more effective in memorizing process compared to Mozart music.

1. Introduction

This paper is an extension of work originally presented in 2016 IEEE EMBS Conference on Biomedical Engineering and Sciences [1]. Listening to background music/sound during performing cognitive tasks can improve the brain functions, develop creative thinking, make people relaxed, relieve mental stress, and increase work efficiency [2]. Working and studying in silence condition for a long time make people easily to feel sleepy, bored, stress, and lost attention. This condition may affect their performance in solving the tasks. Hence, one of the alternatives to make people enjoyable and effective during doing their activities is by listening to auditory stimuli. Auditory stimuli can be defined as events and things that evoke a specific functional reaction in a tissue or organ i.e., the sound that affects our auditory sensations. The auditory stimuli are stimulated when sound cause pressure changes in air or other elastic medium. Usually, the loudness of sound is measured in decibels and travel to the ears as waves. Human ears can detect any sounds within the range of 20 to 20 000 Hz. Then, the detected sounds are converted into neural impulses as a result from transduction process. Deafness and hearing loss can occur if the disruptions happen during transduction process. The sounds can be categorized into many types either comes from man-made or natural. The examples of sounds are birdsong, bell sounds, musical sounds, water sounds, and nightlife sounds. Some of the sounds have a beneficial influence on human emotion, psychology, and

*Corresponding Author: Rubita Sudirman, Faculty of Electrical Engineering, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia
Email: rubita@fke.utm.my

congenital. Thus, many research has been carried out to investigate the effect of sound on human cognition either for children or adult [2-5]. Besides, the development of the biomedical equipment especially the brain modalities i.e. electroencephalography (EEG), magnetic resonance imaging (MRI), computed tomography (CT), magnetoencephalography (MEG), and positron emission tomography (PET) have facilitated the study.

In this globalization era, an education plays the important role for produce a better, independent, successful and knowledgeable people. The knowledge are important and powerful part of human life. Knowledge is an understanding, awareness or information that has been obtained by study or experience by learning, discovering or perceiving, and that is either in a person's mind or possessed by people [6]. Focusing on students life, every day they are learning and expose too much of new knowledge, such as information, facts, descriptions or skills. Examples of student learning activities are memorized the formula, facts, structure of the design, reading, and drawing. All of this requires them to have a better and good memory in order to process the input information. As we know, education system enforced the students to undergo the test, quiz, and examination to identify and evaluate their level of the learning process. The performance of the students in answering the task has depended on their memory ability to interpret the information and knowledge that they have learned and obtained before.

Attention on the student's problem which is some of them is able to recall the required information during the examination, while, some of them are unable to retrieve what they had learned. This situation has relation with their memory capacity to store the information. One of the suspected factors that resulted in this condition is due to wrong chose of study environment. Usually, the students are interested to study in silent or listening to auditory stimuli condition. Thus, this present study aims to help the students by proposing an effective study environment especially for memorizing purpose. The effective study environment is the condition that can give the positive influence on student performance. According to Soderlund et al., 2010 the human cognitive performance can be influenced by the auditory stimuli but depends on the human population, types of auditory stimuli and cognitive task involved [7]. They had done a study to investigate the effect of white noise on inattentive and attentive children when performing episodic verbal free recall test. The test assessment was done in two different conditions: (1) low noise: no noise, and (2) high noise: 78 dB of white noise. Difference influence has been found between these two children groups which are the performance of inattentive children was improved when listened to white noise but not for attentive children. Based on the Soderlund et al., 2010 [7] findings, we are inspired to discover the effect of environmental factors on memory performance for the normal adult population. Knowing from the previous research that the Mozart music and white noise are able to give the positive influence on people's performance, thus, it's were chosen as auditory stimuli in this present study.

Mozart music is a popular auditory stimulus among the researcher because of its highly structured organization which has potential to excite the same cortical firing patterns that used for cognitive processing [8]. It was considered as a highly structured for three different reasons. First, its harmony has approximately 8-bar phrases that separated by definitive cadence points which are easy to distinguish the differences of each section. Besides, the beats were divided equally at a constant tempo. The last reason is it used various types of equipment or voices but often combined into a single. Shaw et al., 1985 reported that the effect of music can be represented by Trion Model [9]. This model suggests that the music has the ability to alter the synaptic weights of neurons in specific patterns due to Hebbian learning principles [8]. In this principle, the brain regions that involve for learning process was explained. They believe that when listening to Mozart music or any music, the neuron firing becomes stronger and each time the new information coming to the memories it's able to process the information actively and decrease the losing of information. White noise can be recognized by 'sh' sound. Some of the researchers reported that listening to it during performing the task able to give a positive influence on cognitive processing [10]. The explanation on how the white noise, improves the cognitive processing represented by the stochastic resonance concept [7]. The simple example to describe this concept is when the weak signal (e.g. visual stimulus) enters the sensory memory it becomes detectable when white noise is added to the signal [7]. The white noise was interacting with the weak stimulus and pushing it or in simple word, it gives motivation to people to give more attention to the required task. Thus, their performance is increase. Stochastic resonance improves the touch, auditory, and visual sensory. The previous works showed that the white noise has improved the human performance in the verbal task, arithmetic task and the spatial task [11-12].

Generally, the human memory can be categorized into three types that are the sensory memory, working memory/short-term memory and long-term memory [13]. The information or stimuli were caught-up by sensory responsiveness and will be stored either in working memory or long-term memory. The human memory easily to be disturbed by external and internal factors such as interference, storage failure, environmental condition, task difficulty, and emotional factors [13-14]. These factors can lead to losing of information. Discovering the effect of sound on memory performance is an interesting work, thus has attracted some researchers to conduct the study [2, 15-16]. The different style of music/sound gives different influence on memory performance as reported by them. In Boyle and Colheart, and Fu and Kuan, study, the subject required to memorize the task under three different condition; no music, listening to gentle music and heavy music. They had found that the gentle music able to give a positive influence on memory performance compare to the others two conditions. The study by Zhang *et al.* in 2009 was purposely to investigate the effect of music familiarity on memorizing task under different difficulty level [2]. Their study used visual working memory task with two difficulty level (easy and difficult) with four different types of background condition (no music, Chinese,

English, and French). The memorizing test score result revealed that the subject had better performance under no noise/silence condition compare to the others. In this present study, the same task assessment as Zhang *et al*., with some modification was used. A visual memory task that consisted image and number is used in order to avoid the effect of language on subject's performance. The number of items that need to remember is reduce from 15 to 10 items only. The memorizing task was involved in the experiment because it is one of the techniques that used during the learning process. This present study preferred to use Mozart music and white noise as auditory background stimuli because limited of the study had investigate its effect on memory performance.

2. Review on Past Studies

This section discuss on the past research that investigate the effect of Mozart music and white noise on human psychology and cognitive performance.

2.1. Mozart music

Mozart music was introduced by Wolfgang Amadeus Mozart in 18 centuries. It is classified as a classical music that played using pianos as the major instrument. Although this music was famous long time ago among the musicians and public, it did not attract the researchers until this 20th century. Rausher is the first person who interested to conduct an experiment to investigate the effect of Mozart music on spatial tasks [17]. They reported that listening to Mozart music for 10 minutes had enhanced the spatial reasoning skills. Due to that, many researchers had conducted the study to prove the findings. However, it's become a controversial issue when some of the researchers unable to prove Rausher et al., findings [17-21]. Nevertheless, the others confirmed that there was a significant difference was found in spatial-temporal performance, such as in pencil-and-paper maze tasks and paper-cutting and folding procedures compare to other environmental condition [22-24]. Up to now, the study on the Mozart effect is still being conducted by researchers as summarized in Table 1. Some of the researchers used the Mozart music as background music for solving or performing certain tasks, whereas some of them used the music alone without any task. Background music can be defined as a condition which the music is played while the listener's primary attention is focused on another task [12] White noise.

Generally, when we talk about the noisy condition, people will imagine the irritating, distracting and uncomfortable sound. The past research has reported that noise attenuates performance on cognitive tasks involving memory, reasoning, vigilance, decision making and vigilance [34-37]. They examine the effect of noise on cognitive performance by using a high volume of noise (85 dB and above). According to Broadbent's listening to the high volume (above 90 dB) of noise can lead to over-arousal, which causes an increase in stress level [34]. They claimed that listening to high-intensity background noise can irritate, disturbs and attenuates vigilance performance [34].

Smith et al., 2003, stated that the effect of noise on cognitive performance is based on three major factors, which is the type of noise, duration, and intensity (loudness) [38]. As an example, when people listening to moderate intensity of white noise (55 dB) during performing a task, there is only a small effect on performance, however, when the moderate intensity of irrelevant speech is used, it disrupted the person's performance [35-36]. Generally, noise can be divided into several types such as ambient noise, pink noise, and white noise. In this study, the white noise is selected as auditory background stimuli for memorizing the task. White noise can be defined as a sound that is artificially created by combining all audible frequencies in equal amounts [40]. According to new research, listening to the 'sh' sound or white noise consistently could boost the memory [41]. The changes in midbrain activity are examined when the people listening to white noise. This brain part is linked to the learning process and reward pathways. Table 2 shows the past research that uses white noise as stimuli.

The gap of knowledge from previous study and contribution of the present study is focused on three major criteria that are:

(i) Selection of auditory background stimuli

Most of the study focused on either one of the sounds, whether using the Mozart music or white noise on experiment except study by Bottiroli et al. that has been discussed before [12]. The intensity of white noise selected in previous research mostly around 70 dB and above. Based on Zhang et al., (2009), they state that the volume greater than 70 dB will increase the heart rate, blood pressure and cause other responses, thus, it does not suitable to listen for a long duration at this intensity level.

Table 1: Past research of Mozart effect on human

Study	Condition and intensity level	Task	Analysis
[25]	Silent and listen to Mozart, 60 dB	Visual oddball	EEG signal
[26]	Listen to Mozart music and Brahms Hungarian, 60-70 dB	Spatio-temporal rotation task	EEG signal and test scores
[27]	Silent and listen to Mozart music (played normally and reversely)	Attentional blink task	Test scores
[28]	Experiment 1: Silent and listening to Mozart Experiment 2: Silent and Chromatic scales	Visual oddball	EEG signal
[29]	Silent and listen to Mozart music	Trigonometry tests	Test scores
[30]	Mozart music	Psychological stress test, tinnitus handicap inventory and visual analog scale	Test scores
[31]	Listening to Mozart music and Koto music, 55 dB	Science quiz	Test scores and hedonicity of music
[32]	Silent and listen to Mozart music, 60-70 dB	-	EEG signal
[32]	Listen to Mozart (pre-music, during music, Post-music), 60-70 dB	-	EEG signal and heart rate variability
[33]	Silent, listening to Mozart music and Beethoven music	-	EEG signal

Table 2: Past research of white noise effect on human

Study	Condition and intensity level	Task	Analysis
[42]	Baseline, recorded ICU noise exposure, ICU noise and mixed-frequency white noise exposure	-	Measures of sleep architecture, total numbers of arousals, arousal indices, noise parameters and sleep quality
[43]	Silent and listen to white noise	Compound remote associates task, Hopkins verbal learning test and Rey-Osterrieth complex figure test	Test scores
[44]	Listen to white noise, novel sound and high frequency tone	Pitch difference between the target and standard stimuli (Easy and hard)	EEG signal
[45]	Silent and listen to white noise	Easy task: verbal task Difficult task: self-performed mini tasks	Test scores
[10]	Experiment 1: Silent and listen to white noise Experiment 2: Silent and listen to white noise at the left, right and both ear sides	Experiment 1: Dichotic listening Experiment 2: Visuospatial working memory task	Test scores
[7]	Silent and listen to white noise	Episodic verbal free recall test	Test scores
[46]	Silent, speech, white noise and Speech+white noise	Visuo-spatial working memory	Test scores
[47]	Silent and Listen to white noise (left side, right side and both sides)	Verbal memory task and visuospatial working memory	Test scores
[48]	Silent, listen to continuous and short bursts of white noise	Oddball task	Test scores
[11]	Listen to white noise	Non-executive and executive function task	Test scores

Contribution 1: The effectiveness between of Mozart music and white noise on memory cognitive task are investigated in this study. The sound intensity is control at medium level (40 to 55 dB) which is safer for long-term exposing to sound especially for study activity.

(ii) Type and difficulty of assessment task

Effect of listening to Mozart music and white noise on task difficulty level (i.e. easy and difficult) does not extremely discover by researchers. Furthermore, the previous research mostly focused on the spatial task ability compares to memorizing process.

Contribution 2: Different difficulty levels (i.e. easy and difficult) of memory task are used in this present study. Thus, the effect of sounds on subject's performance can be investigated either it gives same influence or not for different difficulty level. The memory task is used instead of spatial task because to help the student to find the suitable sound to be listened during memorizing. As we know, memorizing technique is always apply in the study.

(iii) Measurement of memory performance

The effect of sound on memory performance mostly based on cognitive test score. Less of research focused on the effect of the brain activity. Besides, the relation between the score results and activity of the brain was not further discussed and explored.

Contribution 3: The findings of this study are discussed based on two measurements that are memorizing score result and pattern of EEG signal. The EEG machine was used to record the electrical signal of the brain. The activation of the brain rhythm during task-performing are taken account for explaining on how it gives influence on memory ability to store the input information.

3. Theoretical

3.1. Human Brain

Brain is one of the largest, unique and complex organ in the human body that located in the head. The shape of the brain is like a walnut and compose of blood vessels, glial cell and millions of the neuron [49]. It is protected by the cranial bones and the cranial meninges. The weight of the adult brain is approximately three pounds or about 2% of total human's body weight [49]. It is responsible for processing the information, control our behavior, the seat of intelligence and initiator body movement. When receiving input information from the sensory organ, the brain will process the information and then sends the output to muscle. The five sensory responsiveness (touch, smell, visual, auditory, and taste) catch up the stimuli from surrounding and send to the brain to be interpreted before sending to the body for action. The brain can be divided into four main regions that are cerebrum, diencephalon, cerebellum, and brainstem [50]. Each of the regions plays their own roles. The cerebellum can be found behind and below the cerebrum. It is attached to the brain stem. It is responsible for controls motor functions, maintain the body in balance state and interpret the input information from sensory organs [50].

The brain stem is important for basic vital life functions such as controlling and regulation of respiration, cranial nerves, blood pressure, heart rate, breathing, and consciousness in the human body [49]. Besides that, it also functions as a pathway for relaying the information between the brain and the body. It consists of three smaller areas; midbrain, medulla oblongata, and pons. It connects the brain to the spinal cord. Small damage in brain stem areas can cause death. The diencephalon can be found inside the cerebrum. It consists of the thalamus and hypothalamus. It is responsible for maintaining homeostatic balance, sleep cycle, and controlling the motor activities, memories, and arousal. Other than that, it also acts as a relay station for sensory impulses that reach the cerebral cortex, which coming from the brainstem, spinal cord and parts of the cerebral hemisphere. Lastly, the largest region of the brain is

cerebrum or cerebral hemispheres that make up 85 % of human brain weight [49]. The cerebrum separated into the left and right hemisphere by longitudinal fissure. It can be divided into four lobes that are frontal, parietal, temporal and occipital. Each of the lobes represents their own roles. Table 3 describes the function of each brain lobe.

Table 3: Brain lobe functions [49]

Region	Descriptions
Frontal	Controls voluntary muscular functions, aggression, moods, motivation and smell reception
Parietal	Evaluate sensory information of pain, touch, taste, temperature and balance
Temporal	Center for abstract thoughts and judgment decisions, involve in memory processes and evaluates hearing input and smell
Occipital	Receiving and interpreting visual input

In this present study, the cerebral region is interested in evaluating the effect of mental task and sound stimuli on it. The study of the brain has started a long time ago among the scientist and philosophers in order to know deeply how does it work, factors affecting and it functions. Due to the lack of advance and modern modalities/equipment had limited the research. In 19 to 20 centuries, the brain imaging modalities are started introducing by scientists. Thus, it attracts many researchers to conduct the study of the brain.

3.2. Electroencephalography

Electroencephalography has been introduced by Hans Berger in 1920. The scientist believes that human brain generates electrical signal starting between the 17th and 23rd week of prenatal development and throughout life [51]. When the central nervous system is complete and functional there are approximately 10^{11} neurons developing at birth. This assumption has motivated the researchers to apply various type of brain modality in order to know deeply about the brain. Electroencephalography recorded the brain activities by reading the signal that picks up using metal electrode and conductive gel of the scalp. The electrical potential (EEG signal) is generated by the nerve cell of the brain in the cerebral cortex region.

The EEG comes from a combination of three words, which is electro (referring to the registration of brain electrical activity), encephalon (referring to emitting of signals from head) and gram/graphy (referring to writing or drawing) [51]. In research works, the electroencephalography is widely used by researchers compare to other brain modalities due to its ability to record the signal in a short time (milliseconds), low cost, non-invasive technique, not harmless and painful to research subjects.

4. Experimental

The experiment has been conducted at the laboratory with controlled temperature, lighting and auditory intensity level. All equipment and materials such as Neurofax 9200 of EEG acquisition machine, ECI electrode-cap, body harness, ear electrodes, ECI electrode-gel, EEG electrode paste, blunted needle, wet tissue, speaker, laptop, pen and answer sheets were prepared and provided before the experiment started. When the subject entered the laboratory room, they were asked to sit on the chair and took a rested for 5 minutes. The instruction and flow of the experiment was displayed on the laptop screen. Then, the subject was required to fill-up the consent form and survey questionnaire. Subject was prepared with all the EEG equipment and material for recording their brain signal. The simulation program was started when the subject ready to take the experiment. Subject need to undergo three different conditions while memorizing the visual working memory task. The subjects were divided equally into two groups to counterbalance the effect of environmental condition on subject performance [2, 12, 26]. Groups 1 were memorized under silence, listened to Mozart music and listened to white noise condition, whereas group 2 undergo silence, listened to white noise and listened to Mozart music condition. The sound intensity level was controlled at 40 to 55 dB to reduce the effect of loudness on them [52]. Sixty healthy (40 females, 20 males) subjects from Universiti Teknologi Malaysia at the age between 19 to 26 years old with no musical talents were involved in this study. They were selected based on mini-mental state examination (MMSE) to determine their mental conditions. Only subjects that passed the MMSE test were undergo the experiment. The flow of task require to be followed by subjects were as stated below:

Condition 1: Silent

- Memorize easy level task for 2 minutes and brain signals are record at this time.
- Rest for 30 seconds.
- Recall the task for 2 minutes.
- Rest for 1 minutes.
- Memorize difficult level task for 2 minutes and brain signals are record at this time.
- Rest for 30 seconds.
- Recall the task for 2 minutes.
- Rest while look up the scenery pictures for 3 minutes before proceed to next environmental condition.

Condition 2 and 3: Listen to auditory stimuli

- Same flows as condition 1 are used for condition 2 and 3 except during the memorizing time the subjects require to listen to the auditory stimuli.

5. Data Processing and Analysis

The brain signals were recorded during the memorizing period using EEG machine. The raw EEG signals were saved in ASCII format before processed via Matlab software. Usually, the raw EEG signals were composed of the noise signals that come from the muscle movement (high-frequency noise) and eyes blinking noise (low-frequency noise). Thus, the raw EEG signal needs to be pre-processed for filter the unwanted signal. This present study preferred to use the stationary wavelet transform (SWT) method to remove the high frequency and low frequency of noise. The db3 mother wavelet with 5 decomposition level was chosen to remove the high-frequency noise and low-frequency noise from EEG signals [53-55]. Before filtered the noises, there were 10 out of 20 EEG channels was selected to determine the effect of auditory stimuli on visual memory. The selected channels were frontal: Fp1, Fz, F8; temporal: T3, T4, T5, T6; parietal: Pz and occipital: O1; O2. The clean EEG signals were process again using db4 mother

wavelet with 7 decomposition level of discrete wavelet transform (DWT) method for obtained alpha, beta, theta, and gamma rhythm. The relative power of each frequency band was calculated.

6. Result and Discussion

This part is discuss on the results obtain from the assessment test and EEG pattern. The test scores are determine based on the correct response that answer by the subject. Meanwhile, the EEG features are extract using wavelet transform as mentioned earlier. Statistical analysis of sign test and paired samples *t*-test are used to determine the significant difference between control condition and test condition. The results has been tabulated and graphed in the following section.

6.1. Assessment test score

The answered from assessment test were only considered as correct when the filled numbers on figure was same as those were presented. The sign test with 0.05 of alpha (α) level was used to compare the median differences of correct responded between auditory stimuli condition relative to control condition. Two hypothesis has been set as below:

Null hypothesis, H_0: *p*-value > α; the median value of test score does not shows a significant different between the test condition and control condition.

Alternative hypothesis, H_1: *p*-value < α; the median value of test score shows a significant different between the test condition and control condition.

According to Table 4 and Figure 1, the subjects were able to memorize better when the white noise played for easy (mean: 8.561) and difficult (mean: 4.228) level of the task. A significant difference have been found for median test score of white noise (*p*-value: easy level = 0.009; difficult level = 0.01) relative to control condition. Although listening to Mozart music also can improve subject visual memory but it does not give bigger influence on their test score when comparing with the control condition, since no significant difference (p-value: easy level = 0.203; difficult level = 0.061) was found from the sign-test analysis. Looking at these three conditions it revealed that the subjects' memory ability was improved when listening to auditory stimuli compared to silence condition. This happened because of the positive effect of auditory stimuli on subject's mood, arousal, attention and alertness level, and sensory input processing. Bottiroli et al., 2014, has reported that listening to white noise and Mozart music during performing the mental task will affect the arousal (degree of physiological activation), mood (long lasting emotions), and listener's enjoyment which in turn influence the memory performance [12].

Table 4: Mean and *p*-value of test score result

	Easy Level		Difficult Level	
Condition	**Mean**	***p*-value**	**Mean**	***p*-value**
Control	7.281		2.912	
White Noise	8.561	0.009*	4.228	0.010*
Mozart Music	8.070	0.203	3.632	0.061

Considering the effect of task difficulty level on subjects performance, it showed that the increasing of items need to be remember had degraded the ability to memorize. The easy level of visual memory task gave better result compared to difficult task. Subjects were able to memorize better in easy level compared to the difficult level due to the number of items that need to be remembered were different for each level. In easy levels, the subjects were required to remember 2 digits number with pictures only, but for difficult task, they need to memorize the pictures with 4 digits number. In the control condition, the percentage decreased of test score for easy level relative to difficult level was about 60%. Meanwhile, 50.6% and 54.6% were found for white noise and Mozart music. The higher percentage decreased was showed by control condition compared to auditory stimuli condition. From here, it indicated that the auditory stimuli were also improved the memory performance when the difficulty level was increased, especially the white noise.

6.2. Electroencephalography pattern

The frequency domain features were extracted from the EEG signal to determine its characteristics when induced by auditory stimuli during memorizing test. The paired samples *t*-test analysis with α = 0.05 was used to identify the significant different of the relative rhythm power of auditory stimuli condition relative to control condition. Besides, the increasing and decreasing trend of the relative rhythm power of auditory stimuli condition toward control condition also discussed. The hypothesis of paired samples *t*-test is set as below:

Null hypothesis, H_0: *p*-value > α; the mean value of relative rhythm power does not shows a significant different between the test condition and control condition.

Alternative hypothesis, H_1: *p*-value < α; the mean value of relative rhythm power shows a significant different between the test condition and control condition.

Relative gamma power

Gamma rhythm (< 30 Hz) is associated with the movement, sensory processing or solving high mental task [51]. In this case study, we are focused on the role of gamma rhythm in activating sensory processing in order to fulfill the study objectives. Sensory processing is an ability of the brain to receive and response to the coming information from sense organ. The sense organs are connected to the human brain via the peripheral nervous system which allowing us to process the input information. Generally, human have five dominant sense organs which are eyes, ears, nose, tongue and skin that use for see, hear, smell, taste and touch. This sense organs contain receptor that transport the input information through the sensory neurons to the nervous system. The ability of the brain in processing the input information can be observed through the relative gamma power value.

Based on Table 5 and Table 6, the white noise (easy level: 67.522 and 65.925) had the highest relative gamma power with significance difference of easy level: *p*-value = 0.005; and difficult level: *p*-value = 0.007) compared to Mozart music and control condition. However, the Mozart music also showed

significant difference on relative gamma power for difficult task (p-value: 0.003). The relative gamma power was increased for auditory stimuli condition because of the brain need to process information that caught up by the eyes and ears. When listening to Mozart music and white noise, the brain is not only process the upcoming information from the visual but also the auditory stimuli. Different from control condition which is no sound is provided, thus the brain only need to process the information from the visual only. That is the reason why the control condition had lower relative gamma power compared to auditory stimuli conditions.

Next, comparing on difficulty level, it showed that the relative gamma power of control and white noise condition was decreased for difficult task relative to control condition. Meanwhile, the Mozart music showed the highest relative gamma power for a difficult task (64.282) compared to easy task (63.628). So, it indicated that the increases of task difficulty level for Mozart music condition will increase the sensory processing. But, for control and white noise condition, the sensory processing was decreased when the task difficulty level increased.

Relative beta power

Beta rhythm (13-30 Hz) is associated with the attention/concentration level, active thinking, focus on the outside world or solving concrete problems [51]. This study focused on the relative beta power for determining the attention/concentration and active thinking level of subjects during memorizing. Based on Table 5, it shows that the relative beta power was the highest in white noise condition for easy (117.318) and difficult (121.218) level of the task. It means that the subjects attention/concentration and active thinking level were significant improved when listening to the white noise during remembering the items as indicated in Table 6 by paired samples t-test analysis relative to control condition (easy level: p-value = 0.001; difficult level: p-value = 0.003). However, the attention/concentration and thinking level was significantly improve for the easy level (p-value = 0.004) but not for the difficult task (p-value = 0.141) when listening to Mozart music relative to control condition. Among these conditions, the findings showed that listening to Mozart music and white noise able to increase the attention/concentration and thinking level compared to silence condition.

The higher and good attention/concentration level plays an important role when working memory task was involved [56]. In this case study, the subject was only given 2 minutes to memorize all the items, so they need to give full attention on the task in order to memorize it. If they unable to focus on the task during the memorizing time, they will have problem to retrieve back the information later. That is why it is really important to have high beta rhythm during the memorizing process for getting a good test score result. As seen from the result that listening to auditory stimuli makes the subject more attentive and concentrate on the task compared to the silent condition.

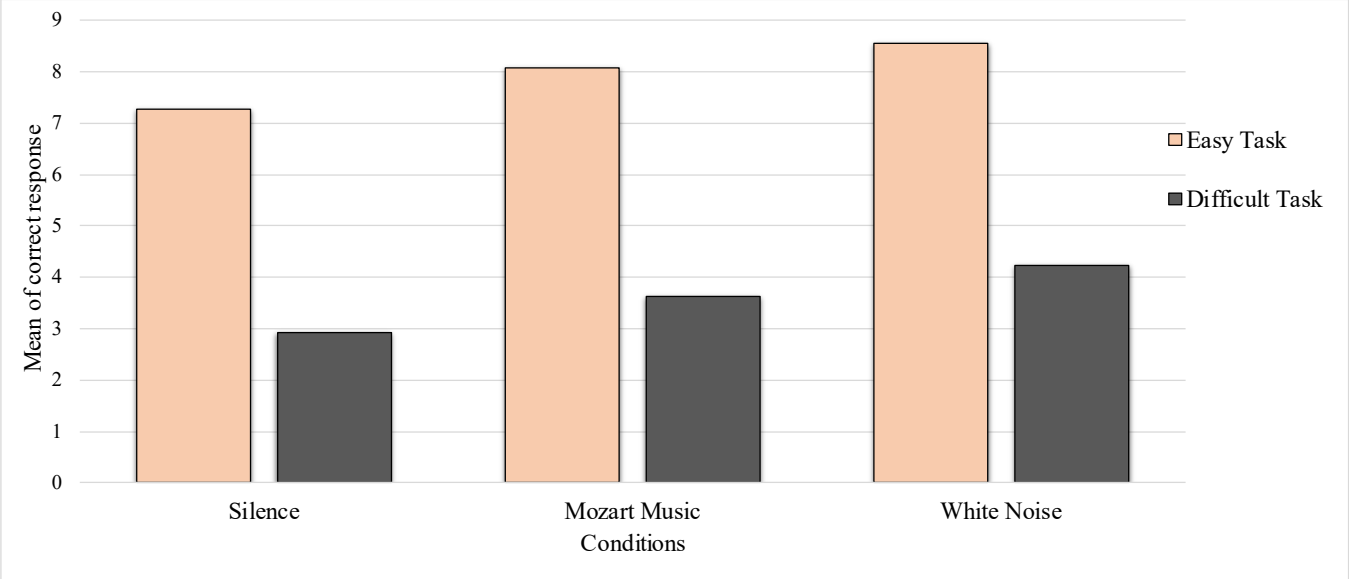


Figure 1: Graph of mean test score at silence, Mozart music and white noise condition

Table 5: Relative power of rhythms for easy and difficult task at silence, listen to Mozart music and white noise condition

	Easy Level			Difficult Level		
	Silence	**Mozart music**	**White noise**	**Silence**	**Mozart music**	**White noise**
Gamma	60.449	63.628	67.522	59.193	64.282	65.925
Beta	111.032	115.640	117.318	113.166	115.842	121.218
Alpha	165.469	167.464	166.372	168.520	169.526	165.092
Theta	225.215	222.400	224.511	227.440	224.280	226.487

Table 6: The *p*-value of paired sample t-test analysis for relative rhythm power of auditory stimuli condition relative to control condition

	Easy level		Difficult level	
	Mozart music	**White noise**	**Mozart music**	**White noise**
Gamma	0.124	0.005*	0.014*	0.007*
Beta	0.004*	0.001*	0.141	0.003*
Alpha	0.257	0.607	0.597	0.068
Theta	0.385	0.820	0.371	0.738

The relative beta power was higher for difficult task compared to easy task for all conditions. The reason was due to subject require giving more attention/concentration level during memorizing the difficult task which will increase the relative beta power. It can be concluded that the more items need to remember, the more attention/concentration level was required.

Relative alpha and theta power

Alpha and theta rhythm associated with the human emotion, mood and arousal state [51]. The frequency of 8-13 Hz (alpha rhythm) represents the relaxed awareness without attention and concentration level. Meanwhile the frequency of 13-30 Hz (beta rhythm) represents the deep meditation, access to unconscious material, and creative inspiration. According to the paired samples *t*-test result in Table 6, there was no significant difference found for relative alpha and theta conditions of auditory stimuli condition relative to control condition. It showed that no significant influence was found on alpha and theta rhythm when listening to white noise and Mozart music. However, the relative alpha power was the highest when in Mozart music condition for easy level (167.464) and difficult level (169.526) of task compared to others (see Table 5). Listening to Mozart music gave a little influence on subject relaxation state but does not give bigger impact on them since the relative alpha rhythm does not have much difference relative to control condition. The relative theta power showed highest value for control condition for all levels. It means that a little influence was found on theta rhythm roles but not too much when compared to auditory stimuli conditions. From the findings, it revealed that listening to the white noise and Mozart music during memorizing does not makes the subject to feel stress and distract.

7. Conclusion

The Mozart music and white noise has different influence on memory performance due to the different types of harmony and beats of the auditory stimuli. White noise has approximately constant harmony and beat, whereas Mozart music has slow, fast, low and high of harmony and beat. Listening to the auditory background stimuli during memorizing the task is able to increase the attention, concentration and thinking level due to the activation of the brain neuron when exposed to it. Before selecting the auditory stimuli to be listened during the learning process, the volume intensity level, type of auditory stimuli, duration, and type of cognitive task should be considered to get positive influence on memory performance. Therefore, in conclusion, this study suggested that the student should listen to white noise at a moderate intensity (40 to 55 dB) of volume during the memorizing time. The limitation of the previous study has been overcome in this present works in terms of types of task and selection of sound. It has been found that the white noise is more effective to be listened, compared to Mozart music for memory and recall ability of visual task. Therefore, the findings in this study same as the expected outcome and set in the objectives of the study.

a) The auditory background stimuli and task difficulty give different influence on the ability of subject in memorizing the visual task and electroencephalography pattern.
b) The increases of relative gamma and relative beta power has increase the level of attention, thinking, and sensory processing. No significant influence was found for relative alpha and theta rhythm for auditory stimuli condition relative to control condition.
c) White noise is effective to listen during memorizing visual working memory task due to it able to improve the memory performance.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The authors would like to express their appreciation to University Teknologi Malaysia for providing the facilities and funding for this study under vot Q.J130000.2523.13H25.

References

[1] S. S. Daud, R. Sudirman, "Discovering Sound Effect on Visual Memory Performance based on Electroencephalography" in IEEE EMBS Conference on Biomedical Engineering and Sciences, Kuala Lumpur Malaysia, 2016.

[2] X. Zhang, L. Chuchu, Z. Jing, M. Xiyu, "A Study of Different Background Language Songs on Memory Task Performance" in IEEE International Symposium on Intelligent Ubiquitous Computing and Education, Chengdu China, 2009.

[3] S. Nishifuji, M. Sato, D. Maino, S. Tanaka, S, "Effect of Acoustic Stimuli and Mental Task on Alpha, Beta and Gamma Rhythms in Brain Wave" in IEEE SICE Annual Conference, Taipei Taiwan, 2010.

[4] T. Tamesue, H. Kamijo, H. Itoh, I. Kazunori, I, "Quantitative Evaluation using EEG for Influence of Meaningful or Meaningless Noise on Participants During Mental Tasks" in IEEE International Symposium on Advanced Intelligent Systems, Kobe Japan, 2012.

[5] B. Geethanjali, K. Adalarasu, R. Rajsekaran, R, "Impact of Music on Brain Function during Mental Task Using Electroencephalography" World Academy of Science, Engineering and Technology, 66, 883–887, 2012.

[6] B. Russell, Human Knowledge: Its Scope and Value, Routledge, 2013.

[7] G. Soderlund, S. Sikstrom, J. M. Loftesnes, E. J. Sonuga-Barke, "The Effects of Background White Noise on Memory Performance in Inattentive School Children" Behavioral and Brain Functions, **6**(55), 1-10, 2010.

[8] J. Konner, "Does Music Predictability Improve Spatial-Temporal Reasoning? Reexamining the Mozart Effect," Bachelor Thesis, University of Colorado Boulder, 2013.

[9] G. L. Shaw, D. J. Silverman, J. C. Pearson, "Model of Cortical Organization Embodying a Basis for a Theory of Information Processing and Memory Recall" Biophysics, **82**(8), 2364-2368, 1985.

[10] G. Soderlund, E. Marklund, F. Lacerda, "Auditory White Noise Enhances Cognitive Performance Under Certain Conditions: Examples from Visuo-Spatial Working Memory and Dichotic Listening Tasks" in Fonetik Swedish Phonetics Conference, Stockholm University Sweden, 2009.

[11] S. K. Helps, S. Bamford, E. J. Sonuga-Barke, G. B. W. Soderlund, "Different Effects of Adding White Noise on Cognitive Performance of Sub-, Normal and Super-Attentive School Children" PLoS ONE, **9**(11), 1-10, 2014.

[12] S. Bottiroli, A. Rosi, R. Russo, T. Vecchi, E. Cavallini, "The Cognitive Effects of Listening to Background Music on Older Adults: Processing Speed Improves with Upbeat Music, While Memory Seems to Benefit from Both Upbeat and Downbeat Music" Frontiers in Aging Neuroscience, **284**(6) 1-7, 2014.

[13] J. Henderson, "Memory and Forgetting" Routledge. 2005.

[14] N. Cowan, "What Are the Differences between Long-Term, Short-Term, and Working Memory?" Progress in Brain Research, 169, 323-338, 2008.

[15] R. Boyle, V. Coltheart, "Effects of Irrelevant Sounds on Phonological Coding in Reading Comprehension and Short-term Memory" The Quarterly Journal of Experimental Psychology, 9, 398-416, 1996.

[16] H. F. Fu, T. M. Kuan, "Under Different Conditions of Learning Memory in the Electroencephalography (EEG) Analysis and Discussion" in IEEE International Conference on Power Electronics and Intelligent Transportation System, Shenzhen China, 352-355, 2009.

[17] F. H. Rausher, G. L. Shaw, "Music and Spatial Task Performance" Nature, **6447**(365), 611, 1993.

[18] J. Newman, J. H. Rosenbach, K. L. Burns, H. R. Matocha, E. R. Vogt, "An Experimental Test of "The Mozart Effect": Does Listening to His Music Improve Spatial Ability" Perceptual and Motor Skills, 81, 1379-1387, 1995.

[19] K. M. Steele, T. N. Ball, R, Runk, "Listening to Mozart Does Not Enhance Backwards Digit Span Performance" Perceptual and Motor Skills, **84**(3), 1179-1184, 1997.

[20] C. F. Chabris, "Prelude or Requim for the Mozart Effect?" Nature, 400, 826-827, 1999.

[21] K. M. Steele, J. D. Brown, J. A. Stoecker, "Failure to Confirm the Rausher and Shaw Description of Recovery of the Mozart Effect" Perceptual and Motor Skills, **88**(3), 843-848, 1999.

[22] F. H. Rauscher, G. L. Shaw, K. N. Ky, "Listening to Mozart Enhances Spatial-Temporal Reasoning: Towards a Neurophysiological Basis" Neuroscience Letters, **185**(1), 44–47, 1995.

[23] B. E. Rideout, C. M. Laubach, "EEG Correlates of Enhanced Spatial Performance Following Exposure to Music" Perceptual and Motor Skills, **82**(2), 427-432, 1996.

[24] E. R. Bruce, S, Dougherty, L, Wernert, L. "Effect of Music on Spatial Performance: A Test of Generality" Perceptual and Motor Skills, 86, 512–514, 1998.

[25] N. Jausovec, K. Habe, K. "The Influence of Auditory Background Stimulation (Mozart's Sonata K. 448) on Visual Brain Activity" International Journal of Psychophysiology, **51**(3), 261-271, 2004.

[26] N. Jausovec, K. Jausovec, I. Gerlic, "The Influence of Mozart music on Brain Activity in the Process of Learning" Clinical Neurophysiology, **117**(12), 2703-2714, 2006.

[27] C. Ho, O. Mason, C. Spence, "An Investigation into the Temporal Dimension of the Mozart Effect: Evidence from the Attentional Blink Task" Acta Psychologica, **125**(1), 117-128, 2007.

[28] W. Zhu, L. Zhao, J. Zhang, X. Ding, H. Liu, E. Ni, Y. Ma, C. Zhou, "The Influence of Mozart's Sonata K.448 on Visual Attention: An ERPs Study" Neuroscience Letters, **434**(1), 35-40, 2008.

[29] J. M. Taylor, B. J. Rowe, B. J. "The Mozart Effect and the Mathematical Connection" Journal of College Reading and Learning, **42**(2), 51-66, 2012.

[30] G. Attanasio, G. Cartocci, E. Covelli, E. Ambrosetti, V. Martinelli, M. Zaccone, M. Cacciafesta, "The Mozart Effect in Patients Suffering from Tinnitus" Acta Oto-Laryngologica, **132**(11), 1172–1177, 2012.

[31] L. Perlovsky, A. Cabanac, M. C. Bonniot-Cabanac, M. Cabanac, "Mozart Effect, Cognitive Dissonance, and the Pleasure of Music" Behavioural Brain Research, 244, 9-14, 2013.

[32] L. C. Lin, C. S. Ouyang, C. T. Chiang, R. C. Wu, H. C. Wu, R. C. Yang, "Listening to Mozart K.448 Decreases Electroencephalography Oscillatory Power Associated with an Increase in Sympathetic Tone in Adults: A Post-Intervention Study" Journal of The Royal Society of Medicine, **5**(10), 1-7, 2014.

[33] W. Verrusio, E, Ettorre, E. Vicenzini, N. Vanacore, M. Cacciafesta, O. Mecarelli, "The Mozart Effect: A Quantitative EEG Study. Consciousness and Cognition, 35, 150-155, 2015.

[34] D. E. Broadbent, E. A. J. Little, "Effects of Noise Reduction in a Work Situation" Occupational Psychology, 34, 133–140, 1960.

[35] A. P. Smith, "A Review of the Effects on Noise on Human Performance" Scandinavian Journal of Psychology, **30**(3), 185-206, 1989.

[36] E. Salas, J. E. Driskell, S. Hughes, Stress and Human Performance, Lawrence Erlbaum, 1996.

[37] S. P. Banbury, W. J. Macken, S. Tremblay, D. M. Jones, "Auditory Distraction and Short-Term Memory: Phenomena and Practical Implication" Human Factors, **43**(1), 12-29, 2001.

[38] D. G. Smith, J. V. Baranski, M. M. Thompson, S. M. Abel, "The Effects of Background Noise on Cognitive Performance during a 70 Hour Simulation of Conditions Aboard the International Space Station" Noise Health, **6**(21), 3-16, 2003.

[39] D. C. L. Compte, "Extending the Irrelevant Speech beyond Serial Recall" Journal of Experimental Psychology Learning, **20**(6), 1396-1408, 1994.

[40] R. Mehta, R. J. Zhu, A. Cheema, "Is Noise Always Bad? Exploring the Effects of Ambient Noise on Creative Cognition" Journal of Consumer Research. **39**(4), 784-799, 2012.

[41] H. R. Vanessa, E. M. Bauch, N. Bunzech, "White Noise Improves Learning by Modulating Activity in Dopaminergic Midbrain Regions and Right Superior Temporal Sulcus" Journal of Cognitive Neuroscience, **26**(7), 1469-1480, 2014.

[42] M. L. Stanchina, M. Abu-Hijleh, B. K. Chaudhry, C. C. Carlisle, C. C. R. P. Millman, "The Influence of White Noise on Sleep in Subjects Exposed to ICU Noise" Sleep Medicine, **6**(5), 423–428, 2005.

[43] A. Hillier, J. K. Alexander, D. Q. Beversdorf, "The Effect of Auditory Stressors on Cognitive Flexibility" Neurocase: Case Studies in Neuropsychology, Neuropsychiatry, and Behavioural Neurology, **12**(4): 228-231, 2006.

[44] L. A. Combs, J. Polich, J. "P3a from Auditory White Noise Stimuli" Clinical Neurophysiology, **117**(5), 1106–1112, 2006.

[45] G. Soderlund, S. Sikstrom, A. Smart, A. "Listen to the Noise: Noise is Beneficial for Cognitive Performance in ADHD" Journal of Child Psychology Psychiatry, **48**(8), 840-847, 2007.

[46] G. Soderlund, S. Sikstrom, J. M. Loftesnes, E. J. Sonuga-Barke, "The Effects of Background White Noise on Memory Performance in Inattentive School Children" Behavioral and Brain Functions, 6(55), 1-10, 2010.

[47] S. Flodin, E. Hagberg, E. Persson, L. Sandbacka, S. Sikstrom, G. B. W. Soderlund, "Lateralization Effects of Auditory White Noise on Verbal and Visuo-Spatial Memory Performance" in Swedish Phonetics Conference, Gothenburg Swedish, 2012.

[48] E. Schroder, "The Effect of Auditory White Noise on a Three- Stimulus Oddball Task in Attentive and Inattentive Participants" Master Thesis, University of Lund, 2013.

[49] D. Rizzo, Introduction to Anatomy and Physiology, Cengage Learning, 2012.

[50] M. S. Sweeney, Brain: The Complete Mind : How It Develops, How It Works, and How to Keep It Sharp, National Geographic, 2009.

[51] S. S. Chambers, A. Jonathan, EEG Signal Processing, Wiley, 2007.

[52] C. Sun, Y. Bao, J. Xu, D. Kong, H. Zhou, Q. Wang, H. Shang, W. Wang, M. Jin, X. Wang, Y. Duan, "The Effects of Different Types of Music on Electroencephalogram" in IEEE International Conference on Bioinformatics and Biomedicine, Shanghai China, 2013.

[53] S. S. Daud, R. Sudirman, "Butterworth Bandpass and Stationary Wavelet Transform Filter Comparison for Electroencephalography Signal" in IEEE International Conference on Intelligence Systems, Modelling and Simulation, Kuala Lumpur Malaysia, 2015.

[54] S. S. Daud, R. Sudirman, "Decomposition Level Comparison of Stationary Wavelet Transform Filter for Visual Task Electroencephalogram" Jurnal Teknologi, 74, 7-13, 2015.

[55] S. S. Daud, R. Sudirman, "Artifact Removal and Brain Rhythm Decomposition for EEG signal Using Wavelet Approach" Jurnal Teknologi, 78, 135-143, 2016.

[56] M. Matsukura, S. J. Luck, S. P. Vecera, "Attention Effects during Visual Short-Term Memory Maintenance: Protection or Prioritization?" Perception and psychophysics, **69**(8), 1422-1434, 2007.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1381-1389 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

A Way for Measuring the Temperature Transients of Capacitors

Zoltan Sarkany[*1], Marta Rencz[1, 2]

[1]*Mentor Graphics, MAD, Budapest, Hungary*

[2]*Budapest University of Technology and Economics, Department of Electron Devices, Budapest, Hungary*

ARTICLE INFO

Article history:
Received: 30 May, 2017
Accepted: 16 July, 2017
Online: 05 August, 2017

Keywords:
Thermal transient measurement
Switched capacitor
Simulation model calibration

ABSTRACT

With the increasing integration level of electronic circuits management of the generated heat became one of the most important design aspects. Beside the semiconductor components capacitors are also affected by the elevated temperature. In this paper a new thermal characterization method is proposed adopting the thermal transient measurement technique for capacitors utilizing the capacitance itself as temperature dependent parameter. The proposed switched capacitor based circuit generates a signal proportional to the capacitance of the component and hence the temperature. The measurement method is demonstrated on self and transfer resistance measurement setups and also verified using CFD simulations. It is shown that the captured thermal transient curve can even be used to fine tune the inaccurate simulation model parameters achieving a more accurate calibrated 3D simulation model.

1. Introduction

The thermal design of electronic systems received high focus in the last decades as the increasing level of integration increases also the power densities that the designers need to face with. The primary sources of heat are often the semiconductor components, and hence there is extensive research and development on how to construct new components with lower thermal resistance, how to arrange the components appropriately in the system and how to find new cooling solutions in order to avoid overheating.

However, not only the semiconductor components suffer from the increased power densities but also the discrete components like capacitors. The elevated temperature can have an immediate effect on their electric parameters. Changing capacitance and dissipation factor values and increased leakage current usually affect negatively the performance of the overall system. In addition there are also long term consequences of the elevated operation temperature, namely the reliability of the component usually decreases as a consequence.

The operating temperature of the capacitors, just like that of other components depends on the environment, i.e. on one hand on the heating of the surrounding components, on the other hand on the internal heat generation of the device itself. If there is a significant power loss on the capacitor then its internal core temperature can even be significantly higher than the circuit board that is hosting it. For this reason there is an increasingly growing demand towards developing a methodology to measure the thermal behavior and the thermal properties of capacitors.

In order to be able to take these effects into account the thermal resistance between the surrounding components and the capacitor (the transfer resistances) and the thermal resistance between the region of the capacitor where the heat is generated and the environment (the self-resistance) should all be characterized

For thermal engineering of discrete capacitors simulation is the primary method of investigation and characterization [1] although analytical calculations based on simplified models are sometimes also provided in application notes [2-4]. Experimental results are only rarely mentioned in relevant papers or datasheets, although the numerical results should always be verified using experiments.

Key problem is that the core temperature of the capacitors is hard to accurately measure in-situ. Most of the published measurement solutions utilize embedded thermocouples [5] or use infrared temperature sensors focused on a hole drilled in the casing of the component. But all these methods do modifications in the device structure. The published nondestructive methods simply measure the surface temperature only and use approximate calculations to guess the internal temperature of the capacitor. The common issue of all these measurement solutions is the high uncertainty and low reproducibility because of the badly defined sensor placement.

This paper is an extension of the work originally presented in EPTC 2016 [6].

[*]Corresponding Author: Zoltan Sarkany, Mentor Graphics, MAD, Budapest, Hungary | Email: sarkanyz@gmail.com

2. Electrical test method

In case of semiconductor components the characterization of the thermal impedance may occur either by using some steady state solutions or more commonly by using the thermal transient measurement [7] method. Although in some cases the junction temperature can be measured using infrared imaging as well, the most commonly used solution is the electrical test method for its simplicity and accuracy.

In case of the electrical test method the inner temperature of the component is measured by utilizing a temperature dependent electrical parameter of the chip itself, like the forward voltage of a constant current biased p-n junction or the threshold voltage of a MOSFET device. This makes the thermal characterization of these components quick and nondestructive and also enables the measurement of the transient response in order to characterize the time domain behavior of the device besides the thermal resistance.

A thermal transient measurement gives information not only about the thermal resistance of the components, but as the time constants of the cooling depends on the building materials and the geometry of the device, it can also reveal structural information.

The output of the thermal transient measurement is usually a cooling curve, showing the inside temperature of the device, continuously from a hot steady state temperature to a cool steady state temperature. Some basic thermal parameters can be calculated from this curve directly. As the heating power is known (P) the junction to ambient thermal resistance can easily be calculated from the peak temperature read at the beginning of the cooling curve (T_J) and the environment temperature that is practically equal to the junction temperature after it has cooled down (T_{Env}):

$$R_{th} = \frac{T_J - T_{Env}}{P} \quad (1)$$

Utilizing the whole captured thermal transient data more sophisticated measurements can be implemented using the transient dual interface method as it is described by the JEDEC JESD51-14 measurement standard.

Finally with appropriate mathematical transformations [8] the time constants of the cooling curve can be identified and a one dimensional equivalent RC network can be constructed, where the thermal resistances and thermal capacitances can be associated to the different physical features in the structure.

In order to do thermal transient measurement three mandatory criteria has to be fulfilled:

- An appropriate temperature sensitive electric parameter (TSP) must be found.
- A known heating power must be applied on the component to elevate its temperature.
- The heating power must be switched off abruptly and the transient response of the TSP has to be captured.

If a measurement setup can be constructed that fulfills the above three criteria, the same methodology could be applied for the thermal characterization of discrete capacitor components. In the following sections a possible implementation of the thermal transient measurement for capacitor components will be discussed.

2.1. Temperature measurement

The capacitor temperature can also be measured with an electrical measurement but it needs to be considered that selecting the appropriate temperature dependent electric parameter is a key for the good measurement results. Most of the electric parameters of the capacitors are typically temperature dependent e.g. capacitance, dissipation factor (tanδ) and leakage current.

Any of the electric parameters of a capacitor could be selected for temperature sensing, but it always needs to be considered that unlike in the case of semiconductor chips, where both the dissipation and the temperature dependent electric parameter are well localized, in case of capacitors some parameters might depend on the temperature of several different sections of the structure that not necessarily have the same temperature. In such cases a weighted average temperature is perceived that may or may not be representative to the real device temperature.

The construction of the various capacitor types (electrolytic, tantalum, ceramic and foil capacitors) are different and accordingly not only the electrical, but their thermal behavior is also different.

Due to the significant temperature dependence of the electrochemical properties of the electrolytes, the electrolytic capacitors are the most affected by the temperature [9]. The increasing component temperature decreases the viscosity of the fluid leading to higher charge transfer capability. As a result the equivalent serial resistance (ESR) the impedance and the dissipation factor all decrease. The leakage current however shows significant increase with the temperature. Manufacturer datasheets reported even an order of magnitude change in the leakage current for 100°C temperature change. The capacitance also shows high temperature dependence, but the breakdown voltage and the serial inductance are not significantly affected by the temperature.

The tantalum electrolytic capacitors show similar thermal behavior, but the temperature dependence of the capacitance is lower. The ESR also decreases with the temperature; this dependence is reverse proportional to the ion concentration in the electrolyte. As a result tantalum capacitors with higher ESR show also higher temperature dependence. The leakage current is significantly dependent on the temperature.

The temperature dependence of ceramic and foil capacitors are both primarily defined by the type of dielectric material, i. e. the polymer and ceramics type. The higher permittivity materials show higher temperature dependence. The Calss-1 ($\varepsilon<200$), capacitors show limited or no temperature dependence, while Class-2 ($\varepsilon \approx$ 200 to 10000) devices show high temperature dependence, often not linear or not even monotonous.

Finally the capacitance of the foil capacitors show either positive (Polyester) or negative (Polystyrene and Polypropylene) temperature dependence.

2.2. Capacitance as a temperature dependent parameter

It is visible that there are numerous options for the selection of the temperature dependent parameter, but we wanted to select one that can be used for various capacitor types. We selected the most obvious parameter that almost always shows significant temperature dependence, this is the capacitance.

The temperature dependence of the capacitance is mainly based on the temperature dependence of the dielectric material

used in the structure, hence its location is well defined in the component. It has to be noted however that due to the relatively low thermal conductivity of the materials in the capacitor relatively high temperature gradients can arise inside the device, thus the measured temperature is an average value.

As a demonstration of the concept a couple of different capacitors have been selected with different capacitance values and voltage ratings and were calibrated. For the calibration the capacitors were put in a climate chamber, the temperature was increased from 10 to 90°C in 10°C steps. At each temperature level the capacitance value was measured with a simple multimeter repeatedly until we did not see further change in its value. When the capacitance value stabilized we could assume that the component stabilized at the environmental temperature and the capacitance value was registered. This low accuracy measurement method was already sufficient to give a rough picture on the temperature dependence of the selected components. In *Figure 1* and *Figure 2* the relative change of the capacitance is plotted as the function of temperature.

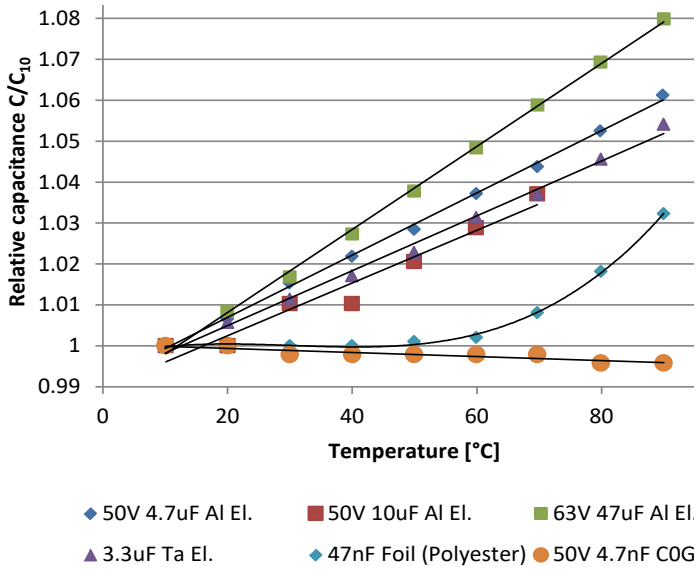


Figure 1 Temperature dependence of various capacitors measured with a simple multimeter, low dependence

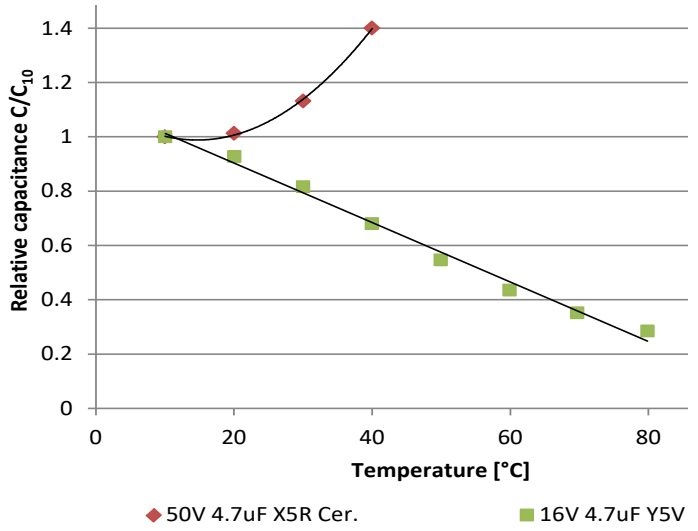


Figure 2 Temperature dependence of various capacitors measured with a simple multimeter, high dependence

The different electrolytic capacitors, both the Aluminum and the Tantalum types had similar magnitude of temperature dependence, the capacitance increased linearly by 5-8% in the 10-90°C range.

The ceramic capacitors showed very different temperature dependence depending on their classification, while the polyester film capacitor showed strong nonlinearity.

2.3. Applying heating power

After the temperature dependent electrical parameter has been selected one must find a way to heat up the component. The heating power is either generated in the component itself (self-heating) by the ripple current, or by an external component (cross-heating). Both cases can be examined with appropriately selected electrical and mechanical setups.

3. Experimental setup

Although the accuracy of a simple multimeter was sufficient for the demonstration of the temperature dependence; in precise measurements a more sophisticated instrumentation is needed. Moreover, in order to be able to measure the thermal transient of the component, the temperature needs to be measured with a relatively high frequency.

On the other hand measuring the absolute value of the capacitance is not required. For our application it is enough to produce a voltage signal that is proportional to the change of the capacitance, hence the change of the device temperature. It has to be noted however that for accurate self-resistance measurement the heating and sensing points must be at the same geometrical location otherwise the results will be distorted. For optimal interfacing to our existing thermal transient measurement system the optimal transient signal has to be in the 0...5 V voltage range.

After considering the available options we found that a switched capacitor system can be used to generate the required signal with high flexibility to adjust to various capacitance values.

The high level block diagram of the measurement circuit is shown in *Figure 3*.

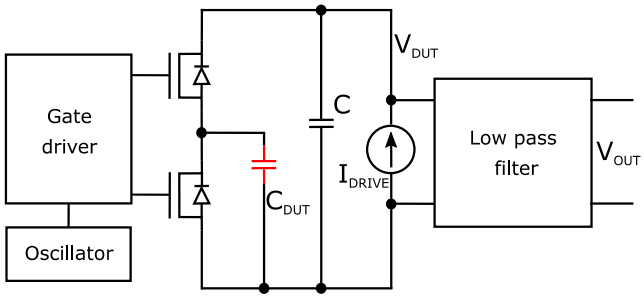


Figure 3 Block diagram of the experimental measurement setup

By driving the MOSFETs with non-overlapping clock signals the tested capacitor (C_{DUT}) is being periodically charged and discharged. The average current transferred towards the measurement ground is proportional to the DUT (Device Under Test) capacitance (C_{DUT}), the switching frequency (f) and the applied voltage drop (V_{DUT}). The switched capacitor can be represented by an equivalent resistance:

$$R_C = \frac{1}{fC_{DUT}} \quad (2)$$

In the arrangement presented in *Figure 3*the capacitor C is charged with a constant current I_{drive}, and discharged by the switched capacitor circuit. In equilibrium state the average charging current is equal to the discharging current:

$$I_{drive} = \frac{V_{DUT}}{R_C} . \quad (3)$$

Hence the voltage drop is inverse proportional to C_{DUT} capacitance:

$$I_{drive} = \frac{V_{DUT}}{f C_{DUT}} . \quad (4)$$

It has to be noted that the pulsating discharge current causes periodic voltage variation on capacitance C. For optimal operation capacitance C must be significantly higher than C_{DUT}, in order to minimize the voltage drop when the charge sharing happens between the two capacitors. Nevertheless it also limits the settling speed of the circuit, so it's appropriate sizing is critical. In the current experiments the value of capacitor C was selected to be 10x C_{DUT}, and an appropriately sized filter has been utilized to provide an acceptable noise level at the output. Finally the output signal V_{out} is proportional to the C_{DUT} capacitance and to measure its temperature dependent value it is connected to one of the measurement channels of a thermal transient tester (T3Ster) equipment.

The circuit parameters can be adjusted to be used with various capacitor values. The actual measurement setup was optimized for capacitances in the 0.1 nF… 1 µF range. The oscillator frequency is freely adjustable in the 10 kHz..1 MHz range and the cut off frequency of the filter was selected to be 5 kHz hence the useful thermal transient should start after 1 ms.

The measurement circuit can be easily adjusted to smaller capacitances, but for higher capacitance values the switching frequency would be too low and the useful bandwidth of the transient signal would be insufficient. For higher capacitances a different measurement approach should be selected.

In this paper the measurement method is presented with three slightly different measurement arrangements, providing examples for transfer and self-resistance measurements, and demonstrating that the measurement of both the heating or cooling transient is possible.

The first measurement setup demonstrates the transfer R_{th} measurement. A 33 nF foil capacitor was selected as DUT that was fixed to a 5 W ceramic power resistor as a heat source. The measurement arrangement is shown in *Figure 4*. In this setup the transfer effect is expected to last long. In order to cancel the effect of the air movement the test structure was put in a still air chamber (a one square feet size standard measurement environment). 10mA current was applied on the capacitor and the switching frequency was adjusted to achieve about 1 V voltage drop at room temperature. At t=0 s 1.5 A heating current was applied on the resistor generating approximately 3 W of heat. In order to be able to calculate the thermal resistance of the structure the thermal transient curve needs to be captured from one steady state to another. In the current measurement setup 1800 seconds were required to achieve complete steady state after the heating switched on.

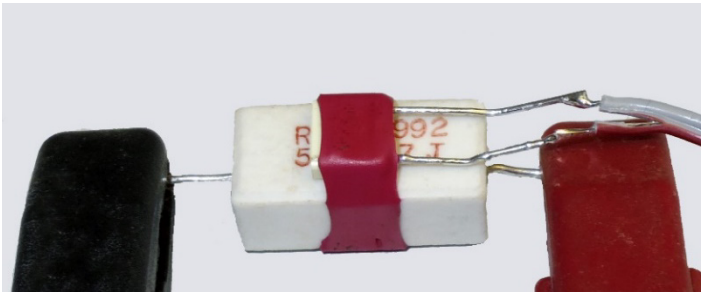


Figure 4 Transfer resistance measurement setup

For the second experiment a 100 nF ceramic capacitor was selected with X7R dielectric material. The leaded capacitor was soldered to a test PCB panel with different pin lengths and the effect of the pin length on the cooling transient was investigated. A photo of the measurement setups is shown in *Figure 5*. For the measurement similar sensing parameters were selected as for the foil capacitor: 10 mA sensing current was applied and the switching frequency was adjusted to obtain a voltage drop of about 1 V. In this case the device was heated up by applying a higher current across the switched capacitor. The higher average current caused a power loss on the ESR of the capacitor and hence heated it up. By applying 60 mA heating current the voltage increased significantly and the dissipation on the capacitor was about 0.5 W. The heating current was applied on the component for 100 seconds, then the heating was switched off and the transient measurement was immediately started and a 120 s cooling transient captured.

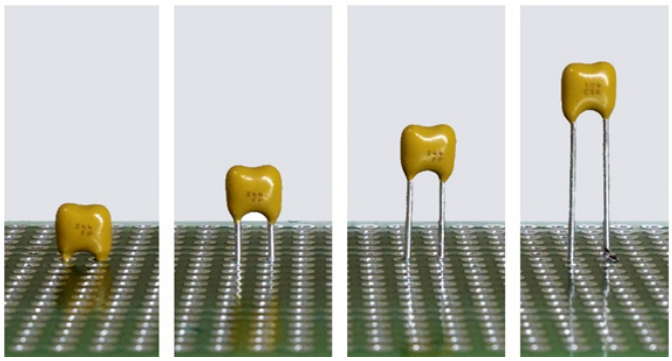


Figure 5 Thermal impedance measurement setups with different pin lengths

In the final experiment the same mechanical arrangement was used as shown in *Figure 5*, but only with the shortest and longest pin configuration. In this case the heating transient was captured instead of the cooling one. 70 mA heating current was selected and the switching frequency was adjusted to set the voltage drop to about 4 V. After switching on the heating a 200 s heating transient was captured. This time was sufficient for the structure to achieve steady state.

Before presenting the results of the experiments we show how the measured voltage values will be converted into temperature values.

3.1. Temperature sensitivity calibration

In order to transform the captured voltage transient into temperature transient the temperature sensitivity of the measured

signal needs to be calibrated. The temperature sensitivity calibration was straightforward in case of the first two experimental setups. The transient change of the capacitances were captured in a low power state, when the applied dissipation was practically negligible; only about 10 mW. In this state the temperature difference between the core of the capacitor and its environment is negligible, hence in steady state we can assume that the core temperature equals to the environment temperature.

For the calibration the capacitors were placed in a climate chamber and the electrical parameters of the driving circuit were set exactly the same as during the measurement. The chamber temperature was adjusted between 15°C and 85°C in 10°C (5°C for the ceramic capacitor) steps, and the output signal was registered at each steps after both the temperature and the output signal have stabilized.

The resulting calibration curves are shown in *Figure 6*. The data points corresponding to the foil capacitor (red dots) fit well on a linear curve (red dashed line) with a slope of 2.16 mV/°C. The temperature dependence of the ceramic capacitor however shows significant nonlinearity. In the measured temperature range this temperature dependence could be fairly approximated with a fourth order polynomial, but this does not guarantee that outside of the characterized range it is still a valid approximation. It also has to be noted that the temperature dependence is monotonous in the measured range in both of the cases. The monotony of the calibration curve is a primary requirement to be able to use the curve for voltage-temperature transformation.

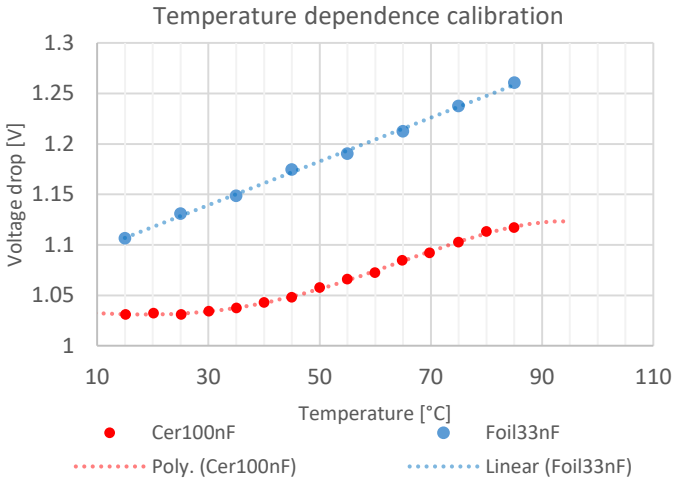


Figure 6 Temperature sensitivity calibration curves of the two capacitors at low power level

The TSP calibration for the third experiment is more challenging. In this case the heating transient of the component was captured, when the applied dissipation was high (0.28 W). Due to the significant junction to environment thermal resistance of the component there is a considerable temperature difference between the core temperature of the component and the regulated environmental temperature. In order to overcome this issue the device was also put in the environmental chamber, but now in unpowered state.

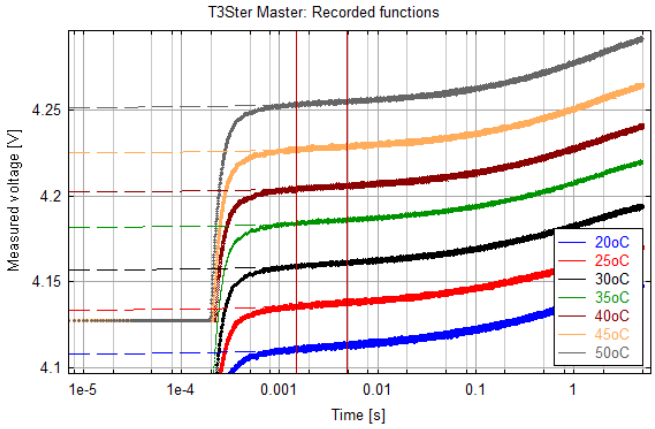


Figure 7 Measured short switch on transients of the ceramic capacitor at several environment temperatures

At each temperature level after the temperature has stabilized the powering of the capacitor was switched on and a short heating transient was captured. Just after switching on, the capacitor temperature is still equal to the chamber temperature, but immediately starts to increase. The captured transient responses are shown in *Figure 7* at a few environment temperature levels. In the initial section an electric transient can be observed, which is not related to the component temperature, only caused by the change of the electrical operating conditions. After about 1 ms the signal value is proportional to the temperature. In order to approximate the voltage drop just after turning on the heating, a square root function was fitted on each transient responses between 1.5 ms and 5 ms. and the voltage extrapolated to t=0 s time. The square root fitting is a widely used initial electric transient compensation method in case of semiconductor devices, where the heat generation is localized to the surface of the semiconductor chip. In case of the capacitor component where the heat is mostly generated in the core volume the validity of this approximation is questionable, but it has only a minor effect on the final results.

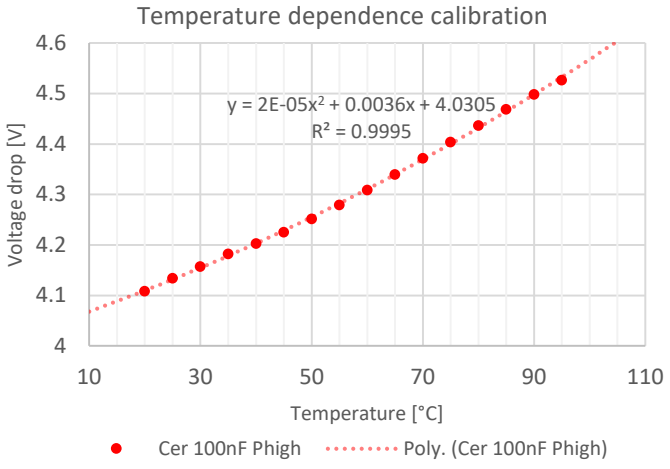


Figure 8 Temperature sensitivity calibration curve of the ceramic capacitor at high dissipation

The calibration was done between 20°C and 95°C in 5°C steps. The resulting calibration points are shown in *Figure 8*. The experienced nonlinearity was small, a second order curve fit well on the measured data points.

The overall time required for the thermal transient measurement of a single capacitor is mainly defined by the calibration time. The measurement time depends on the time constants of the test structure, but in case of good cooling it can be as low as a few tens of seconds. Meanwhile the calibration is a rather time consuming task as the environment temperature needs to be changed, waiting until the device temperature also stabilizes and then the corresponding voltage registered. This process often takes more than an hour for a single calibration. However, the calibration task can be done in parallel on high number of devices achieving a good overall throughput.

4. Measurement results

Another experiment is also presented in this paper in order to demonstrate the transfer resistance measurement, where the heating is coming from external heat source, and the capacitor core temperature is captured using the switched capacitor circuit. After the temperature sensitivity calibration the captured raw data could be transformed into temperature response as it is shown in *Figure 9*. Despite the switching mode operation of the circuit the noise level of the measured signal is reasonably good. The noise RMS (Root Mean Square) was about 0.2°C in all measurements.

As it is expected the initial section of the curve is completely flat, there is no temperature change measurable until 3 seconds. The heat generated by the resistor needs this time to propagate to the capacitor. As the thermal coupling in the experimental setup was much stronger than in normal applications, normally an even longer propagation time can be expected. Consequently if the goal is to measure only the transfer resistances the switching frequency can be significantly reduced to allow the measurement of much larger capacitance values, or the filter frequency can be reduced to achieve lower noise and more accurate temperature reading.

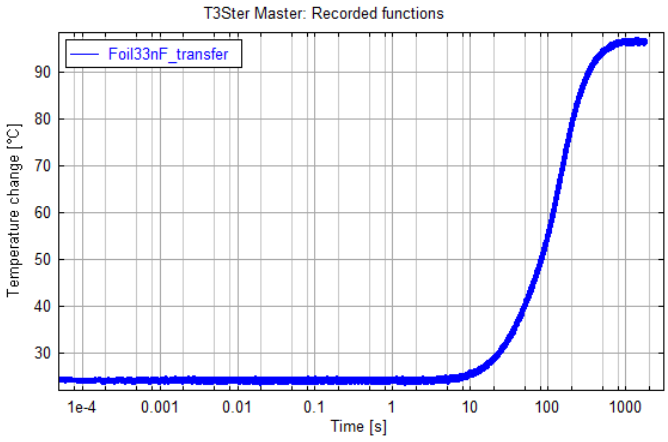


Figure 9 Temperature transient curve for transfer heating of the foil capacitor sample

The goal of the second experiment was to prove that the measured signal is in fact temperature transient and hence it is dependent on the thermal resistance of the structure. The different pin lengths were representing different thermal resistances between the capacitor and the hosting PCB. *Figure 10* shows the measured temperature transient responses. It is clearly visible that the longer the pin was the higher temperature change was measured. This also proves that one of the main heat flow paths in such arrangements is the conduction through the pins to the PCB. It has to be noted that in case of the red curve the calculated temperature was above 90°C, which is beyond the calibration range. As a consequence the calculated temperature in this range is expected to be inaccurate.

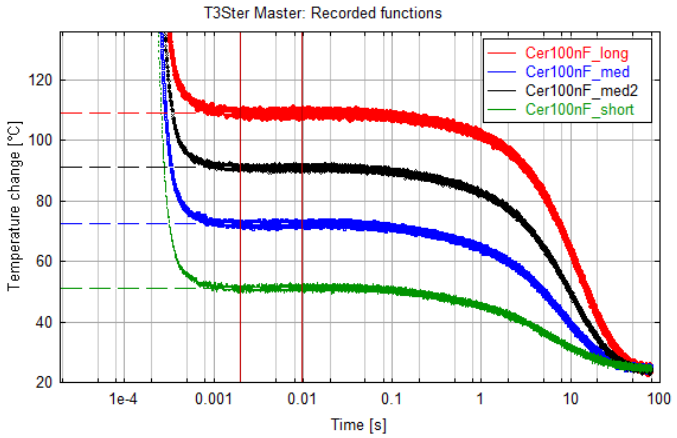


Figure 10 Cooling transient curves measured on the ceramic capacitor with different pin lengths

As opposed to the transfer heating measurement the temperature change starts much earlier in case of self-heating. The end of the electric transient is about 1 ms, and the valuable thermal signal starts soon after it. This shows that if the heated and measured points overlap, then the response speed of the measurement circuit becomes more important. How much information we lose that depends only on the internal thermal resistance and capacitance of the capacitor structure.

In order to obtain a more accurate thermal transient curve, good enough for further post-processing the parameters of the measurement setup were further optimized in the third experiment. We found that instead of the cooling curve that is most commonly used in case of testing semiconductor components, the heating curve provides a better signal. The temperature sensitivity is higher with lower nonlinearities, ensuring better signal to noise ratio.

The captured heating curves for the short and long pin scenarios are shown in *Figure 11*. It can be observed that the initial section of the two curves overlap perfectly, and split up only after about 2 seconds. In the initial section the heat experiences the same physical structure. The difference appears when the heat arrives to the PCB in case of the short pin arrangement but it needs to pass through longer pins representing higher thermal resistance when the pin is long.

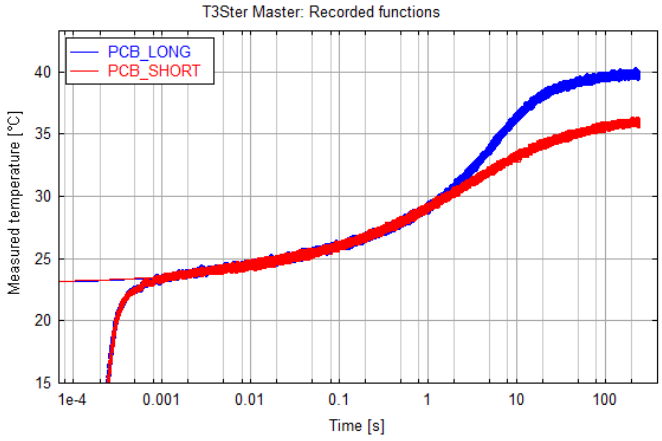


Figure 11 Measured heating transient curves of the ceramic capacitor with long and short legs

With the evaluation of the thermal transient measurements a one dimensional Cauer RC model can be constructed that in case of semiconductor components not only provide a good behavioral model, but structural information can also be extracted from it [8]. The structure function [10] is a graphical representation of the identified equivalent Cauer RC network. It plots the sum of the thermal capacitances in the function of the sum of the thermal resistances from the source of the heat towards the ambient, in cases of one dimensional heat flow that is a very good approximation in most of the practical cases. The resulting curve helps in interpreting the larger data sets. The steep sections in the structure function represent a physical layer with high added thermal capacitance but low thermal resistance that corresponds to a good thermal conductor material. On the other hand the flat sections of the curve represent high resistance with low added capacitance, hence a thermal insulator. In case of real one dimensional heat flow path the subsequent layers can be identified based on the corresponding slope on the curve and hence the structure can be analyzed.

4.1. Verification with simulation

The thermal transient measurement of the capacitor provided realistic transient curves, and the effect of the change of pin length could also be clearly identified. However the validity of the curve still needs to be proven. We decided to build up the detailed 3D model of the capacitor and compare the simulated and the measured thermal transient responses.

As we did not have detailed structural information from the manufacturer we cross sectioned the capacitor to identify the important features and measure their sizes. A cross section picture of the capacitor and the built simulation model can be seen in *Figure 12*.

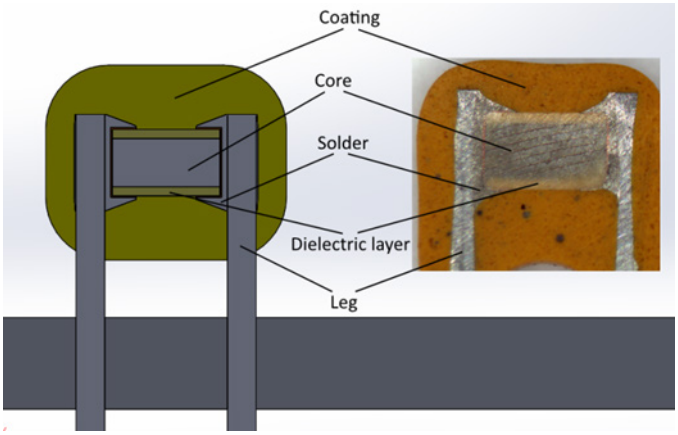


Figure 12 Simulation model of the measured ceramic capacitor

The most important structural elements are the capacitor core, where the electrodes and the X7R dielectric material are providing the layered structure, a dielectric cover layer, where no electrodes are visible, the solder, the copper legs and the coating ceramics.

According to the information found in the literature [4] the BaTiO3 is a commonly used ceramics to create X7R grade ceramic capacitors. As a base approximation we supposed that this material was used as dielectric layer and used the material properties suggested by the referenced paper. In order to simplify the simulation model the layered core has not been modeled in detail, but the weighted average thermal parameters of the BaTiO3 and the chromium electrodes have been used. As the layer thicknesses were unknown we used an arbitrary 8:2 volume ratio as the first approximation. The coating material was also unknown; hence as a first approximation we used the same thermal parameters as for the dielectric material. The solder was assumed to be SAC (Tin-silver-copper) solder. The initial guess material parameters can be read in Table 1.

Table 1. Material parameters, initial guess

Component	Material	λ [W/mK]	C_p [J/kg K]	ρ [g/cm^3]
core	custom	20.5	412	6.16
Dielectric	$BaTiO_3$	2.7	400	5.9
Coating	custom	2.7	400	5.9
Solder	SAC	78.4	306	7.4
Leg	Al	230	1037	2.69
PCB	FR4	0.3	880	1.2

The transient response of the structure was simulated with time steps identical to the sampling rate of the measurement. The time domain curve of the transient response provides much less information than its transformed version the structure function. The simulated and measured curves can be better compared using the structure function as it is shown in *Figure 13*.

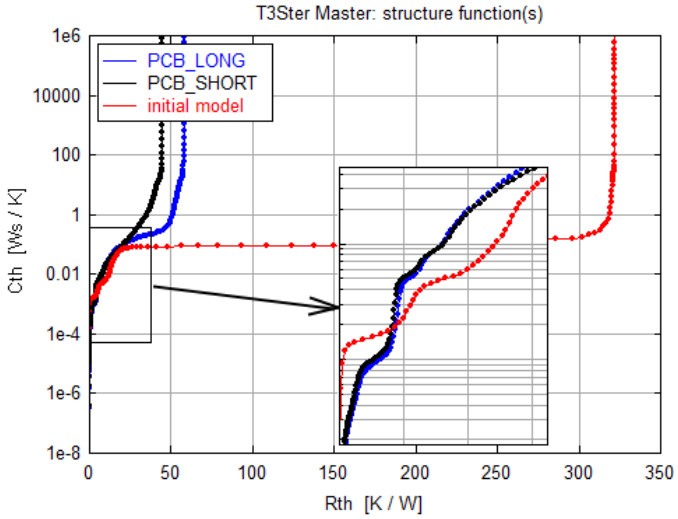


Figure 13 Structure function generated from the simulated transient of the initial "best guess" model compared with the measured curves

The black and blue color curves were generated from the measured transients with short and long pin lengths and the red curve was generated from the simulated transient. At the large scale plot the initial section of all the curves seem very similar, but after the initial section a large excess resistance can be observed in the simulation model. If we take a closer look at the initial part of the curves it becomes apparent that while the two measured curves are identical despite the differences in the pin length the simulated structure functions already show slight differences. In order to understand the differences between the simulation model and the modeled component the structure function has to be analyzed.

4.2. Using the structure function to fine-tune the simulation model

For semiconductor packages a detailed thermal model can be step by step calibrated based on the measured transient curve starting from the chip and moving towards the more remote features [11]. This can be done because there is typically a

dominant heat flow path that can be well described with the one dimensional spreading model and the time constants corresponding to the different elements of the heat flow path are increasing gradually as we are receding from the junction. However in case of the capacitor component the one dimensional spreading from the core is not a good assumption. As an example from the core of the capacitor the heat also spreads towards the pins through the solder contact, and the coating. In such situations what we see in the structure function is the parallel sum of the two thermal resistances (and capacitances).

As a result the same section of the structure function can be affected by more than one model parameters. As it is shown in *Figure 14* the conductivity of the core material (green and blue curve) and the surrounding dielectric layer (red and orange) are also affecting the length of the first flat section of the curve. Other sections are more dependent on one single material parameter and can be more clearly identified on the structure function.

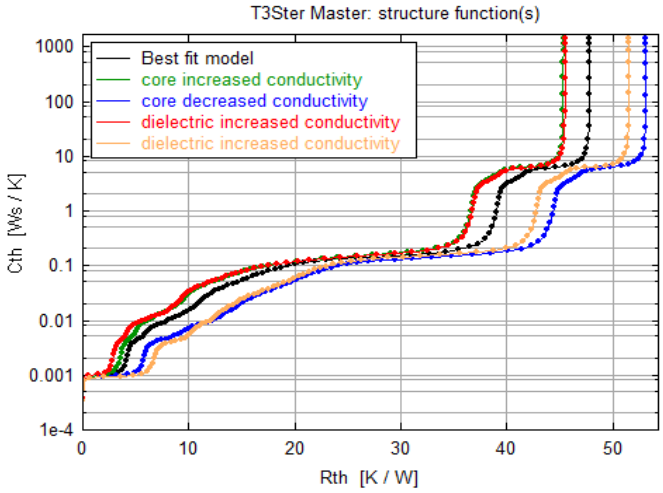


Figure 14 Effect on the structure function of the increased and decreased conductivity of the core (green and blue) and the dielectric layer (red and orange)

The large resistance at the end of the structure function corresponding to the initial model (*Figure 13*) was caused by the low conductivity of the PCB by neglecting the effect of the copper traces. *Figure 15* shows a simplified summary about which sections of the structure function are affected by the various model parameters.

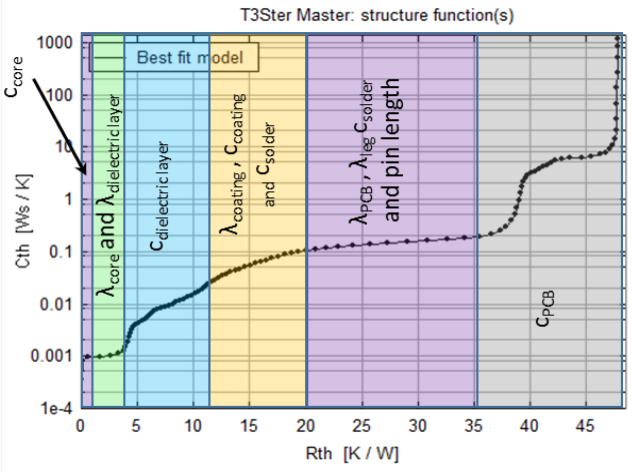


Figure 15 Different areas of the structure function and the various simulation model parameters affecting their shape

Having this information we know which parameters to adjust in order to have the simulated transient fit the measured one. After iterative adjustment of the appropriate thermal parameters of the PCB, capacitor core material, dielectric material, and coating material a very good match between the simulated and the measured transients could be achieved in the important domain up to about 35 K/W. Above this thermal resistance the thermal response is no more dependent of the thermal parameters of the capacitor model, hence it was out of our scope. With further work the whole simulation model could be matched to the experimental results.

Figure 16 shows the structure function of the measured transients and the one simulated after the model calibration, demonstrating very good agreement. The final model parameters are listed in Table 2. Almost perfect match can be achieved until 35 K/W.

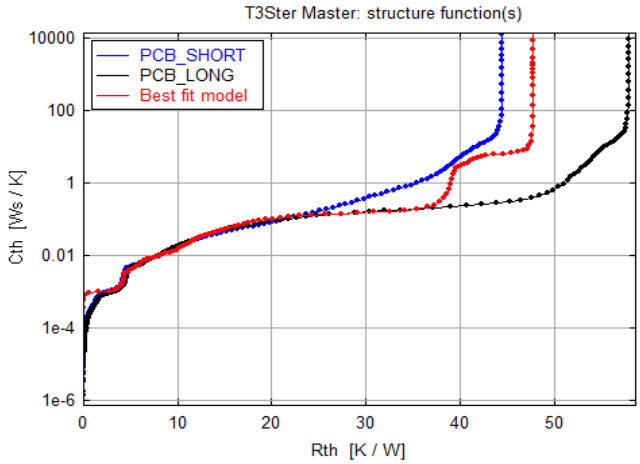


Figure 16 Structure function corresponding to the simulated (red) and measured structure (blue and black) after the calibration

Table 2. Material parameters after calibration

Component	Material	λ [W/mK]	C_p [J/kg K]	ρ [g/cm³]
core	custom	25	300	5
Dielectric	$BaTiO_3$	2.6	1000	5.9
Coating	custom	3	500	6
Solder	Custom	78	2500*	7.4
Leg	Cu	380	416	8.96
PCB	Custom	30	880	4

*The specific heat of the solder was increased to unrealistic levels to avoid geometrical changes (the volume in the model was lower than in the actual capacitor)

Nevertheless a small difference still remains after the calibration in the very early section of the structure function, below 2.5 K/W, where the simulation result and the measured transient do not match. In this section the structure function could not be affected by the adjustment of the model parameters. The simulated structure function shows stepwise capacitance change at practically 0 K/W, as the heat generation was assumed to be evenly distributed in the core of the capacitor. But the lower slope section of the structure function corresponding to the measured transient denotes that the heat generation is more localized to a smaller volume and the heat has to spread in the core to fill its volume.

5. Conclusions

In this paper a new measurement setup for the thermal characterization of capacitors was proposed. This method adopts the thermal transient measurement technology that is widely used for the characterization of semiconductor packages and structures.

The measurement method uses no external temperature sensors, but utilizes a temperature dependent electric parameter of the measured component itself to measure its temperature. The various temperature dependent parameters were reviewed and the capacitance was selected as it shows significant temperature dependence in case of most capacitor constructions.

In order to measure the change of the capacitance with sufficiently high temporal resolution a switched capacitor based measurement solution is proposed. The presented measurement circuit can be adjusted to measure capacitances in a wide range with approximately 1 kHz bandwidth.

The measurement method was demonstrated with three measurement setups. The first setup demonstrated the transfer R_{th} measurement, capturing the heating curve of the capacitor after a closely attached heat source turned on. It is shown that due to the long time required for the heat to reach the capacitor even a strong thermal coupling enables us to significantly reduce the switching speed of the switched capacitor and hence measure significantly higher capacitances.

The second setup demonstrated the self-resistance measurement, heating up the capacitor by its own losses at high ripple current and then measuring the cooling transient. The measurement was repeated with different pin lengths and the effect of the pin length on the thermal resistance of the component could be identified.

In this third setup also the self-resistance was measured, but capturing the heating curve of the capacitor. This setup provided better signal to noise ratio than the cooling measurement.

The detailed 3D simulation model of the capacitor and the PCB to which it was soldered to were built and the heating transient was measured and simulated. The first simulation and measurement results showed acceptable match in the initial section of the transient and the generated structure functions, but further from the heat source the difference was large.

By comparing the structure functions generated from the measured transient and the simulated one, the inaccurate model parameters could be iteratively fine-tuned to achieve an almost perfect match between the measured and simulated structure functions. This demonstrates that the thermal transient measurement can not only be used to measure the thermal resistance of the component, but also helps in calibrating the simulation model.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] Sam G. Parler Jr, “Thermal modeling of aluminum electrolytic capacitor”, IAS Annual Meeting (IEEE Industry Applications Society) 4:2418 - 2429 vol.4 · February 1999 DOI: 10.1109/IAS.1999.799180

[2] Schaubauer, F. M.; Blumkin R.; “Thermal Resistance, Power Dissipation and Current Rating for Ceramic and Porcelain Multilayer Capacitors”; Reprinted from RF Design Magazine, 1981

[3] Parler, S. G. Jr.; Macomber, L. L.; “Predicting Operating Temperature and Expected Lifetime of Aluminum-Electrolytic BusCapacitors with Thermal Modeling”, PCIM November, 1999

[4] P. Guillemet, C. Pascot and Y. Scudeller, "Electro-thermal analysis of Electric Double-Layer-Capacitors," 2008 14th International Workshop on Thermal Inveatigation of ICs and Systems, Rome, 2008, pp. 224-228.doi: 10.1109/THERMINIC.2008.4669913

[5] Huesgen, T., "Thermal resistance of snap-in type aluminum electrolytic capacitor attached to heat sink," Energy Conversion Congress and Exposition (ECCE), 2012 IEEE, vol., no., pp.1338,1345, 15-20 Sept. 2012, doi: 10.1109/ECCE.2012.6342661

[6] Z. Sarkany and M. Rencz, "A way for measuring the temperature transients of capacitors," 2016 IEEE 18th Electronics Packaging Technology Conference (EPTC), Singapore, 2016, pp. 818-822.doi: 10.1109/EPTC.2016.7861594

[7] Szekely, V.; Torok, S.; Nikodemusz, E.; Farkas, G.; Rencz, M., "Measurement and evaluation of thermal transients," Instrumentation and Measurement Technology Conference, 2001. IMTC 2001. Proceedings of the 18th IEEE , vol.1, no., pp.210,215 vol.1, 21-23 May 2001

[8] V.Székely: “A new evaluation method of thermal transient measurement results”, Microelectronics Journal, V.28, No.3, pp.277-292 (1997)

[9] Zongli Dou; Rong Xu; Alfonso Berduque; “The Development of Electrolytes in Aluminum Electrolytic Capacitors for Automotive and High Temperature Applications” ;CARTS Europe 2008 Conference Proceedings; Helsinki, Finland, 20-23 Oct. 2008

[10] M. Rencz, V. Szekely, A. Morelli and C. Villa, "Determining partial thermal resistances with transient measurements, and using the method to detect die attach discontinuities," Eighteenth Annual IEEE Semiconductor Thermal Measurement and Management Symposium. Proceedings 2002 (Cat.No.02CH37311), San Jose, CA, USA, 2002, pp. 15-20. doi: 10.1109/STHERM.2002.991340

[11] A. Vass-Varnai, R. Bornoff, Z. Sarkany, S. Ress and M. Rencz, "Measurement based compact thermal model creation - accurate approach to neglect inaccurate TIM conductivity data," Electronics Packaging Technology Conference (EPTC), 2011 IEEE 13th, Singapore, 2011, pp. 67-72. doi: 10.1109/EPTC.2011.6184388

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1390-1400 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Radiation Hybrid Mapping: A Resampling-based Method for Building High-Resolution Maps

Raed Seetan[*1], Jacob Bible[1], Michael Karavias[1], Wael Seitan[2], Sam Thangiah[1]

[1]*Computer Science Department, Slippery Rock University, 16057, USA*

[2]*School of Educational Sciences, The University of Jordan, 11942, Jordan*

ARTICLE INFO

Article history:
Received: 31 May, 2017
Accepted: 18 July, 2017
Online: 05 August, 2017

Keywords:
Radiation Hybrid Mapping
Clustering
Resampling

ABSTRACT

Abstract— The process of mapping large numbers of markers is computationally complex, as the increase of numbers of markers results in an exponential increase in the mapping runtime. Also, having unreliable markers in the dataset adds more complexity to the mapping process. In this research, we have addressed these two issues and proposed our solution. The proposed approach builds solid maps in two phases: Phase 1 builds an initial map following these steps: 1) Resample the original dataset to generate variant datasets, then cluster all resampled datasets into groups of markers. 2) Merge all groups of markers to filter out unreliable markers. 3) Generate a Map for each group of markers. 4) Concatenate all groups' maps to form the final map. Phase 2, Adds more markers to the initial framework to build a high resolution map as follows: 1) Use Kmeans algorithm to filter out unreliable markers and cluster the remaining markers. 2) Insert the remaining markers in their best positions in the initial framework. To evaluate the performance of the proposed approach, we compare our constructed maps on the human genome with the physical maps. Moreover, we compare our constructed maps with a state-of-the-art tool for building maps. Experiment results show that the proposed approach has a very low computational complexity and produces solid maps with high agreement with the physical maps.

1. Introduction

This paper is an extension of work originally presented in The 15th IEEE International Conference on Machine Learning and Applications [1]. Genome mapping [2] is the process of finding the order of markers within chromosomes; where markers are short fragments of Deoxyribo Nucleic Acid (DNA) sequence most often located in noncoding regions of the genome (regions that do not encode protein sequences). Markers orders can provide researchers with essential information for localizing disease-causing genes in the genome. Radiation Hybrid (RH) Mapping [3] is a statistical mapping technique to order markers along a chromosome, and estimate the physical distances between markers. RH mapping is widely used in many mapping projects, and has several advantages over alternative mapping techniques [4]. In RH Mapping, chromosomes are separated from one another, then high doses of X-rays are used to break the chromosome into several fragments. The main principle of RH mapping is that the probability of separation of two adjacent markers due to radiation breakage increases with the increase in physical distance [4]. The order of markers on a chromosome can be calculated by estimating the frequency of breakage and retention between the markers.

An RH population can be seen as a m X n Boolean matrix, where m is the number of markers, and n is the number of individual organisms in the mapping population. A single RH vector represents one marker across all individuals/Panels. All markers in each panel is assigned a value of 1 or 0, where the value 1 indicates the presence of that marker in that panel, and the value 0 indicates the absence of that marker. Once all markers in all panels are screened and saved, the RH population will be used for the mapping process. Figure 1 shows a sample RH population of 3 markers on 6 panels.

After preparing the RH population, the RH mapping process starts. The Obligate Chromosome Breaks (OCB) is used to indicate the similarity and estimate the distance between two markers in RH population [5]. The basic mapping step is to estimate the OCB between all pairs of markers, where a 1 is followed by 0 or vice versa in the same panel, and place the closest (similar) markers together. For example to map the simple population in Figure 1, we need to find the similarity between these three RH vectors

[*]Corresponding Author: Raed Seetan, Computer Science Department, Slippery Rock University, 16057, Tel: 724-738-2940, USA | Email: raed.seetan@sru.edu

(markers), and then place the similar markers close together in the constructed map. First calculate the OCB for all possible markers permutations (the number of times a 1 is followed by 0 or vice versa in the same panel), and second select the permutation with the minimum number of OCB; the best map of these three markers is {M1, M2, M3} with five breaks. Any other permutation will have more than five breaks.

The toy example in Figure 1 is used just to explain the process of mapping a RH population. Typically, the number of markers in chromosomes range between a hundred to thousands. Thus, calculating the OCB for all markers permutations is unfeasible solution for such large numbers of markers; where for n markers, there are n!/2 permutations to evaluate. In this study, we are going to map a RH dataset of the human genome with hundreds of markers. Heuristic approaches [6-8] have been used to map such larger datasets of markers in a reasonable time. Although the heuristic approaches are relatively fast, they scale exponentially with the number of markers. Figure 2 shows the typical pattern of mapping different datasets of markers using some heuristic algorithms that have been used to map RH populations in many researches [9-12].

The main goal of this study is to propose an alternative approach to map large numbers of markers in a short time. The proposed approach works in two phases. Phase 1, has been originally published in [1], constructs maps as follows: 1) Uses Jackknife resampling method on the original RH population to create slightly variant RH samples. Then, all the generated samples are grouped into clusters based on their LOD (logarithms of odds -base 10) scores. The LOD score [14] was developed as a probabilistic measure for linkage, and has been used consistently throughout the RH literature [9-11]. 2) Merges all RH samples clusters into consensus clusters, and filters out the unreliable markers. 3) Constructs a map for each consensus cluster. 4) Connects all consensus clusters maps into one single map. Phase 2, uses the output map of Phase 1 as a skeleton to map additional markers and improve the resolution of the initial map. Phase 2 works as follows: 1) Uses Kmeans clustering to find the candidate markers to be added to the map. 2) Finds the best position of each candidate marker in the initial map in order to add them to improve the final constructed map.

To demonstrate the effectiveness of our approach, three metrics are going to be used: 1) Accuracy, which indicates the agreement of the constructed maps with the published maps. 2) The number of mapped markers in the constructed maps. 3) The running time to construct the maps. We will compare our constructed maps with the maps we published in our original work in [1]. Also, we will compare our results with a state-of-the-art tool for building radiation hybrid maps. In our approach, we use the Clustering technique to reduce the mapping computational complexity, thus, mapping large datasets of markers in a short time. Also, our approach considers the problem of noisy markers and can filter out unreliable markers to increase the accuracy of the constructed maps.

The rest of this paper is structured as follows: Section 2 presents the related work. In Section 3, our proposed approach is discussed in detail. Section 4 presents the experimental results of the proposed approach, and Section 5 concludes this research.

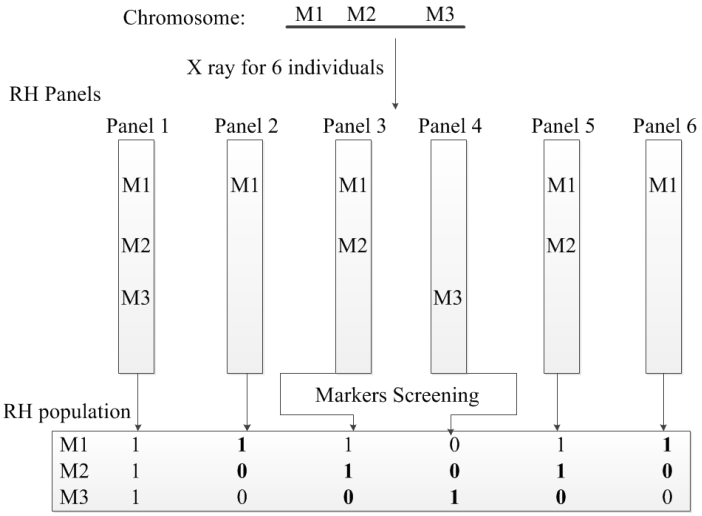


Figure 1. A toy example of an RH population for 3 markers on 6 panels. Breakages are highlighted in bold.

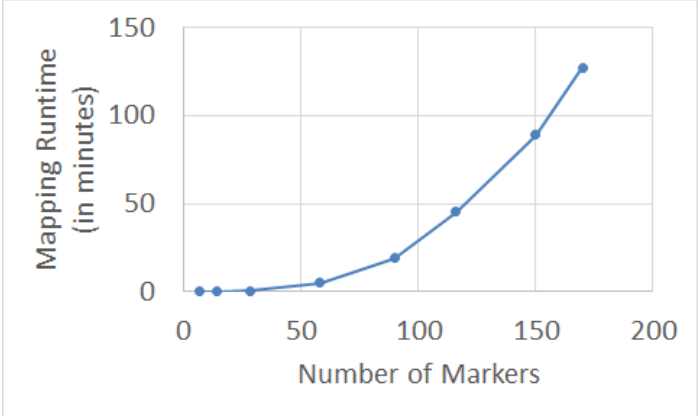


Figure 2. Relationship between No. of markers and mapping complexity using heuristic algorithms implemented in the Carthagene tool.

2. Related Work

Conventional techniques [15, 16] for filtering out unreliable markers depend mainly on resampling [17]. Thus, these techniques are not recommended for filtering out unreliable markers from a large dataset due to their high computational complexity. The filtering process can be summarized in three sequential steps: 1) Resample the original dataset. 2) Use mapping techniques to map each resampled dataset. 3) Construct a matrix to show the reliability of all markers and filter out the most unreliable marker. These three steps are repeated to filter out unreliable markers one at a time. This iterative process is too time consuming for a large dataset, as to filter out only one unreliable marker, there is a need to resample the dataset and map all resampled datasets. The computational complexity of mapping a single resampled dataset scales exponentially with the number of markers to be mapped, as shown in Figure 2. Thus, repeating this complex mapping process for every single resampled dataset to filter out one unreliable marker is not a feasible solution for large datasets. As we discuss our proposed solution in Section 3, we will show that how our approach resampled the dataset and filtered out the unreliable markers without the need of mapping the resampled datasets. Thus, the proposed approach reduces the computational complexity of the whole mapping/filtering process.

RHMapper tool [12] provides another approach to build solid maps with only the reliable subset of markers. The mapping process is divided into two steps: 1) use *find_triples* command to search the data set for well-ordered triples, and save them. 2) Use *assemble_framework1* or *assemble_framework2* commands to assemble the saved well-ordered triples into a final map. The assembly process uses the overlapping between the well-ordered triplets of markers to combine them into a final map, where the *assemble_framework1* command seeks for overlapping with two consecutive markers in each triplet, for example: triplets (A-B-C) and (B-C-D) will be assembled into (A-B-C-D). *assemble_framework2* command can add gaps to assemble triplets, for example: triplets (A-B-C) and (A-C-D) will be assembled into (A-B-C-D). Although, this mapping strategy can build solid maps, mapping a hundred markers takes several hours to finish [13], so this approach does not scale well with the number of markers.

Multimap [13] is another tool to build maps with reliable markers. The mapping process starts with searching for the strongest pair of markers, then iteratively adds one marker at a time to extend the initial map. Apparently, this mapping process does not scale well with large datasets of markers.

Carthagene tool [11] is another well-known package that has been used in many studies to construct solid maps [24, 25]. The *Buildfw* command implements an incremental insertion procedure to build maps with only reliable markers; and it works as follows: first, search for the triplet of markers that maximizes the difference between the likelihood of the best map, and the second best map using only this triplet of markers. Once the best triplet of markers is found, save it as an initial framework map. Second, for each marker not mapped in the framework, try to insert it into its best interval. In order to insert a marker into its best interval in the framework map, two thresholds are evaluated: 1) AddThres and 2) KeepThres. AddThres is used to determine if a marker can be placed in its best interval or not; where if the loglikelihood difference between the best two insertion intervals is greater than the AddThres, then the marker can be inserted in its first best interval in the map. If that marker can be inserted in its best interval in the map, the KeepThres is evaluated, which is used to determine if the new inserted marker can be saved in the framework map or not. If the loglikelihood difference between the best two maps is greater than the KeepThres, then the new inserted marker will be saved in its best interval and will be used to map other markers in the following iterations. Otherwise, the new inserted marker will be removed from the current framework map.

One of the limitations of the incremental insertion procedure in Carthagene is that only a few markers are mapped at the final map, if the recommended values of the AddThres and KeepThres are used [11]. Another limitation is that the mapping process starts with only three markers as an initial map. Thus, adding one marker at a time to the initial map makes the mapping process not suitable for large numbers of markers. Moreover, the incremental insertion procedure cannot be parallelized.

In this research, we compare our proposed approach with the Carthagene incremental insertion approach. Our proposed approach can take advantage of the parallel computing to reduce the computation complexity of the mapping process; where markers are grouped into smaller clusters, and these groups of markers can be mapped in parallel. We expect to outperform the Carthagene approach in terms of the mapping runtime. Also, in terms of mapping accuracy, our proposed approach uses the grouping process over all resampled data to filter out the unreliable markers, and thus uses only the reliable markers to build solid final maps. We expect our approach to build solid maps, as the grouping process will help in reducing the effect of unreliable markers to the local clusters maps, not to the entire map.

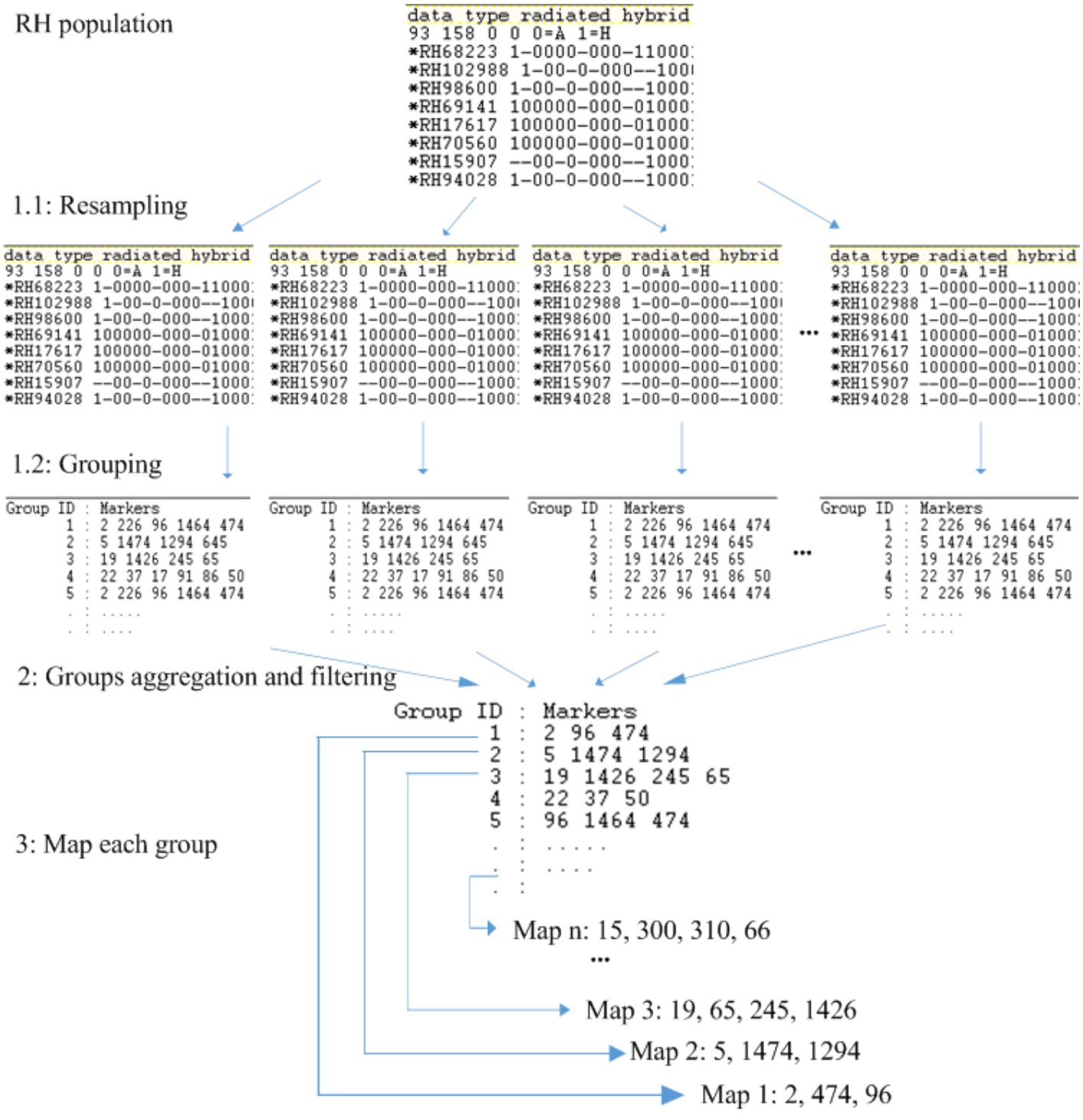
3. Proposed Approach

The proposed approach divides the mapping process into two phases, where each phase consists of several steps. The first phase builds an initial map with the most reliable set of markers, Figure 3 shows the systematic workflow of the first phase. The second phase uses the initial map constructed in phase 1 as a framework to map more markers to the initial framework, which results in building a high resolution map (mapping larger numbers of markers), Figure 4 shows the systematic workflow of the second phase

3.1. Phase 1: Building Initial Framework Map

1) Resampling and grouping: The Jackknife resampling technique is used to generate N variant RH samples of the original RH dataset, where each resampled RH dataset will have N-1 individuals (leave-one-individual-out).f After generating the resampled datasets, the Single Linkage Agglomerative Hierarchical clustering technique [18] is used to group each dataset into different clusters of markers. The *group* command is implemented in the Carthagene tool can be used to create these clusters of markers. The idea of the grouping is to create small groups of markers with low intra-cluster distances and high inter-cluster distances.

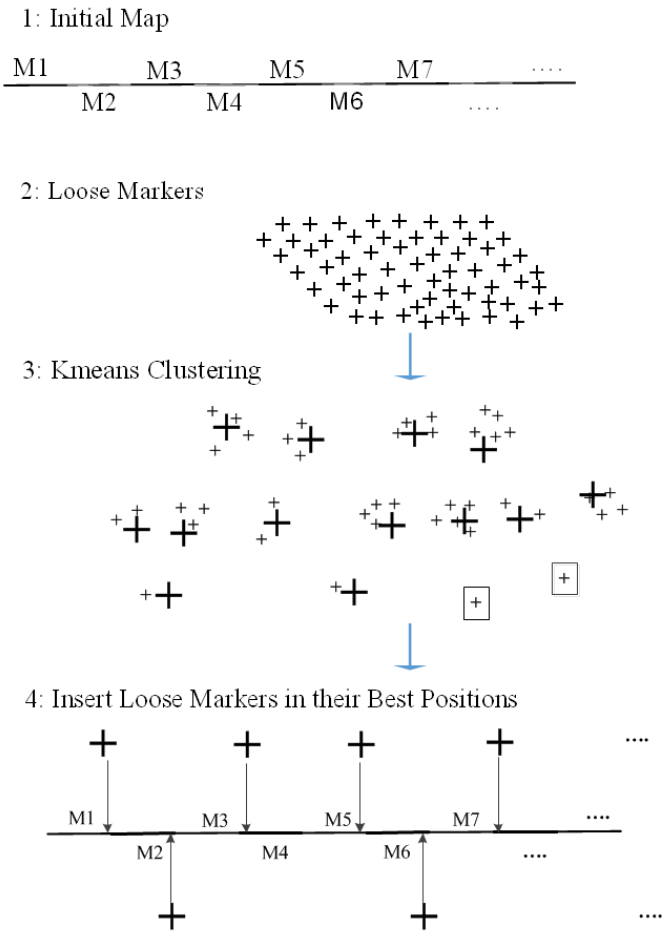
Groups aggregation and filtering: The first step generates N variant RH sampled datasets, and with each resampled dataset many clusters of markers. The goal of this step is to use all clusters of markers of all resampled datasets to filter out unreliable markers and aggregate the clusters of markers to generate solid consensus clusters. A matrix of size *m x m* is used to show the stability of linkage relationship between all pairs of markers, where m is the number of markers in the RH population. The matrix is created and initialized to Zero. Then, we trace all the clusters in all RH sampled datasets to update the values of the elements of the matrix. For each paired ordered markers (Mi, Mj), we increment both indexes (i, j) and (j, i) in the cluster matrix by 1 each time the pair (Mi, Mj) appear in one cluster in all sampled datasets.


RH population
data type radiated hybrid
93 158 0 0 0=A 1=H
*RH68223 1-0000-000-11000
*RH102988 1-00-0-000--100
*RH98600 1-00-0-000--1000
*RH69141 100000-000-01000
*RH17617 100000-000-01000
*RH70560 100000-000-01000
*RH15907 --00-0-000--1000
*RH94028 1-00-0-000--1000
1.1: Resampling
1.2: Grouping
Group ID : Markers
1 : 2 226 96 1464 474
2 : 5 1474 1294 645
3 : 19 1426 245 65
4 : 22 37 17 91 86 50
5 : 2 226 96 1464 474
2: Groups aggregation and filtering
Group ID : Markers
1 : 2 96 474
2 : 5 1474 1294
3 : 19 1426 245 65
4 : 22 37 50
5 : 96 1464 474
3: Map each group
Map n: 15, 300, 310, 66
Map 3: 19, 65, 245, 1426
Map 2: 5, 1474, 1294
Map 1: 2, 474, 96
4: Connect Maps and polish
Map 1
2, 474, 96
Map 2:
1294, 1474,5
Map 3:
19, 65, 245, 1426
....

After that, we start filtering out unreliable markers; two factors are defined for each marker: 1) the Stability-Factor (SF), which indicates the stability of neighbors markers throughout all sampled datasets clusters. For example, the Stability-Factor value of 80 for markers (Mi, Mj) means that over all resampled RH datasets, markers (Mi, Mj) appear 80 times together in the same cluster. 2) The Reliability-Ratio (RR), which is calculated by the count of relationships above the Stability-Factor divided by the total number of relationships for that marker. Figure 5 shows a simple example to illustrate the filtering process. All markers with Reliability-Ratio less than a threshold are considered unreliable markers and can be filtered out. In our experiments, we used RH population of 93 individuals and filtered out markers with SF values less than 80, and RR less than 50%. After filtering out the unreliable markers from all clusters, the remaining markers will be considered reliable markers and will form the consensus clusters.

2) Mapping clusters' markers: In this step, we build a map for each consensus cluster of markers. Several heuristics algorithms can be used to build the maps, in this study, we follow the mapping strategy published in [4, 9, 10] to build a map for each cluster, using the heuristics algorithms implemented in the Carthagene tool [11]. The following steps present the mapping strategy: 1) *Mrkdouble* command to identify and merge identical markers by a single marker. 2) *Build* command to build an initial nice map, where the command finds the best pair of markers, with the strongest linkage, and incrementally tries to add the remaining markers if they satisfy the insertion criteria. 3) *Greedy* and *Annealing* commands are used as improvement steps to find a better map in case a local improvement exists. 4) *Flips* and *Polish* commands are used to apply all possible permutations in a sliding window to check if a better map can be achieved. Running all these commands in that sequence will build a map for each cluster.

3) Connect maps and polish: Building a map for each consensus cluster shows the local order of the markers in each cluster. However, all these small maps need to be concatenated to form a whole chromosome map. In this step, we concatenate the clusters' maps generated in the previous step to form one map for a whole chromosome. This process can be done in an iterative manner. First, we extract the boundaries, the first and the last marker, of each cluster map. Second, we group all boundaries into clusters, the *group* command in Carthagene can be used to make these clusters; we start with a high LOD score T, i.e. T= 18. Boundaries from different maps that fall into one cluster are connected using the closest boundaries to form a bigger map. This step is repeated, where at each iteration the LOD score T needs to be lowered by a factor of x, i.e. x= 3, to let the far away boundaries to be connected to form one single map. A polishing step is used at the end of this step to improve the constructed map by trimming marker at the edges of the final map with LOD score less than 3, which generally indicates that there is no linkage between two markers.



	M1	M2	M3	M4	M5	SF >= 80	RR >=50%	Reliable
M1	--	**90**	**90**	**60**	**26**	**2**	**2/4**	**Yes**
M2	90	--	10	15	2	1	1/4	No
M3	**90**	**10**	--	**80**	**80**	**2**	**3/4**	**Yes**
M4	60	15	80	--	0	1	1/3	No
M5	26	2	80	0	--	1	1/3	No

3.2. Phase 2: Improving Framework Map

1) Loose markers extraction and grouping: After building a framework map with only the most reliable subset of markers, our goal is to use the framework map as an initial map to construct a high resolution map by mapping more markers from the remaining left out markers (Loose Markers) to the initial framework map. In this step, we use the initial constructed framework map as a skeleton to map the loose markers, where loose markers will be placed into the intervals of the initial framework map. This process works as follows: 1) Extract the loose markers from the original dataset. 2) Group the loose markers into clusters; the Kmeans algorithm is used to make the clusters of the loose markers. Kmeans algorithm works well for large datasets. To run the Kmeans algorithm, we need to set the number of clusters (K). In this study, we set the value of K

to the number of markers in the initial framework map plus 1. Assuming that all loose markers are going to be placed in between framework's markers, so if the number of markers in the framework map is n, that means we have n+1 number of intervals to place the loose markers in. Thus, the number of clusters K can be set as n+1, if the number of loose markers is less than the number of markers in the initial framework map, then we consider each loose marker as a single cluster.

2) Insert loose markers in their best positions: The kmeans algorithm will group the loose markers into clusters, where markers within clusters are too close to each other and far away from other clusters' markers. Generally, for large datasets singleton clusters, clusters with only one marker, are considered unreliable clusters and will be filtered out, as they are not attached to any other loose markers. After we group the loose markers into clusters, we choose randomly a loose marker from each cluster and try to place that loose marker into its best position in the initial framework map. To find the best position of that loose marker, the *Buildfw* command in Carthagene is used to report the best position for each loose marker, where the loose marker is inserted in all possible positions in the framework map, and then reports the best position based on the linkage of the loose marker with its immediate neighbors (left and right markers, if possible). After we find the best position of each loose marker, we add them to their best positions. In case there are more than one loose marker in one position, we randomly pick one marker and keep it in that position. Step 1 and Step 2, in Phase 2, can be repeated to include more markers in the framework map, where the final map of each iteration is used as an input for the next iteration to map as many loose markers as needed. Figure 4 shows the systematic workflow of Phase 2.

4. Experimental Results

4.1. Datasets

The three common used human genome radiation hybrid panels are: 1) The G3 [19] and 2) TNG [20] panels produced by Stanford University and 3) The Genebridge 4 [21] panel by the Sanger Center. In this study, we use the Genebridge 4 panel where the number of individuals is 93. We have selected 8 different chromosomes with varying number of markers to show the scalability of our proposed approach over the increasing number of markers. Table 1 shows the selected chromosomes with their numbers of markers. The choice of these chromosomes is determined by the availability of markers in both radiation hybrid data set and physical marker locations. The physical marker locations are extracted from the Ensemble website [22], and the RH dataset are downloaded from the EMBL-EBI website [23].

4.2. Evaluation of the approach

The proposed mapping approach builds maps in two phases: In Phase 1, we construct initial map with the most reliable markers. Once the initial map is constructed, Phase 2 extends the initial map to include more markers in the final map. In this section, we are going to refer to Phase 1 as Clustering Method; and Phase 1 and Phase 2 together as Extended Clustering Method. The Clustering Method has been evaluated in [1], where the reported results show that the constructed maps have high agreement with the corresponding physical maps. To evaluate the Extended Clustering Method, we use the same dataset we have used in [1], the constructed maps reported in [1] will be used as an input to Phase 2 in the Extended Clustering Method.

The running time for the Clustering Method has been reported in [1] and can be seen in Figure 6. The big improvement in the mapping runtime of our Clustering Method over the Carthagene Method can be explained by the grouping of the large numbers of markers into small groups, and taking advantage of parallel processing to map all these groups of markers simultaneously in a short time. The hieratical clustering in the Clustering Method and the Kmeans clustering in the Extended Clustering Method only take a few seconds to complete. On the other hand, the Carthagene Method, uses only three markers as an initial map and then incrementally adds more markers to that initial map one marker at a time to construct the maps. Figure 6 shows the mapping running time for both Carthagene Method and the Clustering Method using the same dataset. The Extended Clustering Method running time depends on two factors; first, the numbers of markers in the initial map, and second, the number of iterations; where the more markers there are in an initial map, the less time it takes to map the remaining markers. Table 1 shows that the Clustering Method can generate initial maps with large numbers of markers compared to the number of markers in the maps constructed using the Carthagene Method.

The accuracy of the constructed maps is measured by Pearson Correlation, where the closer the value is to 1, the stronger the linear correlation (agreement) between the markers positions in the constructed maps, and the markers positions in the corresponding physical maps. Table 1 shows the Pearson Correlation between the physical maps and the constructed maps using 1) Carthagene tool; 2) The Clustering Method; and 3) The Extended Clustering Method. The results show that maps generated using the Clustering Method have higher correlations with the physical maps than the maps generated using the Carthagene tool. Out of 10 chromosomes, the Carthagene tool outperforms the Clustering Method for only one chromosome, Chromosome 21, with a slight difference. While, the Clustering Method outperforms the

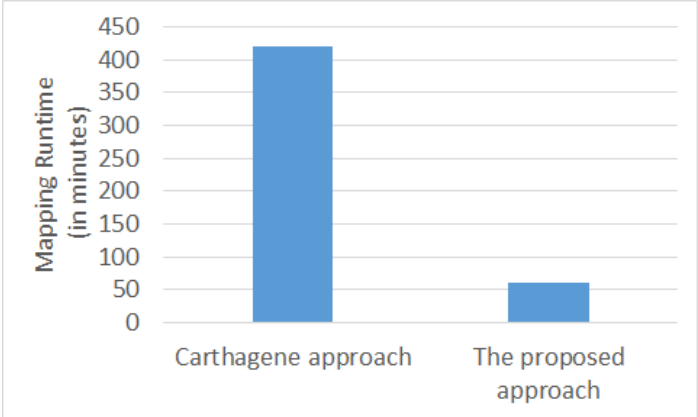


Figure 6. The proposed approach mapping runtime vs the traditional Carthagene approach.

remaining 9 chromosomes. For some chromosomes the improvement is huge, for example, Chromosome 3, the accuracy of the Carthagene Method is 0.45 while the accuracy of the Clustering Method is 0.99. Another example is Chromosome 12, where the accuracy of Carthagene Method is 0.10, while the accuracy of Clustering method is 0.73. For some other chromosomes, the difference in accuracy between the Carthagene Method and Clustering Method is small, for example: Chromosome 10, the accuracy of Carthagene Method is 0.98, while the accuracy of the Clustering method is 0.99. Also in Chromosome 21, the accuracy of Carthagene Method is 0.98, while the accuracy of Clustering method is 0.97.

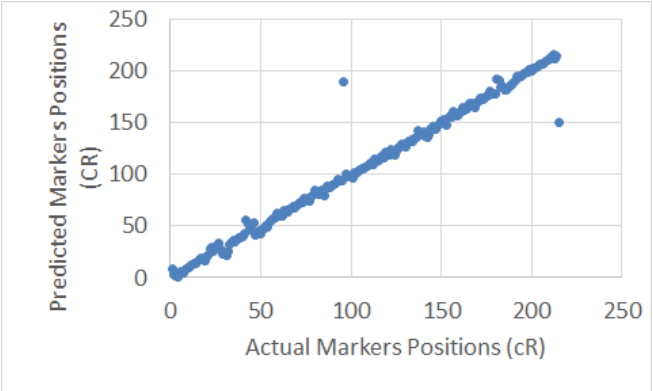
Moreover, Table 1 shows the number of mapped markers in each chromosome using all three methods. The results show that for most chromosomes the constructed maps using the Clustering Method have more markers than the constructed maps using the Carthagene Method. One of the limitation of the Carthagene Method is that the constructed maps have only a few numbers of markers, and this can be seen in Chromosomes 10 and 12. For Chromosome 10, Carthagene maps only 16 markers, where the Clustering Method maps 46 markers. For Chromosome 12, Carthagene maps 42 markers, while the Clustering Method maps 71 markers. For other chromosomes the numbers of mapped markers between the Carthagene Method and The Clustering Methods are similar and this can be seen in Chromosome 16, 21 and 22 where the difference in mapped markers is small. In some cases, Carthagene tool maps more markers than the Clustering Method, for example Chromosome 3 and 7. However, the correlations between the constructed maps and the physical maps for these chromosomes are not as strong as our proposed approach maps for the same corresponding chromosomes.

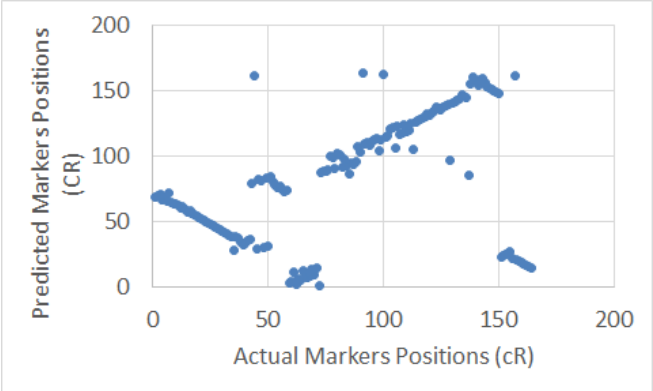
Table 1- Comparison Of The Number Of Mapped Markers And Correlation Between The Constructed Maps And The Physical Maps For Carthagene Method, Clustering Method, Extended Clustering Method.

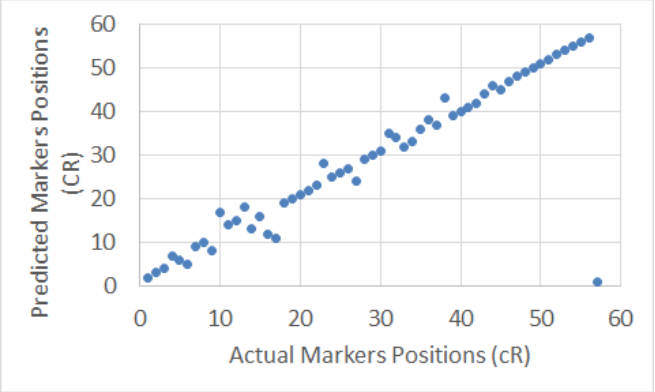
	Input Markers	Carthagene	Clustering	Extended Clustering
Chromosome 3 No. Of Markers In Map Pearson Correlation	1038 -	164 0.45	153 0.99	215 0.99
Chromosome 5 No. Of Markers In Map Pearson Correlation	1071 -	49 0.84	84 0.94	115 0.88
Chromosome 7 No. Of Markers In Map Pearson Correlation	1026 -	165 0.71	78 0.88	100 0.82
Chromosome 10 No. Of Markers In Map Pearson Correlation	846 -	16 0.98	46 0.99	57 0.88
Chromosome 12 No. Of Markers In Map Pearson Correlation	1028 -	42 0.10	71 0.73	81 0.68
Chromosome 15 No. Of Markers In Map Pearson Correlation	619 -	83 0.88	77 0.99	110 0.80
Chromosome 16 No. Of Markers In Map Pearson Correlation	677 -	70 0.53	69 0.95	87 0.89
Chromosome 18 No. Of Markers In Map Pearson Correlation	407 -	65 0.95	84 0.99	103 0.99
Chromosome 21 No. Of Markers In Map Pearson Correlation	173 -	33 0.98	34 0.97	46 0.95
Chromosome 22 No. Of Markers In Map Pearson Correlation	314 -	41 0.64	36 0.71	41 0.59

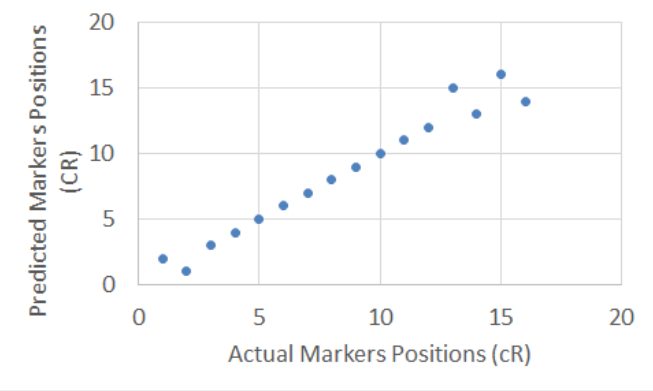
The main goal of the Extended Clustering Method is to map more markers to build high resolution maps. The Extended Clustering Method is an iterative process, where in each iteration some markers are added to the current map to form a new map, and that new map will be used in the next iteration to map more markers, and so on. Table 1 shows the number of mapped markers for each chromosome using the Extended Clustering Method for one iteration. For Chromosome 3, the Extended Clustering Method mapped 215 markers, while the Clustering Method mapped 153 markers. Although, the number of mapped markers is increased in the Extended Clustering Method map, the accuracy of the constructed map remains high 0.99. The same pattern is shown for Chromosome 18. For other chromosomes, Chromosomes 5 and 7, the number of mapped markers is increased and the accuracy of the final maps is decreased slightly. In other cases, for example Chromosome 12, The Extended Clustering Method mapped more markers, but the accuracy dropped from 0.73 to 0.68. Although the accuracy of Chromosome 12 was dropped to 0.68, it is still higher than the accuracy of the Carthagene map, 0.10.

To show graphical representations of the alignment of the constructed maps for some chromosomes using both the Carthagene Method and the Extended Clustering Method, we plot the known markers positions along the x-axis, and the predicted markers position along the y-axis. The plots show how well the constructed maps agree with the corresponding physical maps. The diagonal line in each plot shows the perfect alignment between the predicted markers positions and the actual markers positions. Figures 7 to 22 show the maps for all 10 chromosomes. Our approach is designed to build robust maps. The resampling and clustering techniques are intended to filter our unreliable markers and map only reliable markers. Moreover, mapping markers inside a cluster does not affect the mapping of the other markers outside that cluster. Thus, if there is a noisy marker in a cluster, the effect of that noisy marker will be limited to only the markers inside that cluster, other markers outside that cluster will not be affected; and that can been seen through the local flipping markers' positions errors in our proposed approach constructed maps.



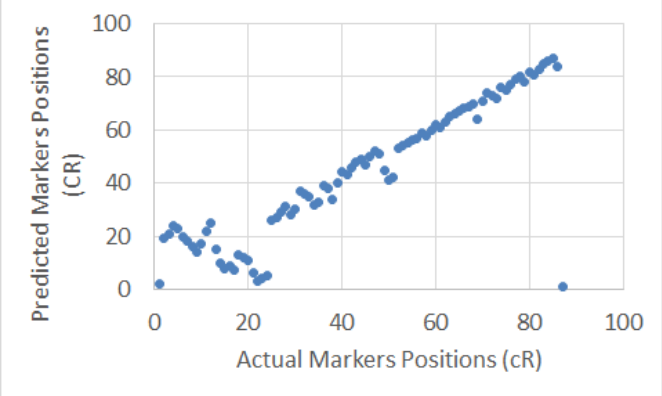


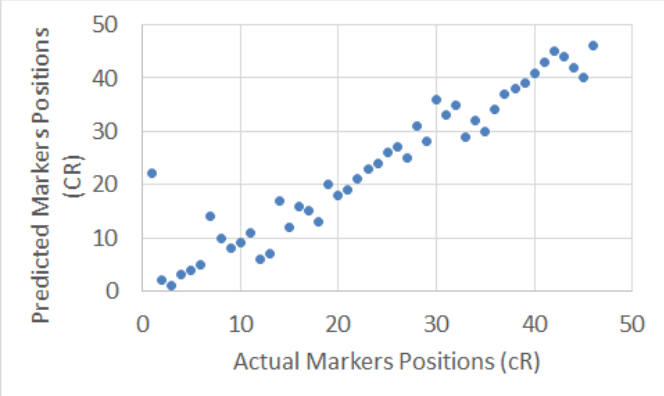


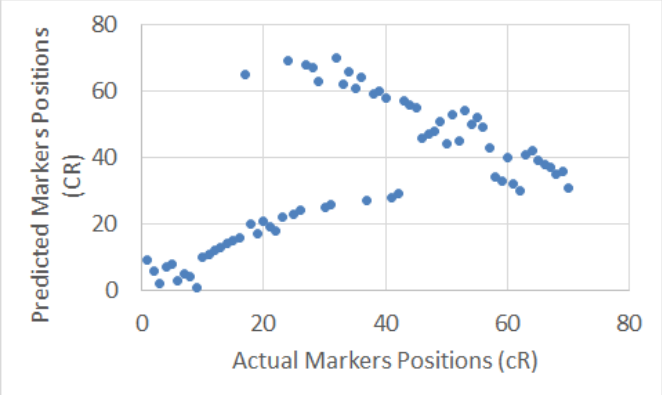


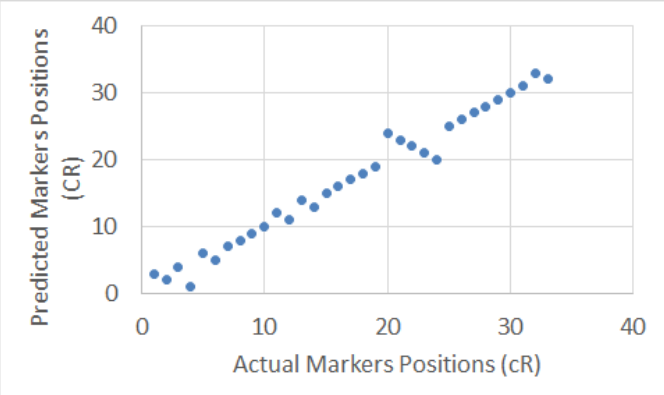
5. Conclusion

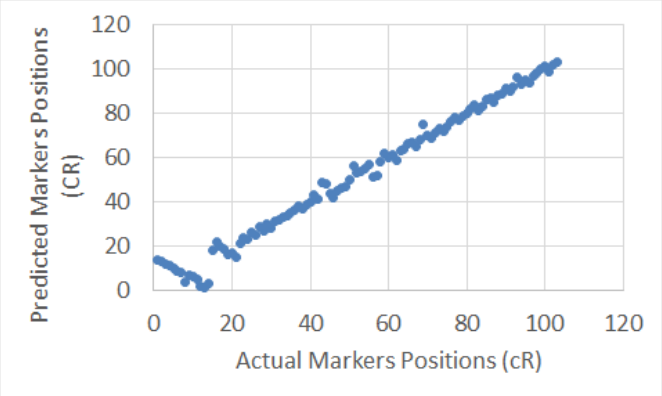
In this research, we have proposed a scalable approach for building high resolution maps. The proposed approach can take advantage of the parallel computing to map large numbers of markers in a short time, thus reduce the computational complexity of the mapping process. The proposed approach can be summarized in two phases: Phase 1, generates resampled datasets, then group all datasets into small clusters to filter out unreliable markers and construct consensus clusters. These clusters are mapped in parallel. Once the initial map is constructed, Phase 2 can be used to iteratively add more markers to the initial map and build high resolution maps. Experiment results on the human genome show that the proposed approach has a very low computational complexity and produces solid maps with high agreement with the physical maps. Also, the results show that our approach outperforms a state-of-the-art tool for building radiation hybrid maps in terms of accuracy of the constructed maps and mapping runtime.

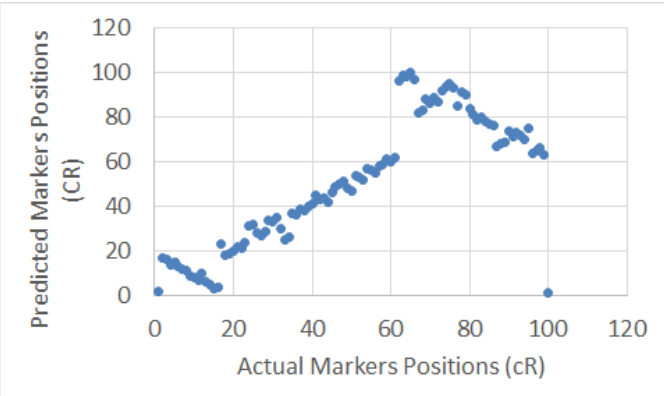

Predicted Markers Positions (CR)
Actual Markers Positions (cR)
0
20
40
60
80
100

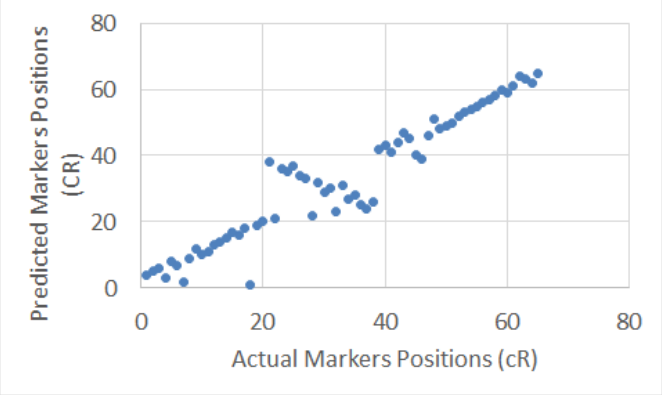

Predicted Markers Positions (CR)
Actual Markers Positions (cR)
0
10
20
30
40
50

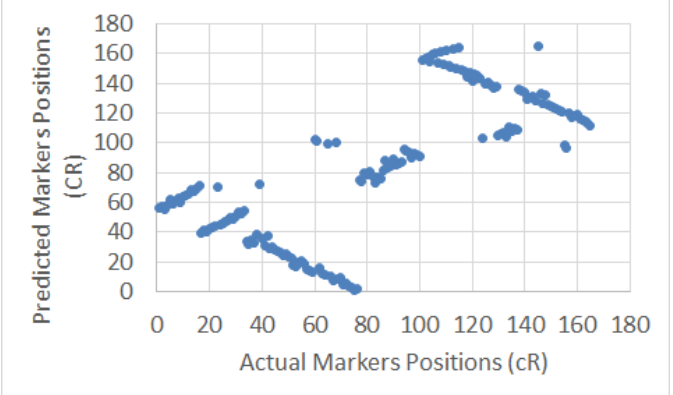

Predicted Markers Positions (CR)
Actual Markers Positions (cR)
0
20
40
60
80

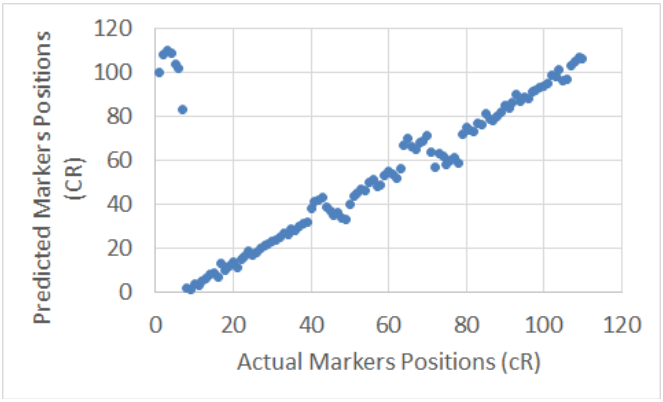

Predicted Markers Positions (CR)
Actual Markers Positions (cR)
0
10
20
30
40

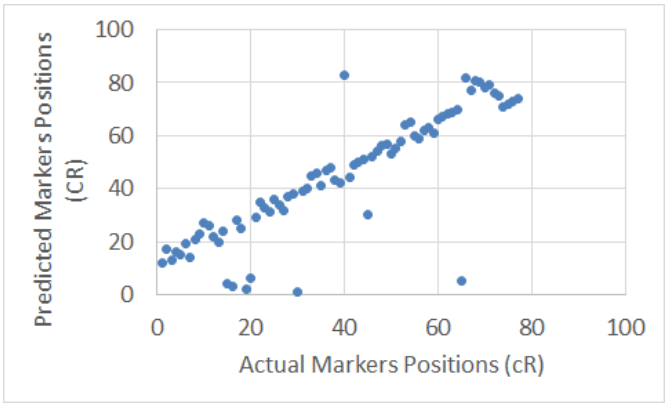

Predicted Markers Positions (CR)
Actual Markers Positions (cR)
0
20
40
60
80
100
120

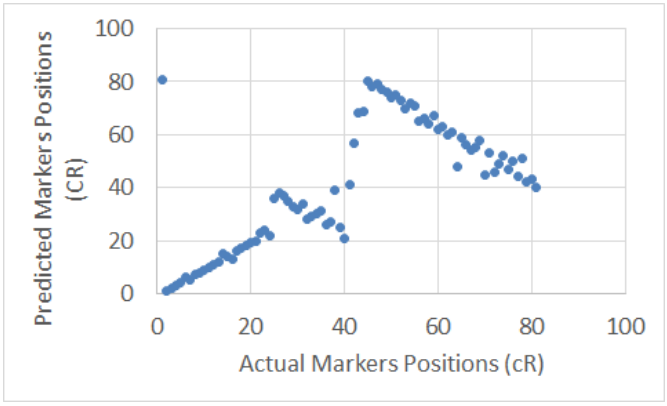

Predicted Markers Positions (CR)
Actual Markers Positions (cR)
0
20
40
60
80
100
120

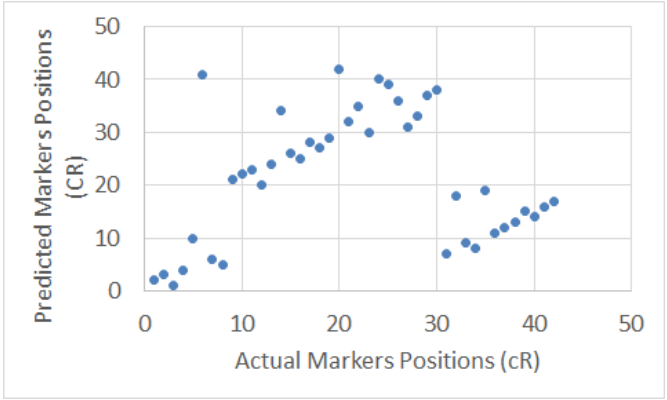

Predicted Markers Positions (CR)
Actual Markers Positions (cR)
0
20
40
60
80


Predicted Markers Positions (CR)
Actual Markers Positions (cR)
0
20
40
60
80
100
120
140
160
180









References

[1] Seetan, R.I., Bible, J., Karavias, M., Seitan, W. and Thangiah, S., 2016, December. Consensus Clustering: A Resampling-Based Method for Building Radiation Hybrid Maps. In Machine Learning and Applications (ICMLA), 2016 15th IEEE International Conference on (pp. 240-245). IEEE. doi: 10.1109/ICMLA.2016.0047

[2] Kalavacharla V, Hossain K, Riera-Lizarazu O, Gu Y, Maan SS, Kianian SF. Radiation hybrid mapping in crop plants. Advances in agronomy. 2009 Dec 31;102:201-22. doi: 10.1016/s0065-2113(09)01006-2

[3] Goss SJ, Harris H. New method for mapping genes in human chromosomes. Nature. 1975 Jun 26;255(5511):680-4. doi: 10.1038/255680a0

[4] Seetan, R.I., Denton, A.M., Al-Azzam, O., Kumar, A., Iqbal, M.J. and Kianian, S.F., 2014. Reliable radiation hybrid maps: an efficient scalable clustering-based approach. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 11(5), pp.788-800. doi: 10.1109/tcbb.2014.2329310

[5] Ben-Dor A, Chor B, Pelleg D. RHO—radiation hybrid ordering. Genome Research. 2000 Mar 1;10(3):365-78. doi: 10.1101/gr.10.3.365

[6] Press WH, Teukolsky SA, Vetterling WT, Flannery BP. Numerical Recipes in Fortran 77, The Art of Scientific Computing, 1992. doi: 10.5860/choice.30-5638a

[7] Cvijovic D, Klinowski J. Taboo search: an approach to the multiple minima problem. Science. 1995 Feb 3;267(5198):664-6.

[8] Helsgaun K. An effective implementation of the Lin–Kernighan traveling salesman heuristic. European Journal of Operational Research. 2000 Oct 1;126(1):106-30. doi: 10.1016/s0377-2217(99)00284-2

[9] Seetan, R.I., Denton, A.M., Al-Azzam, O., Kumar, A., Iqbal, M.J. and Kianian, S.F., 2014, September. A noise-aware method for building radiation hybrid maps. In Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (pp. 541-550). ACM. doi: 10.1145/2649387.2649443

[10] SDM. 2014. Seetan, R.I., Denton, A.M., Al-Azzam, O., Kumar, A., Sturdivant, J. and Kianian, S.F., 2014, April. Iterative framework radiation hybrid mapping. In Proceedings of the 2014 SIAM International Conference on Data Mining (pp. 1028-1036). Society for Industrial and Applied Mathematics. doi: 10.1137/1.9781611973440.117

[11] De Givry S, Bouchez M, Chabrier P, Milan D, Schiex T. CARHTA GENE: multipopulation integrated genetic and radiation hybrid mapping. Bioinformatics. 2004 Dec 14;21(8):1703-4. doi: 10.1093/bioinformatics/bti222

[12] Stein L, Kruglyak L, Slonim D, Lander E. RHMAPPER, unpublished software. Whitehead Institute/MIT Center for Genome Research. 1995. doi: 10.1093/bioinformatics/14.6.538

[13] Matise TC, Perlin M, Chakravarti A. Automated construction of genetic linkage maps using an expert system (MultiMap): a human genome linkage map. Nature genetics. 1994 Apr 1;6(4):384-90. doi: 10.1038/ng0494-384

[14] Morton NE. Sequential tests for the detection of linkage. American journal of human genetics. 1955 Sep;7(3):277. doi: 10.1002/gepi.1370060402

[15] Ronin Y, Mester D, Minkov D, Korol A. Building reliable genetic maps: different mapping strategies may result in different maps. Natural Science. 2010 Jun 30;2(06):576. doi: 10.4236/ns.2010.26073

[16] Mester, D.I., Ronin, Y.I., Korostishevsky, M.A., Pikus, V.L., Glazman, A.E. and Korol, A.B., 2006. Multilocus consensus genetic maps (MCGM): formulation, algorithms, and results. Computational biology and chemistry, 30(1), pp.12-20. doi: 10.1002/gepi.1370030105

[17] Chaubey YP. Resampling methods: A practical guide to data analysis. Springer, 2001. doi: 10.2307/1271090

[18] Sneath PH, Sokal RR. Numerical taxonomy. The principles and practice of numerical classification. 1973. doi: 10.2307/2285473

[19] Stewart EA, McKusick KB, Aggarwal A, Bajorek E, Brady S, Chu A, Fang N, Hadley D, Harris M, Hussain S, Lee R. An STS-based radiation hybrid map of the human genome. Genome research. 1997 May 1;7(5):422-33.

[20] Lunetta KL, Boehnke M, Lange K, Cox DR. Selected locus and multiple panel models for radiation hybrid mapping. American journal of human genetics. 1996 Sep;59(3):717.

[21] Gyapay G, Schmitt K, Fizames C, Jones H, Vega-Czarny N, Spillett D, Muselet D, Prud'Homme JF, Dib C, Auffray C, Morissette J. A radiation

hybrid map of the human genome. Human Molecular Genetics. 1996 Mar 1;5(3):339-46. doi: 10.1093/hmg/5.3.339

[22] Kinsella RJ, Kähäri A, Haider S, Zamora J, Proctor G, Spudich G, Almeida-King J, Staines D, Derwent P, Kerhornou A, Kersey P. Ensembl BioMarts: a hub for data retrieval across taxonomic space. Database. 2011 Jan 1;2011. doi: 10.1093/database/bar030

[23] Amid C, Birney E, Bower L, Cerdeño-Tárraga A, Cheng Y, Cleland I, Faruque N, Gibson R, Goodgame N, Hunter C, Jang M. Major submissions tool developments at the European Nucleotide Archive. Nucleic acids research. 2011 Nov 12;40(D1):D43-7**.** doi: 10.1093/nar/gkr946

[24] Kumar A, Seetan R, Mergoum M, Tiwari VK, Iqbal MJ, Wang Y, Al-Azzam O, Šimková H, Luo MC, Dvorak J, Gu YQ. Radiation hybrid maps of the D-genome of Aegilops tauschii and their application in sequence assembly of large and complex plant genomes. BMC genomics. 2015 Oct 16;16(1):800. doi: 10.1186/s12864-015-2030-2

[25] Kumar A, Mantovani EE, Seetan R, Soltani A, Echeverry-Solarte M, Jain S, Simsek S, Doehlert D, Alamri MS, Elias EM, Kianian SF. Dissection of genetic factors underlying wheat kernel shape and size in an elite× nonadapted cross using a high density SNP linkage map. The plant genome. 2016 Mar 1;9(1). doi: 10.3835/plantgenome2015.09.0081

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1401-1412 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Two-axis Gimbal with Passive Revolute Joints for Motion Isolation of a Stabilizing Mechanism

Erick Moya*, [1], Keizo Araki[2], Masaki Okada[2], Akira Mizuno[2], Kenjiro Tadakuma[3], Riichiro Tadakuma[1]

[1]*Department of Mechanical Systems Engineering, Yamagata University, 992-8510, Japan*

[2]*Equos Research, 101-0021, Japan*

[3]*Department of Mechanical and Aerospace Engineering, Tohoku University, 980-8579, Japan*

ARTICLE INFO

Article history:
Received: 31 May, 2017
Accepted: 20 July, 2017
Online: 05 August, 2017

Keywords:
intertially stabilized platform
mass stabilization
omnidirectional gear
passive gimbal

ABSTRACT

Energy efficiency is a salient design consideration for stabilization mechanisms when assembled on a battery-powered vehicle. In this paper we study the implementation of a two-axis gimbal with passive revolute joints for motion isolation of an inertially stabilized platform (ISP) mounted on a host vehicle. The objective is for the ISP to maintain a steady line of sight (LOS) of a payload placed in an arbitrary position on the ISP workspace by performing mass stabilization. The proposed method eliminates the need to operate gimbal motors throughout the operation of the host vehicle, stabilizing the platform LOS by relying on the gimbal passive joints acting as suspension to isolate the ISP from disturbances imparted by the host vehicle. We describe the principle of operation and a prototype of the proposed stabilizing mechanism is presented.

1. Introduction

An inertially stabilized platform consists of a gimballed electromechanical system that maintains a steady line of sight (LOS) of a payload (imaging instrument, communication instrument, sensor, etc.) with respect to an inertial reference frame. The control system adjusts the gimbal orientation angle to compensate for the changes in orientation of the host vehicle on which it is mounted. The angular motion of the host vehicle is measured with an inertial measurement unit (IMU). The stabilizing algorithm calculates the necessary coordinate frame transformations in order to derive the target orientation of the gimbal in inertial space for the platform to hold its line of sight.

For the payload to maintain its line of sight, the stabilizing system must have a quick response time. Moreover, in conventional applications of gimballed stabilization, the system must actively drive the gimbal actuators throughout the operation of the host vehicle to compensate its changes in orientation and to reject other disturbances. This turns out to be computationally expensive, since the required kinematic calculations and the calculations of target angular velocities of the gimbal motors have to be performed fast enough not to impose a lag in response time of the system. Perhaps more importantly, the active gimbal compensation throughout the operation of the vehicle demands a high consumption of power.

In this research, we study the implementation of a two-axis gimbal with unactuated freely rotating joints for disturbance isolation of an inertially stabilized platform (ISP). A payload is placed by a user in an arbitrary position within the ISP workspace, and the line of sight stabilization of the payload is achieved by displacing a counterweight horizontally and vertically to adjust the center of gravity (COG) of the payload-counterweight system to be at the crossing point of the passive gimbal rotation axes. As a result, undesired moments of inertia applied to the platform by external disturbances are eliminated. The passive gimbal joints act as a suspension that decouples the host vehicle chassis from the ISP, isolating it from disturbances and keeping the payload's orientation stationary in inertial space.

The work presented in this paper is an extension of the stabilization method originally presented in IEEE SII2016 [1]. In the present work, we first discuss the method of elimination of moments of inertia to isolate the payload from external disturbance. We next discuss a method to calculate the position of the payload COG on the surface of the platform and present a theoretical framework to calculate the height of the payload COG by measuring the roll and pitch inclination angles of the passive gimbal. Finally, we present the results of fundamental experiments with the second prototype to evaluate the mechanism.

*Corresponding Author: Erick Moya, 4-3-16 Jonan Yonezawa-shi, (+81)238-26-3893, Japan | Email: tym80862@st.yamagata-u.ac.jp

2. Balancing Principle

The aim of the stabilizing mechanism discussed in this paper is to keep a steady line of sight of a payload of arbitrary weight within the load capacity specifications of the mechanism, and with arbitrary mass distribution, placed in an arbitrary position on the stabilized platform's workspace. To accomplish this by performing mass stabilization it is therefore necessary to measure the payload weight, and to calculate the position of the center of gravity of the payload [2, 3]. With this information about the payload, the three-dimensional target position of the counterweight is calculated to perform mass-stabilization.

Due to the freely rotating joints of the gimbal, placing a payload on the platform in a position offset from the vertical axis of symmetry of the ISP induces a torque on the passive gimbal's axes; this results in the inclination of the platform about the gimbal axes. The stabilizing mechanism aligns the orientation of a payload with respect to the inertial reference frame by adjusting the horizontal and vertical position of a counterweight. An XY Stage mechanism displaces the counterweight center of gravity horizontally on the *x-y* plane and a Z Stage displaces the counterweight vertically along the *z* axis. The position of counterweight is adjusted to locate the center of gravity of the payload-counterweight system at the crossing point of the gimbal rotation axes.

2.1. Static Balance: Gravity

Balance on roll axis: the offset position r_1 of the payload with respect to the platform's vertical axis of symmetry on the *y-z* plane induces a torque on the passive roll axis of the gimbal due to gravity, resulting in an inclination angle σ of the gimbal about its roll axis of rotation, as illustrated in Figure 1. To compensate the torque induced by the payload, the horizontal position of the counterweight is displaced along the *y* axis with the Y Stage up to position r_2 to induce a torque with the same magnitude and opposite direction. For static balance on the roll axis of a payload with weight w_1 and counterweight with weight w_2, the position r_2 of the counterweight on the *y* axis must satisfy (1). This places the center of gravity of the payload-counterweight system on the vertical axis of symmetry of the inertially stabilized platform on the *y-z* plane as illustrated in Figure 2.

$$r_2 = \frac{w_1 r_1}{w_2} \tag{1}$$

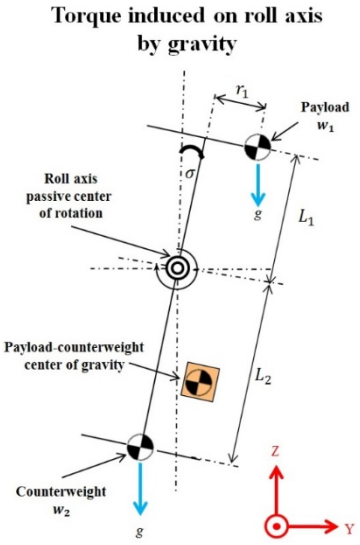


Figure 1. Inclination of the passive gimbal about the roll axis due to load unbalance against gravity on the *y* axis.

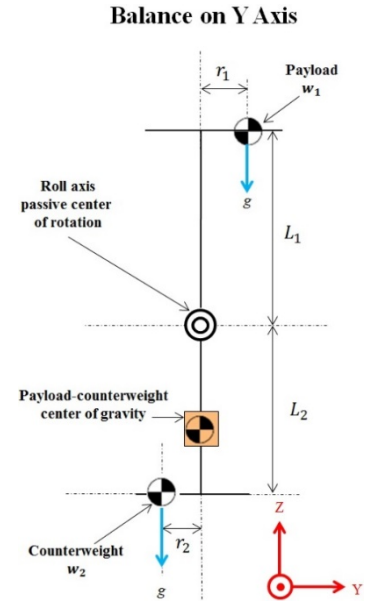


Figure 2. Static balance on the *y* axis against gravity provided by the displacement of the counterweight with the Y Stage.

Balance on pitch axis: the offset position r_3 of the payload with respect to the platform's vertical axis of symmetry on the *x-z* plane induces a torque on the passive pitch axis of the passive gimbal due to gravity, resulting in an inclination angle θ of the gimbal about its pitch axis of rotation, as illustrated in Figure 3. The horizontal position of the counterweight is displaced along the *x* axis with the X Stage to position r_4, which for static balance must satisfy (2). This places the center of gravity of the payload-counterweight system on the vertical axis of symmetry of the inertially stabilized platform on the *x-z* plane, as illustrated in Figure 4.

$$r_4 = \frac{w_1 r_3}{w_2} \tag{2}$$

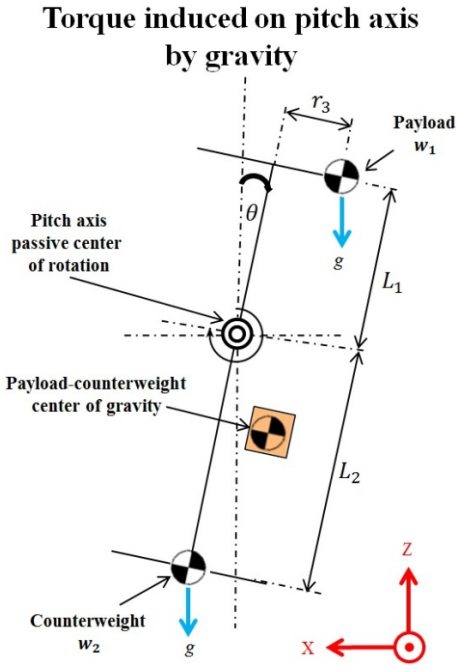


Figure 3. Inclination of the passive gimbal about the pitch axis due to load unbalance against gravity on the *x* axis.

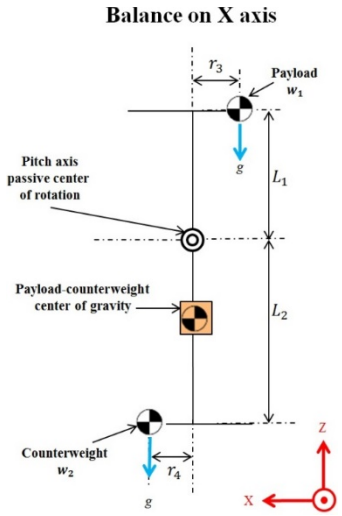


Figure 4. Static balance on the *x* axis against gravity provided by the displacement of the counterweight by the X Stage.

2.2. Static Balance: Disturbances by Host Vehicle

Accelerations of the host vehicle also induce torque on the passive gimbal. Depending on the direction of motion of the host vehicle, this torque is induced on either the pitch axis of the gimbal, the roll axis of the gimbal, or both. Figure 5 illustrates the inclination of the passive gimbal due to a horizontal acceleration a of the host vehicle along the x axis of the stabilizing mechanism, inducing a toque on the pitch axis of the gimbal.

To eliminate the torque induced on the passive gimbal due to accelerations of the host vehicle, the vertical position of the counterweight is displaced along the z axis with the Z Stage. For static balance the along the z axis of a payload with weight w_1 and counterweight with weight w_2 , the distance L_2 from the gimbal axis of rotation to the counterweight must satisfy (3)

$$L_2 = \frac{w_1 L_1}{w_2} \tag{3}$$

where L_1 is the distance from the gimbal axis of rotation to the ISP. After the center of gravity of the payload-counterweight system is displaced horizontally by the XY Stage to place it on the vertical axis of symmetry of the gimbal, the vertical displacement along the z axis places the center of gravity of the payload-counterweight system on the crossing point of the passive gimbal rotation axes, as illustrated in Figure 6.

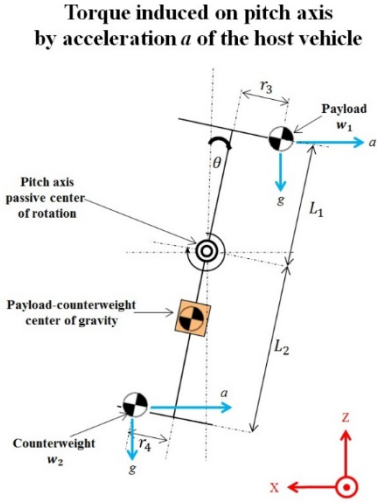


Figure 5. Inclination of the passive gimbal about the pitch axis due to load unbalance against the acceleration of the host vehicle.

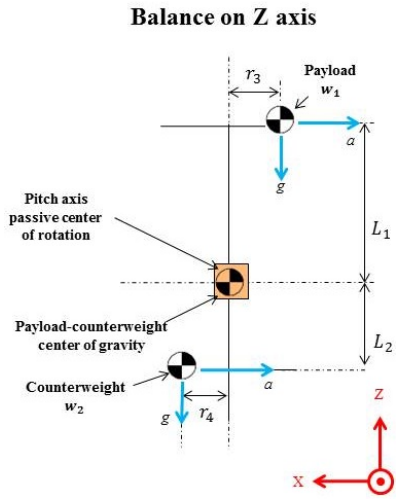


Figure 6. Static balance on *z* axis against host vehicle acceleration provided by the displacement of the Z Stage.

3. Components of the Stabilizing Mechanism

Figure 7 illustrates the components of the proposed stabilizing mechanism. The stabilizing mechanism consists of an XY Stage mechanism with omnidirectional driving gear [4] located below the centers of rotation of the two-axis passive gimbal. The XY Stage acts as counterweight, with adjustable position along the x and y axes. The XY Stage/Counterweight is mounted onto the Z Stage that adjusts its vertical position along the z axis. The stabilized platform is located above the centers of rotation of the passive gimbal roll and pitch axes.

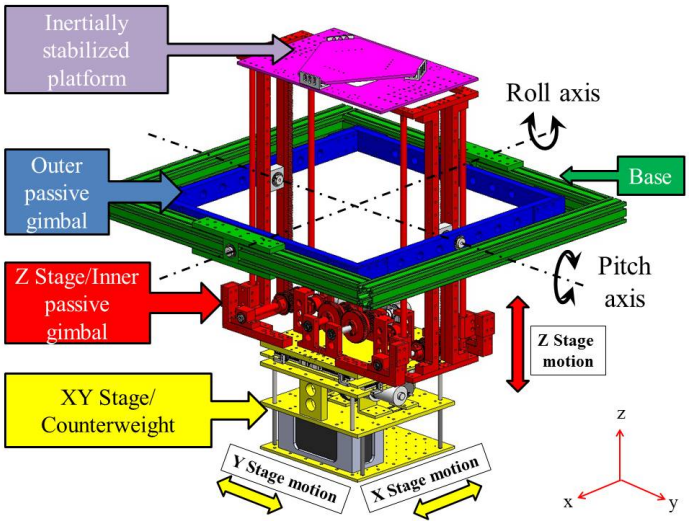


Figure 7. Stabilizing mechanism with passive two-axis gimbal.

The payload weight and the position of its COG on the surface of the ISP are measured with three CZL204E strain gauge load cells located on the platform. We propose the calculation of the height of the payload COG by measuring the roll and pitch inclination angles of the passive gimbal caused by the placement of the payload on the ISP. The passive gimbal's roll and pitch inclination angles are measured with a SEN-10736 inertial measurement unit onboard the inner gimbal of the stabilizing mechanism [5]. The proposed method for calculating the three dimensional position of the payload COG is discussed in chapter 4. Figure 8 illustrates the stabilizing mechanism mounted on the omnidirectional host vehicle.

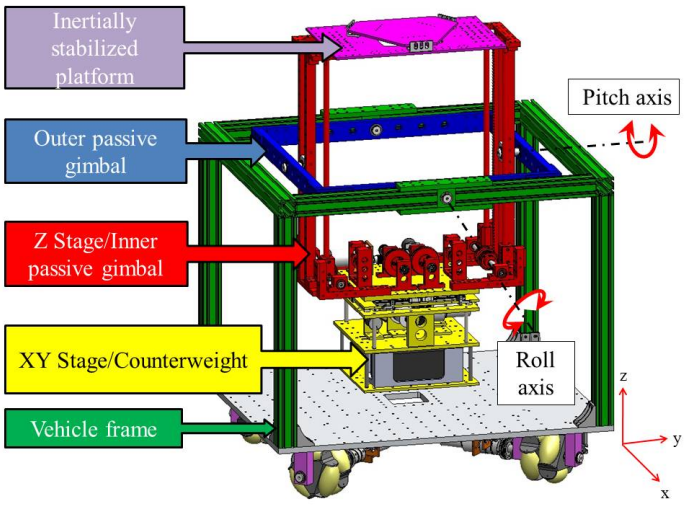


Figure 8. Stabilizing mechanism on the omnidirectional host vehicle.

4. Calculation of the Position of the Payload Center of Gravity on the Inertially Stabilized Platform

In this chapter we discuss the method of calculating the position of the payload's center of gravity (COG) in three dimensions once it has been placed on the inertially stabilized platform. We first describe the coordinate frame assignments to the system, we next describe the method to calculate the position of the payload COG on the surface of the platform ((*x,y*) coordinates of the payload COG on the surface of the platform) employing the multiple-point weighing method [6]. Finally, we propose a method to calculate the height of the payload COG from the platform surface (*z* coordinate of the payload COG from the surface of the platform) by exploiting the passive nature of the two-axis gimbal.

The algorithm to calculate the position of the payload center of gravity in three dimensions is the following:

1. Assign three coordinate frames on the stabilizing mechanism. A coordinate frame Σ_0 as the reference frame, a coordinate frame Σ_1 on the surface of the ISP, and a coordinate frame Σ_2 on the counterweight.
2. Measure the weight of the payload and calculate the horizontal position of its COG on the platform's surface (on ISP coordinate frame Σ_1) after the payload has been placed on the platform by the user.
3. Measure the roll and pitch inclination angles of the passive gimbal caused by the placement of the payload on the ISP surface. Then describe the position of the inclined ISP and the payload COG with respect to the reference frame Σ_0.
4. Define a cross-sectional plane in which both the pitch and roll inclination angles of the ISP are implicit.
5. Describe the position of the inclined ISP and payload COG with respect to the cross-sectional plane.
6. Perform a statics analysis of moments on the cross-sectional plane to calculate the height of the payload COG from the surface of the ISP.

4.1. Coordinate Frame Assignments

The stabilizing mechanism's vertical axis of symmetry lies orthogonal to the plane of intersection of the passive gimbal's roll and pitch axes as shown in Figure 9. Three coordinate frames are assigned to the stabilizing mechanism, illustrated in Figure 10.

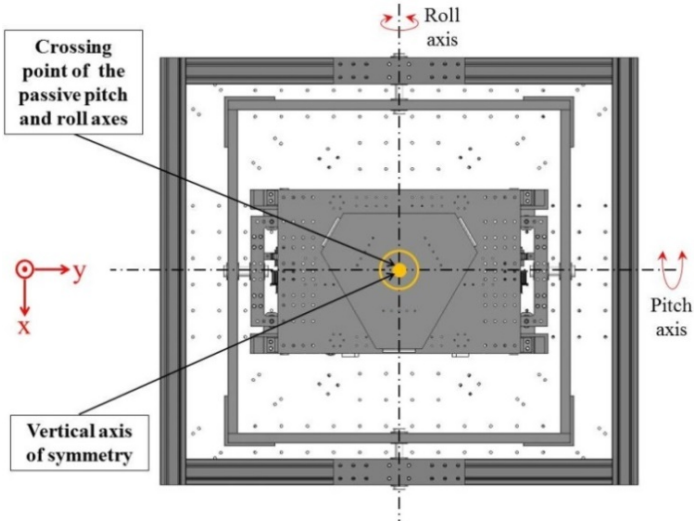


Figure 9. Top view of the stabilizing mechanism on the host vehicle.

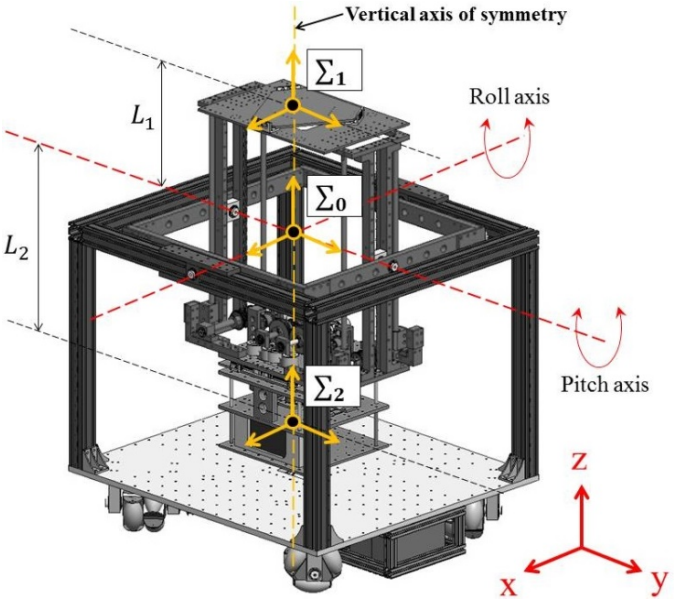


Figure 10. Coordinate frame assignment on the stabilizing mechanism.

A fixed reference coordinate frame Σ_0 is assigned with its origin at the intersection of the passive gimbal's roll and pitch rotation axes and its vertical axis collinear with the stabilizing mechanism's vertical axis of symmetry. Positions and rotations of the ISP and the counterweight are defined with respect to this reference frame.

A coordinate frame Σ_1 is assigned to the surface of the inertially stabilized platform with its vertical axis collinear with the mechanism's vertical axis of symmetry and with its origin at a vertical distance L_1 from the origin of the reference frame Σ_0, above the passive gimbal's pitch and roll rotation axes.

A coordinate frame Σ_2 is assigned to the counterweight with its vertical axis collinear with the balancing mechanism's vertical axis of symmetry and with its origin at a vertical distance L_2 from the origin of the reference frame Σ_0, beneath the passive gimbal's pitch and roll rotation axes.

4.2. Calculation of the Horizontal Coordinates of the Payload Center of Gravity on the Surface of the Inertially Stabilized Platform

In the proposed stabilizing mechanism, the inertially stabilized platform's effective workspace is a tray resting on top of three CZL204E strain gauge load cells, with which the weight of the payload placed on the tray is measured. The layout of the load cells is shown in Figure 11.

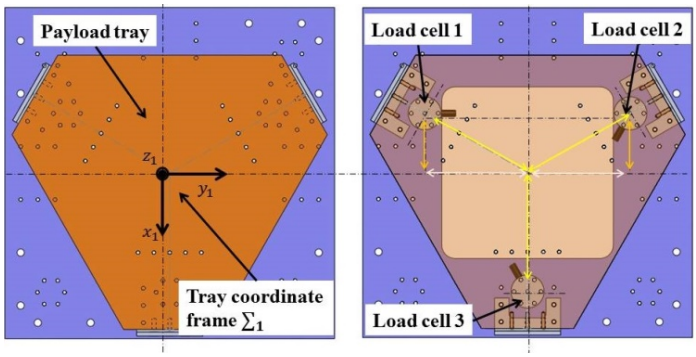


Figure 11. Load cell layout for weight measurement of the payload on the inertially stabilized platform.

The position of the COG of the payload placed on the tray is calculated from the relationship between force measurements of the three CZL204E strain gauge load cells using the multiple-point weighing method. The position coordinates of the payload COG are described with respect to the platform coordinate frame Σ_1, where r_1 is the y coordinate of the position of the payload COG on the platform and r_3 is the x coordinate of the position of the payload COG on the platform, as shown in Figure 12.

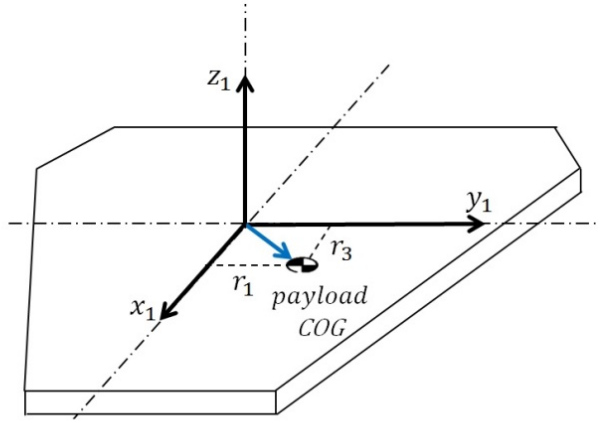


Figure 12. Illustration of the position coordinates r_1 and r_3 of the payload center of gravity on the surface of the inertially stabilized platform.

Figure 13 illustrates the multiple-point weighing method for calculating the position of the payload center of gravity on the surface of the ISP. The weight w_1 of the payload is the sum of the force readings A, B, C of load cell 1, load cell 2, and load cell 3, respectively, described with (4).

$$w_1 = A + B + C \tag{4}$$

The distance r_1 corresponding to the y coordinate of the payload COG on the ISP coordinate frame is calculated with (5).

$$r_1 = \frac{(B)77.942 - (A)77.942}{w_1} \tag{5}$$

The distance r_3 corresponding to the x coordinate of the payload COG on the ISP coordinate frame is calculated with (6).

$$r_3 = 90 - \left[\frac{(A+B)135}{w_1}\right] \tag{6}$$

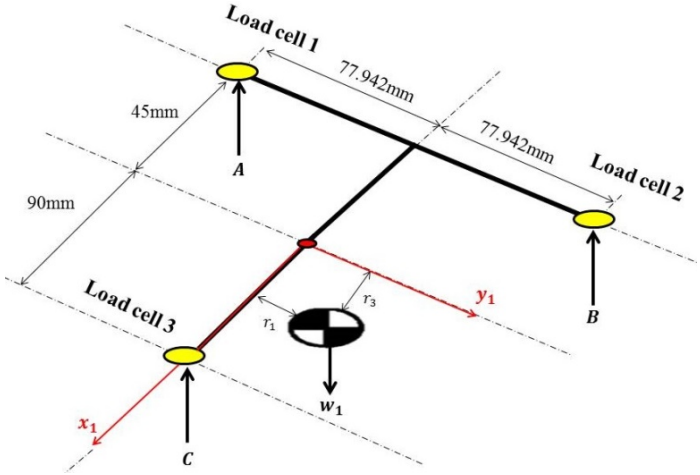


Figure 13. Multiple-point weighing method for calculating the position coordinates of the center of gravity of a payload on the surface of the inertially stabilized platform.

The position coordinates of the payload COG on the surface of the platform are defined with respect to platform coordinate frame Σ_1 with vector ${}^1P_{COG}$ (7).

$${}^1P_{COG} = \begin{bmatrix} r_3 \\ r_1 \\ 0 \end{bmatrix} \tag{7}$$

4.3. Calculation of the Vertical Coordinate of the Payload Center of Gravity from the Surface of the Inertially Stabilized Platform

After calculating the horizontal position coordinates of the payload COG on the surface of the platform as described in the previous section, the next step is to calculate the height h of the payload COG with respect to the platform surface, namely its z position coordinate, as seen on Figure 14.

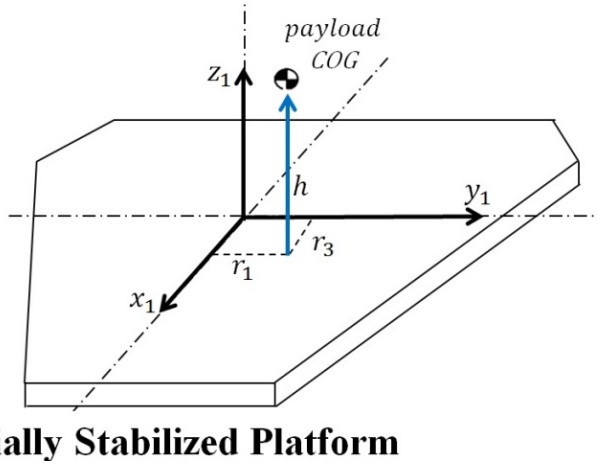


Figure 14. Illustration of the vertical position h of the payload center of gravity from the surface of the inertially stabilized platform.

A payload with weight w_1 initially placed on the platform workspace offset from the stabilizing mechanism's vertical axis of symmetry by distances r_1 and r_3 will cause the passive gimbal to incline along the roll axis by an angle σ and along the pitch axis by an angle θ, as illustrated in Figure 15. After placing the payload on the surface of the ISP, we describe the positions of all the coordinate frames with respect to the reference frame Σ_0. We proceed to calculate the height of the payload COG from the surface of the ISP by analyzing the roll and pitch inclination angles of the passive gimbal.

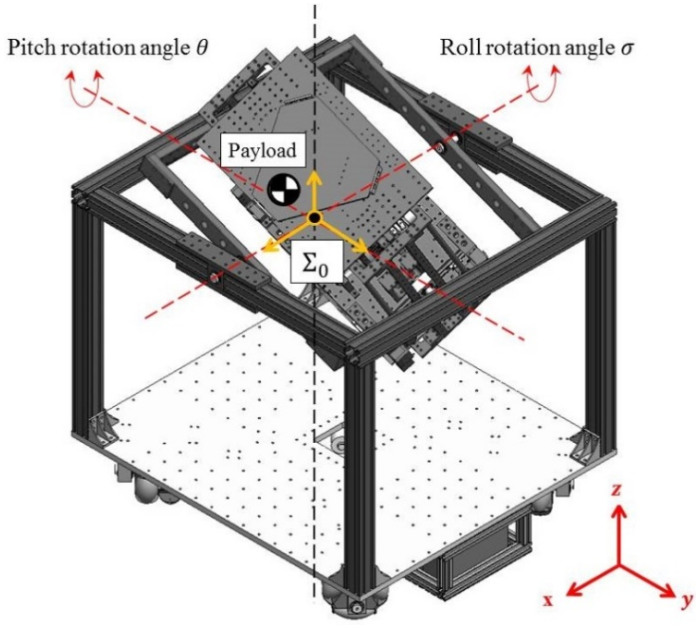


Figure 15. Inclination of the passive gimbal caused by initial placement of a payload on the inertially stabilized platform.

The position coordinates $({}^0P_{origin1x}, {}^0P_{origin1y}, {}^0P_{origin1z})$ represent the position of the origin of ISP coordinate frame Σ_1 with respect to the fixed reference frame Σ_0 resulting after the pitch and roll rotation of the passive gimbal caused by the placement of the payload on the ISP. These coordinates are described by vector ${}^0P_{origin1}$ (8).

$$ {}^0P_{origin1} = \begin{bmatrix} \cos\theta & \sin\sigma\sin\theta & \cos\sigma\sin\theta \\ 0 & \cos\sigma & -\sin\sigma \\ -\sin\theta & \sin\sigma\cos\theta & \cos\sigma\cos\theta \end{bmatrix} * \begin{bmatrix} 0 \\ 0 \\ L_1 \end{bmatrix} = \begin{bmatrix} {}^0P_{origin1x} \\ {}^0P_{origin1y} \\ {}^0P_{origin1z} \end{bmatrix} \quad (8) $$

The position coordinates $({}^0P_{origin2x}, {}^0P_{origin2y}, {}^0P_{origin2z})$ represent the position of the origin of the counterweight coordinate frame Σ_2 with respect to the fixed reference frame Σ_0 resulting after pitch and roll rotation of the gimbal. These coordinates are described by vector ${}^0P_{origin2}$ (9).

$$ {}^0P_{origin2} = \begin{bmatrix} \cos\theta & \sin\sigma\sin\theta & \cos\sigma\sin\theta \\ 0 & \cos\sigma & -\sin\sigma \\ -\sin\theta & \sin\sigma\cos\theta & \cos\sigma\cos\theta \end{bmatrix} * \begin{bmatrix} 0 \\ 0 \\ L_2 \end{bmatrix} = \begin{bmatrix} {}^0P_{origin2x} \\ {}^0P_{origin2y} \\ {}^0P_{origin2z} \end{bmatrix} \quad (9) $$

We now perform a transformation of the position of the payload COG on the surface of the ISP from its description with respect to the ISP local frame (7) to a description with respect to the reference frame Σ_0. The coordinates $({}^0P_{COGx}, {}^0P_{COGy}, {}^0P_{COGz})$ represent the position of the payload COG with respect to the reference frame Σ_0 resulting after pitch and roll rotation of the passive gimbal. They are described by vector ${}^0P_{COG}$ (10).

$$ {}^0P_{COG} = \begin{bmatrix} \cos\theta & \sin\sigma\sin\theta & \cos\sigma\sin\theta & {}^0P_{origin1x} \\ 0 & \cos\sigma & -\sin\sigma & {}^0P_{origin1y} \\ -\sin\theta & \sin\sigma\cos\theta & \cos\sigma\cos\theta & {}^0P_{origin1z} \\ 0 & 0 & 0 & 1 \end{bmatrix} * \begin{bmatrix} r_3 \\ r_1 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} {}^0P_{COGx} \\ {}^0P_{COGy} \\ {}^0P_{COGz} \\ 1 \end{bmatrix} (10) $$

To perform the calculation of the height h of the payload COG from the surface of the ISP, we carry out a statics analysis of the inclined stabilizing mechanism. For this purpose we derive the spherical coordinates (r, β, α) of the ISP coordinate frame Σ_1 with respect to the reference frame Σ_0. The spherical coordinates describe the ISP's radial distance, azimuth angle and elevation angle with respect to the reference frame Σ_0. The spherical coordinates are used to select a cross-sectional plane in which both the pitch and roll inclination of the ISP are simultaneously accounted for. The inclination of the ISP in roll and pitch is simultaneously described by the elevation angle α of the spherical coordinates of the ISP coordinate frame.

The ISP spherical coordinates r, β, and α are defined by (11), (12), and (13) respectively; they are derived from the ISP Cartesian coordinates described by (8) and are all described with respect to the reference coordinate frame Σ_0

$$ r = \sqrt{\left({}^0P_{origin1x}\right)^2 + \left({}^0P_{origin1y}\right)^2 + \left({}^0P_{origin1z}\right)^2} \quad (11) $$

$$ \beta = \tan^{-1} \frac{{}^0P_{origin1y}}{{}^0P_{origin1x}} \quad (12) $$

$$ \alpha = \sin^{-1} \frac{{}^0P_{origin1z}}{r} \quad (13) $$

where:

r = radial distance from the origin of the reference frame Σ_0 to the origin of the ISP coordinate frame Σ_1.

β = azimuth angle of the point ${}^0P_{origin1}$ (origin of the ISP coordinate frame Σ_1).

α = elevation angle of the point ${}^0P_{origin1}$ (origin of the ISP coordinate frame Σ_1).

We proceed to select a cross sectional plane using the description of the position of the ISP in spherical coordinates derived above. A coordinate frame Σ_3 is assigned to the cross-sectional plane, with its origin coincident with the origin of reference frame Σ_0, its x_3 axis collinear with the projection of radial distance r on the x_0, y_0 plane, its z_3 axis collinear with the z_0 axis of the reference frame Σ_0, and with its $y_3$axis orthogonal to axes x_3 and z_3 as illustrated in Figure 16.

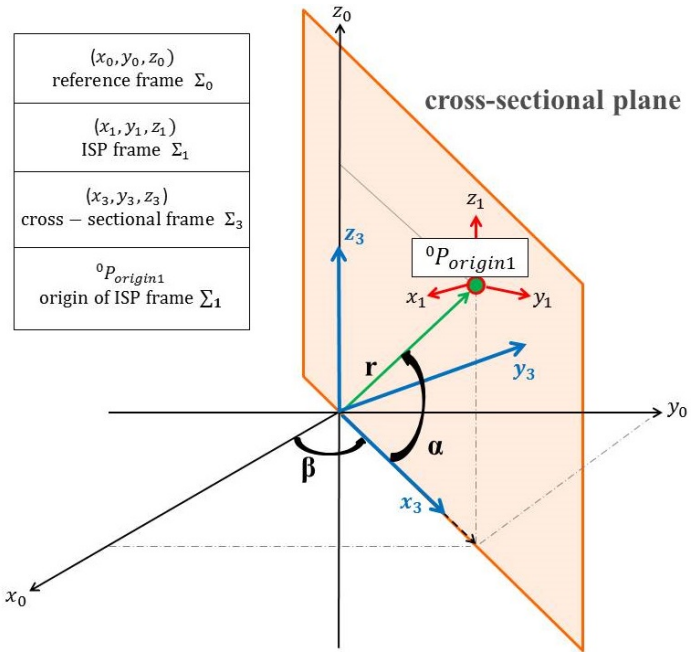


Figure 16. Image of the cross-sectional plane selected from the inertially stabilized platform's spherical coordinates, and the corresponding coordinate frame assignment Σ_3.

We now proceed to describe the position of the origin of the ISP coordinate frame Σ_1, the counterweight coordinate frame Σ_2, and the position ${}^0P_{COG}$ of the payload COG all with respect to the cross-sectional plane frame Σ_3.

The transformation of the position of the origin of the ISP coordinate frame Σ_1 from its description with respect to reference frame Σ_0 (8) to its description with respect to the cross-sectional plane Σ_3 is defined by vector ${}^3P_{origin1}$ (14).

$$ {}^3P_{origin1} = \begin{bmatrix} \cos\beta & \sin\beta & 0 \\ -\sin\beta & \cos\beta & 0 \\ 0 & 0 & 1 \end{bmatrix} * \begin{bmatrix} {}^0P_{origin1x} \\ {}^0P_{origin1y} \\ {}^0P_{origin1z} \end{bmatrix} = \begin{bmatrix} {}^3P_{origin1x} \\ {}^3P_{origin1y} \\ {}^3P_{origin1z} \end{bmatrix} \quad (14) $$

The transformation of the position of the origin of the counterweight coordinate frame Σ_2 from its description with respect to reference frame Σ_0 (9) to its description with respect to the cross-sectional plane Σ_3 is defined by vector ${}^3P_{origin2}$ (15)

$$ {}^3P_{origin2} = \begin{bmatrix} \cos\beta & \sin\beta & 0 \\ -\sin\beta & \cos\beta & 0 \\ 0 & 0 & 1 \end{bmatrix} * \begin{bmatrix} {}^0P_{origin2x} \\ {}^0P_{origin2y} \\ {}^0P_{origin2z} \end{bmatrix} = \begin{bmatrix} {}^3P_{origin2x} \\ {}^3P_{origin2y} \\ {}^3P_{origin2z} \end{bmatrix} \quad (15) $$

The transformation of the position of the payload COG from its description with respect to reference frame Σ_0 (10) to its description with respect to the cross-sectional plane Σ_3 is defined by vector ${}^3P_{COG}$ (16).

$$ {}^3P_{COG} = \begin{bmatrix} \cos\beta & \sin\beta & 0 \\ -\sin\beta & \cos\beta & 0 \\ 0 & 0 & 1 \end{bmatrix} * \begin{bmatrix} {}^0P_{COGx} \\ {}^0P_{COGy} \\ {}^0P_{COGz} \end{bmatrix} = \begin{bmatrix} {}^3P_{COGx} \\ {}^3P_{COGy} \\ {}^3P_{COGz} \end{bmatrix} \quad (16) $$

We now proceed to describe on the cross-sectional plane the distance ρ from the ISP's coordinate frame origin to the position of the payload COG. The position of the ISP coordinate frame described with respect to the cross-sectional plane Σ_3(14) and the position of the payload COG described with respect to the cross-sectional frame Σ_3 (16) shown in Figure 17. Here α is the elevation angle of the ISP and ρ is the distance from the ISP coordinate frame origin to the position of the payload COG.

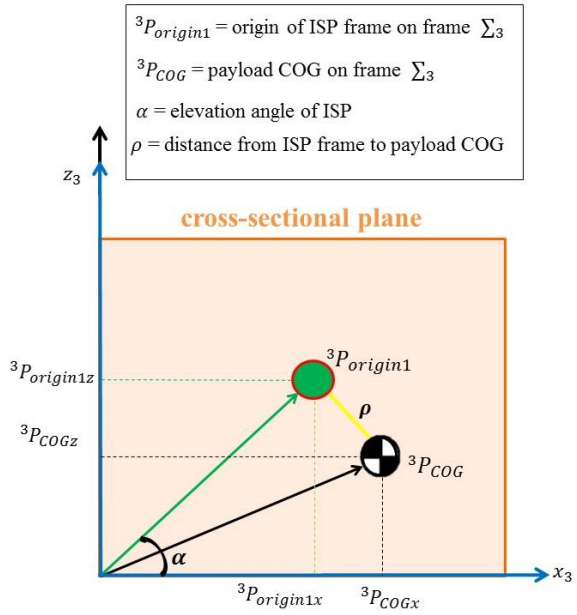


Figure 17. View normal to the cross-sectional plane with the location of the inertially stabilized platform frame and the payload center of gravity.

From the view normal to the cross sectional plane and the coordinate descriptions of the position of the ISP (14) and the payload COG (16) , distance ρ is defined by (17).

$$ \rho = \sqrt{({}^3P_{COGx} - {}^3P_{origin1x})^2 + ({}^3P_{COGz} - {}^3P_{origin1z})^2} \quad (17) $$

Finally, we proceed to take moments under static condition about the crossing point of the gimbal pitch and roll rotation axes (coincident with the origin of the cross-sectional frame Σ_3) to calculate the height h of the payload COG as per Figure 18

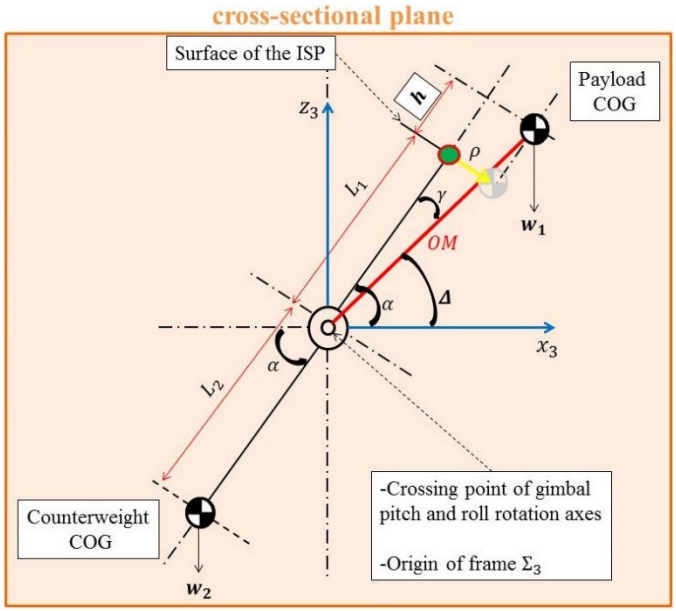


Figure 18. View normal to the cross-sectional plane for the calculation of the height h of the payload center of gravity.

where:

L_1= distance from crossing point of the gimbal rotation axes to the ISP coordinate frame Σ_1.

L_2= distance from the crossing point of the gimbal rotation axes to the counterweight coordinate frame Σ_2.

w_1= weight of the payload.

w_2= weight of the counterweight .

OM = distance from the crossing point of the gimbal rotation axes to the payload COG.

α = elevation angle of the ISP.

h = height of the payload COG from the ISP surface.

$$\gamma = \tan^{-1}\frac{\rho}{L_1} \quad (17)$$

$$\Delta = \alpha - \gamma \quad (18)$$

$$\mathrm{OM} = \sqrt{(L_1 + h)^2 + \rho^2} \quad (19)$$

As illustrated in Figure 18, we use the view of the cross-sectional plane to take moments about the crossing point of the gimbal pitch and roll axes under static condition (20). From this equation the height *h* of the payload COG from the surface of the ISP is calculated with (21).

$$\sum \tau = \sqrt{(L_1 + h)^2 + \rho^2} * cos\,\Delta * w_1 - L_2 * \cos\alpha = 0 \quad (20)$$

$$h = \frac{\sqrt{(\cos\alpha)^2 * {L_2}^2 * {w_2}^2 - (\cos\Delta)^2 * \rho^2 * {w_1}^2} - (\cos\Delta) * L_1 * w_1}{w_1 * \cos\Delta} \quad (21)$$

5. Control

The pitch and roll inclination angles of the passive gimbal are measured by the onboard SEN-10736 inertial measurement unit. The pitch angle measured is used as feedback in a PID control scheme to actuate the X Stage, displacing the counterweight along the *x* axis until the passive gimbal reaches a setpoint angle of zero degrees of inclination about the pitch axis [7]. The roll angle measured is used as feedback in a PID control scheme to actuate the Y Stage, displacing the counterweight along the *y* axis until the passive gimbal reaches a setpoint angle of zero degrees of inclination about the roll axis [7]. As a result, the payload line of sight is oriented to a horizontal position with respect to the inertial reference frame Σ_0.

Both the X Stage and the Y Stage are each equipped with a Copal RE30E500-213 incremental optical encoder for position feedback. Figure 19 shows the block diagram for the X Stage control and Figure 20 shows the block diagram for the Y Stage control.

For the control of the Z Stage, the control system calculates the target vertical position L_2 of the counterweight that satisfies (3) using the weight of the payload measured by the CZL204E strain gauge load cells located on the platform. Once the calculation according to (3) is done, the control system displaces the Z Stage vertically to reach this setpoint. Figure 21 shows the block diagram for the Z Stage control.

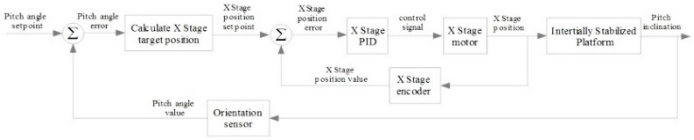


Figure 19. Block diagram for X Stage control.

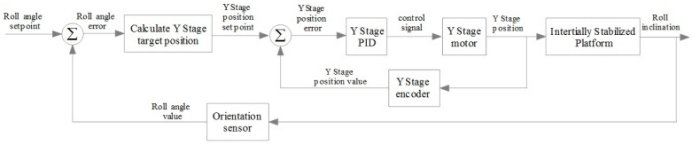


Figure 20. Block diagram for Y Stage control.

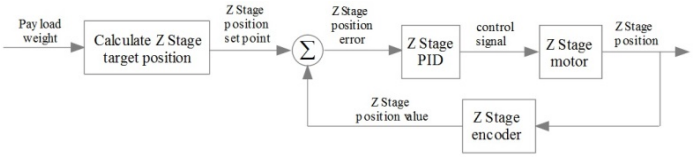


Figure 21. Block diagram for Z Stage control.

6. Experimental setup

The second prototype of the inertially stabilized platform with passive gimbal as well as the host vehicle are here presented. Fundamental experiments were conducted to measure the linear speed of the X Stage, Y Stage, and Z stage. Moreover, a fundamental experiment was conducted to test the accuracy of the calculation of the position of the payload COG on the surface of the inertially stabilized platform.

The stabilizing mechanism mounted on the omnidirectional host vehicle is shown in Figure 22 and its dimensions are shown in Figure 23.

A close-up of the stabilizing mechanism's counterweight comprised of the XY Stage and the Z Stage is shown in Figure 24.

Figure 25 shows an overview of the motions of the XY Stage and Z Stage of the stabilizing mechanism.

An overhead view of the inertially stabilized platform is shown in Figure 26, and a bottom view of the vehicle's omnidirectional mobile base is shown in Figure 27.

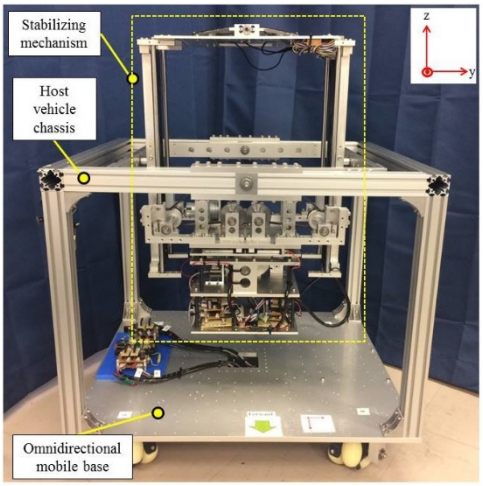


Figure 22. Second prototype of the stabilizing mechanism on the host vehicle.

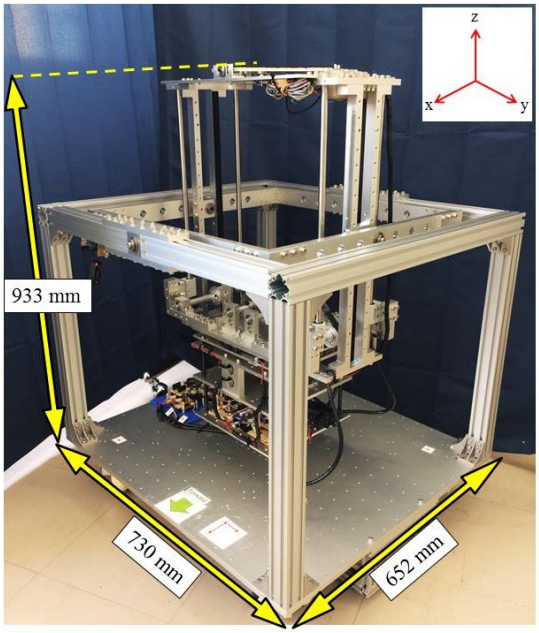


Figure 23. Dimensions of the second prototype.

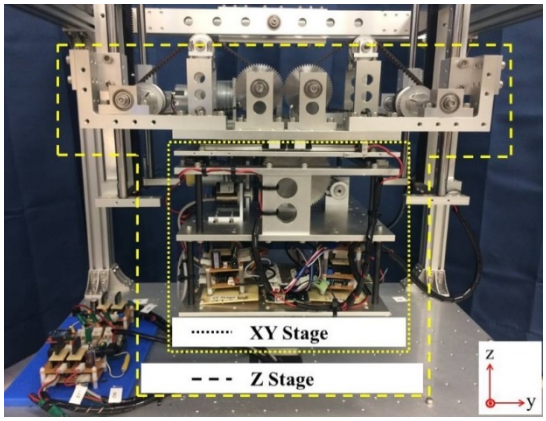


Figure 24. Close-up of the counterweight consisting of the XY Stage and the Z Stage.

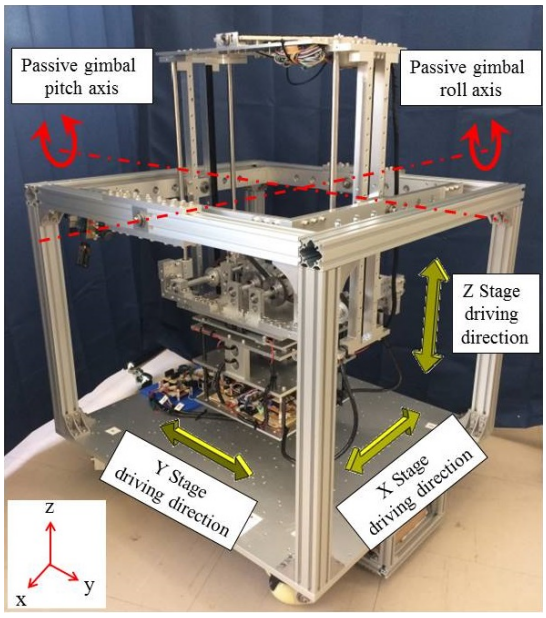


Figure 25. Overview of the motions of the XY Stage and the Z Stage.

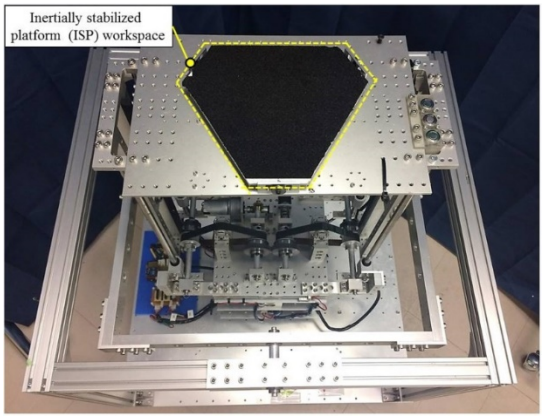


Figure 26. Workspace of the inertially stabilized platform.

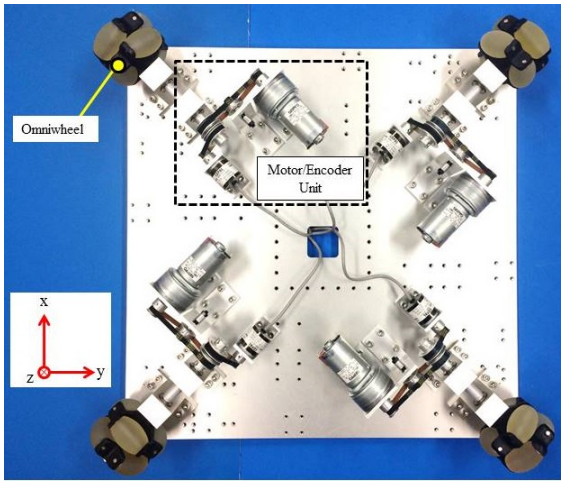


Figure 27. Omnidirectional mobile base of the host vehicle.

The specifications of the inertially stabilized platform are shown in Table 1.

Specifications of the XY Stage and the Z stage are detailed in Table 2 and Table 3, respectively.

Table 1. Specifications of the inertially stabilized platform.

Inertially Stabilized Platform (ISP)	
Workspace area	69,350 mm
Max. payload capacity	18 kg
Max. pitch inclination angle	40 degrees
Max. roll inclination angle	40 degrees
Counterweight	16.53 kg
Inertial measurement unit (IMU)	SEN-10736 nine degrees of freedom inertial measurement unit Voltage: 3.5 – 16 VDC
Weight measurement of payload	CZL204E strain gauge load cell (3x) Max. capacity: 50 kg Excitation voltage: 9-12 VDC

Table 2. Specifications of the XY Stage.

XY Stage	
Dimensions	250mm x 250mm x 124.5mm
Material	Frame: A5052 Planar omnidirectional gear: A7075 X Stage spur gear: SUS304 Y Stage spur gear: SUS304
Motion range on X axis	81 mm
Motion range on Y axis	115 mm
X Stage motor	Nidec DMN37K50G72B brushed DC Torque: 0.98 Nm Speed: 52.3 rpm Voltage: 24 VDC Gear ratio: 72:1
X Stage encoder	Copal RE30E500-213 incremental optical encoder Resolution: 500 pulse/rev
Y Stage motor	Nidec DMN37K50G72B brushed DC Torque: 0.98 Nm Speed: 52.3 rpm Voltage: 24 VDC Gear ratio: 72:1
Y Stage encoder	Copal RE30E500-213 incremental optical encoder Resolution: 500 pulse/rev

Table 3. Specifications of the Z Stage.

Z Stage	
Dimensions	542mm x 453mm x 318mm
Material	Frame: A5052 Rack gear: (2x): S45C Spur gear: S45C
Motion range on Z axis	390mm
Z Stage motor	Nidec DMN37K50G72B brushed DC Torque: 0.98 Nm Speed: 52.3 rpm Voltage: 24 VDC Gear ratio: 72:1
Z Stage encoder	Copal RE30E500-213 incremental optical encoder Resolution: 500 pulse/rev

6.1. Linear Speed of the XY Stage

Fundamental experiments were performed to measure the linear speed of the X Stage as it is displaced along its motion range, and the linear speed of the Y Stage as it is displaced along its motion range. The tests to measure the linear speed of each stage were performed by driving the corresponding stage motor at full capacity from the start of its motion range to the end of its motion range, repeating the motion 5 times for each stage.

Figure 28 shows the velocity profile of the X Stage as it is displaced from rest at 0mm to its maximum motion range of 81mm. The X Stage had an average linear speed of 200.3 mm/s, taking 0.35 seconds to move along its motion range.

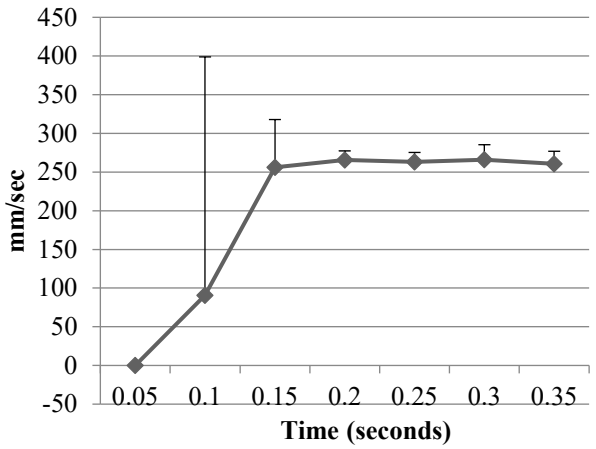


Figure 28. Linear velocity of the X Stage.

Figure 29 shows the velocity profile of the Y Stage as it is displaced from rest at 0mm to its maximum motion range of 115mm. The Y Stage had an average linear speed of 210.2 mm/s, taking 0.5 seconds to move along its motion range.

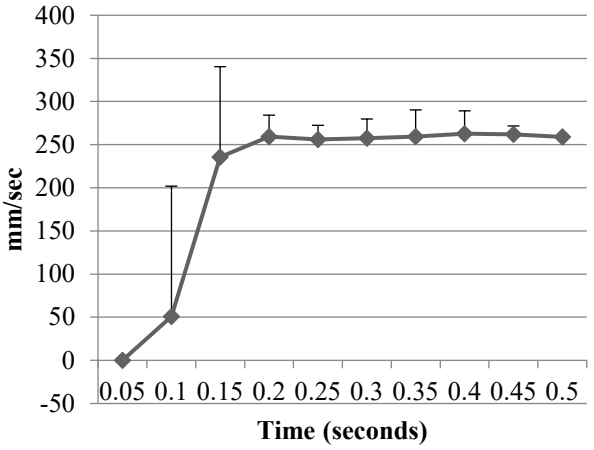


Figure 29. Linear velocity of the Y Stage.

6.2. Linear Speed of the Z Stage

Initially, the deployment of the Z Stage driving motor on the second prototype actuated the Z Stage via a direct attachment of the motor shaft to the driving gear shaft. Experiments revealed that the motor driving the Z Stage lacked enough braking power

to maintain the Z stage at a fixed position (setpoint position) on the Z axis. When the Z stage reached its vertical target position and the motor was braked, the downward weight of the XY Stage/Counterweight produced a torque on the motor shaft higher than its braking torque, and as a result, the Z Stage moved down. This caused the control system to drive the Z Stage upward again trying to maintain the target position, resulting in large oscillations of the Z Stage during operation time. Figure 30 shows the previous deployment of the Z Stage motor.

To tackle this problem, we modified the method to drive the Z Stage to a worm gear drive with a reduction ratio of 20:1 This modification improves the lifting torque of the Z Stage to elevate the XY Stage/Counterweight and takes advantage of the non-backdrivability of the worm gear to provide a mechanical break that prevents the Z Stage from moving down once it has reached its vertical target position. Figure 31 shows the modified deployment with worm gear drive to improve lifting torque and provide a mechanical break to prevent a descent of the Z Stage due to gravity.

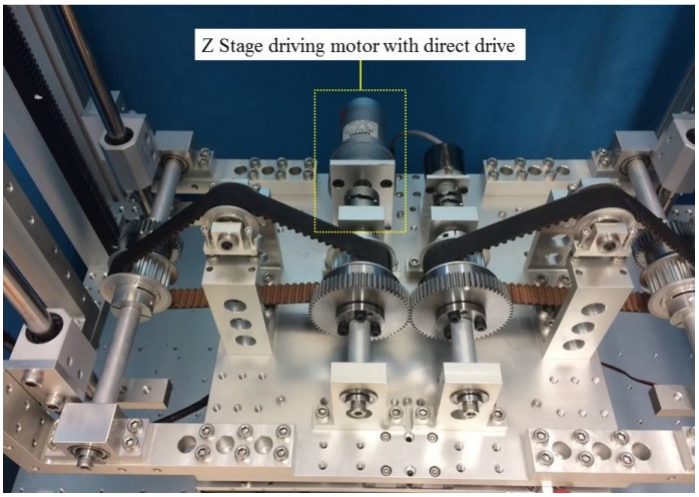


Figure 30. Previous deployment of the Z Stage driving motor.

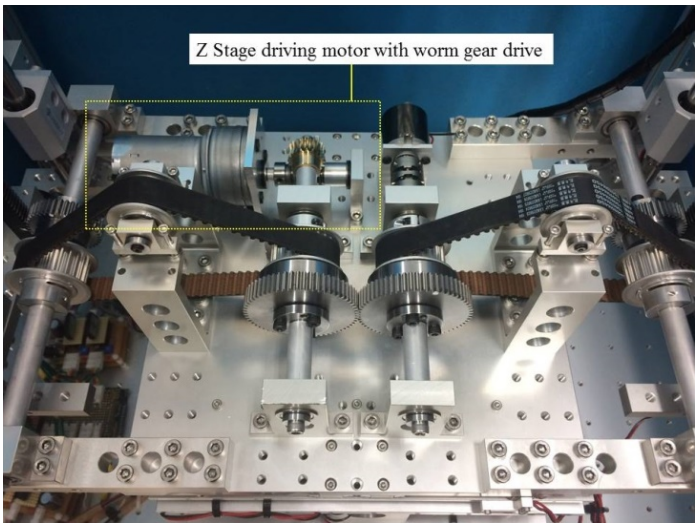


Figure 31. Current deployment of the Z Stage driving motor with worm gear.

With the current deployment of the Z Stage driving motor with worm gear drive, a fundamental experiment was performed to measure the linear speed of the Z Stage as it is displaced along its motion range. The tests to measure the linear speed were performed by driving Z Stage motor at full capacity from the start of its motion range, to the end of its motion range, repeating the motion 5 times.

Figure 32 shows the velocity profile of the Z Stage as it is displaced from rest at 0mm to its maximum motion range of 390mm. The Z Stage had an average linear speed of 5.75 mm/s, taking 67.9 seconds to move along its motion range.

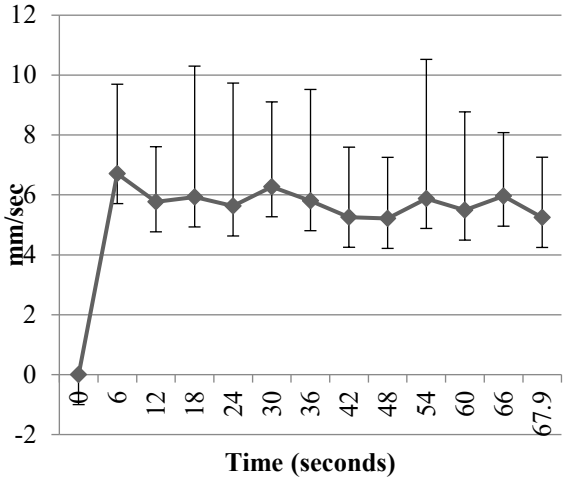


Figure 32. Linear velocity of the Z Stage.

6.3. Calculation of the Position of the Payload Center of Gravity on the Surface of the Inertially Stabilized Platform

For this experiment, a grid with 5mm markers was aligned with the axes of the ISP coordinate frame Σ_1 (Figure 33), and a 1.5 kg calibration weight was placed on eight different positions on the surface of the ISP as payload, corresponding to eight (x,y) coordinates on the ISP coordinate frame Σ_1. The coordinate position of the payload on the surface of the platform was measured with the multiple-point weighing method discussed in Chapter 4, section 4.2.

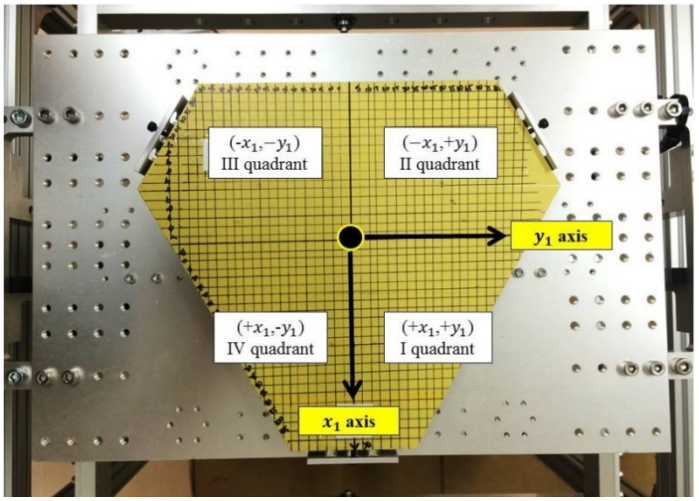


Figure 33. The inertially stabilized platform with a grid for the experiment to calculate the horizontal position of a payload's center of gravity.

Table 4 shows the results of the measurements of the horizontal position of the payload COG corresponding to its physical position coordinates on the ISP surface. Figure 34 illustrates the results shown in Table 4.

Table 4. Results of the measurements for the positions of the payload center of gravity.

Location	Physical COG Coordinates in millimeters	Measured COG Coordinates in millimeters	Standard Error in Measured COG Coordinates	
	(x,y)	*(x,y)*	*x*	*y*
1	(-50.0, 0.0)	(-51.16, 0.57)	0.19	0.09
2	(-40.0, 60.0)	(-33.97, 63.50)	8.39	0.05
3	(-40.0, -60.0)	(-42.09, -59.02)	0.49	0.37
4	(0.0, 0.0)	(4.89, 0.44)	0.06	0.09
5	(20.0, -50.0)	(24.56, -44.74)	0.11	0.22
6	(20.0, 50.0)	(25.76, 51.61)	0.09	0.12
7	(70.0, 30.0)	(70.73, 26.29)	0.15	0.08
8	(70.0, -30.0)	(73.84, 24.27)	0.16	0.10

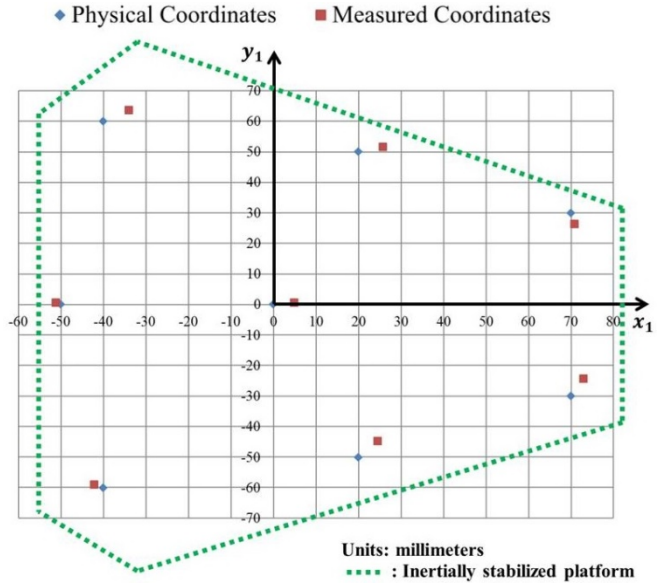


Figure 34. Physical positions of the payload center of gravity on the surface of the ISP and the corresponding measured values.

Future steps

The next step in this research is to experimentally test the proposed theoretical framework for the calculation of the height of the payload center of gravity presented in Chapter 4, section 4.3.

Applications

The stabilizing strategy and motion isolation method with passive gimbal axes discussed in this paper rely on adjusting the mass properties of the system. Since the calculations and motor actuations to accomplish payload stabilization are performed at startup, once the system is balanced no motor output or computational calculations are required during the rest of operation time of the system. These feature reduces the power requirements and the computational resources needed during the operation time of the stabilizing system, making it suitable for battery powered-applications and applications where computational capability is restricted.

Furthermore, the ability of the proposed stabilizing mechanism of having a payload placed in an arbitrary position on the platform and stabilizing its LOS makes its use intuitive and simple, suitable for applications involving human interaction without requiring skill or assembly.

Conclusion

In this paper, a method for disturbance isolation of an inertially stabilized platform using a passive gimbal was described. In the method discussed, the inertially stabilized platform is attached to a gimbal suspended on the host vehicle chassis by freely rotating (passive) joints, with no torque coupling between the gimbal and the host vehicle chassis. As a result, the gimbal is isolated from the vehicle chassis angular motion.

A consequence of the use of a passive gimbal is that in order to stabilize the line of sight of a payload, the payload's weight and the position of its center of gravity are required to be known. A method of calculating the three-dimensional position of the center of gravity of a payload once it is placed on the inertially stabilized platform was presented. The calculation of the (*x,y*) coordinates of the payload COG was tested experimentally and demonstrated to be adequate.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgement

This research was supported by Equos Research Co., Ltd.

References

[1] Erick Moya, Kenichi Shimakawa, Akira Satou, Adrian Zambrano, Kenjiro Tadakuma, Riichiro Tadakuma, "Study on the Inertial Stabilization of a Payload by Center of Gravity Displacement", Proceedings of 2016 IEEE/SICE International Symposium on System Integration (SII), pp.266-271.

[2] Daniel R. Otlowski, Kurt Weiner, Brandon A. Rathburn, "Mass properties factors in achieving stable imagery from a gimbal mounted camera", Proceedings SPIE, Airborne Intelligence, Surveillance, Reconnaissance (ISR) Systems And Applications Wadsworth, vol. 69460B, 2008.

[3] J.M. Hilkert, "Inertially Stabilized Platform Technology Concepts and Principles", IEEE Control Systems, vol. 28, 2008, pp. 26-46.

[4] Kenjiro Tadakuma, Riichiro Tadakuma, Kyohei Ioka, Takeshi Kudo, Minoru Takagi, Yuichi Tsumaki, Mitsuru Higashimori, Makoto Kaneko, "Omnidirectional Driving Gears and their Input Mechanism with Passive Rollers", Proceedings of 2012 IEEE/RSJ International Conference on Intelligent Robots, pp.2881-2888, 2012.

[5] P. J. Kennedy, R. L. Kennedy, "Direct Versus Indirect Line of Sight (LOS) Stabilization", IEEE Transactions on Control Systems Technology, vol. 11, 2003, pp.3-15.

[6] Myer Kutz, Handbook of Measurement in Science and Engineering, vol. 1, Wiley, 2013, pp.277-279.

[7] Michael Masten, "Inertially Stabilized Platforms for Optical Imaging Systems", IEEE Control Systems, vol.28, 2008.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 2, No. 3, 1422-1431 (2017)
www.astesj.com
Special Issue on Recent Advances in Engineering Systems

ASTES Journal
ISSN: 2415-6698

CAM-Infrastructure: a novel solution for service advertisement in Cooperative Intelligent Transportation Systems

Badis Hammi[*,1], Houda Labiod[1], Gérard Segarra[2], Alain Servel[3] Jean Philippe Monteuuis[1,3],

[1] *Telecom ParisTech,46 Rue Barrault, 75013 Paris, France*

[2] *VICI, France*

[3] *Groupe PSA,Grande Armee 75, Avenue de la Grande Armee 75116 PARIS, France*

ARTICLEINFO

Article history:
Received: 31 May, 2017
Accepted: 16 July, 2017
Online: 10 August, 2017

Keywords :
C-ITS
Service advertisement
CAM
DENM
SAM
WSA
FSAM
ETSI

ABSTRACT

Cooperative Intelligent Transportation Systems (C-ITS) are gaining ground and currently are almost part of our everyday life. A C-ITS environment can provide numerous services that soon will become essential to roads' users. The latter resides in improvement of road safety, entertainment, and commercial services. However, in order to provide such services, the C-ITS environment needs an advertisement and dissemination service for them. Indeed, users have to be aware of the available services in order to request them if needed. Current standards of service announcement show their limits, especially regarding the interoperability between communication profiles. For this reason, in this paper, we describes our new service advertisement solution called CAM-Infrastructure. The latter is compliant with ETSI standards and is deployed in SCOOP@F, a nationwide scale project.

1 Introduction

In the few last years, automobile manufacturers gave more interest into Collaborative Intelligent Transportation Systems (C-ITS). Indeed, they provided numerous prototypes of vehicles equipped with sensors, dedicated computing hardware and dedicated short-range radio (DSRC) for communication with other nearby ITS Station Vehicles (ITSS-V) or with road side infrastructure (ITS Station Road side unit (ITSS-R)). Furthermore, numerous deployment project were achieved like SCMS [1], SEVECOM [2], PRESERVE [3] [4], CORRIDOR [5], Eco-AT [6] and SCOOP@F [7].

Within an ITS environment, numerous services such as improvement of road safety, driver assistance, entertainment and commercial services are provided. In order to achieve these services, their advertisement and dissemination are mandatory. In a classical scenario, a service provider which can be an application provider or even a commercial advertisement provider sends out the service advertisement message through an ITSS-R, and nearby receiving, ITSS-Vs start to disseminate those advertisement messages by forwarding them to other vehicles in their neighborhood [8].

Numerous works and studies were proposed and conducted in the goal of providing efficient and secure service advertisement protocols. However, beyond the different organisms' architectures and requirements, only ISO and IEEE proposed standards: Fast service advertisement protocol (FSAP) described by ISO 24102-5 standard [9]; and Wave Service Advertisement (WSA) defined in IEEE 1609.3 standard [10]. Nonetheless, the latter are compatible only with architectures proposed by their respective organisms. ETSI proposes Service Announcement Message (SAM) and Extended SAM (ESAM), which is -while writing this paper- not yet standardized [11] [12]. In addition, SAM is suitable only for systems using Geonetworking protocol on G5. We recall that using Geonetworking protocol for information carrying have the shortcoming of huge time delays, since the packets go by each hop.

Our work is part of SCOOP@F project. The latter is a french Cooperative ITS pilot deployment project (extended to Europe) that aims at deploying C-ITS in a nationwide scale. 2000 kms of road are to be equipped with Road Side Units (ITSS-R) and 3,000

[*] Badis Hammi, Telecom ParisTech, France. & badis.hammi@telecom-paristech.fr

https://dx.doi.org/10.25046/aj0203178

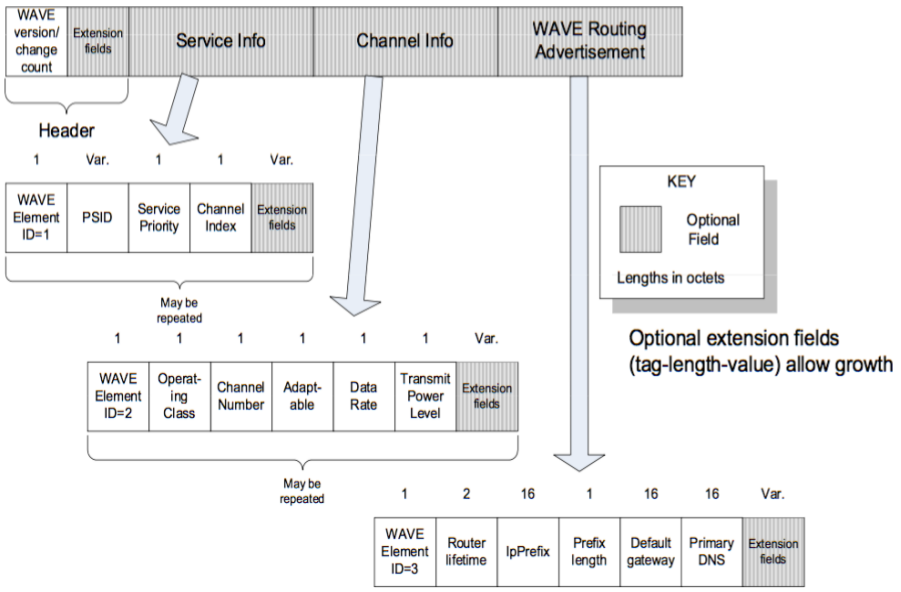


Figure 1: WSA message format [13]

ITSS-V with On-Board Units to enable communication between the infrastructure (the road operator) and the user (the driver). Ile de France, Bordeaux and its ring, Bretagne and Isere counties, in addition to Paris-Strasbourg highway represent the fields of the first deployments . SCOOP@F architecture relies on ETSI published standards. Thus, WSA is not compatible with our needs. Furthermore, SAM does not cover all our needs and requirements. Indeed, in multiple scenarios, we use one hop (face-to-face) advertisement, which is not compatible with SAM.

In order to remedy SAM limits, we implemented a new advertisement solution called Cooperative Awareness Message Infrastructure (CAM-I). Its main advantage is that CAM-I is compliant with the Cooperative Awareness Message (CAM) existing standard [14]. We describe the details of CAM-I structure and functioning in this work which represent an extension of [15].

The rest of the paper is organized as follows: Section 2 gives an overview of the existing advertisement standards and explain our motivation for this work. Section 3 details our CAM-I solution. Section 4 describes a comparison of the proposed solution CAM-I with existing solutions. Finally, Section 5 concludes this paper and introduce our future works.

2 Related Works and motivation

Numerous research proposals and studies on services advertisement were conducted. However, since our project concerns a real nationwide deployment project, we will discuss only standardization working groups and consortia works.

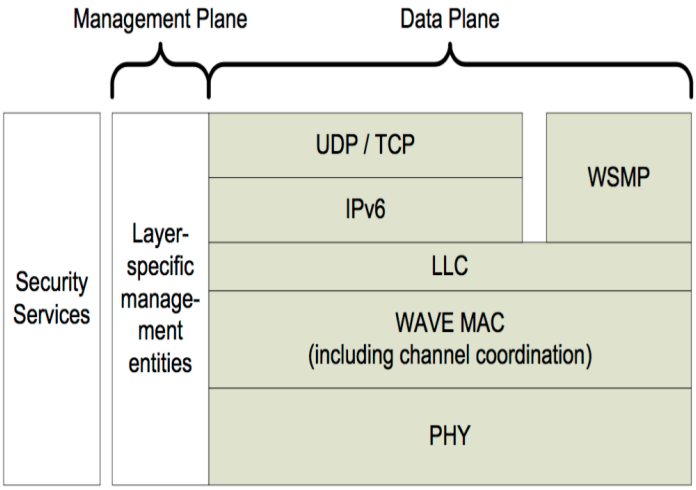


Figure 2: IEEE communication stack [16]

2.1 IEEE WAVE Service Announcement (WSA)

IEEE proposes WAVE Service Announcement (WSA), described in IEEE 1609.3 [10]. In a system implementing it, all stations are required to monitor the multi-channel radio Control Channel (CCH). In 1609.3 provider mode, the station transmits a WSA message on the CCH during the CCH interval. Consequently, since all ITSS are monitoring this channel at that time, they all receive the WSA. The WSA contains a list of the services that the provider will provide during the Service Channel (SCH) interval. It also provides the SCH channel number that they will be using. The services are identified by a code number known as a Provider Service Identifier (PSID). If an ITSS in user mode receives a WSA that contains a PSID of interest, it will switch to the appropriate SCH

during the SCH interval and will make use of that service [17]. Figure 1 describes the structure of a WSA message.

WSA is only compatible with architectures that deploy IEEE WAVE communication stack described by Figure 2. Thus, this solution does not fit needs of SCOOP@F project and more generally european projects which rely on ETSI-based architecture.

2.2 ISO Fast Service Announcement Protocol (FSAP)

ISO proposes Fast Service Announcement Protocol (FSAP) [9] [18] in order to inform peer stations about available services by means of Fast Service Announcement Messages (FSAM). FSAM structure is very similar to WSA messages. This service advertisement distinguishes at least the following roles: (1) Service Advertisement Manager which contains a Server manager that transmits FSAMs and a Client manager, responsible for FSAMs' reception; (2) Service Provider which ensure provision of ITS services; and (3) Service User which consumes ITS services. FSAM messages are encoded according to UPER encoding rules [19].

FSAP relies on two types of unique identifiers in its functioning:

- **ITS Application IDentifier (ITS-AID):** specified in ISO 17419 [20] as an unsigned integer which have the same function as the IEEE PSID. It also shares a common number space with PSID. According to [20] the ITS-AID of FSAP is equal to 2.113.664.
- **ITS Port Number (ITS-PN):** also specified in ISO 17419 [20]. It is used at the OSI transport layer in order to identify source and destination port numbers of advertised application and services.

[18] introduces Harmonized Service Advertisement Message (HSAM), a hypothetical service advertisement message, having almost the same structure as FSAM and that can be extended to different forms. Figure 3 describes its structure details.

2.3 ETSI Service Announcement Message

ETSI proposes Services Announcement Message (SAM), which represents an ITS message that advertises available services. SAM is broadcasted from an ITSS-R on G5 on channel CCH using the Geonetworking protocol. Within a SAM, an ITSS-R must provide some specific information such as the list of the available services, the type of target for each services and the communication profile. More specifically, the available services are a list of Service-IDs. The target of the service has the value (All) or a Community-ID (CID) which associates a specific group of ITSS to an ID. The communication profile is identified by a Communication Profile Identifier (CPID).

SAMs can be sent simultaneously until a maximum threshold of sent messages which depends on the risk of congestion of CCH. At reception, SAMs are received in transport level packets. The Network & Transport layer will send the SAMs to the Facilities Layer. At this level, a first function called "Messages Management" decodes the SAMs and sends the list of services and other data to a second function called "Services Announcement" which will provide several checks. For each service, the Service Announcement function will check if the ITSS is part of "All" or a "community" thanks to its CID. If yes, it will need to verify if the service has been subscribed or not. If the community or ITSS did not subscribed for the announced service then it should not be considered. The ITSS will create a list with all the relevant services checked from the list of the advertised services proposed in the SAM. The service announcement function will select one or several relevant services from the list depending on the channel capability so the service can be executed.

2.4 ETSI Extended Service Announcement Message

Recently, ETSI proposed an extended version of SAM called Extended SAM (ESAM) [12]. It defines three roles: (1) a Service Provider whose role is to register, update or deregister the provided services. (2) Service Announcer which announces the services provided by the Service Provider. The service provider and the service announcer can be the same ITSS or two distinct ITSSs. (3) Service User which represent ITSS-Vs that use the service.

The communication between each role is defined following the communication stack in Figure 4. At the application layer, the type of service, the requested actions regarding this service and other fields depending on the previously mentioned roles are defined to the management entity through the MA-SAP interface. At facility layer, ESAM needs to be encoded or decoded then managed by the management entity for each message transmission where UPER encoding rules are applied. One principal function of the management entity is to define the ITSS Communication Profile (ITS-SCP) used for the message transmission through the MA-SAP interface. Security processes between the Facility Layer and the Security Entity are achieved through the interface SF-SAP. At Networking and Transport (NT) Layer, the ESAM is delivered as a payload to the facility layer through the NF-SAP interface.

One novelty of the ESAM is the inclusion of ExtendedChannelInfos field and other needed extensions in its definition. For instance, the ExtendedChannelInfos is used in order to define new communication technologies which can be used to announce the ESAM such as IPv6. This feature was missing from the previous SAM definition which supported only G5 communication as discussed in [15]. However, no

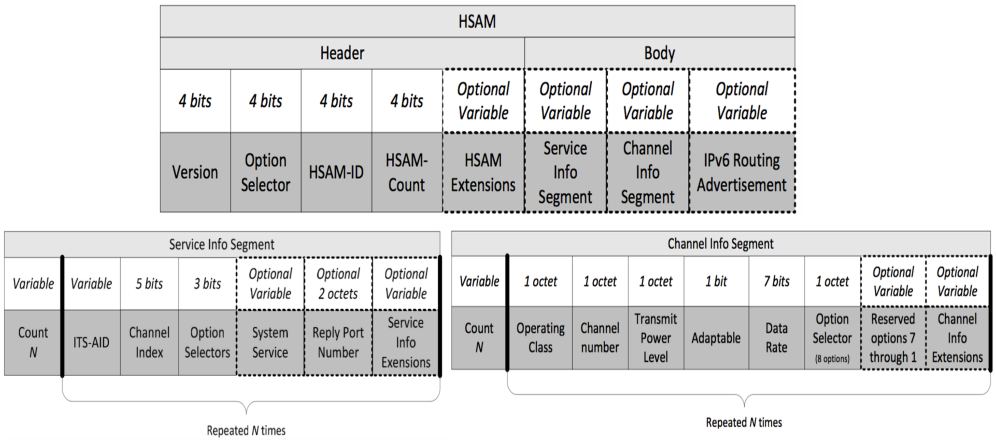


Figure 3: HSAM structure [18]

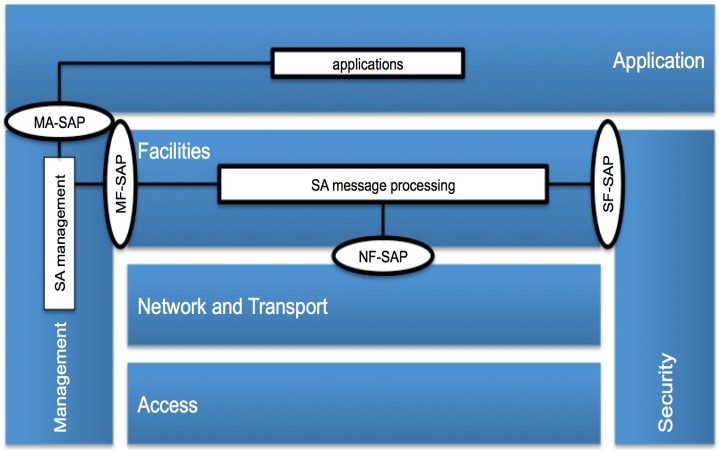


Figure 4: SA in the context of the ETSI ITSS reference architecture [21]

clear definition on the use of new protocol type different than GN is defined yet.

SAM and ESAM solutions are suitable only for systems using Geonetworking protocol on G5 (plus IPv6 for ESAM), which relies on broadcast of messages. The latter does not suits to SCOOP@F needs, in which it does exist multiple scenarios where ITSSs communicate in face-to-face way (one hop) using IP protocol. In addition SAM and ESAM still being drafts from years and are not considered as standards yet.

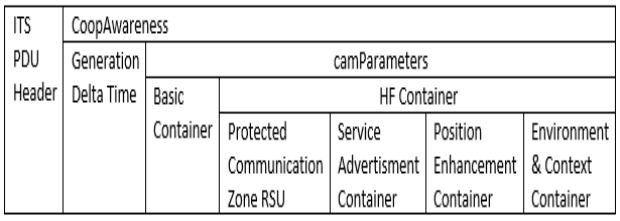


Figure 5: CAM-I structure

To remedy these limits and to achieve SCOOP@F needs, we propose a new solution for service advertisement called CAM-I, a solution compliant with ETSI CAM standard [14].

3 New solution for services advertisement: CAM-I

In this section we describe the details of the advertisement solution that we propose.

3.1 Global structure

In ETSI architecture, periodic messages called CAM are diffused continuously. In our solution, the available services are advertised via the periodic broadcasting of a specific CAM message by the ITSS-R, called CAM-Infrastructure (CAM-I) message. The latter has almost the same structure as a CAM message sent by an ITSS-V.

As described in [14], a CAM is composed of one common ITS PDU header and multiple containers. The ITS PDU header is a common header that includes the information of the protocol version, the message type and the ITSS ID of the originating ITSS.

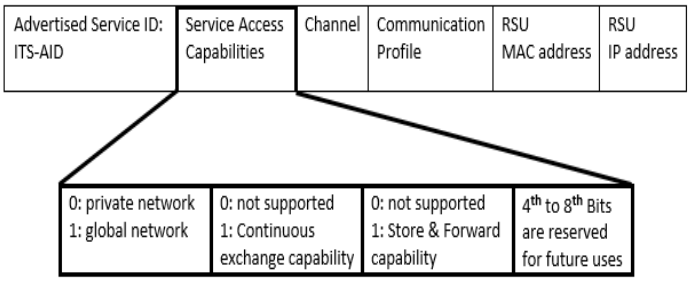


Figure 6: Service Advertisement Container structure

For an ITSS-V a CAM shall comprise one basic container and one high frequency container, and may also include one low frequency container and one or more other special containers:

- Basic container: includes basic information related to the originating ITSS.
- High frequency container: contains highly dynamic information of the originating ITSS.
- Low frequency container: contains static and not highly dynamic information of the originating ITSS.
- Special vehicle container: contains information specific to the vehicle role of the originating vehicle ITSS.

As indicated in the CAM standard [14], all CAMs generated by an ITSS-R shall include a basic container and optionally more containers. Consequently, the proposed CAM-I is composed of an ITS PDU Header, a basic container and a High Frequency container, which in turn, is composed of 4 containers. Described by Figure 5, CAM-I structure is similar to the one of CAM. Its header (ITS PDU Header) and Basic Container are compliant with the standard ETSI EN 302 637-2 [14]. The ITS PDU header includes the protocol version, the message type, the ITSS-R identifier of the originating ITSS-R. In addition, a generation delta time of the message is included. The Basic Container includes the following fields: (1) The type of the emitting station (ITSS-R): 15; (2) The geographic position of the ITSS-R. While the PDU header and the low frequency are the same as in CAM, we modified the High Frequency (HF) container. More specifically, we defined a HF Container composed of the following containers:
(1) Service Advertisement Container;
(2) Position Enhancement Container;
(3) Environment & Context Container;
(4) Protected Communication Zone RSU Container.

Table 1 describes the structure of the three last containers. Service Advertisement container is detailed in next section.

3.2 Service Advertisement Container

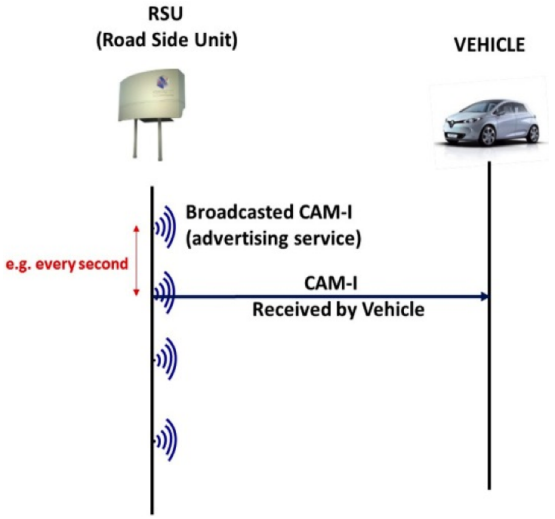


Figure 7: CAM-I Service Advertisement scenario

As described by Figure 6, Service Advertisement Container includes the following fields:
(1) Advertisement Service ID: this byte provides the ID of the advertised service.
(2) Service Access Capabilities (SAC): The second byte details access capabilities to (1) a private network such as a network of the road operator that offers an access reserved to its staff, or (2) to a global network like Internet. The choice of access policy for SCOOP@F vehicles depends on the global access policy selected by each road operator. In SAC:

- The first bit indicates if a private access or a global access is available. If it is equal to 1, then the ITSS-R provides access to a private network and if it is equal to 0 ITSS-R provides access to a global one.
- The second bit, when having the value 1, it indicates the capacity of the ITSS-R to establish a continuous exchange with a remote server via the private/ global network. If this it has the value 0, then the ITSS-R does not possess this capacity.
- The third bit, when it is equal to 1, it indicates

Container	Data Elements	Type	Size in Bytes	Description
ITS PDU Header	Protocol version	same as in [14]		Version 1
	Message ID	same as in [14]		CAM
	Station ID	same as in [14]		ITSS-R identifier
Generation Delta Time	Generation Delta Time	same as in [14]		
Basic Container	Station Type	same as in [14]		ITSS-R
	Reference Position	same as in [14]		Accurate position of the ITSS-R established at installation time
Protected Communication Zone RSU Container	Data elements for Toll collect protection	same as in [14]		for tolling collection
Position Enhancement Container	GPS Position Delta Latitude	same as in [14]		Optional, enable to correct positioning error by comparing the RSU accurate position with the given RSU GPS position
	GPS Position Delta Longitude	same as in [14]		Same as for GPS Latitude
	GPS Position Delta altitude	same as in [14]		Same as for GPS Latitude
	Satellite constellation locally available	Integer	1 byte	Enable the use of correction if the vehicle detects the same satellite constellation
	Trace leading to the ITSS-R	same as in [14]		Enable the improvement of the vehicle position and the computation of the time window for RSU exchange
Environment and context container	local meteorological data	Binary	1 byte	Provide local meteorological data for environmental characterization
	Road environment	Integer	byte	Provide the road environment type in which the RSU is positioned
	Traffic condition	Integer	1 byte	Provide the current traffic condition

Table 1: Global Structure of the proposed CAM-I

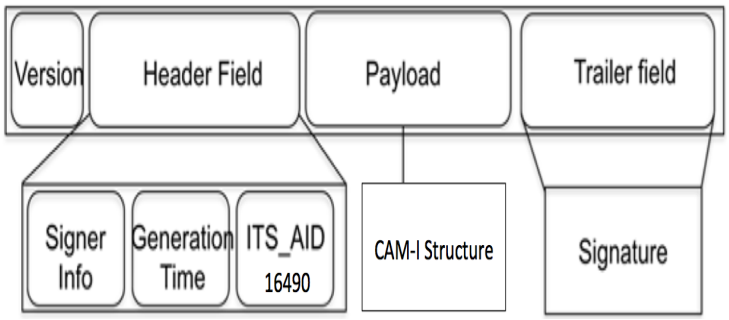


Figure 8: Secured Message transporting CAM-I structure

Octet	Description
0	SSP version control
1 to 2	Service-specific parameter
3 to 30	Reserved for future usage

Table 2: Octet Scheme for CAM-I SSPs

the capacity of the ITSS-R to store locally a message and to transmit it as soon as possible towards a remote server (store and forward) via the private/ global network, as soon as this one is available. If this bit has the value 0, the ITSS-R does not possess this capacity.

- The bits N° 4, 5, 6, 7 and 8 are reserved for a future usage.

(3) Channel: This field contains the channel used for communication
(4) Communication Profile (CP): The fourth byte indicates the communication profile.
(5) RSU MAC Address (RM): This field indicates the MAC address of the ITSS-R.
(6) RSU IP Address (RIP): This field indicates the IP address of the ITSS-R. If IPv4 is supported the field contains the IPv4 address of the RSU. If IPv6 is supported the field contains the IPv6 address of the RSU. If both IPv4 and Ipv6 are supported, the field contains both addresses.

Figure 7 describes a typical scenario where an ITSS-R broadcasts a CAM-I, an ITSS-V receives the message and requests a needed service.

CAM-I messages are encoded in ASN.1 UPER encoding and integrated as the payload of a secured message compliant with the structure described in the standard ETSI TS 103097 [22] as described by Figure 8. For security purposes, the secured message is signed using the private key of the sender through Elliptic Curve Digital Signature Algorithm (ECDSA). The secured message contains also a field named ITS Application ID (AID). The latter describes the type of message transported by the secured message and its value should belongs to a standardized list defined by ISO organism. For example the ITS AID value of CAM is 36, of DENM is 37. A request to ISO was sent to provide an ITS AID for CAM-I. Consequently, pending a response, we use 16490 value, which belongs to test range values. CAM-I is broadcasted via CCH channel with a frequency f as follows: $1\ hertz \leq f \leq 10\ hertz$. this frequency is chosen according to the environment and context needs. As for CAM messages and in compliance with ETSI 103097 standard, for each second, the first sent message contains the sender's certificate. For the rest, just the digest of the certificate, computed using SHA 256 algorithm is sent. Detailed ASN.1 structure of the advertisement container is described by Listing 1.

Listing 1: ASN1 definition of the Service Advertisement Container

```
ServiceAdvertismentContainer ::= SEQUENCE {
  advertisedServiceItsAid
      AdvertisedServiceItsAid,
  serviceAccessCapabilities
      ServiceAccessCapabilities,
  channelUsedByTheAdvertisedService
      ChannelUsedByTheAdvertisedService,
  communicationProfileUsedForTheService
      CommunicationProfileUsedForTheService,
  rsuMacAddress RsuMacAddress,
  rsuIpAddress RsuIpAddress
  }
AdvertisedServiceItsAid ::= INTEGER(0..
    4294967295)
ServiceAccessCapabilities ::= BIT STRING {
  globalNetwork (0),
  continuousExchangeCapability (1),
  storeForwardCapability (2)
  } (SIZE(8))
ChannelUsedByTheAdvertisedService ::= ENUMERATED
    {
  cch(0),
  sch1(1),
  sch2(2),
  sch3(3),
  sch4(4),
  sch5(5),
  sch6(6)
  }
```

Octet Position	Bit Position	Permission Items	Bit Value
1	0 MSBit	CenDsrcTollingZone/ ProtectionCommunicationZoneRSU	0: certificate not allowed to sign 1: certificate allowed to sign
1	1	PKI/ ServiceAdvertisementContainer	0: certificate not allowed to sign 1: certificate allowed to sign
1	2	Log/ ServiceAdvertisementContainer	0: certificate not allowed to sign 1: certificate allowed to sign
1	3	UseCaseToDefine1/ PositionEnhancementContainer	0: certificate not allowed to sign 1: certificate allowed to sign
1	4	UseCaseToDefine2/ ServiceEnvironmentAndContextContainer	0: certificate not allowed to sign 1: certificate allowed to sign
1	5	Reserved for future usage	Not used, set to 0
1	6	Reserved for future usage	Not used, set to 0
1	7 LSB	Reserved for future usage	Not used, set to 0

Table 3: CAM-I SSPs details

```
CommunicationProfileUsedForTheService ::=
    ENUMERATED {
  btpgeonet(0),
  tcpipv4(1),
  tcpipv6(2)
  }
RsuMacAddress ::= OCTET STRING (SIZE(6))
RsuIpAddress ::= CHOICE {
  rsuipv4Andv6Address Ipv4Andv6,
  rsuiPv4Address IPv4Address,
  rsuiPv6Address IPv6Address
  }
Ipv4Andv6 ::= SEQUENCE {
  rsuiPv4Address IPv4Address,
  rsuiPv6Address IPv6Address
  }
```

3.3 Security material for CAM-I users

in order to use CAM-I service, the ITSS-R should have the authorization for. The latter is provided within its certificate. Indeed, ETSI certificates contains (1) a field called ITS AID, which includes the list of the services that the station is authorized to access and use; and (2) a field called ITS AID Service Specific Permissions (SSP), which indicates specific sets of permissions within the overall permissions indicated by the ITS-AID. In other words the SSP is a field that includes the different authorized options for each ITS AID. For example, if an ITSS-R can have authorization to use CAM-I service but not authorized to use PKI requests sub-service.

SSPs for CAM message are defined using 3 byte as described in ETSI EN 302 637-2 v1.3.2 [14]. SSPs for DENM message are defined using 4 bytes as presented in ETSI EN 302 637-3 v1.2.2 [23]. For CAM-I, we choose to use the octet scheme for CAM-I SSPs described by Table 2.

In Table 3 we give the SSPs adopted in SCOOP@F project for CAM-I usage.

4 Comparison of CAM-I with existing standards

In this section we produce, through Table 4, a comparison of the proposed CAM-I with WSA and SAM . One can note that CAM-I represents the advantage of using face-to-face communication instead of Geonetworking, which can performs enormous time savings. Furthermore, it is compliant with all standards, thus, CAM-I can be used in other deployment projects, even

[1]CAM-I uses a proprietary access control protocol named NACS. Once the advertisement message CAM-I is intercepted by ITSS-Vs

Criterion	WSA	SAM	ESAM	FSAM	CAM-I
Compatible architectures	IEEE	ETSI	ETSI	ISO	ETSI
Advertising channel	CCH	CCH/SCH	CCH/SCH	CCH/SCH	CCH
Advertised service channel	Any SCH (except CH 172)	SCH2-SCH4	SCH2-SCH4	SCH2-SCH4	Any SCH
Message emission Frequency	$1\ Hz \leq f$	$1\ Hz \leq f \leq 10\ Hz$	$1\ Hz \leq f \leq 10\ Hz$	$1\ Hz \leq f$	$1\ Hz \leq f \leq 10\ Hz$
Security mechanism	Digital Signature: ECDSA	Digital Signature: ECDSA	Digital Signature: ECDSA	Digital Signature: ECDSA	Digital Signature: ECDSA
Access control	MAC layer	MAC layer	MAC layer	MAC layer	MAC layer NACS[1]
Communication profile (Access/ Network/ Transport;)	WAVE/ LLC/ IPV6/ TCP; WAVE/ LLC/ IPV6/ UDP; WAVE/ LLC/ WMSP;	G5/ Geonet/ BTP;	G5/ Geonet/ BTP; G5/ IPv6;	G5/ Geonet/ BTP; G5/ IPv6;	G5/ IPv6/ TCP; G5/ IPv6/ UDP; G5/ IPv4/ TCP; G5/ IPv4/ UDP;

Table 4: Comparison of advertisement services

if they are not based on ETSI standards.

5 Conclusion and future work

In this paper we have proposed a new solution for service advertisement in Intelligent Transportation Systems' environments called CAM-I. The latter is compliant with existing ETSI standards. Furthermore, it remedies SAM and ESAM limits for interoperability with communication protocols. CAM-I is deployed in SCOOP@F project, which represents a french deployment project that intends to implement a full ITS environment in a nationwide scale. Till the day we are writing this paper, three advertised services are deployed: (1) ITSSs' provisioning with certificates from PKI; (2) upload of logs generated by ITSSs; and (3) Data Exchange with specific applications.

For our future works, at short term we plan to study the impact of this solution on the network's overhead and the study of its evolution regarding the infrastructure scalability. For long term perspectives, we plan to submit CAM-I proposal to ETSI organism in the goal to standardize it.

Conflict of Interest The authors declare no conflict of interest.

Acknowledgment This work is supported by SCOOP@F project [2].

References

[1] CAMP. Security Credential Management System Proof-of-Concept Implementation. EE Requirements and Specifications Supporting SCMS Software Release 1.1. Standard, Vehicle Safety Communications 5 Consortium, May 2016.

[2] B. Wiedersheim, M. Sall, and G. Reinhard. Sevecom x2014; security and privacy in car2car ad hoc networks. In *2009 9th International Conference on Intelligent Transport Systems Telecommunications, (ITST)*, pages 658–661, 2009.

[3] PRESERVE. PRESERVE deliverable 1.3. v2x security architecture v2. *PREparing SEcuRe VEhicle-to-X Communication Systems IST-269994*, page 106, 2014.

[4] PRESERVE. PRESERVE deliverable 1.1. security requirements of vehicle security architecture. *PREparing SEcuRe VEhicle-to-X Communication Systems IST-269994*, page 69, 2011.

[5] van Sambeek Marcel, Ophelders Frank, Bijlsma Tjerk, van der Kluit Borgert, Turetken Oktay, Eshuis Rik, Traganos Kostas, and Grefen Paul. Architecture for Cooperative ITS Applications in the Netherlands. Technical report, DITCM Innovations, April 2015.

[6] ECO-AT. Eco-AT European Corridor - Austrian Testbed for Cooperative Systems. Technical report, Eco-AT, 2015.

[7] Hasnaâ Aniss. *Overview of an ITS Project: SCOOP@F*, pages 131–135. Springer International Publishing, 2016.

in the neighborhood of the ITSS-R, the ITSS-V must go through an access control/authorization phase to the network (global or private) before executing the selected service. This access control protocol is named NACS (Network Access Control for C-ITS Services)

[2] https://ec.europa.eu/inea/en/connecting-europe-facility/cef-transport/projects-by-country/multi-country/2014-eu-ta-0669-s

[8] Suk-Bok Lee, Gabriel Pan, Joon-Sang Park, Mario Gerla, and Songwu Lu. Secure incentives for commercial ad dissemination in vehicular networks. In *Proceedings of the 8th ACM International Symposium on Mobile Ad Hoc Networking and Computing*, MobiHoc '07, pages 150–159. ACM, 2007.

[9] ISO. ISO 24102-5:2013 Intelligent transport systems - Communications access for land mobiles (CALM) - ITS station management - Part 5: Fast service advertisement protocol (FSAP). *ISO*, page 25, July 2013.

[10] IEEE. IEEE Standard for Wireless Access in Vehicular Environments (WAVE) – Networking Services. *IEEE Std 1609.3-2016 (Revision of IEEE Std 1609.3-2010)*, pages 1–160, April 2016.

[11] ETSI. Draft ETSI DTS 102890-2: Intelligent Transport Systems (ITS), Facilities Layer Function, Part 2: Services Announcement . *ETSI DTS 102890-2*, February 2010.

[12] ETSI. Draft ETSI TS 102 890-1 V0.0.12: Intelligent Transport Systems (ITS); Facilities layer function; Part 1 : Services Announcement (SA) specification. *ETSI TS 102 890-1*, January 2017.

[13] IEEE. Ieee standard for wireless access in vehicular environments security services for applications and management messages. *IEEE Std 1609.2-2013 (Revision of IEEE Std 1609.2-2006)*, pages 1–289, 2013.

[14] ETSI. ETSI EN 302 637-2 V1.3.2: Intelligent Transport Systems (ITS), Vehicular Communications. Basic Set of Applications, Part 2: Specification of Cooperative Awareness Basic Service. *ETSI EN 302 637-2 V1.3.2*, November 2014.

[15] Houda Labiod, Alain Servel, Gerard Seggara, Badis Hammi, and Jean Philippe Monteuuis. A new service advertisement message for etsi its environments: Cam-infrastructure. In *New Technologies, Mobility and Security (NTMS), 2016 8th IFIP International Conference on*, pages 1–4. IEEE, 2016.

[16] TrafficCom Kapsch, Hjelmare Anders, and Moring John. Service announcement using ieee wave service advertisement as a model. In *6th ETSI ITS Workshop 2014*, 2014.

[17] T. Weil. Service management for its using wave (1609.3) networking. In *2009 IEEE Globecom Workshops*, pages 1–6, Nov 2009.

[18] ISO. ISO/TS 16460:2016 Intelligent transport systems – Communications access for land mobiles (CALM) – Communication protocol messages for global usage. Technical report, International Organization for Standardization, September 2016.

[19] ITU-T. ITU-T X.691 - Information technology - ASN.1 encoding rules: Specification of Packed Encoding Rules (PER). Standard, International Telecommunication Union, 2015.

[20] ISO. Intelligent transport systems – Cooperative systems – Classification and management of ITS applications in a global context. Technical report, International Organization for Standardization, April 2014.

[21] ETSI. ETSI EN 302 665 V1.1.1: Intelligent Transport Systems (ITS), Communication Architecture. *ETSI EN 302 665 V1.1.1*, September 2010.

[22] ETSI. ETSI TS 103 097 V1.2.1: Intelligent Transport Systems (ITS), Security header and certificate formats. *ETSI TS 103 097 V1.2.1*, June 2015.

[23] ETSI. ETSI EN 302 637-3 V1.2.2: Intelligent Transport Systems (ITS), Vehicular Communications. Basic Set of Applications, Part 3: Specifications of Decentralized Environmental Notification Basic Service. *ETSI EN 302 637-3 V1.2.2*, November 2014.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 2, No. 3, 1432-1442 (2017)
www.astesj.com
Special Issue on Recent Advances in Engineering Systems

ASTES Journal
ISSN: 2415-6698

Use of machine learning techniques in the prediction of credit recovery

Rogerio Gomes Lopes[*,1], Marcelo Ladeira[2], Rommel Novaes Carvalho[2]

[1]*Bank of Brazil, IT Department, Brazil, rglopes@bb.com.br*

[2]*University of Brasilia, Department of Computer Science, Brasilia, Brazil, mladeira@unb.br, rommel.carvalho@gmail.com*

ARTICLEINFO

Article history:
Received: 04 June, 2017
Accepted: 28 July, 2017
Online: 10 August, 2017

Keywords :
machine learning
data mining
credit recovery
h2o.ai

ABSTRACT

This paper is an extended version of the paper originally presented at the International Conference on Machine Learning and Applications (ICMLA 2016), which proposes the construction of classifiers, based on the application of machine learning techniques, to identify defaulting clients with credit recovery potential. The study was carried out in 3 segments of a Bank's operations and achieved excellent results. Generalized linear modeling algorithms (GLM), distributed random forest algorithms (DRF), deep learning (DL) and gradient expansion algorithms (GBM) implemented on the H2O.ai platform were used.

1 Introduction

This paper is an extension of the work originally presented at the International Conference on Machine Learning and Application (ICMLA 2016) [1], which presented the first results of a Brazilian bank research to reduce its losses with defaulting clients. That study covered only a sample of 22.764 transactions, representing a homogeneous group of bank customers. We extend our previous work by adding all operations from individual costumers which were in arrears in July 2016.

The Figure 1 shows that there was a slight decrease in the number of debtors in June 2016, but increased again in the following months.

The Bank had nearly 54 million active credit agreements with individuals at the end of July 2016. Of this amount, approximately 8.6 million were delayed for 15 days or more, accounting for 15.9% of the contracts. These delinquent contracts amounted to more than R$20.8 billion (US$6.4 billion in July 2016), accounting for approximately 5.8% of the Bank's individuals loan portfolio, an increase of 1.2 percentage points over December 2014. That is, in 21 months the financial volume of overdue loans contracted by individuals increased by 26%.

The Brazilian Central Bank (BACEN) regulation requires financial institutions to classify their credit operations and perform a Provision for Doubtful Accounts (PDA), according to a risk classification. The main criteria for the classification is the number of days in arrears of each individual credit agreement. The Table 1 shows the days-in-delay ranges considered to determine a risk classification and therefore the minimum percentage PDA that financial institutions must reserve. As an operation increases the number of days in arrears, there is a non-linear increase of PDA, which may allocate 100% of the outstanding balance of the contract. For example, an operation with a debit balance of R$ 1,000, with 15 days in arrears, must reserve a minimum provision of R$ 10. The amount of the provision may reach R$ 1,000 if the arrear reach 180 days.

Table 1: Days in arrears x Provision

Days in arrears	Minimum Risk	PDA %
15-30	B	1
31-60	C	3
61-90	D	10
91-120	E	30
121-150	F	50
151-180	G	70
over 180	H	100

At the time of the credit granting, financial institutions assume the credit risk and make the corresponding provisions in accordance with the current Central Bank regulation. Acting in this way, in a possible default of the customer, the financial institution and the stability of the financial system will be protected. However, as a customer delays its operations, the natural reaction of financial institutions is to restrict credit to them, increasing the chances of these

[*]Rogerio Gomes Lopes, Braslia, Brazil, rglopes@bb.com.br.

https://dx.doi.org/10.25046/aj0203179

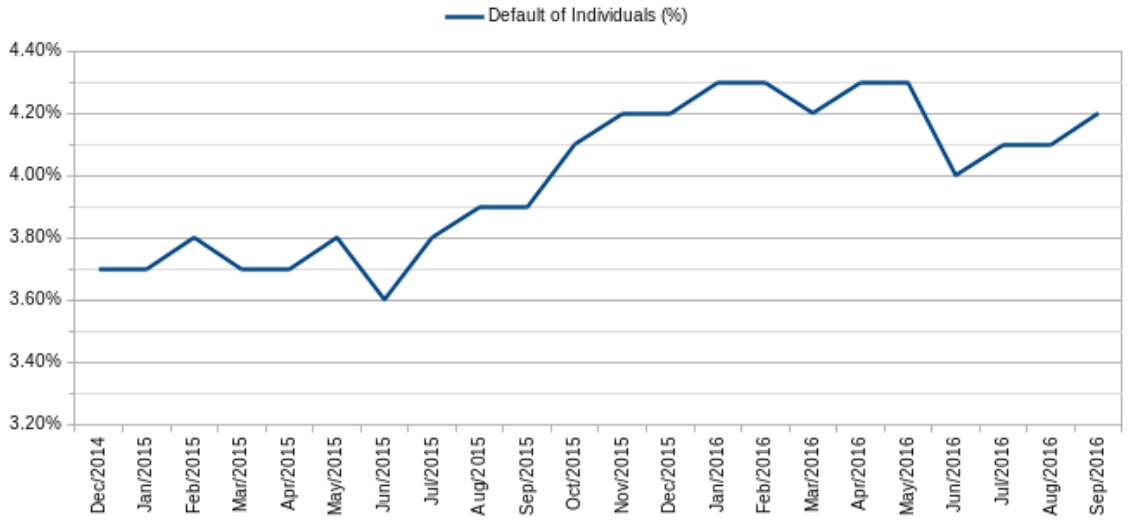


Figure 1: Default of individuals.

customer's evasion to other institutions, since they will not be able to carry out new credit operations with the original institution.

With the increase in delinquency, a mobilization of account managers of the bank began in order to mitigate the evasion of its clients by approaching the customer in arrear and proposing alternatives that could fix the delayed payments. Hence, solving the default situation and the possible loss of the customer of its portfolio, as well as reducing the financial amount allocated to (PDA).

Provided that the selection of the clients is a time and resource consuming task, the main objective of this study was to apply machine learning techniques to predict the recovery probability of credit transactions, providing a list of delinquent clients with the greatest potential for regularization of their operations.

Models were developed using Generalized Linear Models (GLM), Gradient Boosted Methods (GBM), Distributed Random Forest (DRF) and Deep Learning (DL)[1] . The models were compared using the recall indicator, which will be explained on section 3. The models were developed using the R language and H2O machine learning platform, considering its parallel processing capabilities. Further details on section 3. [2]

This paper is organized as follows: Section 2 presents the credit scoring state of the art. Section 3 presents the methodology used in this study. Section 4 presents the modeling and evaluation of the generated models for each method. Section 5 presents the conclusion and future works.

2 State of the Art

The default numbers observed in Brazil, from December 2014 to September 2016, indicate that financial institutions need a tool to support their credit granting decisions. Although there are several studies to identify the customer credit risk, qualifying them as good or bad payers, helping to make a decision to grant credit, there is few research studying the credit recovery, when the delinquency occurs. [2]

In [3], the author conducted a study evaluating 41 publications on the award of credit since 2006, all of them using classifiers to categorize customers as good or bad payers. Those works were organized into three categories of classifiers: individuals; homogeneous ensemble; and heterogeneous ensemble classifiers. Most of the algorithms used were implemented through logistic regression and decision trees, with their use of boosting, bagging and forest variants.

The Table 2 lists the eight datasets that were used in [3] to verify the performance of each of the 41 models proposed, evaluating them from the standpoint of 6 indicators: Area Under the Receiver Operating Curve (AUC), percentage correctly classified (PCC), partial Gini index, H-measure, Brier Score (BS) and Kolmogorov-Smirnov (KS).

Table 2: Datasets used in [3].

Name	Samples	Features	Debtors %
AC	690	14	44.5
GC	1000	20	30.0
Th02	1225	17	26.4
Bene 1	3123	27	66.7
Bene 2	7190	28	30.0
UK	30000	14	4.0
PAK	50000	37	26.1
GMC	150000	12	6.7

[1] Documentation available at http://docs.h2o.ai/h2o/latest-stable/index.html
[2] H2O is an open source machine learning platform, available at www.h2o.ai

In [4], the author presents AUC as an indicator that represent how well classified were the data, independent of its distribution or misclassification costs. PCC is an overall accuracy measure that indicates the percentage of outcomes that were correctly classified.[5]

A score was assigned to each algorithm, referring to the classification received in the comparison between them within the same performance measure . For example, the algorithm K-means was in 12th place considering the AUC indicator, while the KNN was in 29th place. Thus, the scores attributed to them were 12 and 29, respectively. Then, the algorithms were ordered by the average of all metrics, where the 1st place were the algorithm that obtained the lowest score.

The heterogeneous multi-classifiers presented a better performance, although the performance between the three categories was very similar.

The Table 3 presents the results of the benchmark, indicating that the HCES-Bag algorithm obtained the highest AUC result, while the AVG-W and Gasen algorithms reached 80.7% of the PCC.

Table 3: State of Art - Models Comparison - Adapted from [3]

	Algorithm	AUC	PCC
Heterogeneous Ensemble	HCES-Bag	0.932	80.2
	AVG W	0.931	80.7
	GASEN	0.931	80.7
Homogeneous Ensemble	RF	0.931	78.9
	BagNN	0.927	80.2
	Boost	0.93	77.2
Individual	LR	0.931	70.84
	LDA	0.929	78.4
	SVM-Rbf	0.925	79.9

3 Methodology and Infrastructure Setup

This section presents the methodology used in this study, which was segmented in stages according to the phases proposed by CRISP-DM [6]. The result of each phase is described in the next Section.

Training model environment - The models were trained on the H2O.ai platform, in a cluster formed by 5 virtual machines on the same subnet and with the same configuration. Their operating system was Red Hat Enterprise Linux 6.8 64 bits, with 34 cores and 80 GB of RAM. It were used H2O.ai version 3.10.4.5 and R version 3.3.0. It were allocated 44 GB of RAM and all cores of each machine, reaching a total of 170 cores and 220 GB of RAM.

The training dataset consisted of about 40 million copies, requiring a robust platform to be made available for the processing of this data.

The Figure 2 shows the CPU meter of the H2O.ai cluster in action at the moment of the training models. It shows the percentage of use of the processors of each machine, identified by the final number of its IP address (174 to 178) and the port number where the service was running (54321). The intensive use of the 170 available cores shown in the Figure 2 reinforces the need for a robust platform.

Each vertical bar represents 1 core and the colors represent the type of process executed: idle time (blue), user time (green) and system time (red).

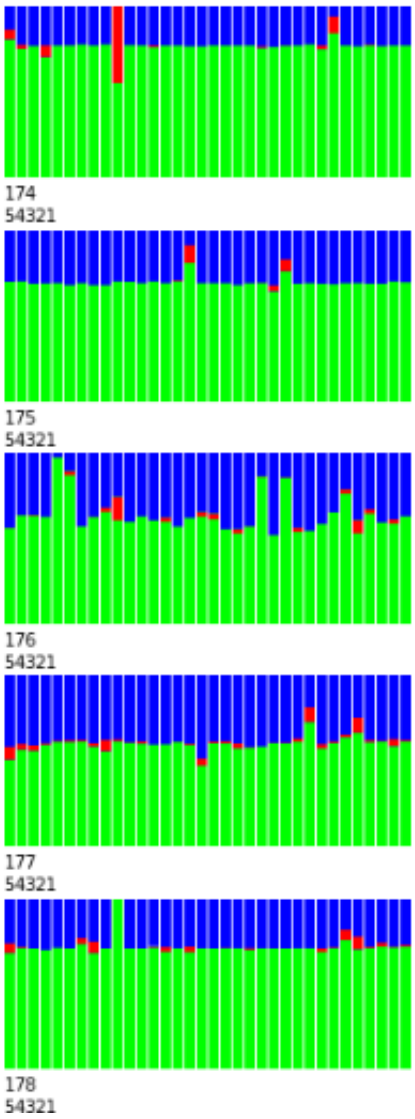


Figure 2: Cluster H2O in action

4 Results

In this sections, the results of the CRISP-DM phases are detailed: Data Understanding, Data preparation, Modeling, Evaluation and Implementation.

4.1 Data Understanding

The dataset was obtained by the extraction of information from legacy systems and customers relationship data marts. It has information about customers accounting,demographic and financial data. The dataset had 28 features and 1 label that indicates the recovery of the respective credit operation. The

Tables 4 and 5 present these 28 characteristics organized by categorical and numerical features.

Table 4: Numeric features

Features	Description
V1	Number of days of delinquency.
V2	Number of days remaining for the end of the contract.
V3	Contract value.
V4	Amount of the outstanding balance.
V5	Amount PDA provisioned for the contract.
V6	Percentage loss expected for the contract.
V7	Quantity of products owned by the customer.
V8	Time of customer relationship with the Bank.
V9	Customer age.
V10	Customer income.
V11	Customer total contribution margin amount.
V12	Value of Gross Domestic Product per capita

Table 5: Categorical features

Features	Description
VC1	Customer portfolio type.
VC2	Customer behavioral segment.
VC3	Product.
VC4	Product modality.
VC5	Structured operation indicator.
VC6	Management level that approved the operation.
VC7	Transaction risk credit.
VC8	Range of past delays.
VC9	PDA lock indicator.
VC10	Customer relationship with the bank.
VC11	Client instruction level.
VC12	Customer gender.
VC13	Nature of customer occupation.
VC14	Customer registration status.
VC15	Customer's age group.
VC16	Age group of relationship time.

For the data understanding, the analysis began in July 2016 containing all credit operations contracts, regardless of the contracted product, with more than 14 days in arrears. In addition, transactions with the highest risk were considered as already lost contracts by our business specialists and removed from our dataset.

For definition of the label, the delay reduction indicator, the following operation was performed, considering that the data of the delayed operations were used in July 2016:

- Delay Reduction Indicator = 1, for all transactions that showed a reduction in the number of days overdue in the subsequent month, that is, in August 2016, or that their debit balances have been reduced.

- Delay Reduction Indicator = 0, otherwise, that is, presenting a delay or debit balance in August 2016 equivalent to or greater than that observed in July 2016.

The Table 6 presents the summary of transactions in the month of July 2016, which resulted in a base with 4,514,029 contracts. Of this total, only 271,193 (6.01%) were recovered.

Table 6: Dataset July 2016

Not recovered	Recovered
4,242,836	271,193
93.99%	6.01%
Samples	4,514,029

The bank has several strategies for credit recovery, according to the customer profile and the category of the credit operation, grouping them with distinct trading rules. Existing segments are divided into massive and individual strategies. Massive strategies are implemented for segments that have a known behavior pattern, whereas individual strategies cover operations that have atypical or special characteristics which require a case-by-case analysis to perform a collection and recovery.

Based on this information, the dataset was splitted into segments compatible with the institution's recovery strategies, grouping similar products and customer segments with characteristics in common removing from the study the segments that have an individualized trading strategy. The Table 7 lists the 11 segments that will be worked on in this study, in addition to the Individualized Strategy segment, which was removed from the study.

4.2 Data Preparation

In this study, the analysis were performed only in the first 3 segments, Mortgage Loan I, II and III. The remaining segments are in the final analysis phase and will be presented at a later time.

Then, the data preparation was started, analyzing each one of the segments, preparing the data sets for the modeling phase.

The Tables 8, 10 and 12 present the summary of descriptive analysis of the numerical features of segments Mortgage Loan I, II and III, respectively. In these tables the data of quartiles and Kendall's Tau [7] of each feature are presented.

The Tables 9, 11 and 13 present the summary of the descriptive analysis of the categorical variables, listing the Kendall's Tau and the number of levels of each feature.

Table 7: Credit Operations Segments

Segment	Credit Recovered				Samples
	No		Yes		
	Qty	%	Qty	%	
Mortgage Loan I	41,398	70.45	17,365	29.55	58,763
Mortgage Loan II	400	73.94	141	26.06	541
Mortgage Loan III	3,537	78.11	991	21.89	4,528
Vehicle Financing I	12,115	87.90	1,667	12.10	13,782
Vehicle Financing II	32,357	86.63	4,993	13.37	37,350
Agribusiness	258,618	98.84	3,021	1.16	261,639
Social Business	137,474	93.53	9,504	6.47	146,978
Credit Card I	17,124	98.92	187	1.08	17,311
Credit Card II	454,864	98.56	6,661	1.44	461,525
Other Operations Income I	186,572	96.53	6,714	3.47	193,286
Other Operations Income II	2,668,890	92.96	201,977	7.04	2,870,867
Individualized Strategy	429,487	95.98	17,972	4.02	447,459

Table 8: Mortgage Loan I - Numerical Features

Feature	Min	1QT	Median	Avg	3QT	Max	Kendall's Tau
V1	15	20	51	87.43	112	624	-0.29
V2	0	10,180	10,420	10,290	10,670	11,620	-0.03
V3	0	0.1	0.13	0.17	0.27	0.67	-0.02
V4	0	1	2	3.94	5	26	0.06
V5	17	25	29	31.28	36	73	0.03
V6	1	3	4	4.45	5	37	0.08
V7	14,790	74,400	87,460	86,270	97,470	164,800	0.01
V8	-124,400	-390.8	156.7	-1,397	256.3	183,400	0.38
V9	0	1,349	3,221	3,712	5,525	46,040	0.04
V10	0	888.1	2,670	19,090	20,960	173,700	-0.17
V11	0	1,586	1,700	1,877	2,000	20,000	0.03
V12	0.68	74,920	88,720	87,250	99,510	173,500	0.00

4.3 Modeling

For each dataset, 4 predictive models were elaborated, using the H2O platform integrated to the R, using the algorithms Generalized Linear Models (GLM), Gradient Boosting Method (GBM), Random Forest (DRF) and Deep Learning (DL). The first three algorithms were chosen because they represent the techniques most used in the calculation of credit risk, which performs a classification task very similar in [8]. The algorithm DL was used to verify its behavior in a knowledge area not yet explored, but with expectation of good suitability due to the use of a great amount of variables. [9]

The datasets of the Mortgage Loan I and III segments were splitted into 3 parts: 70% for training, 20% for validation and 10% for testing. Due to the small number of observations in the Mortgage Loan II, this dataset was splitted only in training and validation in a proportion of 80% and 20%, respectively. The next subsections present the evaluation results for each segment.

4.3.1 Mortgage Loan I

- GLM - This algorithm obtained an AUC = 0.7774755 and a PCC of 66.53%, as shown in the Figure 3 and in the Table 14

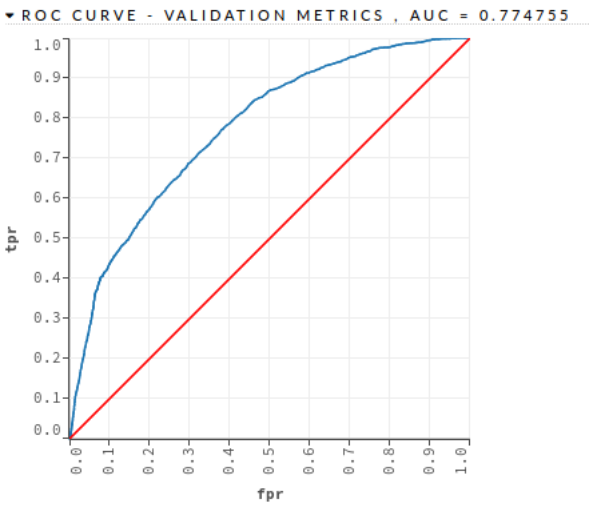


Figure 3: Mortgage Loan I - GLM - Validation Dataset AUC

Table 9: Mortgage Loan I - Categorical Features

Feature	Kendall's Tau	Number of levels
VC1	0.23	6
VC2	0.02	5
VC3	-0.09	3
VC4	-0.21	9
VC5	0.03	12
VC6	0.06	7
VC7	-0.01	2
VC8	0.03	5
VC9	-0.04	16
VC10	0.00	4
VC11	0.05	8
VC13	-0.21	9
VC14	0.01	4
VC15	-0.02	18
VC16	0.00	2

Table 10: Mortgage Loan II - Numerical Features

Feature	Min	1QT	Median	Avg	3QT	Max	Kendall's Tau
V1	15	21	48	89	113	507	-0.15
V2	0	1,626	2,928	3,037	4,076	6,776	-0.14
V3	0	0	0	0	0	1	0.09
V4	2	6	6	10	13	31	0.03
V5	30	42	48	49	55	85	0.02
V6	1	4	6	7	9	36	-0.04
V7	4,400	28,000	45,000	59,560	70,560	240,000	-0.05
V8	-31,770	-148	91	-650	371	25,070	0.30
V9	0	4,990	6,190	6,663	9,099	15,260	0.08
V10	0	173	650	8,744	10,230	116,000	-0.20
V11	0	1,598	2,965	5,082	5,553	128,900	-0.11
V12	0	9,316	20,500	36,670	47,120	215,200	-0.14

- DRF - This algorithm was implemented with 500 trees and a maximum depth of 7. The DRF algorithm obtained an AUC = 0.880589 and a PCC = 75.85%, as shown in the Figure 4 and in the Table 14

- DL - Deep Learning This algorithm was implemented with 2 hidden layers with 200 neurons each one. The DRF algorithm obtained an AUC = 0.898203 and a PCC = 79.22%, as shown in the Figure 5 and in the Table 14.

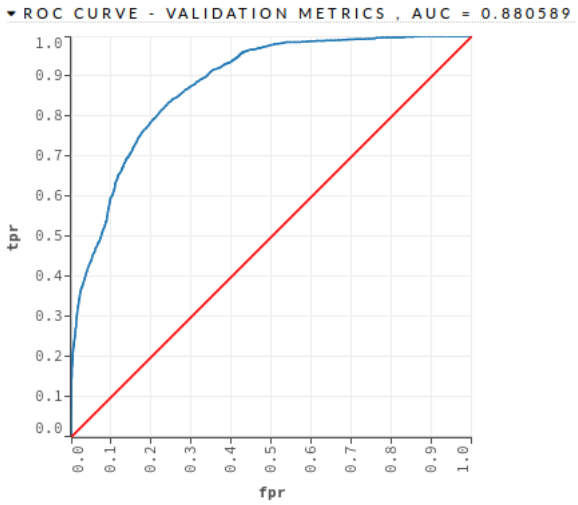


Figure 4: Mortgage Loan I - DRF - Validation Dataset AUC

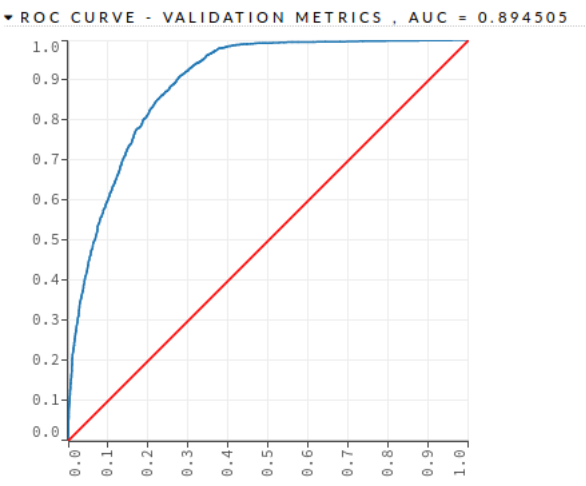


Figure 5: Mortgage Loan I - DL - Validation Dataset AUC

Table 11: Mortgage Loan II - Categorical Features

Features	Kendall's Tau	Number of levels
VC1	0.11	4
VC2	0.13	3
VC4	-0.19	8
VC5	0.01	10
VC6	0.03	6
VC7	-0.04	2
VC8	0.10	5
VC9	0.06	10
VC11	-0.15	5
VC13	-0.19	8
VC14	0.10	4
VC15	-0.06	13

Table 12: Mortgage Loan III - Numerical Features

Feature	Min	1QT	Median	Avg	3QT	Max	Kendall's Tau
V1	15	30	81	131	181	511	-0.31
V2	0	4,876	7,530	6,616	8,203	10,910	-0.01
V3	0.00	0.02	0.05	0.06	0.07	0	0.00
V4	0	5	9	11	15	54	-0.02
V5	20	35	43	44	52	78	-0.06
V6	2	6	8	10	11	71	0.05
V7	20,000	100,000	142,500	188,700	213,800	3,000,000	-0.07
V8	-257,600.00	-5,476.00	-930.00	-9,087.00	257.70	199,600	0.36
V9	0	2,769	4,660	4,968	6,485	46,040	0.01
V10	0	3,317	14,200	51,540	53,470	1,212,000	-0.20
V11	0	2,280	5,542	10,700	11,130	337,600	-0.01
V12	250	88,900	134,500	177,500	203,200	3,084,000	-0.08

- GBM - This algorithm was implemented with 500 trees and a maximum depth of 7. The GBM algorithm obtained an AUC = 0.988574 and a PCC = 93.90%, as shown in the Figure 6 and in the Table 14

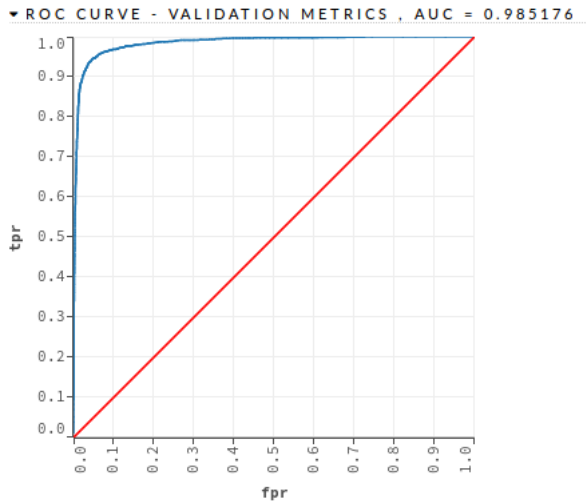


Figure 6: Mortgage Loan I - GBM - Validation Base AUC

4.3.2 Mortgage Loan II

Because of the small number of records, this dataset was splitted only in training and testing, in the ratio of 80:20, and validation was performed through cross validation with 10 folds.

- GLM - This algorithm obtained an AUC = 0.848474 and a PCC = 65.51%, as shown in the Figure 7 and in the Table 15.

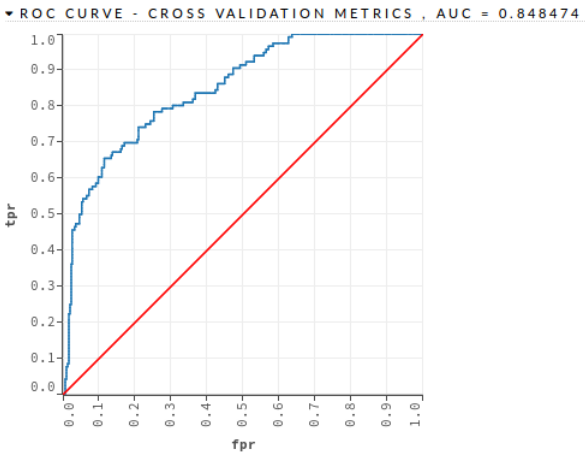


Figure 7: Mortgage Loan II - GLM - Validation Dataset AUC

Table 13: Mortgage Loan III - Categorical Features

Feature	Kendall's Tau	Number of levels
VC1	0.16	6
VC2	0.00	4
VC3	-0.24	2
VC4	-0.21	33
VC5	-0.06	12
VC6	-0.02	7
VC7	-0.03	2
VC8	0.00	5
VC9	-0.02	16
VC10	-0.10	5
VC11	0.03	7
VC12	0.00	2
VC13	-0.21	9
VC14	0.00	5
VC15	-0.04	17
VC16	0.17	2

Table 14: Mortgage Loan I - Confusion Matrix

Algorithm		0	1	Err %	PCC
GLM	0	5003	3050	37.87	
	1	776	2603	22.96	
	Total	5779	5653	33.46	66.52
DRF	0	6638	1415	17.57	
	1	816	2563	24.14	
	Total	7454	3978	19.51	75.85
DL	0	6290	1763	21.89	
	1	517	2862	15.39	
	Total	6897	4625	19.94	79.22
GBM	**0**	**7770**	**283**	**3.51**	
	1	**238**	**3141**	**7.04**	
	Total	**8008**	**3424**	**4.55**	**93.90**

Table 15: Mortgage II - Confusion Matrix

Algorithm		0	1	Err %	PCC
GLM	0	270	35	11.47	
	1	40	76	34.48	
	Total	310	111	17.81	65.51
DRF	0	302	3	0.98	
	1	17	99	14.65	
	Total	319	102	4.75	85.34
DL	**0**	**297**	**8**	**2.62**	
	1	**10**	**106**	**8.62**	
	Total	**307**	**114**	**4.27**	**91.37**
GBM	0	301	4	1.31	
	1	16	100	13.79	
	Total	317	104	4.75	86.20

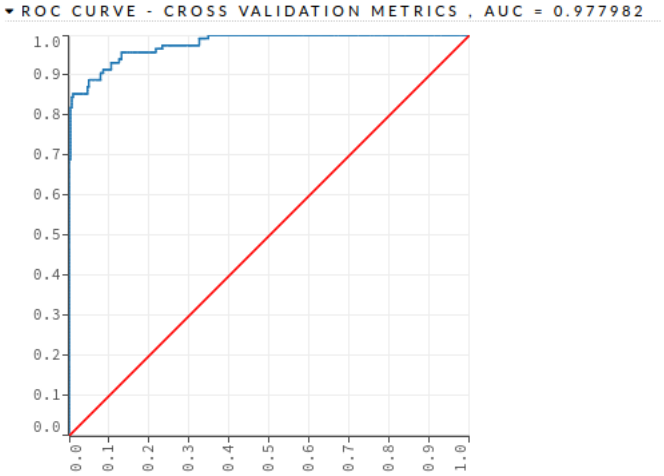


Figure 8: Mortgage Loan II - DRF - Validation Dataset AUC

- DRF - This algorithm was implemented with 500 trees and a maximum depth of 7. The DRF algorithm obtained an AUC = 0.977982 and a PCC = 93.10%, as shown in the Figure 8 and in the Table 15

- DL - This algorithm was implemented with 2 hidden layers with 200 neurons each one. The DRF algorithm obtained an AUC = 0.956868

and a PCC = 91.37%, as shown in the Figure 5 and in the Table 15.

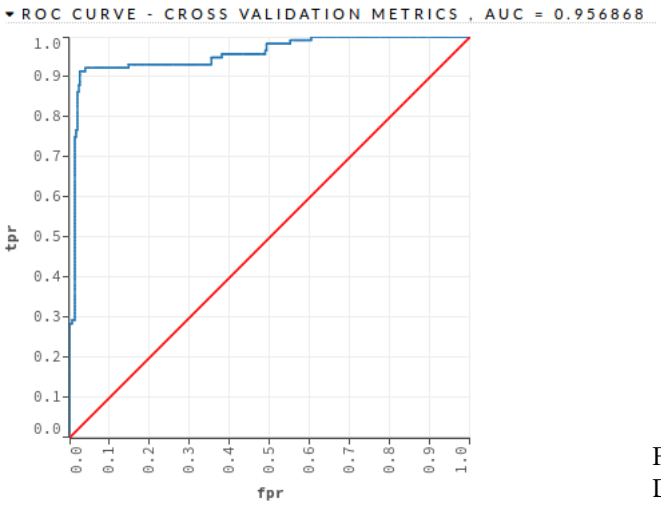


Figure 9: Mortgage Loan II - DL - Validation Dataset AUC

- GBM - This algorithm was implemented with 500 trees and a maximum depth of 7. The GBM algorithm obtained an AUC = 0.972640 and a PCC = 86.20%, as shown in the Figure 10 and in the Table 15

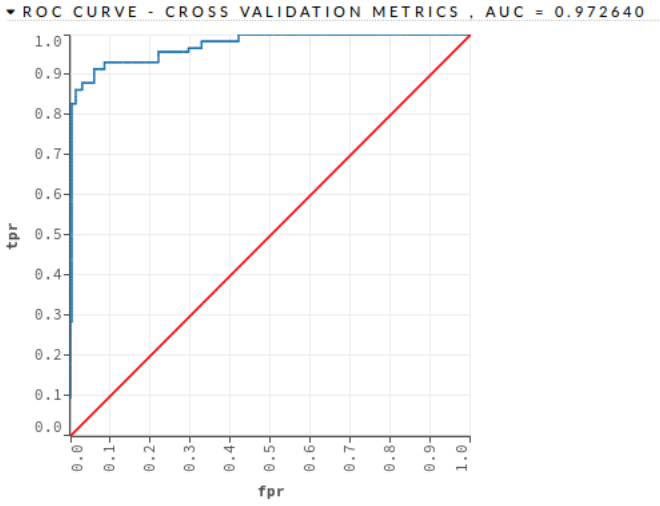


Figure 10: Mortgage Loan II - GBM - Validation Dataset AUC

4.3.3 Mortgage Loan III

- DRF - This algorithm was implemented with 500 trees and a maximum depth of 7. The DRF algorithm obtained an AUC = 0.950718 and a PCC = 83.51%, as shown in the Figure 11 and in the Table 16

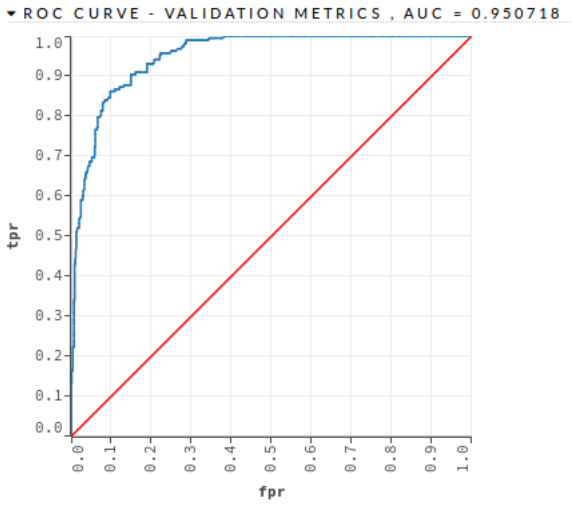


Figure 11: Mortgage Loan III - DRF - Validation Dataset AUC

- DL - This algorithm was implemented with 2 hidden layers with 200 neurons each one. The DRF algorithm obtained an AUC = 0.939082 and a PCC = 78.20%, as shown in the Figure 12 and in the Table 16.

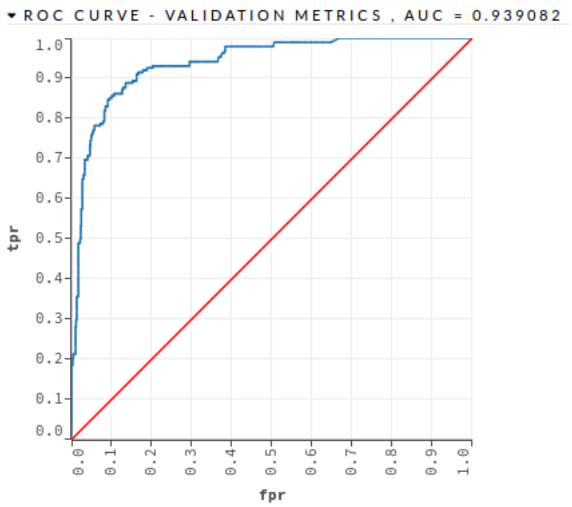


Figure 12: Mortgage Loan III - DL - Validation Dataset AUC

- GBM - This algorithm was implemented with 500 trees and a maximum depth of 7. The GBM algorithm obtained an AUC = 0.955728 and a PCC = 98.93%, as shown in the Figure 13 and in the Table 16

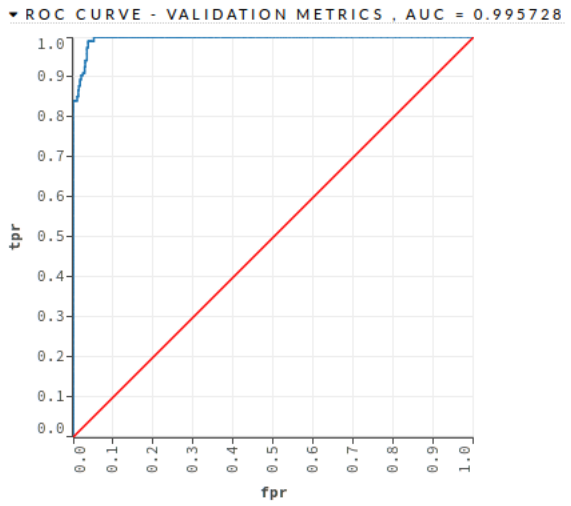


Figure 13: Mortgage Loan III - GBM - Validation Dataset AUC

- GLM - This algorithm obtained an AUC = 0.814560 and a PCC = 60.10%, as shown in the Figure 14 and in the Table 16

Table 16: Mortgage Loan III - Confusion Matrix

Algorithm	0	1	Err %	PCC
GLM	601	95	13.64	
	75	113	39.89	
	676	208	19.23	60.10
DRF	640	56	8.04	
	31	157	16.48	
	671	213	9.84	83.51
DL	656	40	5.74	
	41	147	21.80	
	697	187	9.16	78.20
GBM	**670**	**26**	**3.73**	
	2	**186**	**1.06**	
	672	**212**	**3.16**	**98.93**

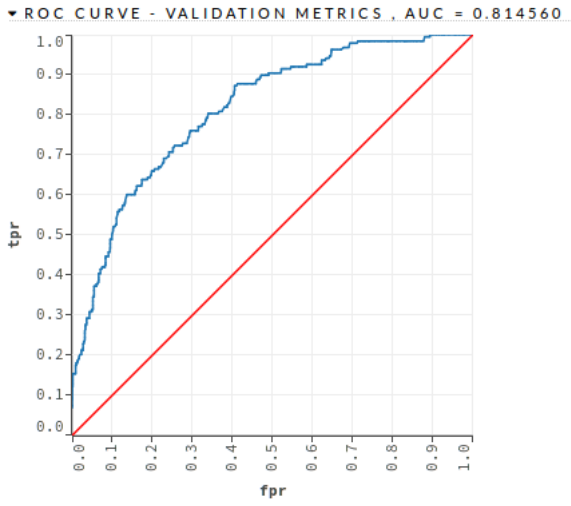


Figure 14: Mortgage Loan III - GLM - Validation Dataset AUC

5 Conclusion

The main objective of this study was to apply machine learning techniques to predict the probability of recovery of credit transactions, providing a list of defaulting clients with greater potential for regularization of their operations.

Studies were carried out on 3 segments of credit operations, which have different recovery strategies, the Mortgage Loan segments I, II and III. With the machine learning, it was possible to elaborate predictive models with great contribution to assist the Managers in the approach to their clients with operations in arrears.

Mortgage Loan I - The model with the highest recall was obtained with the GBM algorithm. In a total of 11,342 contracts in default, there were 3,424 contracts recovered. The model was able to correctly predict 3,141 contracts, reaching a recall of 92.86%. Using the prioritization list generated by the model, the work of Bank Managers would be more assertive. In addition, the model correctly predicted 7,770 (97.02%) contracts, out of 8,008 contracts that would not be recovered.

Mortgage Loan II - The model with the highest recall was obtained with the DL algorithm. In a total of 421 delinquent contracts, there were 116 contracts recovered. The model was able to correctly predict 106 contracts, reaching a recall of 94.38%. In addition, the model correctly predicted 297 (96.74%) contracts out of 307 contracts that would not be recovered.

Mortgage Loan III - The model with the highest recall was obtained with the GBM algorithm. In a total of 884 delinquent contracts, there were 212 contracts recovered. The model was able to accurately predict 186 contracts, reaching a recall of 98.94%. In addition, the model correctly predicted 670 (99.70%) contracts out of 672 contracts that would not be recovered.

The predictive models obtained from the analysis of the first three segments, out of a total of 11, have already shown a potential great benefit to the bank, effectively assisting its customers with delayed operations and avoiding unnecessary efforts in attempts in attempts of negotiation in contracts with low probability of recovering.

5.1 Future Works

The results obtained so far strengthen initiatives for the development of predictive models using machine learning techniques in the Bank studied.

With the increase in the efficiency of credit recovery, the Bank will benefit from the reduction in Allowance for Loan Losses (PDA), directly promoting positive results, with the reversal of provisions already made.

Thus, the study will be expanded to the 8 segments that have not yet been modeled, increasing the use of models obtained through machine learning techniques in credit recovery. In addition, the models al-

ready obtained can be improved with the use of ensemble models.

References

[1] Rogerio G. Lopes, Rommel N. Carvalho, Marcelo Ladeira, and Ricardo S. Carvalho. Predicting Recovery of Credit Operations on a Brazilian Bank. pages 780–784. IEEE, December 2016.

[2] Sung Ho Ha. Behavioral assessment of recoverable credit of retailer's customers. *Inf. Sci.*, 180(19):3703–3717, October 2010.

[3] Stefan Lessmann, Bart Baesens, Hsin-Vonn Seow, and Lyn C. Thomas. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1):124–136, 2015.

[4] Wouter Verbeke, Karel Dejaeger, David Martens, Joon Hur, and Bart Baesens. New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *European Journal of Operational Research*, 218(1):211 – 229, 2012.

[5] Karel Dejaeger, Frank Goethals, Antonio Giangreco, Lapo Mola, and Bart Baesens. Gaining insight into student satisfaction using comprehensible data mining techniques. *European Journal of Operational Research*, 218(2):548 – 562, 2012.

[6] Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rudiger Wirth. Crisp-dm 1.0 step-by-step data mining guide. Technical report, The CRISP-DM consortium, August 2000.

[7] Stephan Arndt, Carolyn Turvey, and Nancy C Andreasen. Correlating and predicting psychiatric symptom ratings: Spearmans r versus kendalls tau correlation. *Journal of psychiatric research*, 33(2):97–104, 1999.

[8] Bart Baesens, Tony Van Gestel, Stijn Viaene, Maria Stepanova, Johan Suykens, and Jan Vanthienen. Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the operational research society*, 54(6):627–635, 2003.

[9] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 2, No. 3, 1443-1452 (2017)
www.astesj.com
Special Issue on Recent Advances in Engineering Systems

ASTES Journal
ISSN: 2415-6698

Monocular Height Estimation Method with 3 Degree-Of-Freedom Compensation of Road Unevennesses

Alex Masuo Kaneko [*], Kenjiro Yamamoto

Hitachi, Ltd. Research & Development Group. Center for Technology Innovation - Mechanical Engineering. Robotics Research Department. Japan.

ARTICLEINFO

Article history:
Received: 08 June, 2017
Accepted: 26 July 2017
Online: 10 August, 2017

Keywords:
height estimation
autonomous robots
monocular camera
3 DOF compensation

ABSTRACT

Height estimation of objects is a valuable information for locomotion of autonomous robots and vehicles. Even though several sensors such as stereo cameras have been applied in these systems, cost and processing time have been motivating solutions with monocular cameras. This research proposes two new methods: i) height estimation of objects using only a monocular camera based on flat surface constraints and ii) 3 degree-of-freedom compensation of errors caused by roll, pitch and yaw variations of the camera when applying the Flat Surface Model. Experiments outdoors with the KITTI benchmark data (4997 frames and 436 objects) resulted in improved accuracy of the estimated heights from a maximum error of 1.51 m to 1.12 m and reduced number of estimation failures by 4 times, proving the validity and effectiveness of the proposed method.

1 Introduction

This paper is an extension of work originally presented in the 2016 IEEE/SICE International Symposium on System Integration [1]. This previous work proposed a monocular height estimation method by chronological correction of road unevenness, which basic experiments in laboratory and asphalt correcting camera pitch variations proved the validity of the method. However, since only pitch variations were considered and experiments were conducted on asphalt in one environment, several items remained as future work. Therefore, we extended our work by the following:

- Analysis and 3 degree-of-freedom (3DOF) compensation of errors caused by roll, pitch and yaw variations.
- Extended experiments on asphalt with several conditions of road and objects, permitting further analysis of external disturbances.

Cameras have been applied in many fields such as robotics and autonomous driving for localization and object recognition [2]-[7]. Even though stereo cameras can provide the depth to obstacles, their higher cost and required processing have motivated several studies to estimate depth or height with monocular cameras. Height estimation permits the robot to detect and avoid potential obstacles on the road, becoming a valuable information of the surrounding environment. Studies [8] and [9] estimate height of objects with a steady camera. While the first relies on a previous calculation of the vanishing point, the latter relies on a known object height in the scene. On the other hand, other studies focus on height estimation using a moving camera. Study [10] estimates height by computing the focus of expansion in the scene and segmenting ground and plane by sinusoidal model fitting in reciprocal-polar space. The method proposed in [11] estimates height of objects on the road by obstacle segmentation and known camera displacement from odometric measurements, refining the measurements with several frames. Although there are many promising height estimation methods, they still strongly rely on extra extraction of information from the scene or external sensors. Moreover, the presented methods assume that the ground is flat and no discussions or analyses of eventual pose variations of the camera were mentioned. In this context, we propose in this work a height estimation method that requires no previous information of the scene, nor information from

[*]Corresponding author: Alex Masuo Kaneko. Hitachi, Ltd. Research & Development Group. Center for Technology Innovation - Mechanical Engineering. Robotics Research Department. 832-2, Horiguchi, Hitachinaka, Ibaraki, Japan.
E-mail: alexmasuo.kaneko.bs@hitachi.com

https://dx.doi.org/10.25046/aj0203180

external sensors. Furthermore, even though we assume flat surface, we analyze and compensate the effects of roll, pitch and yaw variations of the camera.

The rest of this paper is organized as follows. Section 2 introduces the existing issues of monocular cameras due to depth ambiguity and roll, pitch and yaw variations. Section 3 explains the proposed monocular height estimation method and compensations of roll, pitch and yaw variations. The conducted experiments are detailed in section 4. The obtained results are discussed in section 5. Lastly, the conclusions and future work are provided in section 6.

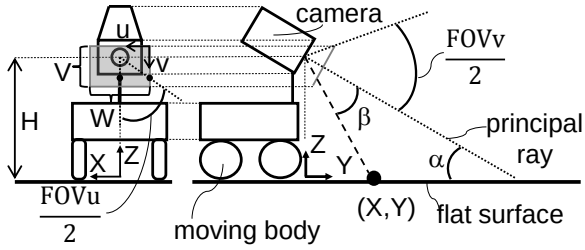


Figure 1: Conversion between pixels and meters using the Flat Surface Model.

2 Issues of Monocular Cameras

In this section, two main issues of monocular cameras are introduced: A) depth ambiguity and B) effects of roll, pitch and yaw variations.

2.1 Depth Ambiguity

In order to explain this limitation, we first briefly introduce the Flat Surface Model (FSM), a technique that permits easy relation of pixels (in camera coordinates) and meters (in real world coordinates), and is commonly applied with monocular cameras [12]-[14]. Figure 1 shows a moving body, a camera and flat surface. The camera is attached to the moving body at a known height H and angle α in relation to the flat surface. The line connecting the camera center and the flat surface with angle α is called principal ray. The coordinate system of the camera is defined by pixels (u,v), with a known and fixed vertical length V and a horizontal length W. The maximum angle seen by the camera (field of view) in both u and v directions are fixed and defined as angles $FOVu$ and $FOVv$. The coordinate system defined by (X,Y,Z) is fixed on the moving body. If we assume that the surface is flat, the relation between a pixel (u,v) and its real position (X,Y,Z) in meters is given by (1) to (3), where β is the angle in relation to the principal ray.

$$\beta = atan(\frac{(2v-V)}{V} tan(\frac{FOVv}{2})) \quad (1)$$

$$Y = \frac{H}{tan(\alpha+\beta)} \quad (2)$$

$$X = Y \frac{cos(\beta)}{\cos(\alpha+\beta)} \frac{(2u-W)}{W} tan(\frac{FOVu}{2}) \quad (3)$$

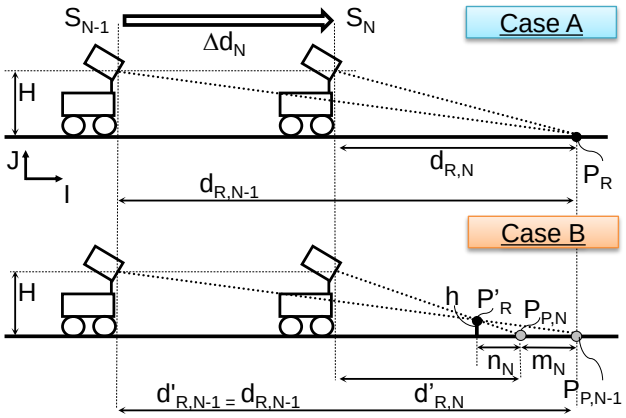


Figure 2: Example of application (A) and limitation (B) when applying the Flat Surface Model.

One common application of the FSM is to estimate the camera displacement by Visual Odometry (VO) [12]-[14], which is briefly explained in "Case A" of Figure 2. First, consider the moving body and camera described in Figure 1 running on a flat surface. In an initial position S_{N-1}, the camera takes a frame and shoots a point P_R on the ground, computing $Y_{N-1} = d_{R,N-1}$. Next, consider that the moving body moves Δd_N in direction I of a coordinate system (I,J) fixed on the ground, reaching position S_N. In this new position, it takes another frame, tracks point P_R and $Y_N = d_{R,N}$ is obtained by the FSM. Using the computed information Y_{N-1} and Y_N from the two positions S_{N-1} and S_N, the real camera displacement ΔCam_N can be correctly estimated by (4). By repeating the previous steps, the displacement of the moving body can be estimated on the following positions. Notice that we showed a simple case with only one point P_R and direction I, but many points and directions can be considered in VO.

$$\Delta Cam_N = d_{R,N-1} - d_{R,N} = \Delta d_N \quad (4)$$

The depth ambiguity can be visualized in "Case B" of Figure 2, which contains an object of height h on the ground. Let's assume that in position S_{N-1} the camera shoots a point P'_R on the top of the object, and the projection on the flat surface is point $P_{P,N-1}$, exactly on the same location as point P_R in "Case A". Although points P_R and P'_R belong to different (X,Y,Z) in the world, the FSM can't distinguish this ambiguity and computes $Y_{N-1} = d'_{R,N-1} = d_{R,N-1}$. However, when the camera moves Δd_N in direction I, it tracks point P'_R and consequently, computes $Y_N = d'_{R,N}$ at position S_N. Finally, the displacement $\Delta Cam'_N$ seen by the camera becomes as (5), what shows that the real displacement Δd_N is not correctly estimated due to the presence of the object.

$$\Delta Cam'_N = d'_{R,N-1} - d'_{R,N} \neq \Delta d_N \quad (5)$$

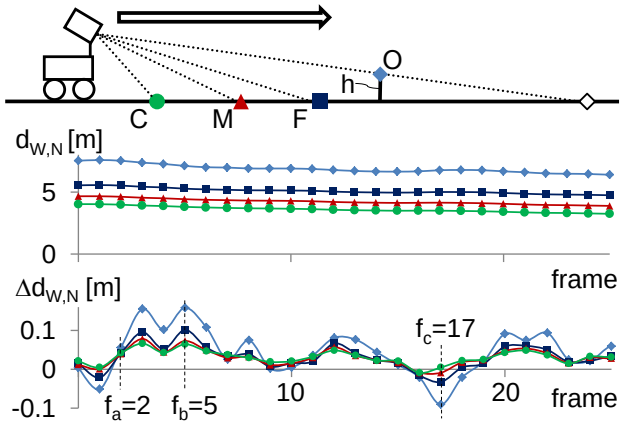


Figure 3: Influences of pose variation on the Flat Surface Model.

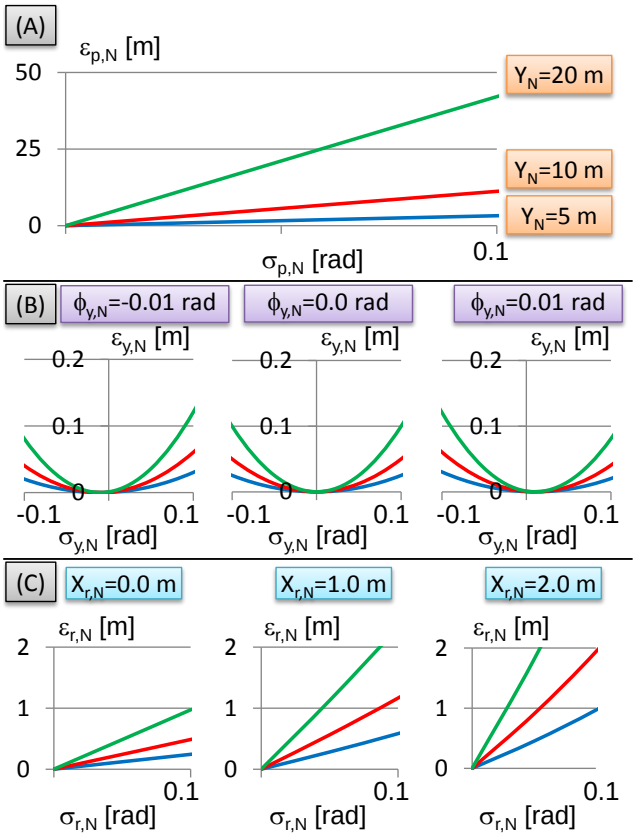


Figure 4: Expected errors of the calculated distances by the FSM in function of pitch (A), yaw (B) and roll (C).

2.2 Effects of Roll, Pitch and Yaw Variations

Even though the FSM assumes that the surface is perfectly flat, in fact small unevennesses exist. Such influences are detailed in Figure 3, with a simple analysis done on a real asphalt. A camera was fixed on a moving body and moved tracking four points: i) on a static object with height h (point O), ii) on the ground far from the camera (point F), iii) on the ground close to the camera (point C) and iv) on the ground with a distance between F and C (point M). In each frame, the distances $Y_N = d_{W,N}$ obtained by the FSM were computed. In order to better visualize the influences of unevennesses, the corresponding displacements $\Delta d_{W,N} = d_{W,N} - d_{W,N-1}$ are also displayed in the figure. If the surface of the asphalt was really flat, then the displacements $\Delta d_{W,N}$ of each point C, M and F were expected to be the same in each frame. However, while this happened in frame $f_a = 2$ for example (suggesting that the pose of the camera was exactly the one expected by the FSM), in frames $f_b = 5$ and $f_c = 17$ the displacements were different (suggesting that the pose of the camera was different from the one expected by the FSM). Moreover, Figure 3 shows another important pattern, which points closer to the camera have smaller magnitudes of $\Delta d_{W,N}$: the magnitude of $\Delta d_{W,N}$ calculated with C in a determined frame is smaller than the one calculated with M, which is smaller than the one calculated with F, which is smaller than the one calculated with O in the same frame. This example clearly shows the effect of unevennesses on the FSM. A further analysis according to roll, pitch and yaw variations is presented in Table 1 and Figure 4.

2.2.1 Pitch

Figure (a) of Table 1 shows the influence of pitch variations on the FSM. Consider that in a certain frame N a moving body that is shooting a point O_R on a static object on the surface has its pitch changed (represented by $\sigma_{p,N}$) by an unevenness. This variation shifts height H to $H_{p,N}$ and therefore the camera computes the distances by the FSM in relation to a new wrong flat surface, $FS_{p,N}$. In this wrong surface, the projection of point O_R becomes $P_{p,N}$ and consequently, $Y_N = d_{p,N}$ is calculated. However, the correct distance in such configuration is the one computed with G_R, the projection of O_R on the real surface on the ground, resulting in $Y_N = d_{R,N}$. The relation between the wrong ($d_{p,N}$) and correct ($d_{R,N}$) distances is shown by (6), (7) and (8), where l_p is the axis of rotation and $\gamma_{p,N}$ is the angle between the camera height H and l_p. It is important to notice that the presented equations don't explicitly contain the object height h. It happens because the computed $Y_N = d_{p,N}$ itself contains this information, since it is function of O_R, $P_{p,N}$ and h. We define the error caused by pitch variation $\sigma_{p,N}$ as $\varepsilon_{p,N}$, which is the difference between the wrong estimated distance ($d_{p,N}$) and correct one ($d_{R,N}$), according to (9). Figure 4 (A) quantifies this error $\varepsilon_{p,N}$ by adopting as example $H = 1.65$ m, $l_p = 1.03$ m, several camera pitch variations ($\sigma_{p,N}$) and apparent distances to a point by the FSM ($Y_N = d_{p,N}$). The adopted values of the constant parameters H and l_p are the same as the ones used in the experiments section for easier analysis of the obtained results. We can observe that $\varepsilon_{p,N}$ increases with the increase of $\sigma_{p,N}$ and Y_N. Even for a pitch variation of $\sigma_{p,N} = 0.1$ rad, the error $\varepsilon_{p,N}$ can reach nearly 50 m when $Y_N = 20$ m, what shows that $\varepsilon_{p,N}$ is very sensitive even to small $\sigma_{p,N}$.

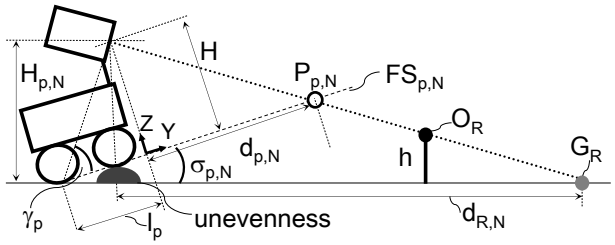
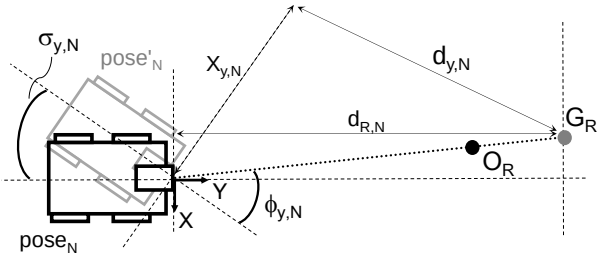
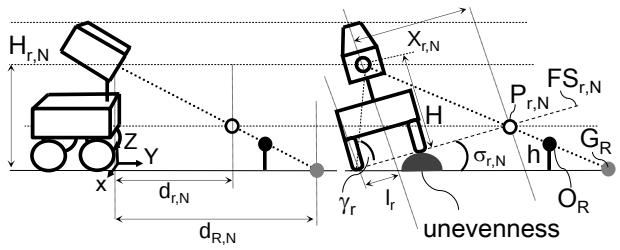
Variation	Compensation
 (a) pitch	$\gamma_p = \arctan(\frac{H}{l_p})$ (6) $H_{p,N} = \sqrt{l_p^2 + H^2} sin(\sigma_{p,N} + \gamma_p)$ (7) $d_{R,N} = \frac{H_{p,N}[(d_{p,N} + l_p)cos(\sigma_{p,N}) - \sqrt{l_p^2 + H^2}cos(\sigma_{p,N} + \gamma_p)]}{H_{p,N} - (d_{p,N} + l_p)sin(\sigma_{p,N})}$ (8) $\varepsilon_{p,N} = \lvert d_{R,N} - d_{p,N} \rvert$ (9)
 (b) yaw	$\phi_{y,N} = atan(\frac{X_{y,N}}{d_{y,N}})$ (10) $d_{R,N} = \sqrt{X_{y,N}^2 + d_{y,N}^2} cos(\sigma_{y,N} - \phi_{y,N})$ (11) $\varepsilon_{y,N} = \lvert d_{R,N} - d_{y,N} \rvert$ (12)
 (c) roll	$\gamma_r = \arctan(\frac{H}{l_r})$ (13) $H_{r,N} = \sqrt{l_r^2 + H^2} sin(\sigma_{r,N} + \gamma_r)$ (14) $d_{R,N} = \frac{H_{r,N} d_{r,N}}{H_{r,N} - (l_r + X_r)sin(\sigma_{r,N})}$ (15) $\varepsilon_{r,N} = \lvert d_{R,N} - d_{r,N} \rvert$ (16)

Table 1: Error compensation of the calculated distances by the FSM caused by variations of roll, pitch and yaw.

2.2.2 Yaw

Figure (b) of Table 1 describes the influence of yaw variations on the FSM. Consider that in a certain frame N, a moving body that is shooting a point O_R on a static object on the surface has its yaw changed ($\sigma_{y,N}$). In the former pose ($pose'_N$), the camera computes the distance to point G_R as $Y_N = d_{y,N}$ and $X_N = X_{y,N}$. However, the correct distance in Y direction in the new pose ($pose_N$) is $d_{R,N}$. Equations (10) and (11) show the relation between the wrong and correct distances, where $\phi_{y,N}$ is the angle in which the camera sees the object in the former pose. The error $\varepsilon_{y,N}$ caused by $\sigma_{y,N}$ is defined as (12). Figure 4 (B) shows error $\varepsilon_{y,N}$ in function of yaw variation ($\sigma_{y,N}$), $\phi_{y,N}$ and apparent distance to the point ($Y_N = d_{y,N}$). First, when $\phi_{y,N} = 0$ rad, we can observe that $\varepsilon_{y,N}$ increases with the increase of Y_N and absolute value of $\sigma_{y,N}$. For similar conditions of Y_N and $\sigma_{y,N}$, when $\phi_{y,N} \neq 0$ the errors are shifted to the left or right according to the value of $\phi_{y,N}$. From this analysis we can observe that small yaw variations such as $\sigma_{y,N} = 0.1$ rad results in $\varepsilon_{y,N} > 0.1$ m for the adopted range of the parameters in the example, what shows that yaw variations cause smaller errors in the distances computed by the FSM comparing to the pitch variations.

2.2.3 Roll

Finally, Figure (c) of Table 1 shows the influence of roll variations on the FSM. Consider a camera tracking a point O_R on a static object of height h on the surface in a certain frame N. In the same frame, the moving body is affected by an unevenness on the flat surface, varying its roll ($\sigma_{r,N}$). This variation changes height H to $H_{r,N}$ and the camera computes the distances by the FSM in relation to a wrong flat surface, $FS_{r,N}$. In this wrong surface, the projection of point O_R becomes $P_{r,N}$ and $Y_N = d_{r,N}$ is calculated. However, the correct distance is the one computed with G_R, the projection of O_R on the real surface on the ground, resulting in $Y_N = d_{R,N}$. The relation between the wrong ($d_{r,N}$) and correct ($d_{R,N}$) distances is shown by (13), (14) and (15), where l_r is the axis of rotation

and $\gamma_{r,N}$ is the angle between the camera height H and l_r. We define the error caused by roll variation $\sigma_{p,N}$ as $\varepsilon_{p,N}$, which is the difference between wrongly estimated distance ($d_{r,N}$) and correct one ($d_{R,N}$), as shown in (16). Figure 4 (C) shows error $\varepsilon_{r,N}$ in function of camera roll variation ($\sigma_{r,N}$), X_r and apparent distance to the point ($Y_N = d_{r,N}$). We can observe that $\varepsilon_{r,N}$ increases with the increase of $\sigma_{r,N}$, X_r and Y_N. For roll variations of $\sigma_{r,N} = 0.1$ rad, $\varepsilon_{r,N} > 2.0$ m can occur for the adopted range of the parameters in the example, but those errors are still smaller than the ones caused by pitch variation.

3 Proposed Method

This section is divided into three parts: A) proposed height estimation method, B) compensation of roll, pitch and yaw variations and C) proposed algorithm.

3.1 Proposed Height Estimation

Although the FSM has the ambiguity limitation when computing VO, in fact, the difference of the obtained displacements caused by the object (ΔCam_N, $\Delta Cam'_N$) contains useful information. First, the presence of objects in the scene influence the apparent displacements in pixels seen from the camera. Such difference in apparent displacements is explored in [12] to find irregularities in the optic flow and detect precipices. On the other hand, no further information can be extracted by existing techniques. Here, our method is based on the principle that since the FSM assumes known H and α, then we can assume that the resulting projections are also function of these dimensions. If we further observe the geometrical relations caused by the FSM and triangulation in two positions (Case B in Figure 2), we can in fact obtain geometrical relations in function of H and α, as shown in (17), (18) and (19). From these equations and (4), we obtain (20) to (22). Several relations can be observed from the equations. First, the object causes an extra amount of apparent displacement, defined as m_N, and it is proportional to the object height h and camera height H. Second, the object height is function of the correct camera displacement ΔCam_N and wrong apparent displacement $\Delta Cam'_N$. As afore mentioned, ΔCam_N can be estimated by traditional VO and even though this technique will be applied, it is not the focus of this work.

$$\frac{H}{d_{R,N-1}} = \frac{h}{m_N + n_N} \tag{17}$$

$$\frac{H}{d'_{R,N}} = \frac{h}{n_N} \tag{18}$$

$$d'_{R,N} = d_{R,N} - m_N \tag{19}$$

$$\Delta Cam'_N = d_{R,N-1} - d_{R,N} + m_N = \Delta Cam_N + m_N \tag{20}$$

$$m_N = \frac{h(d_{R,N-1} - d'_{R,N})}{H} = \frac{h(d'_{R,N-1} - d'_{R,N})}{H} \tag{21}$$

$$h = H(1 - \frac{\Delta Cam_N}{\Delta Cam'_N}) \tag{22}$$

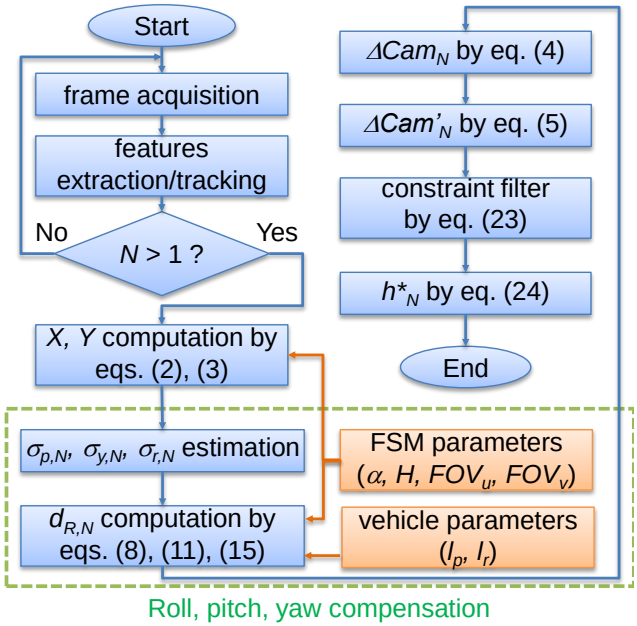


Figure 5: Flowchart of the proposed height estimation method and compensations.

3.2 Compensation of Roll, Pitch and Yaw Variations

The equations displayed in Table 1 show that the compensation of the influences caused by roll, pitch and yaw are straightforward and can be easily done if $\sigma_{p,N}$, $\sigma_{y,N}$ are $\sigma_{r,N}$ are known. Thus, for each acquired frame by the camera, we compute these variations and substitute them with the constant and known parameters of the vehicle (l_p, l_r) and the FSM (α, H, FOV_u, FOV_v) in (8), (11) and (15), obtaining the compensated value $d_{R,N}$. The main steps of the proposed compensations are illustrated in the bottom part of Figure 5.

3.3 Proposed Algorithm

Figure 5 summarizes the flow of the proposed algorithm. Frames are acquired, features are extracted and tracked in two frames. Next, the corresponding values of X and Y are computed by the FSM. Roll, pitch and yaw variations ($\sigma_{p,N}$, $\sigma_{y,N}$, $\sigma_{r,N}$) are estimated by VO and we compensate their influences using (8), (11) and (15), as explained in the previous section. The camera displacement (ΔCam_N) and compensated apparent displacement of the object ($\Delta Cam'_N$) are calculated by the FSM. Here, we propose to apply two filters. First, a filter based on displacement constraints to verify the applied compensations. Even though

the compensations strongly rely on the estimated roll, pitch and yaw, wrong estimations on the contrary lead to high errors. The filter works according to the observed in Figure 3. If a tracked point belongs to an object above the surface, then its apparent displacement must be bigger than the camera displacement ($\Delta Cam_N < \Delta Cam'_N$). Therefore, we check this condition by calculating the apparent displacement of the object with ($\Delta Cam'^{c}_N$) and without ($\Delta Cam'^{nc}_N$) compensations. If one of them satisfies the condition, then we adopt this displacement as $\Delta Cam'_N$. If both or none of them satisfies the condition, then we use the average of the two displacements to estimate the height, as show in (23). Finally, the median (h^*_N) of the previous estimations ($h_1, h_2, ..., h_{N-1}, h_N$) is also applied to filter eventual noises, as (24).

$$\Delta Cam'_N = \begin{cases} \Delta Cam'^{c}_N & if\ \Delta Cam'^{c}_N > \Delta Cam_N > \Delta Cam'^{nc}_N \\ \Delta Cam'^{nc}_N & if\ \Delta Cam'^{c}_N < \Delta Cam_N < \Delta Cam'^{nc}_N \\ \frac{\Delta Cam'^{nc}_N + \Delta Cam'^{c}_N}{2} & \text{otherwise} \end{cases} \quad (23)$$

$$h^*_N = median(h_1, h_2, ..., h_{N-1}, h_N) \quad (24)$$

4 Experiments

The proposed method was evaluated with data from the KITTI Vision Benchmark Suite (called hereon as "KITTI") [15] and processed with a computer Intel(R) Core(TM) i7-4600, 2.10 GHz, operating system Ubuntu (TM) 14.04, Eclipse (TM) development environment and OpenCV libraries [16]. Here, we want to verify the validity and effectiveness of the proposed height estimation and the proposed compensations of errors caused by roll, pitch and yaw variations. Thus, all estimated heights were done with two methods for later comparison: i) with roll, pitch and yaw compensation and ii) without compensation.

4.1 Applied VO

In order to estimate the camera displacement ΔCam_N in each frame, we adopted a simple VO with rotation estimation by Nister's 5-point algorithm [17] and translation by the FSM using features in a ground region close to the camera (details in the Appendix section), similarly to [12]. The parameters (FOV_u, FOV_v, W, V, H, etc) necessary for the experiments were adopted according to the provided by the KITTI. The camera inclination in relation to the ground was not directly provided, but we estimated that $\alpha = 1.1^o$ using the provided velodyne data. The feature extractor applied was FAST [18]. The features were automatically extracted when their number was bellow 1500 features. In order to evaluate the proposed monocular height estimation method, only the left images of the grayscale camera of the KITTI were used.

4.2 Evaluation Criteria

The evaluation was based on the error (ε_N) between the ground truth ($h_{GT,N}$) and the estimated height (h^*_N) in each frame N, according to (25). The ground truth adopted was mainly the height provided by the velodyne, available in the KITTI. Further details can be found in the Appendix section.

$$\varepsilon_N = |h_{GT,N} - h^*_N| \quad (25)$$

We also adopted a criterion to consider if the height estimation in a frame failed or not. For example, since we considered only objects above the ground surface, the estimated heights must be higher than 0 m (i.e., $h_{min} = 0$). Furthermore, since the FSM considers objects below the horizon line, all objects used in the experiment should have maximum height equal to the camera height ($h_{max} = H$). All estimated heights outside this maximum and minimum were considered as failure, as detailed in (26).

$$\text{height estimation} = \begin{cases} \text{success} & h_{min} \leq h^*_N \leq h_{max} \\ \text{failure} & \text{otherwise} \end{cases} \quad (26)$$

4.3 Results

Experiments were conducted with 10 video sequences of the KITTI, which contained many static objects (parked cars, poles, fences, houses, people, boxes, etc) in the scene. In total, height was estimated 4561 times with 436 objects. Objects with height $0 \leq h \leq H$ and distance from 4 to 31 m from the camera were used. The obtained results are summarized in Table 2, Table 3 and Figure 6.

Data	Video sequence (KITTI)	Valid frames	Valid objects
1	00	1103	85
2	05	384	35
3	07	229	18
4	08	373	29
5	09	228	23
6	13	347	43
7	15	471	32
8	16	361	35
9	18	407	32
10	19	1094	104
total	-	**4997**	**436**

Table 2: Summary of the conducted experiments.

	without comp.	**with comp.**
comp. parameter	-	pitch, yaw, roll
average ε_N [m]	0.23	0.20
maximum ε_N [m]	1.51	1.12
number of failures	283	65

Table 3: Comparison of the obtained errors with and without the proposed compensations.

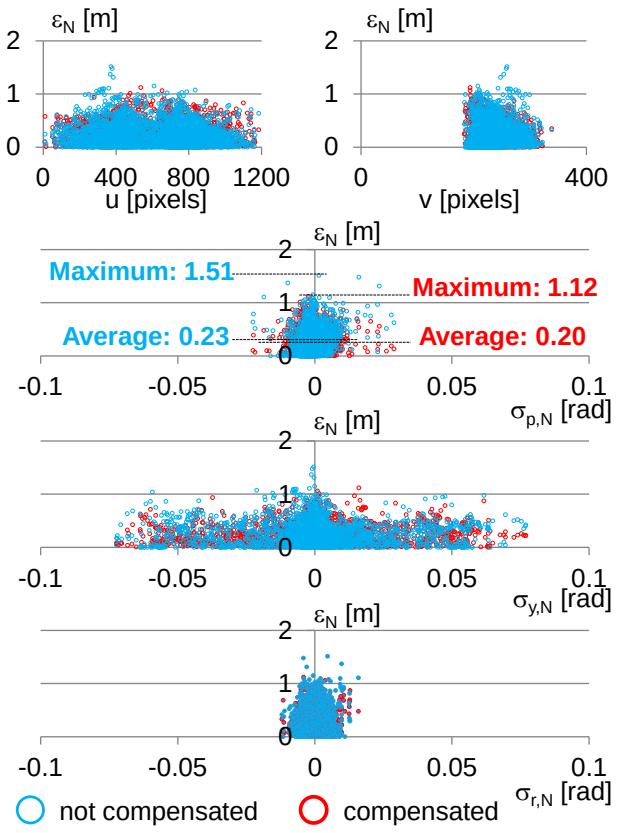


Figure 6: Obtained errors of the estimated heights of all data in the experiments. The results are displayed according to pixels (u,v) and unevennesses ($\sigma_{p,N}$, $\sigma_{y,N}$, $\sigma_{r,N}$).

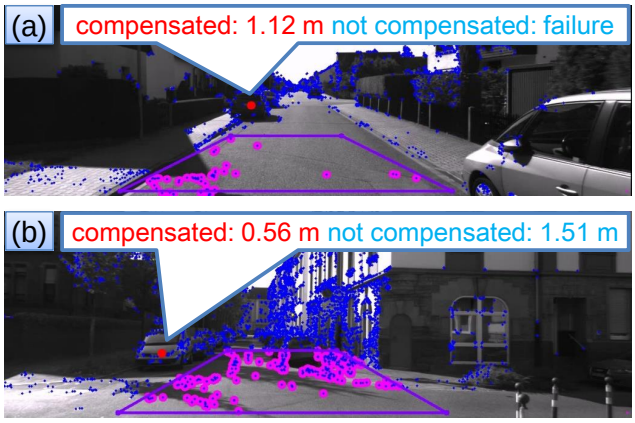


Figure 7: Maximum error cases with and without compensation.

5 Discussions

Table 2 shows the 10 data and corresponding video sequence belonging to the KITTI. The number of valid frames and objects used in each data are also displayed. The estimated heights of all objects chosen within this data are presented in relation to the pixel positions (u,v) in Figure 6. We can observe that the objects were well distributed along pixel u direction, covering many possible positions during the experiments. On the other hand, due to the geometrical limitations of the FSM only pixels below the horizon line (i. e., $v > 200$ pixels) were used, but we can also observe objects distributed along this interval. Next, the used data is displayed in relation to $\sigma_{p,N}$, $\sigma_{y,N}$ and $\sigma_{r,N}$. Variations of roll ($\sigma_{r,N}$) and pitch ($\sigma_{p,N}$) in the used data were smaller than those of yaw ($\sigma_{y,N}$): while the magnitude of roll and pitch variations were within 0.03 rad, the magnitude of the yaw variations were over 0.05 rad. The average error of all used data resulted in 0.20 m with the 3DOF compensations and in 0.23 m when no compensations were done. The maximum error resulted in 1.12 m for the compensated case an 1.51 m for the non-compensated one. Since both average and maximum errors were improved with the compensations, we can affirm that the proposed method is valid and effective. These cases of maximum errors are illustrated in Figure 7. In (a), the case of maximum error with compensations is displayed. The point was chosen too close to the horizon line, becoming very sensitive to noises and causing the high error. However, for the same case, the method without compensation failed to estimate the height. In (b), the case of maximum error without compensation occurred when the camera estimated the object height while climbing a slope. As presented in the previous sections, pitch variations cause higher errors comparing to yaw and roll, and a high error of 1.51 m was expected. Nevertheless, when the proposed compensations were applied in the same data, the estimation error dropped to 0.56 m. Furthermore, the distribution of $\sigma_{p,N}$, $\sigma_{y,N}$ and $\sigma_{r,N}$ in the experiments (Figure 8) shows that even though the data was taken on public roads, most of the pose variations were below 0.01 rad: 98.8% of the $\sigma_{p,N}$, 76.6% of the $\sigma_{y,N}$ and 99.7% of the $\sigma_{r,N}$. We can observe that the errors were higher for higher values of pitch variation $\sigma_{p,N}$ when no compensation was done, as expected by Figure 4. Even though in average the compensated and non-compensated methods had similar errors (0.20 and 0.23 m), the difference of both errors was higher for higher variations of $\sigma_{p,N}$, $\sigma_{y,N}$ and $\sigma_{r,N}$. During the experiments objects with distances further than 20 m were used, what was enough to cause more than 5 m error according to Figure 4. Since the obtained errors were below this expected ones, we can affirm that the obtained results were satisfactory. Examples of cases with higher yaw variations are shown in the Appendix section. We can observe a significant difference of the proposed method in terms of number of successful height estimations. During the experiments, the compensated method failed to estimate height 65 times, while the non-compensated one failed 283 times (4 times more). Such failures require further analysis in future work, but the effectiveness of the proposed compensations became clear.

Although this work relied on VO, it didn't focus on improving its accuracy. However, we estimated that the applied VO had around 13% error per frame and influenced directly the results, what means that the proposed method becomes more accurate with the improvement of VO itself. Such VO errors led to wrong

estimation of camera pose, generating higher height estimation errors and examples are shown in the Appendix section. Even though the proposed method improved the average and maximum error of the estimated heights, some limitations still exist. The small camera pose variation per frame suggests that further evaluation with higher variations is necessary, by for example, increasing the moving body's velocity and analyzing the relation between camera frame rate and obtained height estimation errors. Finally, the experiments made evident another necessary correction: we considered angular variations (rotation) during the proposed compensations, however translations in Y also occurred. According to the FSM, such translations change the camera height H and must be considered when computing the distance to objects. We believe that this consideration can further increase the accuracy of our height estimation method.

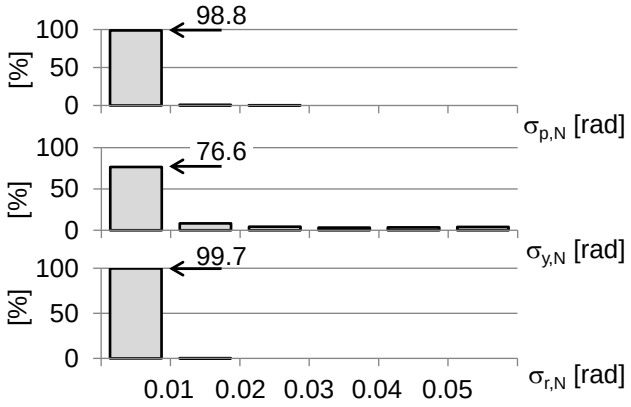


Figure 8: Distribution of the pose variation detected in the used data.

6 Conclusions and Future Work

A novel method of monocular height estimation with 3DOF compensation of roll, pitch and yaw variations was proposed. The method can estimate height of objects with only two frames of a monocular camera. Experiments outdoors with the KITTI benchmark data (4997 frames and 436 objects) resulted in improved accuracy of the estimated heights from a maximum error of 1.51 m to 1.12 m and reduced number of estimation failures by 4 times, proving the validity and effectiveness of the proposed method. This method can be enhanced by improving monocular visual odometry techniques and considering translational variations of the camera during height estimation. Further investigation about influences of frame rate, moving velocity and robustness are planned in the future.

Conflict of Interest The authors declare no conflict of interest.

References

[1] A. M. Kaneko, K. Yamamoto. Monocular Height Estimation by Chronological Correction of Road Unevenness. 2016 IEEE/SICE International Symposium on System Integration (SII2016). Sapporo, Japan. 2016. https://doi.org/10.1109/SII.2016.7843971

[2] A. M. Kaneko, K. Yamamoto. Landmark Recognition Based on Image Characterization by Segmentation Points for Autonomous Driving. SICE Multi-Symposium on Control Systems 2016 (MSCS2016). Nagoya, Japan. 7-10 March, 2016. https://doi.org/10.1109/SICEISCS.2016.7470160

[3] M. Aly. Real Time Detection of Lane Markers in Urban Streets. IEEE Intelligent Vehicles Symposium. Eindhoven. 4-6 June, 2008. https://doi.org/10.1109/IVS.2008.4621152

[4] T. Wu, A. Ranganathan. A Practical System for Road Marking Detection and Recognition. IEEE Intelligent Vehicles Symposium (IV). Alcala de Henares. 3-7 June, 2012. https://doi.org/10.1109/IVS.2012.6232144

[5] P. Foucher, Y. Sebsadji and J. P. Tarrel, P. Nicolle. Detection and Recognition of Urban Road Markings Using Images. IEEE International Conference on Intelligent Transportation Systems (ITSC). Washington DC, USA. 2011. https://doi.org/10.1109/ITSC.2011.6082840

[6] D. Scaramuzza. 1-Point-RANSAC Structure from Motion for Vehicle-Mounted Cameras by Exploiting Non-holonomic Constraints. International Journal of Computer Vision, Volume 95, issue 1, 2011. https://doi.org/10.1007/s11263-011-0441-3

[7] M. W. Maimone, Y. Cheng, L. Matthies. Two years of Visual Odometry on the Mars Exploration Rovers. J. Field Robotics 24(3): 169-186 (2007). https://doi.org/10.1002/rob.v24:3

[8] M. K. Momeni C. H. S. Diamantas, F. Ruggiero, and B. Siciliano. Height Estimation from a Single Camera View. In Proceedings of the International Conference on Computer Vision Theory and Applications (VISIGRAPP 2012)ISBN 978-989-8565-03-7, pages 358-364. DOI: 10.5220/0003866203580364

[9] P. Viswanath, I.A. Kakadiaris and S. K. Shah. A Simplified Error Model for Height Estimation using Single Camera. 2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops. 2009. https://doi.org/10.1109/ICCVW.2009.5457466

[10] N. Pears, B. Liang, Z. Chen. Mobile Robot Visual Navigation Using Multiple Features. EURASIP J. Adv. Sig. Proc. 2005(14): 2250-2259. 2005. https://doi.org/10.1155/ASP.2005.2250

[11] A. Wedel, T. Schoenemann, T. Brox, D. Cremers. WarpCut - Fast Object Segmentation in Monocular Video. 29th DAGM Symposium, Heidelberg, Germany. Proceedings, pp. 264-273. 2007. http://doi.org/10.1007/978-3-540-74936-3_27

[12] J. Campbell, R. Sukthankar et al. A Robust Odometry and Precipice Detection System Using Consumer-grade Monocular Vision. Proceedings of the 2005 IEEE International Conference on Robotics and Automation, 2005. ICRA 2005. Pp 3421-3427. 2005. https://doi.org/10.1109/ROBOT.2005.1570639

[13] S. Lovegrove, A. J. Davison, J. I. Guzman. Accurate Visual Odometry from a Rear Parking Camera. IEEE Intelligent Vehicles Symposium (IV). Pp 788-793. 2011. https://doi.org/10.1109/IVS.2011.5940546

[14] N. Nourani-Vatani, et al. Practical Visual Odometry for Car-like Vehicles. IEEE International Conference on Robotics and Automation (ICRA). Kobe, Japan. 2009. https://doi.org/10.1109/ROBOT.2009.5152403

[15] A. Geiger, P. Lenz and R. Urtasun. Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite. Conference on Computer Vision and Pattern Recognition (CVPR). 2012. http://doi.org/10.1109/CVPR.2012.6248074

[16] Bradski, G. Opencv. Dr. Dobbs Journal of Software Tools. 2000.

[17] D.Nister. An Efficient Solution to the Five-Point Relative Pose Problem. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 26, No. 6. 2004. https://doi.org/10.1109/TPAMI.2004.17

[18] E. Rosten, R. Porter and T. Drummond. FASTER and Better: A Machine Learning Approach to Corner Detection. IEEE Trans. Pattern Analysis and Machine Intelligence. 2010. https://doi.org/10.1109/TPAMI.2008.275

[19] H. Hirschmuller. Accurate and Efficient Stereo Processing by Semi-Global Matching and Mutual Information. International Conference on Computer Vision and Pattern Recognition, 2005. https://doi.org/10.1109/CVPR.2005.56

Appendices

In this section, we provide further details of the experimental conditions and examples of compensations.

A Experimental Conditions

We further explain the experimental conditions adopted in the paper. The experiments were conducted according to Figure 9. The algorithm executed VO automatically for each frame (a), which were shifted manually by a human operator. Among all the extracted feature points, the operator chose with the mouse any point on a desired object (b). Here, since the FSM has geometrical restrictions, only points below the horizon (approximately in $v = \frac{V}{2}$) were chosen. The chosen point (in red) was tracked over the frames until the last frame possible (c), and its height and ground truth were computed and stored in each frame. The estimated position, stored data and extracted points of the current object were reseted after the heights were estimated (d) and the process repeated for all used data.

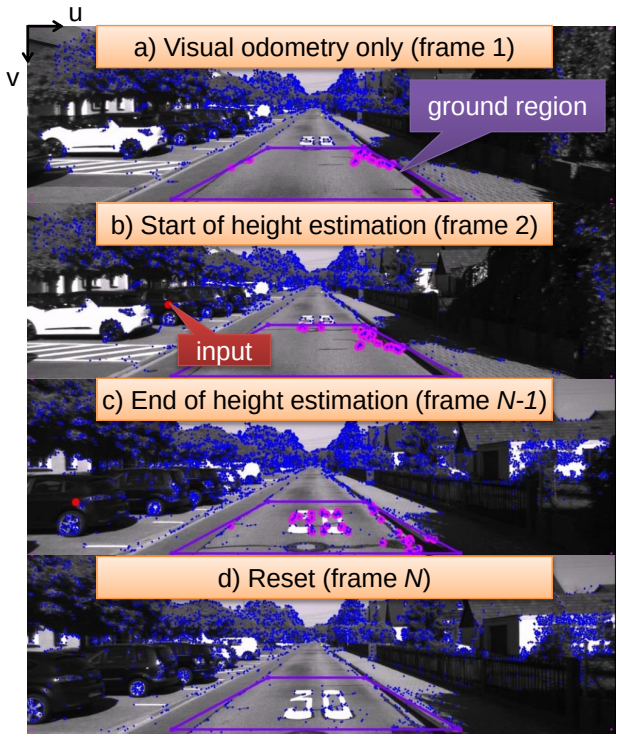


Figure 9: Adopted sequence to estimate the height of an object during the experiments.

First, all estimated heights and relevant data were saved and later verified for validity. Frames whose ground region contained few or no features (due to light conditions for example) or had a moving object were considered invalid (Figure 10). Tracking was also verified (Figure 11). Although we needed only two frames per estimation, we focused on features consistent in many frames (a minimum of 3 estimations per object). Thus, in case the tracking failed before the fourth frame, that object and estimations were considered invalid. If the tracking failed after the fourth frame, only the estimations before that were considered valid.

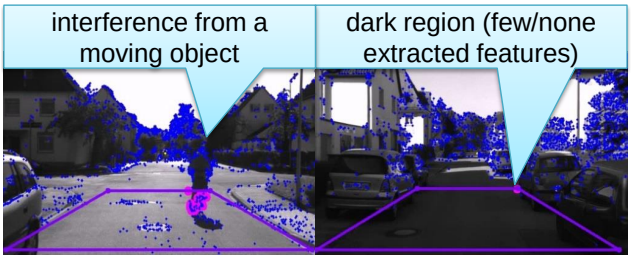


Figure 10: Examples of frames considered invalid: dark regions (right) and moving object (left).

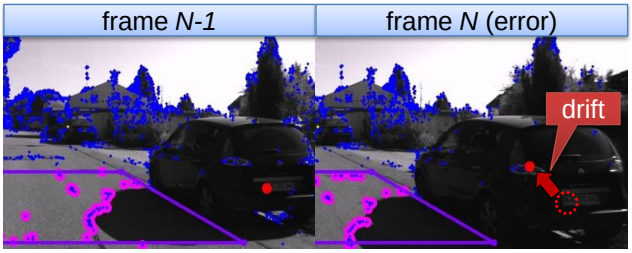


Figure 11: Example of tracked feature considered invalid.

We adopted as ground truth the height from the velodyne. The necessary information for converting the camera-velodyne coordinates were also provided by the KITTI. Due to limitations of the sensor resolution, we chose as ground truth the closest pixel with available velodyne data. This search was also automatically done by the algorithm in each frame. However, since the velodyne fails in some situations, we also computed the height by the disparity from the left and right images of the camera, using Semi Global Block Matching [19] in OpenCV. For each point chosen by the operator, its corresponding velodyne data was recorded. Since the camera was moving forward and approaching the object, we considered that the depth from the velodyne to the object on the next frame should become smaller than the one in the previous frame. In case it wasn't, we considered that the height estimated by the velodyne was also not consistent, and the one calculated by the stereo camera was used instead (Figure 12). In case both depths from the sensors were not smaller than the previous one, then the current estimation and frame were considered invalid. Since we considered the initial frame of each object as reference, variations in Y of the camera itself were also compensated when computing the ground

truth.

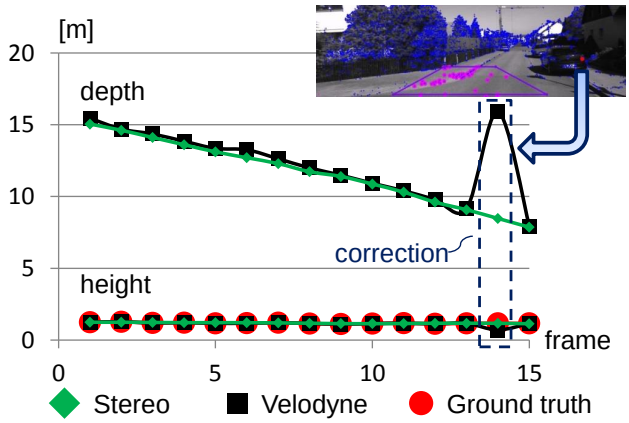


Figure 12: Example of adopted ground truth and correction. Data from the stereo camera was used when the velodyne data was considered inconsistent.

B Further Examples of Compensations

Even though there were few occurrences of yaw variations higher than 0.01 rad per frame during the experiments, we can further observe the benefits of the compensations in the example in Figure 13, which exemplifies the estimated heights during a curve. In the first frames, the camera was mostly moving forward so that ε_N was small for both compensated and not compensated cases. When the yaw variation started to increase the errors also increased. However, in frame 10 we can see ε_N decreasing to a minimum due to the decrease of $\phi_{y,N}$, increasing again when $\phi_{y,N}$ started increasing, as foreseen by Figure 4.

Figure 14 presents other examples of compensation. The cases in Figure 14 (iii) to (vi) had smaller average errors per frame after the compensations. On the other hand, even though the proposed method improved the average and maximum error of the estimated heights, some limitations as in Figure 14 (i) and (ii) still exist. We believe that such errors were caused by the errors of the adopted VO itself (13 % error) and improvements will be further investigated in future work.

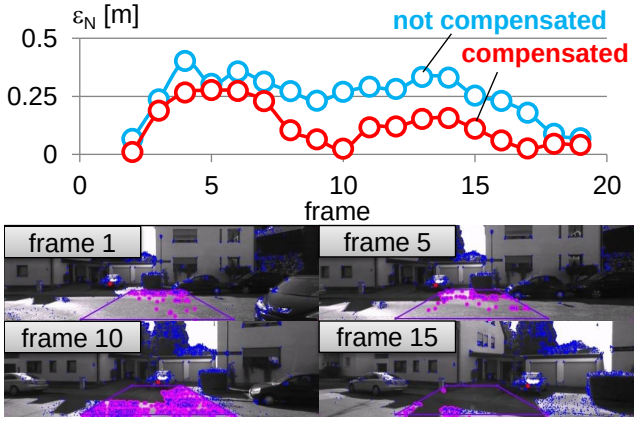


Figure 13: Example of yaw compensation.

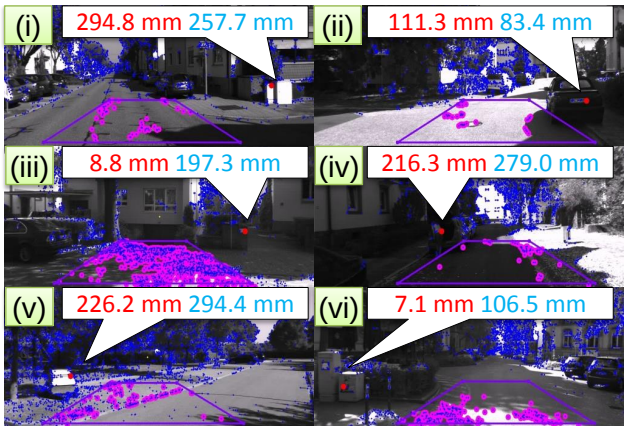


Figure 14: Examples of objects and average ε_N per frame. The numbers in blue (right) are the non-compensated cases and the numbers in red (left) are the compensated ones.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1453-1460 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Scalability Dilemma and Statistic Multiplexed Computing -- A Theory and Experiment

Justin Yuan Shi*, Yasin Celik

Department of Computer and Information Sciences, Temple University, Pennsylvania 19122, USA

ARTICLE INFO

Article history:
Received: 21 May, 2017
Accepted: 20 July, 2017
Online: 10 August, 2017

Keywords:
Extreme Scale Computing
Statistic Multiplexed Computing

ABSTRACT

The For the last three decades, end-to-end computing paradigms, such as MPI (Message Passing Interface), RPC (Remote Procedure Call) and RMI (Remote Method Invocation), have been the de facto paradigms for distributed and parallel programming. Despite of the successes, applications built using these paradigms suffer due to the proportionality factor of crash in the application with its size. Checkpoint/restore and backup/recovery are the only means to save otherwise lost critical information. The scalability dilemma is such a practical challenge that the probability of the data losses increases as the application scales in size. The theoretical significance of this practical challenge is that it undermines the fundamental structure of the scientific discovery process and mission critical services in production today.

In 1997, the direct use of end-to-end reference model in distributed programming was recognized as a fallacy. The scalability dilemma was predicted. However, this voice was overrun by the passage of time. Today, the rapidly growing digitized data demands solving the increasingly critical scalability challenges. Computing architecture scalability, although loosely defined, is now the front and center of large-scale computing efforts.

Constrained only by the economic law of diminishing returns, this paper proposes a narrow definition of a Scalable Computing Service (SCS). Three scalability tests are also proposed in order to distinguish service architecture flaws from poor application programming. Scalable data intensive service requires additional treatments. Thus, the data storage is assumed reliable in this paper. A single-sided Statistic Multiplexed Computing (SMC) paradigm is proposed. A UVR (Unidirectional Virtual Ring) SMC architecture is examined under SCS tests. SMC was designed to circumvent the well-known impossibility of end-to-end paradigms. It relies on the proven statistic multiplexing principle to deliver reliable service using faulty components as the infrastructure expands or contracts.

To demonstrate the feasibility of such a theoretical SCS, an organized suite of experiments were conducted comparing two SMC prototypes against MPI (Message Passing Interface) using a naive dense matrix multiplication application. Consistently better SMC performance results are reported.

1. Introduction

For the last three decades, end-to-end computing has been the de facto paradigm for distributed and parallel programming. MPI (Message Passing Interface), RPC (Remote Procedure Call), OpenMP (share memory) and RMI (Remote Method Invocation) are all end-to-end programming paradigms. A myth persisted since the 1990's that while the data communication community is well served by the end-to-end paradigm, in 1997, direct use of the end-to-end protocol in distributed computing was cited as a fallacy [1]. The computing service scalability dilemma was predicted.

A reexamination of the 1993 proof of the impossibility of implementing reliable communication in the face of crashes [2] exposed the root cause of the alleged fallacy: the robust communication protocols are ineffective when either the sender or the receiver in the end-to-end communication could crash arbitrarily. No error detection and recovery methods can reverse this impossibility.

Direct use of end-to-end reference model in computing applications leaves massively many potential end-point crashes when the application runs (Figure 1). Although the probability of each end-point crash is very small, the growth in application size (thus the processing infrastructure) is guaranteed to increase the planned and unplanned service down times and data loss probabilities. The reproducibility of a large scale application

*Corresponding Author: Justin Yuan Shi, Science Education Research Center 310, Temple University, +1 205-204-6437, USA | Email: shi@temple.edu

(service) becomes increasingly more difficult [3]. These risks are tolerable when the computing clusters are small. They have become increasingly unbearable as the computing services facing increasingly expanding loads.

Incidentally, reliable communication is possible using faulty components. This was also proved in early 1980's [4]. There are three assumptions in the proof: a) all data packets must decoupled from transmission devices, b) both the sender and the receiver must be reliable, and c) there are infinite supplies of resources. In literature, this is often called the "best effort" service. Theoretically, the best effort communication guarantees 100% reliability as long as the infrastructure affords the minimal survivable resource set at the time of needs. This was the theoretical basis of the OSI (Open System Interconnection) reference model [5].

It then follows that the "best effort" computing should also be possible by removing the reliable receiver assumption in the basic program-program communication protocols. In other words, while it is reasonable for a data communication application to assume reliable sender and receiver at the time of need, it is not reasonable to assume reliable receivers in any clustered computing services, since the growing service infrastructure will make it increasingly unlikely.

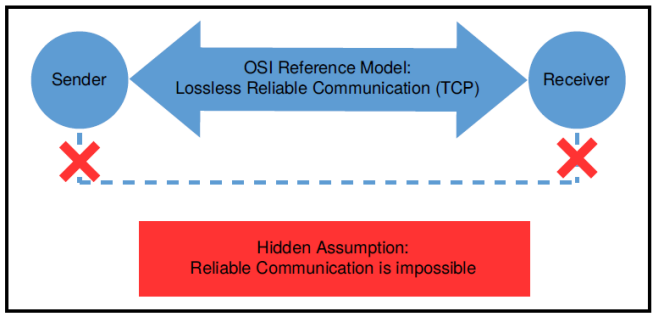


Figure 1. OSI Standards and Its Hidden Assumption

This has led to the design of a single-sided parallel computing paradigm. The new paradigm must multiplex all resources or the application fault tolerance is not attainable (Figure 2). This framework is called Statistic Multiplexed Computing (SMC) or Stateless Parallel Processing. Today, as the single processor speed approaches to a plateau, it seems that unconstrained clustered computing is the only likely future computing architecture. No single big-CPU machine, even if quantum class machines, would be able to meet the exponentially growing data processing needs.

This paper is organized as follows. Section 2 introduces the assumptions and a narrow definition of a Scalable Computing Service (SCS). It also includes three "Reproducible Architecture Tests" as the necessary and sufficient conditions for SCS. In this paper, the data storage is considered stable and lossless. Data intensive SCS will need additional requirements. A single-sided Statistic Multiplexed Computing (SMC) architecture is presented. Section 3 provides a high level examination of the single-sided SMC or Stateless Parallel Processing (SPP) paradigm using the proposed scalability tests. Section 4 illustrates the need for granularity tuning for optimal parallel processing. Section 5 evaluates MPI, Hadoop, Spark and two SMC prototypes: Synergy (SPP) and AnkaCom (SMC) using the scalability tests. Section 6 documents the experiment design and objectives. Section 7 reports the computational results. Section 8 contains the summary.

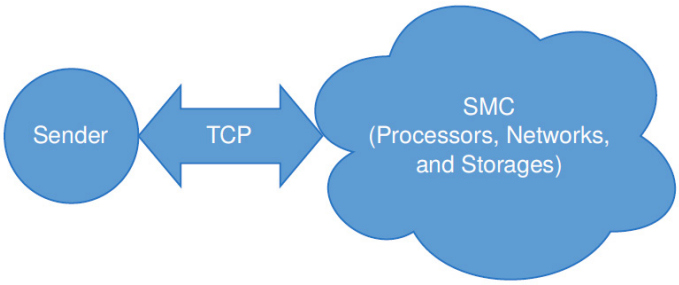


Figure 2. Single-Sided SMC Paradigm

2. Assumption, Definition and Tests

Scalability is the capability of a system to handle a growing amount of work. It is well understood in an economic context, where a company's scalability undermines the potentials economic growth of the company. Since the future societal economic growth depends on all available computing services, the importance of the computing service scalability is self-evident.

The economic law of diminishing returns states that in all productive processes, adding more of one factor of production, while holding all others constant, will at some point yield lower incremental per-unit returns. Parallel computing using multiple processors and networks obey the same law.

There are many possible dimensions in scalability measures. For compute intensive services, assuming infinite supplies of processors and networks, the scalable computing research challenge is to devise a computing service architecture that can scale indefinitely in order to meet the growing demands. The Internet and SMP (Symmetric Multiprocessing) systems are such service architectures. Delivering robust service for growing demands, however, has met the seemingly insurmountable challenge of the scalability dilemma. The infinite resource assumption was rooted in practice. This assumption also makes the theoretical discussions possible.

The dependency on reliable processors and networks for delivering reliable service is the root cause of the scalability dilemma. Since electronic components can suffer random failures, a robust computing architecture must be able to exploit all useable components at the time of need. Thus the *infinite resource assumption* can be translated into the *minimal survivable resource set* assumption in practice. That is, all discussions on scalable performance, reliability and service quality are based on the assumption that the processing architecture affords the ``minimal survivable resource set'' at the time of need. The architecture design challenge is to build a mechanical structure capable of leveraging all available resources at the time of need without reprogramming the application.

In this paper, we also assume that the storage is reliable. Data intensive SCS requires additional treatments following the same principles [6].

Definition: A Scalable Computing Service (SCS) is a computing service that allows unlimited infrastructure expansion without reliability and service quality degradation. Application performance should subject to the same definition until the law of diminishing returns applies.

Parallel applications are the integral parts of the scientific discovery processes. It is thus important to ensure that extreme

scale service architecture does not hinder the scientific discovery process. The service reproducibility requirement is implied. Further, it would be difficult to distinguish a computing architecture design flaw from a poorly composed non-scalable application (crossing the point of diminishing returns prematurely), the following reproducible SCS tests are proposed:

a. **Share-nothing Test**. No processors or networks can be assumed reliable while still delivering full range of services. Not passing this test breaks the scalable reliability definition.

b. **Sublinear Cost Test**. The runtime management overhead should be bound within a sublinear factor with regard to the infrastructure size. Failing this test means diminishing benefits when the application up-scales. It breaks the scalable performance definition.

c. **Reproducible Node Test.** Each node in the architecture must be able to reproduce identical semantically acceptable results given identical inputs. These include both deterministic and non-deterministic programs. Failing this test will break the service quality definition.

Test (a) requires a single-sided programming paradigm (such as Figure 2) and a multiplexing runtime architecture. Tests (a) and (b) are the necessary and sufficient conditions for SCS. Tests (a), (b) and (c) are the necessary and sufficient conditions for a reproducible SCS, since the infrastructure ensures reliable services in any scale using faulty components.

All modern computers running reasonable software can pass the *Reproducible Node Test*. The *Share-nothing* and *Sublinear Cost Tests* are more difficult.

Services built using the end-to-end reference model fail the *Share-nothing Test*. The end-to-end reference model requires explicit receiver addresses in an IP-addressable network. This seemingly innocent requirement causes the entire application to depend on the reliability of all processors and networks. This is the root cause of the scalability dilemma [1].

The following computing systems can pass the *Share-nothing Test*: a) Dataflow machines [7], b) Tuple Space machines [8], and c) services by Content Addressable Networks [9]. Similarly, the search engines and SMP (Symmetric Multiprocessing) systems also qualify. These systems have two common features: a) API (Application Programming Interface) has no fixed destination (single-sided), and b) every task can be processed by any processing elements.

Dataflow and Tuple Space machines are elegant automated parallel systems but suffer the poor performance stigma. Content Addressable Networks are designed for the next-generation Internet services. It is not typically used for parallel computing. Search engines and SMP systems are designed to service multiple clients effectively. They are ineffective for solving the single application's scalability dilemma.

However, putting these systems together brought a unique opportunity: Stateless Parallel Processing (SPP) or Statistic Multiplexed Computing (SMC) [10]-[15]. Specifically, it is possible to build a "Service Content Addressable Network" using the Tuple Space semantics and dataflow principle thus making the "best effort computing" practical.

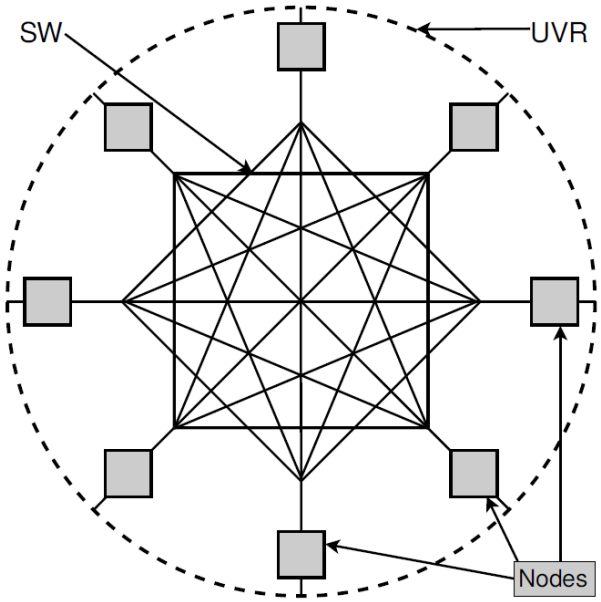


Figure 3. Statistic Multiplexed Computing Architecture

Figure 3 illustrates such a multiplexing computing service under a UVR (Unidirectional Virtual Ring) architecture.

3. Single-sided Statistic Multiplexing

Passing the *Share-nothing Test* for a computing architecture requires the absence of all component reliability assumptions (except for the storage for this paper). The service communication architecture must be capable of exploiting all possible network topologies. In Figure 3, SW is a collection of network switches in any topology. Each node is a standalone computer with any number of cores, local memory, storage and multiple network interfaces. UVR (Unidirectional Virtual Ring) implements a Service Content Addressable Network (or Tuple Switching Network (TSN) [16]) using all available processing and networking components. This allows zero single point failures regardless the number of networks and processors for a given application.

The UVR architecture ensures that except for the Master (the client of the parallel computing service), there is no need for explicit IP addresses for data exchanges.

Unlike traditional parallel Masters, the SMC Master is part of UVR architecture in that it is responsible for tuple retransmission discipline and redundancy handling. Actual data exchanges are implemented using all available network links directly. Services built using the TSN do not subject to the end-to-end impossibility. The high level UVR computing concept passes the *Share-nothing Test*. The "best effort computing" idea could become feasible, if the implementation also passes the same test.

Once the application runs, the dataflow semantics allow the computing infrastructure to automatically form SIMD (Single Instruction Multiple Data), MIMD (Multiple Instruction Multiple Data) and pipeline clusters at runtime. These effects are identical to early dataflow machines [7].

However, passing the *Sublinear Cost Test* needs more work. On the surface, it seems impossible to traverse $\boldsymbol{P}$ nodes paying less than linear traversal cost. However, $\boldsymbol{k}$-order multicast can cut the

UVR traversal complexity to $O(lg_kP)$ [17]. Further, once a tuple is matched against its processor, actual data exchange will be done directly. Imposing an order on the tuple values can further force the tuple matching overheads to $O(1)$ using the ideas from Consistent Hashing [18]. For practical applications, the one-time $O(lg_kP)$ traversal cost seems reasonable for deploying millions of nodes.

4. Single-sided Statistic Multiplexing

All compute intensive applications are typically partitioned into parallel tasks that can be processed using SIMD, MIMD and pipeline parallelisms. Since data exchange costs time, the partition size or *processing granularity* determines the ultimately deliverable performance by the service architecture. The best performance is delivered when all tasks terminate at the exactly the same time. A small granularity allows for higher concurrency but at the expense of higher data exchange costs. A bigger granularity risks lower concurrency and longer synchronization times due to differences in processing capabilities. The optimal granularity delivers the best possible parallel performance for the application running on a given processing environment.

The optimal *processing granularity* (G) of an application defines the Termination Time Equilibrium [19], [20] for that application. Finding the optimal G will allow the SMC applications to overcome the poor performance stigma of earlier dataflow and Tuple Space machines.

The principle of finding the optimal G is analogous to the use of Snell's Law for solving the Brachistochrone problem in physics and mathematics [21]. Figure 4 illustrates that the solution to the fastest descent under the influence of uniform gravitational field

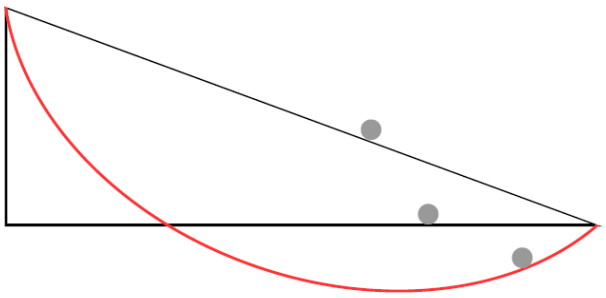


Figure 4. The Brachistochrone Problem.

and friction-less surfaces is not the straight line that connects the source and the destination, but the Brachistochrone curve that optimizes the gravitational and normal forces. It is a cycloid. Using fixed application partitioning is like the straight line (which happened to be the slowest descent). The astonishing feature of the Brachistochrone curve is that it is also called "Tautochrone" that regardless where you start on the curve, all will reach the destination at the same time. In parallel processing, the Brachistochrone curve represents the optimal G's where the overall computing time equals to the overall data exchange time. As will be shown, if we tune G carefully, there are indeed multiple optimal G's (modulated under a communication overhead curve) for every compute intensive kernel [19].

Further, the optimal G is only discoverable without reprogramming under the single-sided paradigms. For end-to-end computing programs, unless the task distribution is pooled [22] or using the single-sided "scatter_v" call in MPI, once compiled, changing processing granularity requires reprogramming.

Other single-sided systems include Hadoop [23] and Spark [24]. They have demonstrated substantially better reliability and performance than similar end-to-end distributed computing systems. But performance against bare metal HPC programs have not being rigorously investigated.

The remaining of this paper reports computation experiments comparing two SMC prototypes: Synergy and AnkaCom against MPI (Message Passing Interface). The Brachistochrone effects are also demonstrated.

5. Computing Service Architectures and Scalability Tests

The MPI parallel processing architecture is embedded in its application programs. The (node) operating systems are responsible for the basic data communication and task execution functions, the programmer has explicit controls of the parallel machines. The explicit end-point requirement makes it impossible to pass the *share-nothing test*. Even with a "task pool" implementation, unless the implementation is completely distributed, thus forming a "application content addressable network", passing the *share-nothing test* remains negative.

The Hadoop system uses the single-sided ***<key, value>*** API (Application Programming Interface) for reliable large scale distributed processing. It can pass the *share-nothing test* at the concept level. Unfortunately, it fails the same test at the implementation level due to the use of the RPC (Remote Procedure Call) protocol. This results in the "**single namenode architecture**". The "namenode" is the single-point failure of the entire system. However, even though Hadoop cannot pass the *Share-nothing Test*, its single-sided API allowed much better runtime fault tolerance than other systems. Many very large scale successful experiments are completed within the *100M* file design limitation.

The Spark system relies on the Hadoop File System (HDFS) but leverages high speed in-memory processing. It's scalability test is identical to Hadoop.

The Ethereum project [25] is a decentralized distributed computing platform that runs smart contracts: applications that run exactly as programmed without any possibility of downtime, censorship, fraud or third party interference. It can pass the *share-nothing test* easily. These applications run on a custom built blockchain, a shared global infrastructure that can move value around and represent the ownership of property. This enables developers to create markets, store registries of debts or promises, move funds in accordance with instructions given long in the past (like a will or a futures contract) and many other things that have not been invented yet, all without a middle man or counterparty risk. However, its Blockchain implementation cannot pass the *Sublinear Cost Test* due to linear overhead increases when the chain expands.

The Synergy system is also a distributed parallel computing system that uses named Tuple Space Servers to simulate the TSN. Although it can pass the *Share-noting Test* at the concept level, the actual implementation fails the test due to the use of the named Tuple Space Servers with fixed IP addresses [9].

The AnkaCom system [26] is a fully distributed peer-to-peer TSN implementation. It can pass the *share-nothing test* at the concept level and the implementation level. The fundamental

departure point of the AnkaCom implementation from others is the absence of the single-ACK/NAK assumption as compared to the RPC protocol. It can also pass the *Sublinear cost test* by implementing the one-time ***O(lg$_k$P)*** UVR traversal protocol.

6. Experiment Design

Matrix multiplication is frequently used in scientific simulations and engineering applications. Dense matrix multiplication has high computing and communication requirements that can saturate resources quickly. It is selected as the benchmark for this study. Matrix multiplication is a "regular" parallel applications that fixed application partition seems reasonable for using a dedicated cluster of bare metal processors. The MPI parallel matrix multiplication program is implemented in C. It is optimized for locality for the three nested *(i,j,k)* loops. The Synergy program is also implemented in C and similarly optimized. The AnkaCom program is implemented in Java. The loop order is also optimized.

The MPI and Synergy runtime systems are implemented in C. The AnkaCom runtime is implemented in Java.

The computing platforms include the NSF (National Science Foundation) Chameleon bare metal cluster [27] at TACC (Texas Advanced Computing Center) and the owlsnest.hpc.temple.edu traditional bare metal cluster [28] at Temple University. The benchmark application is a naive dense square matrix multiplication application. The SMC prototypes are Synergy 3.0+ and AnkaCom 1.0. OpenMPI versions 1.4.4 and 1.10.0 are used in the experiments.

- *Owlsnest cluster* has multiple IB (Infiniband) support. The Chameleon cluster only has single IB support at the time of experiments.}
- The *Chameleon* and *Owlenest* bare metal clusters have same number of cores (48) and sufficient memory for many-core experiments (256 GB for Chameleon) per node and (543GB for Owlsnest).

The Synergy implementation will only sustain to a small number of cores before saturating the single-threaded Tuple Space server. Thus, every test using large number of cores is also a sense of reliability test of the UVR concept.

The AnkaCom implementation will sustain to any number of nodes and cores due to the implementation of distributed TSN. The Java runtime overheads should be much higher than the C-MPI combination. For AnkaCom, the IB support comes from the built-in Java library. The C-MPI implementation has a custom IB driver.

There are three groups of experiments:

- *Single-Node Single-Core.* This group of tests examines the runtime's ability to leverage multiple hardware components in parallel by overlapping single-core computing, communication and disk activities.
- *Single-Node Multiple-Core.* This group of tests examines the parallel runtime system's ability to use up to 48 cores concurrently in addition to other communication and storage components.
- *Multiple-Node Multiple-Core.* This group of tests examines the parallel runtime system's ability to harness multiple multicore nodes effectively. In this study, we used twenty 12 core nodes on Owlsnest with dual IB support.

The single-sided API affords SMC applications a flexible parallel task pooling capability that is not supported by direct message passing systems such as MPI. The SMC workers can be programmed to compute for tasks of different sizes without re-programming. In contrast, the MPI applications rely on fixed ***N/P*** partitioning.

7. Computational Results

7.1. Single-Node Single-Core Results

. The Java compiler does not have an optimization option.

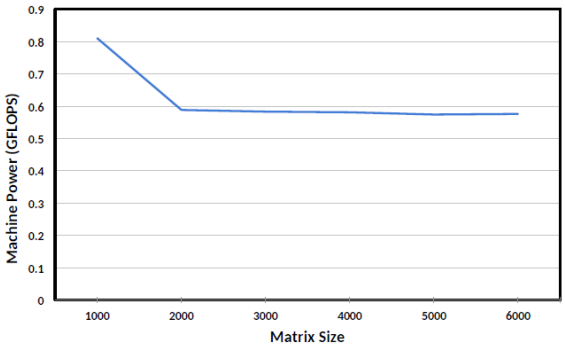


Figure 5. Single-Node Single-Core (Sequential)

The baseline experiment is provided by a sequential C-program running on a single core. Figure 6 shows the results in GLOPS on Owlenest running matrix of sizes 1000 - 6000.

The significance of the Figure 5 curve is that it depicts the typical "Cache, Memory, Swap and Die" (CMSD) single core performance behavior. Understanding this behavior enables further quantitative application scalability analysis [19].

A partitioned parallel program running on a single node with a single core will exhibit different behavior than a sequential program running in the same environment. The partitioned programs can exploit hidden concurrencies even in the single node single core environment. Figure 6 reports the single-core single worker tests for both MPI and Synergy against the sequential program (in yellow). Due to the restriction of end-to-end communication, the MPI (Open MPI 1.4.4) program was not able

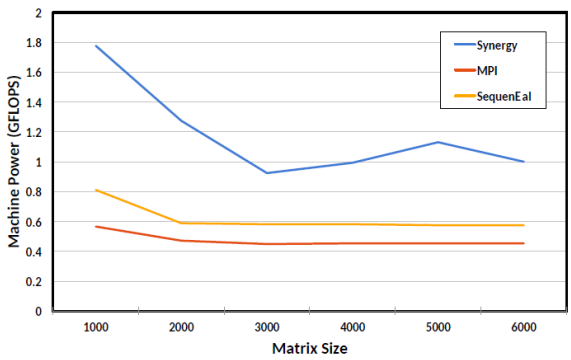


Figure 6. Single-Node Single-Core (Parallel)

to compete against the sequential code in different matrix sizes (1000-6000) consistently. The Chameleon results are similar.

7.2. Single-Node Multiple-Core Results

Single node multi-core is best suited for MPI and Synergy architectures since the inter-program data exchanges can be very fast. Figure 7 reports the granularity tuning results for MPI and Synergy on Chameleon for *N=9000*, *P=45* (factor of 9000). In Figure 7, the optimal granularity sizes are: 11, 22, and 33. The relationship between these sizes are given by the volatility power indices of the 45 cores at their peaks: 303, 409, and 818 [20]. At these optimal points, the synchronization overhead is near zero (note the Brachistochrone analogy). Further quantitative scalability analysis becomes possible using the optimized performance and simpler time complexity models [20]. The MPI (Open MPI 1.10.0) on Centos 7 delivered consistently worse performance than the worst tuned Synergy performance (5.1GFLOPS).

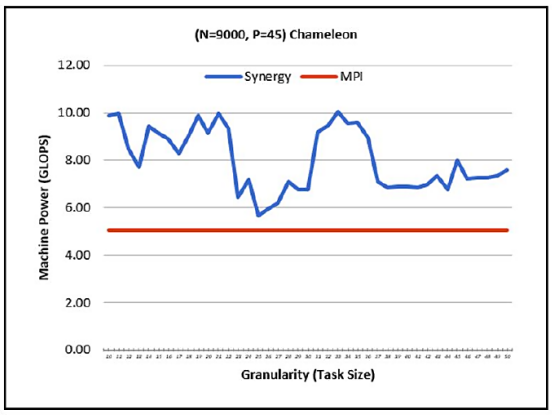


Figure 7. Single-Node Multiple-Core Granularity Tuned

Figure 8 reports a broader range of tests for different matrix sizes without granularity tuning (*G=20* fixed). Synergy still outperforms MPI when the matrix size exceeds 3,200. These results are consistent with Figure 6: the bigger the problem sizes, the better the Synergy performance.

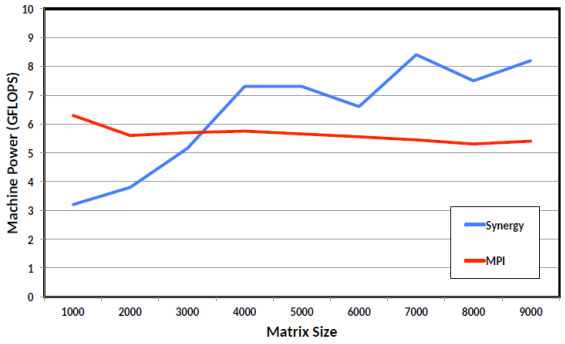


Figure 8. Single-Node Multiple-Core (Broader Range Tests)

7.3. Multiple-Node Multiple-Core Results

the granularity tuning effects in the multi-node multi-core environment. The Brachistochrone effects are shown again by the multiple optimal synchronization times at specific granularity points. These points are modulated by the increasing communication overheads (left to the optimal granularity (***70***) and increasing synchronization (waiting) time (right to the optimal point ***70***).

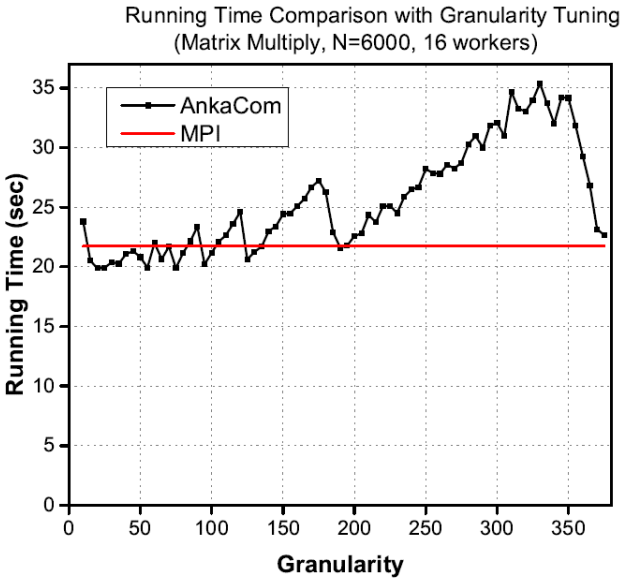


Figure 9. Multi-Node Multi-Core Granularity Tuned

Figure 10 reports AnakaCom performances against MPI using 120 cores.

AnkaCom implements multi-threaded distributed Tuple Space daemons following the UVR architecture. As discussed, this effectively decouples application programs and data completely from processors and networks. Each distributed Tuple Space server can replicate its contents elsewhere ($R > 0$). In this reported experiment, $R = 1$. This means that each tuple is replicated with 1 copy elsewhere. This experiment was designed to reveal the effects of "diminishing of return" as well as the power of granularity tuning.

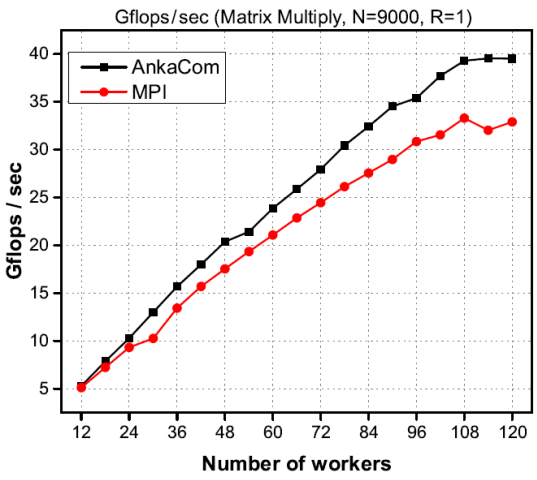


Figure 10. Multi-Node Multi-Core Tests

Each AnkaCom run is optimized by picking the best performing granularity through multiple runs. The recorded MPI run is the best of three runs with identical task size N/P.

Figure 10 shows consistently better AnkaCom performances over MPI. The performance differences grew bigger as the number of cores increased. The MPI performance started to trend down at P=108 while AnkaCom kept strong for yet another 12 more cores until the end where N/P became merely 75 (9000/120). Note that the Intel XEON clock cycle is 2.2GHZ. The 75 rows (9000 columns) of dot-products seems the break-even point for sustained performance of this particular application. After this point, both MPI and AnkaCom's communication overheads took over. This was a demonstration of the economic law of "diminishing return".

The C-MPI runs are via the "mpirun –np X -mca btl ib" option triggering the use of custom Infiniband driver. The AnkaCom Infiniband support is by the standard Java library.

8. Summary

Solving the scalability dilemma is of primary importance for developing robust computing services. End-to-end computing paradigms are suitable for small scale data processing needs. In the era of big data, with the Moore's Law approaching its end, extreme scale clustered computing is the only possible path for the future internet-scaled computing. The prior large scale HPC experiments uncovered the scalability dilemma, energy efficiency and reproducibility challenges in delivering robust computing services. It is time to eliminate the known fallacy in the making of distributed computing services.

According to the impossibility theory [2] and the possibility of delivering reliable service using faulty components [4], addressing the scalability challenges requires a paradigm shift: from IP-addressable programming to content addressable programming. This paper reports a single-sided Statistic Multiplexed Computing (SMC) paradigm in order to circumvent the impossibility and to fully leverage the possibility of statistic multiplexing. A narrow Scalable Computing Service definition and three reproducible scalable service tests are proposed. MPI, Hadoop, Spark, Synergy and AnkaCom are all examined using the same scalability tests.

In order to gain a sense of deliverable performance and reliability of the proposed service architecture in multi-node and multi-core environments, computational experiments are conducted in three groups: single node and single core, single node multiple core and multiple node and multiple core. The single-node single-core group provides the baseline for performance comparisons.

The computational results revealed the following:

- Low communication overhead is *not* a necessary condition for the optimized parallel performance. Figures 2-9 demonstrated overwhelmingly that parallel performance optimization by tuning the processing granularity can easily out-perform the low overhead MPI implementation.
- The choice of fixed *N/P* partition for MPI was deliberate. It is possible to build a task-pool layer in MPI or use single-sided "scatter_v" to allow dynamic granularity tuning without reprogramming. There will be significant implementation challenges considering passing the *Share-nothing* and *Sublinear Cost Tests*. However, when the problem size, the number of processors and the dynamic runtime load distribution happen to hit the "sweat-spot" (the fixed granularity is the optimal for the given runtime dynamics), MPI program will out-perform Synergy and AnkaCom.
- Both Synergy and AknaCom were running with built-in fault tolerance. Including checkpoints in the MPI program will push the performance to be much worse than reported.

The application reliability of the single-sided paradigm was also tested. Connection failures did occur during Synergy 48 core tests. The prototype SMC protocols worked flawlessly allowing all tests to complete without any check pointing.

The single-sided SMC application demonstrated consistently better performance in all experiments. Since SMC applications are natively protected by the TSN, the SMC application's performance and reliability have far exceeded the end-to-end computing MPI programs.

The SMC principle is also applicable for transactional and storage services [5]. Solving the data intensive SCS problem requires a solution to the CAP Theorem [29].

This paper tries to convince the reader that the robust service scalability dilemma of distributed and parallel computing is indeed solvable. The proposed single-sided SMC paradigm is a feasible solution for delivering efficient and robust computing services using faulty components. Correct implementation of the Statistic Multiplexed Computing principle ensures the reliability of large scale distributed services. Extracting the optimal processing performance is similar to finding the Brachistochrone curve in physics and mathematics. The SMC goal of building non-stoppable services is similar to the Etherem project [25]. The implementation of SMC paradigm, however, can promises high information security without linearly increasing overheads.

In addition to extreme scale parallel computing, the single-sided SMC service architecture is also suited for mission critical applications. It promises true non-stop computing service as long as the infrastructure affords the minimal survivable resource set, a task that can be automated completely in practice. This development will impact all existing mission critical distributed and parallel services. These include Enterprise Service Buses (ESB), mission critical transaction processes, mission critical storage systems, smart grid controllers and software defined network (SDN) control plane implementations. Finally, the use of service content addressable network for mission critical services makes all IP-based cyberattacks, such as DDOS (Distributed Denial of Service), ineffective. These developments will redefine the state of the Internet.

About Authors

The corresponding author is responsible for the scalability definition and tests, the UVR architecture design and the first prototype implementation Synergy. He is the original proposer of the Stateless Parallel Processing concept and the SMC framework. Author Yasin Celik is currently a PhD student completing a dissertation at Temple University. Yasin is responsible for the practical rendition of the SMC framework (AnkaCom) and validation tests. His dissertation includes both compute and data intensive SMC implementations and scalability studies.

Acknowledgment

The reported study is supported in part by the 2016 NSF REU Program at Temple University Site, NSF MRI Grant #CNS0958854 (owlsnest.hpc.temple.edu), and computing resource grant #CH-817746 from the NSF Chameleon Computing Cloud Testbed.

The authors thank two NSF REU (Research Experience for Undergraduate) program students, Kimberly Kosman (Purdue University) and Traves Evans (Louisiana Polytechnique Institute) contributed in independent validation of prior performance results in the summer of 2016.

Support from the administrative staff of the Owlsnest.hpc.temple.edu and the CIS Department, College of Science and Technology are also acknowledged.

References

[1] Peter Deutsch. "Eight fallacies of distributed computing". [online] https://blogs.oracle.com/jag/resource/Fallacies.html .

[2] Alan Fekete, Nancy Lynch, Yishay Mansour, and John Spinelli. "The impossibility of implementing reliable communication in the face of crashes". *J. ACM*, 40:1087–1107, November 1993.

[3] XSEDE 2014 reproducibility workshop report, "Standing together for reproducibility in large-scale computing". [online] https://www.xsede.org/documents/659353/d90df1cb-62b5-47c7-9936-2de11113a40f .

[4] John M. Spinelli. "Reliable data communication in faulty computer networks", June 1989. PhD thesis, MIT.

[5] OSI Reference Model [online] https://en.wikipedia.org/wiki/OSI_model

[6] Justin Y. Shi, Yasin Celik, "AnkaStore: A Single-Sided Statistic Multiplexed Transactional Distributed Storage System", unpublished manuscript, 2017.

[7] R. S. Nikhil Arvind. "Implicit parallel programming in pH", In Introduction to Management Science (10th Ed), Morgan Kaufmann, 2011.

[8] David Gelernter and Nicholas Carriero. "Coordination languages and their significance". In *Communications of ACM*. ACM, 1992.

[9] Sylvia Ratnasamy, Paul Francis, Mark Handley, Richard Karp, and Scott Shenker. "A scalable content-addressable network". In *SIGCOMM 2001*, San Diego, CA, USA, 2001. ACM.

[10] Yuan Shi. "*Stateless Parallel Processing Prototype: Synergy"*, 1996. [online] https://github.com/jys673/Synergy30.

[11] Justin Y. Shi. "System for high-level virtual computer with heterogeneous operating systems". *U.S. Patent #5,381,534*, 1995.

[12] Justin Y. Shi. "Multi-computer system and method". U.S. Patent #5,517,656, 1996.

[13] Justin Y. Shi and Suntian Song. "Apparatus and method of optimizing database clustering with zero transaction loss". *U.S. Patent Application: #2008018969*, 2008.

[14] Justin Y. Shi. "Fault tolerant self-optimizing multiprocessor system and method thereof". *U.S. Patent Application #20090019258*, 2007.

[15] Justin Y. Shi. "System and method for fault tolerant scalable computing". *U.S. Patent Application #605725*, November 2016.

[16] Justin Y. Shi, Moussa Taifi, Abdallah Khreishah, and Jie Wu. "Tuple switching network – when slower maybe better". *International Journal of Parallel and Distributed Computing*, 72, 2011.

[17] Ion Stoica, Robert Morris, David Karger, M. Frans Kaashoek, and Hari Balakrishnan. "Chord: A scalable peer-to-peer lookup service for Internet applications". *SIGCOMM Comput. Commun. Rev.*, 31(4):149–160, August 2001.

[18] David Karger, Eric Lehman, Tom Leighton, Rina Panigrahy, Matthew Levine, and Daniel Lewin. "Consistent hashing and random trees: Distributed caching protocols for relieving hot spots on the worldwide web". In *Proceedings of the Twenty-ninth Annual ACM Symposium on Theory of Computing*, STOC '97, pages 654–663, New York, NY, USA, 1997. ACM.

[19] Y. Shi. "A distributed programming model and its applications to computation intensive problems for heterogeneous environments". In *1992 Earth and Space Science Information Systems, AIP Conference Proceedings*, Pasadena, CA, 1992.

[20] Justin Y. Shi, Moussa Taifi, Abdallah Khreishah, Aakash Predeep, and Vivek Anthony. "Program scalability analysis for HPC cloud: Applying Amdahl's law to NAS benchmarks". In *International Workshop on Sustainable HPC Cloud/Supercomputing Conference 2012*. IEEE, 2012. Salt Lake City, Utah.

[21] Brachitoschrone curve. [online] https://en.wikipedia.org/wiki/Brachistochrone.

[22] Judith Hippold and Gudula Runger. "Task pool teams for implementing irregular algorithms on clusters of SMPs". Nice, France, 2003. IEEE Press.

[23] Apache hadoop. http://hadoop.apache.org.

[24] Archley Kattt. "Spark – lightning-fast cluster computing". Berkeley, CA, USA, 2011. University of California.

[25] Ethereum Homestead Documentation – "A Blockchain Application Platform" [online] http://www.ethdocs.org/en/latest/ .

[26] Yasin Celik, Aakash Pradeep, and Justin Y. Shi. Ankacom: "A development and experiment for extreme scale computing". In *CIT/IUCC/DASC/PICom*, pages 2010–2016. IEEE, 2015.

[27] Chameleon. "A configurable experimental environment for large-scale cloud research". https://www.chameleoncloud.org/ , 2017.

[28] Temple University HPC Facility: Owlsnest. http://www.hpc.temple.edu/owlsnest/OwlsnestUserGuide.html.

[29] Gilbert and Lynch. "Brewer's conjecture and the feasibility of consistent, available, partition-tolerant web services". In *ACM SIGACT News (2002)*, volume 33(2), page 59. ACM, 2002.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1461-1466 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Embedded C Programming Using FRDM to Enhance Engineering Students' Learning Skill

Zulzilawati Jusoh[*1], Mohamad Taib Miskon[1], Hasnorhafiza Husni[1], Nur Hafizah Rabi'ah Husin[2]

[1]*Electrical Engineering Faculty, Universiti Teknologi MARA, 23000, Dungun, Terengganu, Malaysia*

[2]*Academy of Language Studies, Universiti Teknologi MARA, 23000, Dungun, Terengganu, Malaysia*

ARTICLE INFO

Article history:
Received: 21 May, 2017
Accepted: 23 July, 2017
Online: 10 August, 2017

Keywords:
C programming
Training of microcontroller
FRDM
MBED compiler
Mass storage device
Flash programming interface

ABSTRACT

Computer programming course that utilizes languages such as C/C++ is always packed with dreary syntax details that consume most of the students' learning time to obtain 'grammatically' correct source code. Consequently, it is difficult for most of the students to apply the theory they have learned in a real life context. Thus, this project proposed a hardware based learning approach for C programming curriculum and reports the effectiveness of using microcontroller board named FRDM-KL05Z to assist teaching and learning activities. The USB-powered microcontroller board is very easy to use and is programmable using C programming language. Students will have the opportunity to learn selection statement with real sensors, touse repetition statement to blink LEDs and utilizing function as well as structure to control actual input and output peripherals. In general we evaluated the students' response in five criteria namely the students attributes, lecturer's profile, implementation, facilities and students' understanding. From the survey, the results in exit survey are higher compared to entrance survey for all criteria. This shows that the students are satisfied with the implementation of the module which has increased their understanding in learning C Programming.

1. Introduction

Many educators realize the urgency of education reform to respond to the current needs and future trends of the society. The urge of re-engineering the curriculum is needed for new approaches in designing the curriculum and active learning methods for encouraging students to be more creative. Responding to these calls for education reform, this study was conducted as an extension of work originally presented in IEEE 8th International Conference on Engineering Education (ICEED2016) [1]. Project-based learning (PBL) and "learning by doing"approach have been recognized as one of the effective means for students to obtain the problem-solving capability [2-3].

Another method in learning approach for engineering students is embedded system education. Teaching embedded system design is a challenging undertaking because a teacher cannot assume that all students enrolled in a class have solid prerequisite knowledge across all these areas. To speed up the learning process and motivate students to learn actively, the project-based learning approach [4-7] is applied in this embedded system design laboratory.

Introduction to the C programming is one of the basic elements in Electrical engineering course. The subject currently is offered in semester 2 [8], and it contains 7 chapters. Each chapter introduces the student with basic command of C programming to execute specific task such as compiling, linking, control flow statement, function and array [9]. The students will be evaluated based on the laboratory task which is only a computer-based laboratory, test that covers theoretical part of the syllabus and mini project. The main objective of the mini project is to gauge the understanding of students in the overall C programming course.

C programming is one of the most important subjects in Electrical Engineering course. However, the process of teaching and learning a programming subject is not an easy task as stated by many studies [10]. Students were reportedly having problem in grasping this subject due to the complex nature of this subject. The development of a program involves similar step like any problem-

[*]Zulzilawati Jusoh, Electrical Engineering Faculty, Universiti Teknologi MARA, Dungun, Terengganu, Malaysia | Email: zulzilawati@tganu.uitm.edu.my

solving task. The processes include definition of the problem, planning of the solution, coding the program, testing the program and documenting the program. Thus, the ability to execute the whole flow of the process is very critical for a student in mastering the subject well.

Various methods in teaching and learning programming courses have been reported. As an example, several researchers have come with game-based digital learning for teaching C programming course [11-12]. Other method includes active learning by students, which may include in-class activities like peer learning, games and mnemonics [13-14]. A more advance approach as discussed in [15] described how LEGO Mindstorms robot was used as hands-on educational technology in teaching introductory programming courses. This proves that C programming is a complex subject and it requires creative way of delivering it to students.

More recently, Conceive - Design - Implement – Operate (CDIO) initiative are gaining significant popularity among educators around the world [16]. Authors in [17-20] reformed their teaching approach in delivering C/C++ curriculum based on CDIO approach.CDIO concept emphasis on the knowledge development by integrating engineering skills such as team cooperation, problem solving, communication andknowledge-application ability of students in real-life context. Results indicated that such initiative can effectively improve the students' ability to comprehend basic knowledge in programming and their applicationin engineering system analysis.

Throughout this course, the main drawback that has been observed by the lecturer is that students having difficulty to implement the theoretical part of the subject into practical. This may be caused by lack of experimental design task exposure. Majority of the students do not have the skill to expand their learning of C programming after finishing this course. This situation can be seen in their final year project course (FYP) where most of the students will use microcontroller in their project and majority of them do not have the ability to apply the C programming knowledge to program the microcontroller to execute the desired task. Therefore, a module has been proposed to help the lecturer to resolve this situation. The module will be included in the students' learning process throughout the semester and it does not change the main objective of the syllabus. This module will only be used as a tool to help the student to gain more knowledge and skill in C programming language.

In fact, the microcontrollers are commonly utilized in numerous industrial applications such as remote control devices, office machines, electrical gadgets, automobiles, motor control systems, robots and other industrial fields. For that reason, it is taught as an applied course at universities, especially in the departments such as engineering of electrical and electronics engineering. Freescale has designed the FRDM-KL05Z in collaboration with MBED for prototyping all sorts of devices, especially those requiring the size and price point offered by Cortex-M0+. It is packaged as a development board with connectors to break out to strip board and breadboard, and includes a built-in USB FLASH programmer.

The aim of this study is to measure the effectiveness of the proposed module which is C programming based on embedded system approach, and the students' perception of this course is also evaluated so that there will be more effective changes that can be made to this course to enhance the student's ability in the real application after their graduate.

2. Methodology

In order to enhance the students' understanding and interest in Introduction to C Programming (ECE126) subject, a one-day training module was conducted to expose them with the implementation of C-programming as shown in Figure 1.0. During the training, participants were given a set of training materials listed in Table 1.0 so as to boost students' engagement throughout the session, harness their interest and motivate them to use the knowledge learnt in ECE126 curriculum in real-life context. A laptop was used to write a program code in C language. Next, the code was transferred to FRDM board and after that user will run the program from the board. The execution of the code was monitored using a smart phone application in real time via a Bluetooth connection.

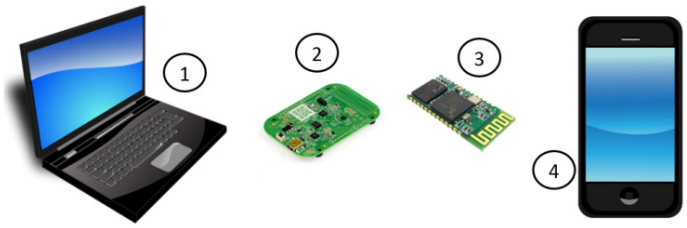


Figure 1: Examples of a real-life implementation of C-programming

Table I. List of training materials

No	Item
1	Personal Computer/laptop
2	FRDM-KL05Z micro controller board
3	HC-05 Bluetooth Module
4	Smart phone

Figure 2 depicts the overall programming process carried out during the training session. Participants were taught to write a program code using a web-based MBED compiler. The software can be used to write and edit a program code using high-level languages such as C and C++. Next, a linking process was carried out in which the written code was linked to FRDM-KL05Z controller board or any hardware platform supported by the compiler. After that, the code was compiled and a binary machine code was generated should there be no syntax error. Next, the compiled code was transferred to the board via mass storage device (MSD) flash programming interface. Subsequently, a communication link between the FRDM board and a smart phone was established via HC05 Bluetooth module so that students can observe their program simply through their own smart phones.

This approach is slightly different from the current practice where students were normally taught to create a console application that produced an executable file, which can only be used within the Windows environment [21].

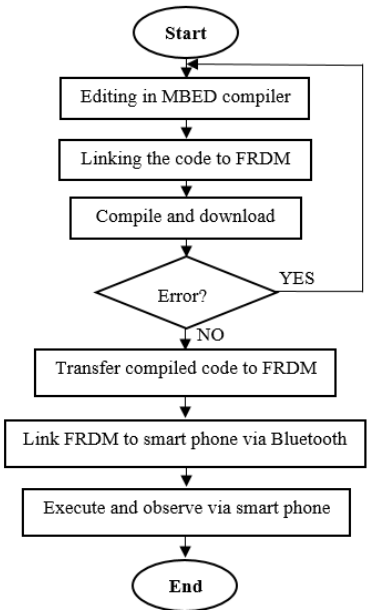


Figure 2: The programming process

2.1 *FRDM KL05Z controller board*

The FRDM-KL05Z microcontroller board depicted in Figure 3 was used in this project. It is one of the ARM Mbed supported platform based on 32-bit ARM Cortex-M processor [21]. It is a powerful tool to build an embedded system from prototyping stage to mass production. One can also write their own software on top of the board's operating system, Mbed OS, which offers an easy way of hardware control, interaction and integration with other tools. It offers convenient on-board sensors such as a touch sensor, an accelerometer and a Red-Yellow-Green (RGB) Light Emitting Diode (LED) that allow new users to practice and test their code without any need of additional hardware interface.

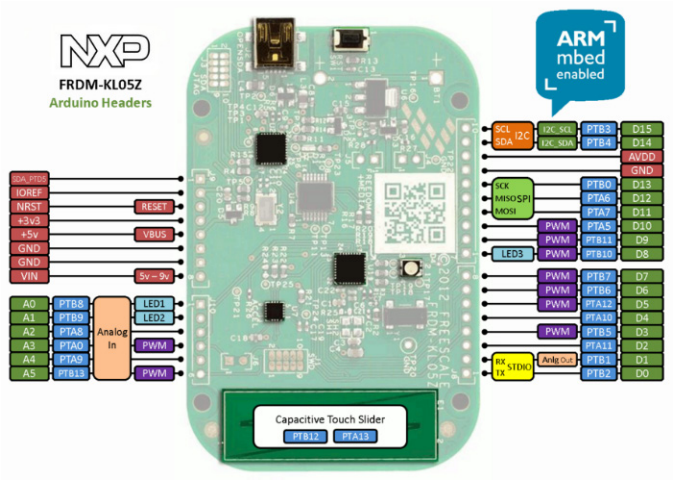


Figure 3: The FRDM-KL05Z development board [22]

2.2 Training Modules

There were seven topics related to ECE126 subject that have been covered during the training session. They include selection statement, repetition, functions, arrays, pointer and structure. These topics were delivered in seven different modules utilizing the on-board peripherals such as the LED, accelerometer and touch sensor. The modules used during the training were listed in Table II.

Table II. Training modules

No	Topic	Module
1	Basic input/output	Hello world!
2	Selection	LED control
3	Repetition	LED blinking
4	Functions	LED pattern
5	Arrays	Sensor Reading
6	Pointer	User login
7	Structure	LED dimmer

In the first module, the participants were taught to observe the execution of their program via smartphones using a Bluetooth serial application known as "BT Term" developed by Futaba Inc. A printf() statement was used to output a "hello world!" sentence via the Bluetooth application as depicted in Figure 4.

Figure 4: Output from the BT Term application

Besides, the second, third and fourth module manipulate the behavior of the on-board RGB LED using several sets of program codes. Students were taught to apply their knowledge on the selection statement to decide on the activation of the LED. Other than that, repetition statements such as while, do-while and for were also employed to create replications of the LED action. Participants were also assigned to blink the LED, in which they need to determine the suitable increment, decrement or assignment operators. Besides, they also learnt how to use break and continue statement to stop the repetition or to skip some iteration within the loop structure. In addition, functions were also been used in module 4 so that the participants can create several set of LED actions according to their preferences. Figure 5 illustrates the use of function to activate the red LED upon receiving value from the user. The program indicates that if a user specifies an input value is equal to 1, the red LED will be activated while red LED will be deactivated if other input value is given. The response of the LED is shown in Figure 6.

```
#include <mbed.h>
#include <frdm.h>
int main(){
scanf("%d",&value);
    if (value == 1)
        RED_LED_ON();
    else
        RED_LED_OFF();
}
```

Figure 5: A function to control the red LED

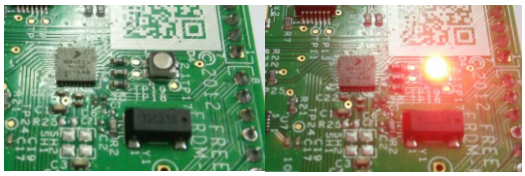


Figure 6: LED OFF (left) and LED ON (right)

Besides, one of the most important topic in C programming such as arrays were also been given much attention during the training session. It has always been useful to store information about certain things under a single variable name. For example, in module 5, students were taught to use such concept to store measured data obtained from the accelerometer sensor every one second into an array variable. Next, they can manipulate the content of each array to do additional action such as activating the LED if certain condition is met. This allows students to think, try and explore more about the concept as well as drawing connections between what have been taught and real-world contexts.

Meanwhile, module 6 focuses on the workings of pointer variable, how to use them as well as the common mistakes faced by a programmer when dealing with it. In this exercise, pointer was used to access a memory location that contains user information. Participants were required to create a login interface that prompt user to input their login ID and password. The program will evaluate the input entered via the serial Bluetooth application before triggering the green LED if correct input is given and red LED if otherwise.

The final module attempted to enlighten the participants about the basic concept of structure in C programming language and how it can be used to hold several information within the same category that share common attributes. Figure 7 demonstrates the use of structure concept to control the brightness of the LED. A structure with a tag named LED was declared in this context and three members named "bright", "medium", and "dim" were defined. The member attributes can be accessed by using "." operator such as "RED.bright" which will provide the variable "LAMP" with 1.0 value.

```
#include "mbed.h"
#include "jenggott.h"
struct LED{
    float dim;
    float medium;
    float bright;
    }
int  main( )
{
    LED RED = {0.2,0.5,1.0};
    LAMP = RED.bright;
}
```

Figure 7: Brightness control using structure

In order to evaluate the approach's acceptability and its effect on learning C programming within the training period, a statistical studies were carried before and after the training program. The results of the survey will be discussed in the following section.

3. Results

Figure 8 shows the bar graph descriptions for Criteria 1 to measure the students attribute towards C Programming subject. As shown in the figure, the mean scores for all questions in exit survey are higher compared to the entrance survey. The students responded that they have better programming skills and sufficient knowledge and understanding in C Programming after the module was implemented.

From the survey, we also measured lecturer's ability in teaching the subject from the students' point of view and the statistic is depicted as in Figure 9 below. The mean scores in exit survey are higher than the entrance survey. Majority of the mean scores for all questions in this criteria are higher than 4.00. This indicates that the students felt that the lecturer is capable of teaching this subject.

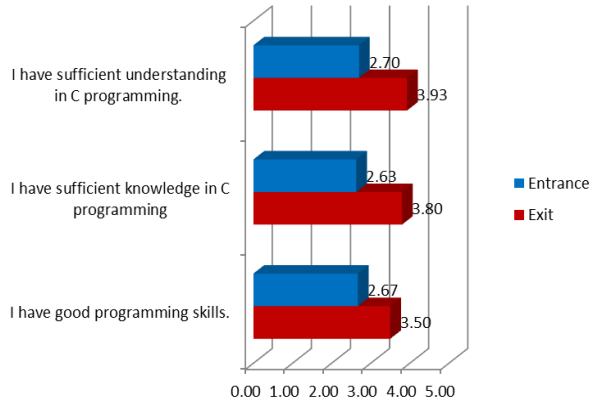


Figure 8: Criteria 1 Students Attributes

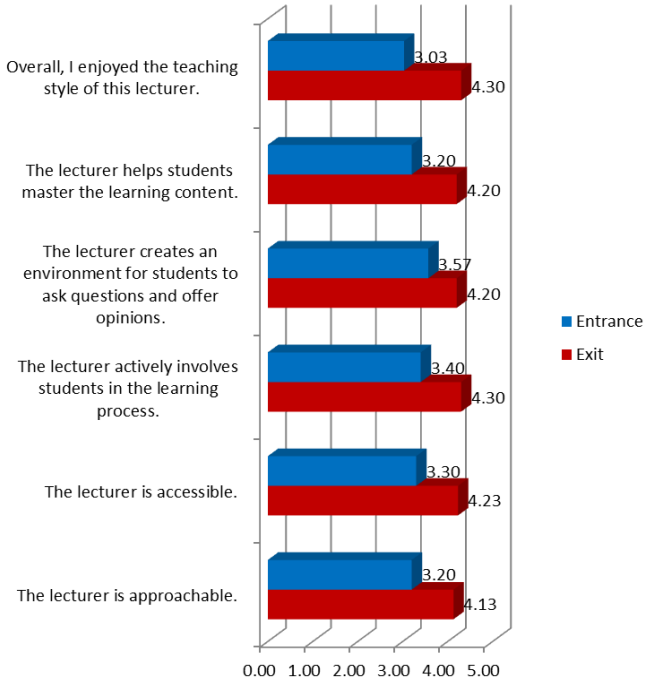


Figure 9: Criteria 2 Lecturer's Profile

We also evaluated the effectiveness on how the lecturer implemented the module as shown in Figure 10 below. The mean scores for all questions are higher in exit survey than entrance survey, which are above 4.00. This shows that the students are satisfied with the implementation of the module.

Other than that, the students gave positive response towards the facilities provided during the module session. Figure 11 below shows the statistic of the criteria. The mean scores for equipment space (mean = 4.13) and functionality (mean = 4.10) are higher in the exit survey compared to entrance survey.

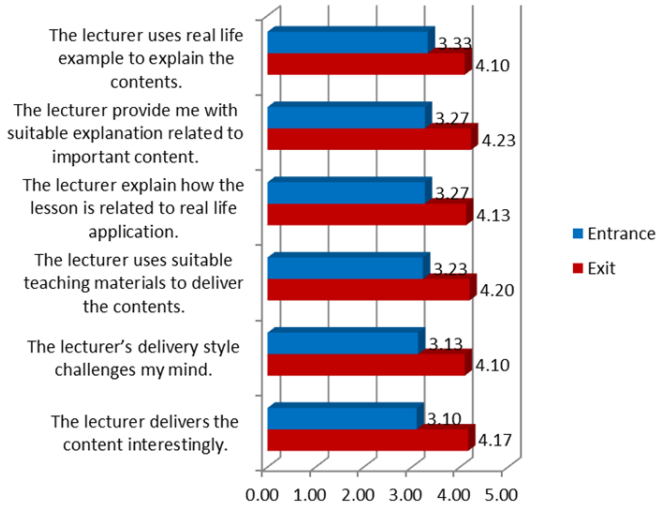


Figure 10: Criteria 3 Implementation

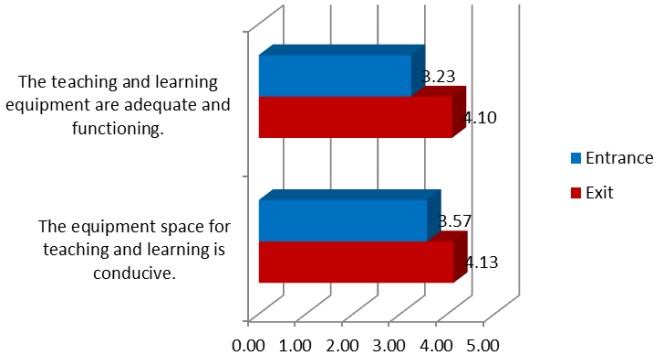


Figure 11: Criteria 4 Facilities

The survey also tested the students understanding of the module as can be seen in Figure 12 below. From the figure, mostly students gave positive response that the module enhances their understanding of C Programming.

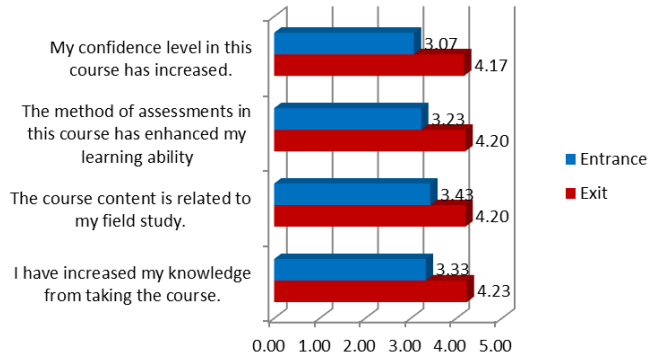


Figure 12: Criteria 5 Students' Understanding

This research study was further extended to evaluate the students' response reflecting the module execution and implementation. Figure 13 to 16 below depict the students' responses regarding the overall module assessment.

Figure 13 shows the bar graph for Criteria Motivation and Students' interest toward the C Programming. The highest score from this criterion was 'Agree' for all three questions and this shows that they are motivated and interested in learning C Programming.

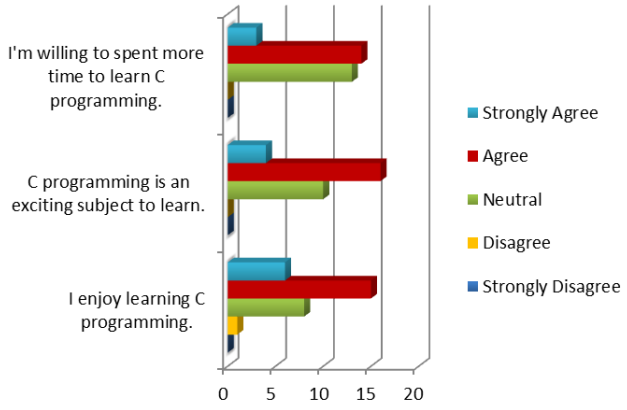


Figure 13: Criteria Motivation and interest

Figure 14 describes the difficulty of the course from the students response. Majority of the students found that the course was quite challenging which they were struggling to understand and complete the task given.

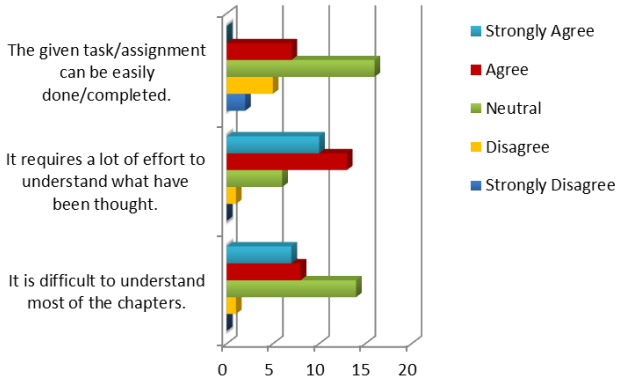


Figure 14: Criteria Difficulties

Figure 15 displays the third criterion which measured the Expectancy for Success among the students. The results shows positive feedback from the students which they believe this course will improve their academic performance.

Figure 16 depicts the students response on the implementation of the course. They found that this course was enjoyable and they highly recommend that the course should be more fun and interesting which interactive games can be integrated and real life application can be interfaced.

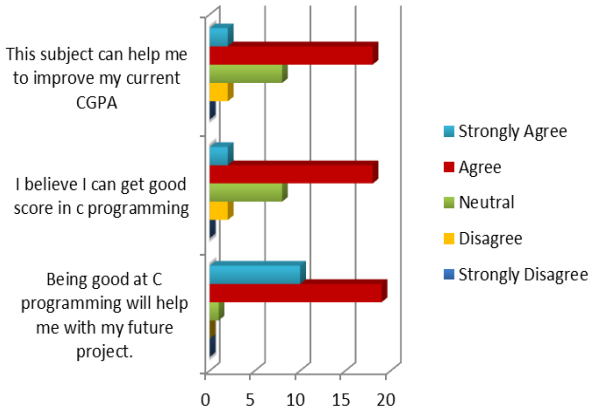


Figure 15: Criteria Expectancy for success

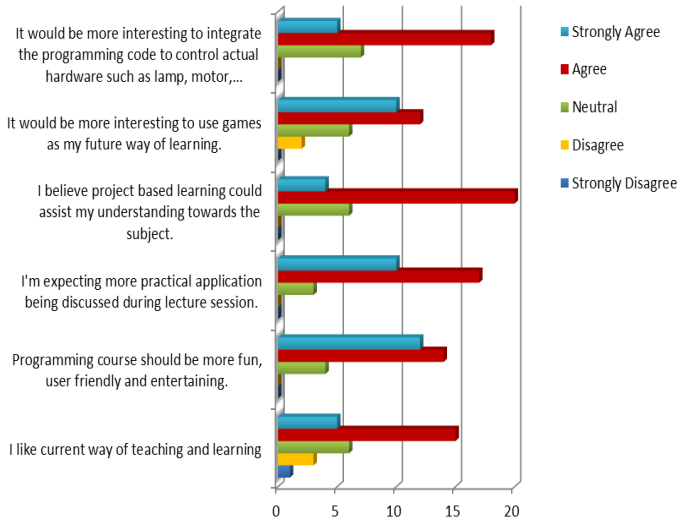


Figure 16: Criteria Deliveries

4. Conclusions

This paper reports the effectiveness of the implementation of the C Programming module. In general we evaluated the students' response in five criteria namely the students attributes, lecturer's profile, implementation, facilities and students' understanding. From the survey, the results in exit survey are higher compared to entrance survey for all criteria. This shows that the students are satisfied with the implementation of the module, which has increased their understanding in learning C Programming.

Acknowledgment

This work is financially supported by Institute of Research Management & Innovation, UniversitiTeknologi MARA Grant Scheme Academic & Research Assimilation (600-IRMI/DANA 5/3/ARAS (0124/2016)) and (600-RMI/DANA 5/3/ARAS (52/2015)).

References

[1] Miskon, Mohamad Taib, et al. "Embedded C programming using FRDM hands-on learning system design education." Engineering Education (ICEED), 2016 IEEE 8th International Conference on. IEEE, 2016.

[2] Bell, Stephanie. "Project-based learning for the 21st century: Skills for the future." *The Clearing House* 83.2 (2010): 39-43.

[3] Conole, Gráinne. Designing for learning in an open world. Vol. 4. Springer Science & Business Media, 2012.

[4] Ajao, Lukman Adewale, et al. "Learning of Embedded System Design, Simulation and Implementation: A Technical Approach." *American* Journal of Embedded Systems and Applications 3.3 (2016): 35-42.

[5] Chu, Edward T-H., and Chi-Wei Fang. "CALEE: A computer-assisted learning system for embedded OS laboratory exercises." *Computers & Education* 84 (2015): 36-48.

[6] Rao, PVV Rama, G. Durga Prasad, and N. Venkata Ramana. "Project Based Learning Method using Simulation Tools and Hardware Exposure." Journal of Engineering Education Transformations (2016).

[7] Malinowski, Aleksander, and Hao Yu. "Comparison of embedded system design for industrial applications." *IEEE Transactions on* Industrial Informatics 7.2 (2011): 244-254.

[8] Study Plan, Faculty of Electrical Engineering.

[9] "*Introduction to C programming (ECE126)*", 2013, Fakulti Kejuruteraan Elektrik course information.

[10] Kelleher C, Pausch R. Lowering the barriers to programming: A taxonomy of programming environments and languages for novice programmers. ACM Computing Surveys (CSUR). 2005 Jun 1;37(2):83-137.

[11] Chen, Woei-Kae, and Yu Chin Cheng. "Teaching object-oriented programming laboratory with computer game programming." IEEE Transactions on Education 50.3 (2007): 197-203..

[12] Leutenegger, Scott, and Jeffrey Edgington. "A games first approach to teaching introductory programming." ACM SIGCSE Bulletin 39.1 (2007): 115-118.

[13] Enzai, Nur Idawati Md, et al. "Harnessing creativity in teaching and learning introductory programming." Research and Development (SCOReD), 2011 IEEE Student Conference on. IEEE, 2011.

[14] Poindexter, Sandra. "Assessing active alternatives for teaching programming." *Journal of Information Technology Education* 2 (2003): 257-265.

[15] Williams, Andrew B. "The qualitative impact of using LEGO MINDSTORMS robots to teach computer engineering." IEEE Transactions on Education 46.1 (2003): 206.

[16] Jianzhong, Zha. "On CDIO model under learning by doing strategy." Research in Higher Education of Engineering 3 (2008): 1-6.

[17] Wang, Jianmin, and Shuling Di School. "The C/C++ Integrative Teaching Based on CDIO Education Model." IERI Procedia 2 (2012): 299-302.

[18] Huang, Xiaodong, Qin Wang, and Lishang Zhu. "The Research on Curriculum Group of Computer Programming Based on CDIO." DEStech Transactions on Computer Science and Engineering icte (2016).

[19] CAO, Yingying, Licai ZHU, and Chang'an CAI. "Design and Implement of Cases Used in Programming Courses Based on CDIO." Computer 6 (2011): 016.

[20] Jusoh, Z et al. "Hands-on and Training on Embedded System." 2015 IEEE 7th International Conference on Engineering Education (ICEED) 2015: 151–154.

[21] "*Introduction to C programming (ECE126)*", 2013, Fakulti Kejuruteraan Elektrik, Laboratory Manual.

[22] https://developer.mbed.org/platforms/frdm-kl05z/

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1467-1471 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

A Modified Patch Antenna with Square-Open Loop Resonator Slot for Improved Bandwidth Performance in WiFi Applications

Rida Gadhafi*, Mihai Sanduleanu

Department of Electrical And Computer Engineering, Masdar Institute, A part of Khalifa University of Science and Technology, 54224, United Arab Emirates

ARTICLE INFO

Article history:
Received: 29 May, 2017
Accepted: 21 July, 2017
Online: 10 August, 2017

Keywords:
Square patch antenna
Open loop resonator slot
Bandwidth enhancement

ABSTRACT

A novel, modified square patch antenna with square open loop resonator slot, operating at 5.5 GHz, for improved bandwidth performance is proposed in this article. A complementary square open loop resonator is introduced on the patch to enhance the bandwidth more than twice that of the conventional square patch antenna operating at the same frequency. A prototype of the antenna operating at 5.5 GHz is designed and fabricated. Results show that 2:1 VSWR bandwidth of the antenna is 7% whereas the conventional square patch antenna offers only 3% bandwidth. The proposed antenna has a compact size of 1.9 cm x 2 cm. It offers a uniform gain of 4.25 dB and a directivity of 6.1 dB at the frequency of operation.

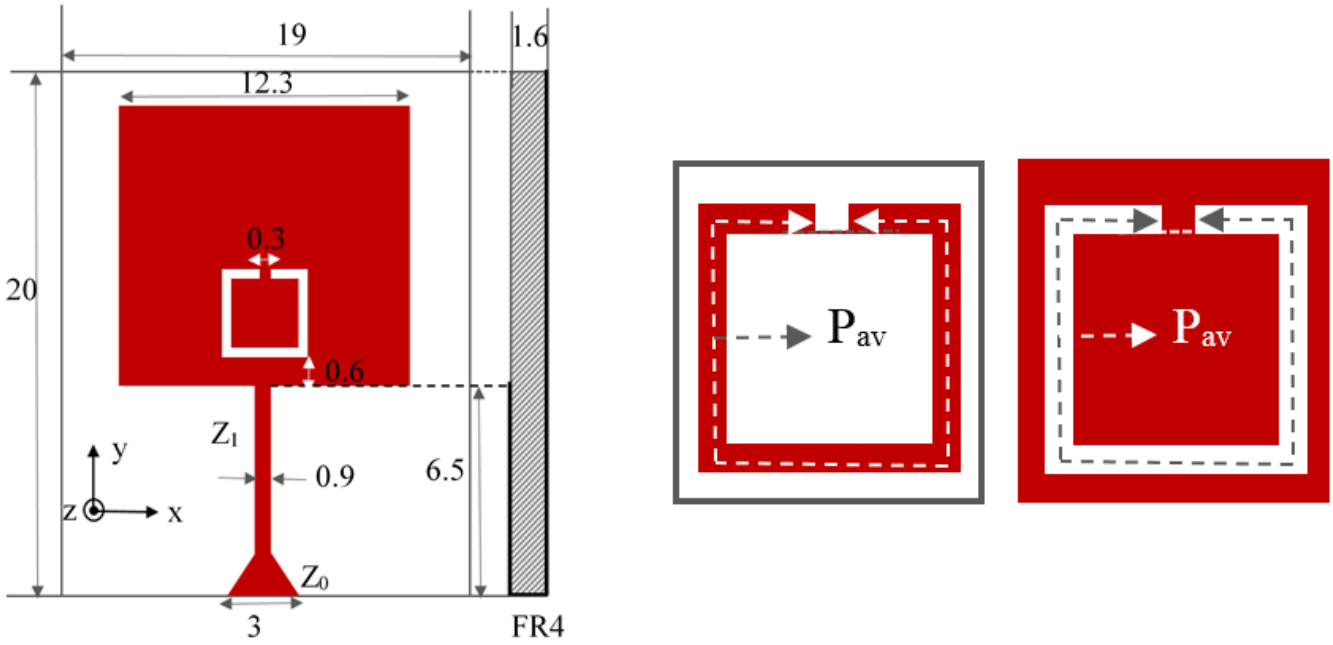
1. Introduction

Owing to their small size, low cost/weight, conformability, and ease of manufacturing, microstrip patch antennas are chosen in many designs applied to standard mobile services and satellite systems [1-2]. However, the main barrier of considering these antenna types for a large range of applications is related to their limited bandwidth [3-5]. The 'Long Term Evolution' (LTE) and 'Long Term Evolution-Advanced' (LTE-A) new standards for 4G, require large bandwidths (up to 200MHz) and it is expected, that newer standards will require even higher bandwidths [6-9]. In that respect, several techniques were discussed in the literature, for bandwidth improvement, including changing the feed type, stacking the patch, modifying the geometry of the patch, and changing the feed location [10-14]. Each method has its own trade-offs, particularly involving the size of the antenna. The use of resonant slots for increasing the bandwidth was proposed in [1]. The resonant frequency of the slot could be chosen in such a way that its operating frequency is different than the resonant frequency of the conventional antenna. This leads to the staggering of two frequencies which, in turn, will increase the bandwidth significantly.

This paper is an extension of work originally presented in the International Conference on Electronic Devices, Systems and Applications [15]. In this extended version, the design of novel square patch antenna with a complementary square open loop resonator (C-SOLR) is explained with the support of measurement results. A square patch antenna with and without C-SOLR is explained. In the proposed design, C-SOLR utilizes proximity coupling. At first, a patch antenna without C-SOLR operating at 5.5 GHz is designed. C-SOLR can reject the corresponding frequency enabling bandwidth enhancement. C-SOLR has been introduced on the patch without affecting the radiation characteristics of the antenna. The proposed antenna offers 400 MHz, 2:1 VSWR bandwidth which is significantly higher than that of the bandwidth obtained by other techniques like changing the feed of the patch proposed in [2], [14]. The proposed antenna offers simulated directivity and gain of 6.1 dB and 4.25 dB respectively.

The remaining sections of this article is organized as follows: Section II explains the design of C-SOLR antenna; Section III, presents simulation results followed by the measurement validation. Finally, the paper is concluded in Section IV.

*Corresponding Author: Rida Gadhafi, Department of Electrical And Computer Engineering, Masdar Institute, A part of Khalifa University of Science and Technology, 54224, United Arab Emirates | Email: rgadhafi@masdar.ac.ae



(a) (b) (c)

Figure 1: Geometry of the antenna and the classical loop resonator. a) antenna with C-SOLR. (All dimensions are in mm). b) Square open loop resonator (S-OLR), c) Complementary Square open loop resonator (C-SOLR)

2. Antenna Design

The geometry of the proposed antenna is shown in Figure 1 (a). The substrate used is FR-4 with relative permittivity of 4.4, loss tangent of 0.02 and thickness of 1.6 mm. The proposed antenna is a square patch with a complementary square open loop resonator (Figure 1 (c)). The square patch used in this design is the conventional square patch antenna. In order to achieve the broadband operation of the antenna, a resonating slot is introduced. Square open loop resonator (S-OLR) has been chosen here because of their compact size and ease of design [15]. S-OLR can resonate at half the frequency as that of the closed loop resonator with the same dimension. S-OLR are also considered as a microstrip line, loaded with folded open stubs at both ends. Figure 1(b) shows the structure of classical square open loop resonator (S-OLR). The fundamental resonance frequency f_0 of S-OLR can be given by:

$$f_0 = \frac{c}{2(P_{av}-s)\sqrt{\varepsilon_{eff}}} \quad (1)$$

In (1), $\boldsymbol{P_{av}}$ is the average perimeter, s is the gap between the open ends and $\boldsymbol{\varepsilon_{eff}}$ is the effective permittivity of the substrate [16]. Inspired by its compactness, we are introducing the 'negative square open loop resonator' which can also be called 'complementary square open loop resonator (C-SOLR) in analogy to the concept of complementary split ring resonator (CSRR) (see Figure 1(c)). From simulation study, we have observed that C-SOLR behavior is governed by eq. (1). When C-SOLR is coupled to a transmission line, the wave propagation is rejected in the vicinity of the resonant frequency. This property of the C-SOLR is exploited here which enables a larger bandwidth compared to the antenna without C-SOLR. The antenna has a compact dimension of 19 mm x 20 mm. The width of the square patch ($\boldsymbol{W_P}$) is 12.3 mm and can resonate at 5.5 GHz. Ansoft High Frequency Structure Simulator (HFSS) is used to perform simulations. The optimum gap $\boldsymbol{g}$ between the C-SOLR and feed line uses proximity coupling A parametric study has been done and the optimum gap is found to be 0.6 mm. This, ensures better coupling and impedance matching as shown in Figure 2. The square patch is matched to a transmission line of characteristic impedance Z_0 (50 Ω line) by using a quarter wavelength transmission line transformer of characteristics impedance Z_1 as shown in Figure 1 (a). The optimum width of the quarter wavelength transmission line is $\boldsymbol{W_T}$=0.9 mm. The 50 Ω line is mitered to alleviate the small mismatch between the high and low impedance line by chopping off the small amount of capacitance introduced due to the sudden transition from the high impedance to low impedance line. Thus, it can also reduce the unwanted reflections due to the abrupt change. The width of the C-SOLR is determined from a parametric study and the optimum value is found to be 0.3 mm for better impedance matching and gain as shown in Figure 3 (a) & (b). Table I summarizes the optimized parameters of the C-SOLR antenna.

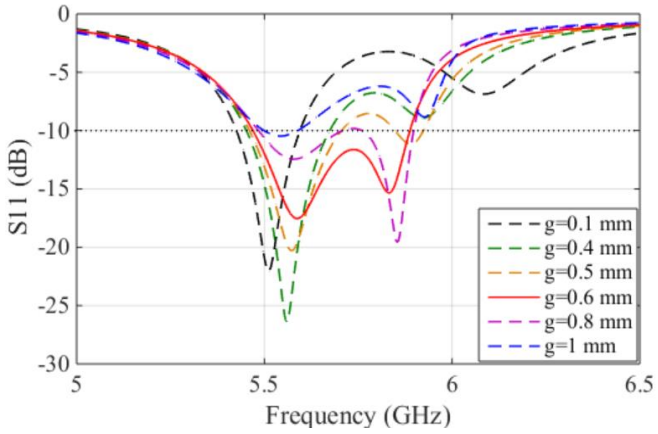
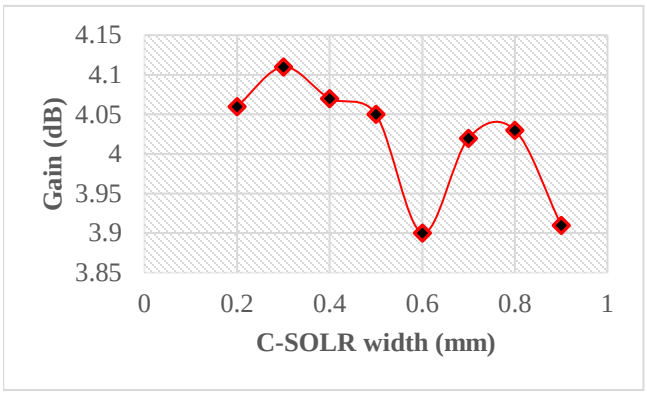
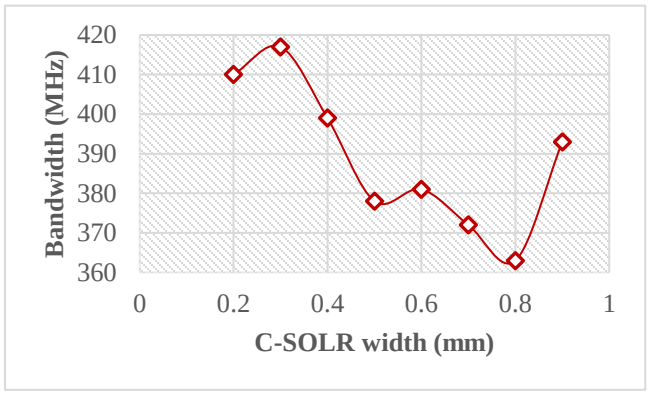


Figure. 2: Influence of coupling gap *g* on the reflection coefficient.



(a)



(b)

Figure 3: Influence of C-SOLR width on gain and bandwidth. a) C-SOLR width versus gain b) C-SOLR width versus bandwidth.

3. Results and Discussion

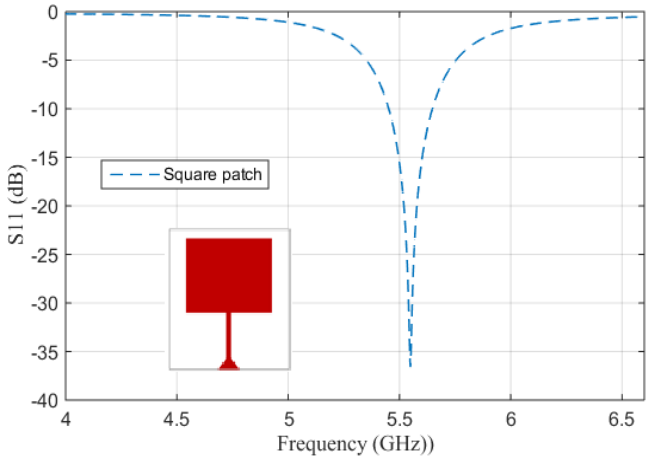
Simulation and measurement of proposed antenna is discussed in this section. Antenna measurements were conducted using Agilent Network Analyzer (PNA-X N5242A).

Firstly, a square patch antenna without resonating slot has been simulated. Figure 4 (a) shows the simulated S-parameters of the square patch antenna without any slot. As shown, it produces a narrow bandwidth of 170 MHz. Figure 4 (b) shows the corresponding E-plane and H-plane pattern. As can be seen, the patterns are unidirectional. The square patch antenna offers a gain and directivity of 4.3 dB and 6.1 dB respectively.

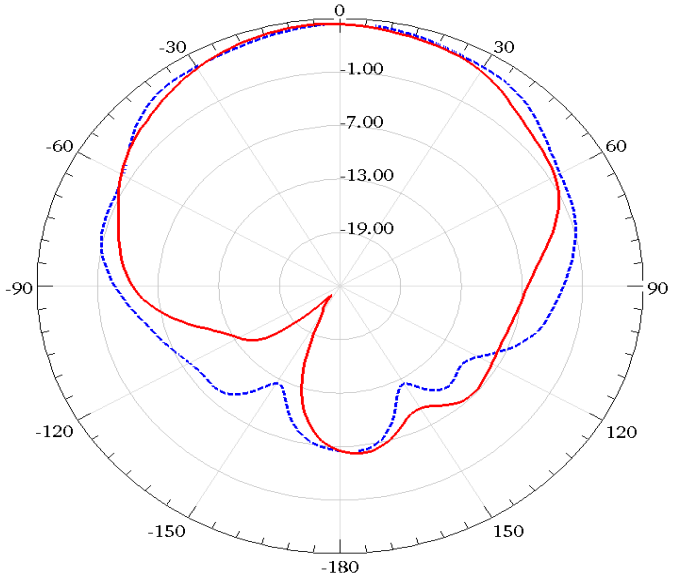
Table I Optimal parameter values of the proposed antenna

Parameters	***g***	***Pav***	W_T	***Wp***
Value (mm)	0.6	15.2	0.9	12.3

By properly placing the C-SOLR on the patch, the prototype has been transformed to the C-SOLR patch antenna as depicted in Figure 1(a). From simulations it was found that the parameter ***g*** places an important role in the impedance matching and coupling to the feed line as well. Figure 5 shows the influence of *g* on VSWR bandwidth. Thus, the optimum value of ***g*** has been found as 0.6 mm (see Table I). The resonance frequency of the C-SOLR has been chosen as a frequency near to the resonance frequency of the square patch without any slot.



(a)



(b)

Figure 4: Simulated reflection and 2-D radiation characteristics of a conventional patch antenna. a) simulated S-parameters, b) E-plane (solid line) and H-plane (dashed line) pattern.

The designed antennas were fabricated on standard FR4 substrate with relative permittivity of 4.6 and a dielectric height of 1.6 mm. Figure 6 (a) shows the fabricated antenna. The simulated and measured reflection characteristics are depicted in Figure 6 (b). As shown in the figure, they are in good agreement. From simulations, we have observed that C-SOLR operates in accordance with eq. (1). Moreover, more than twice bandwidth of the conventional antenna can be obtained using the C-SOLR antenna. The proposed antenna offers 7%, 2:1 VSWR bandwidth whereas, the conventional antenna offers only 3%, 2:1 VSWR bandwidth.

Figure 7 shows the 2-Dimensional radiation characteristics of the proposed antenna. The antenna provides a directive radiation pattern towards the patch surface. Notably, the C-SOLR has been introduced without disturbing the radiation characteristics of the antenna. This is the main advantage of our proposed methodology.

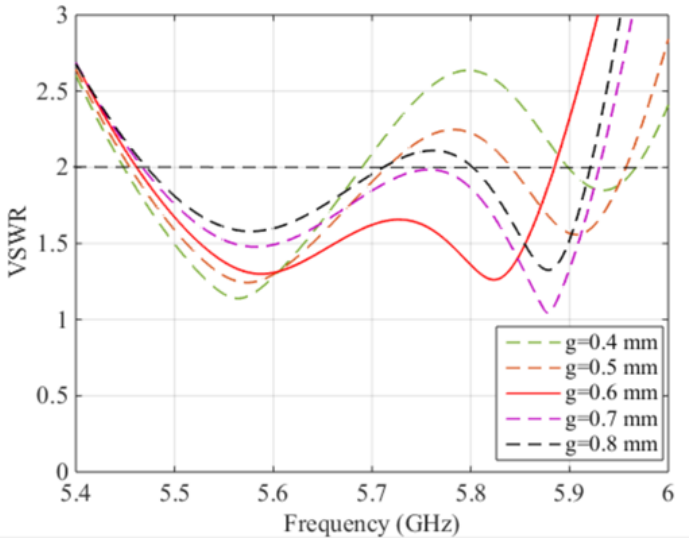
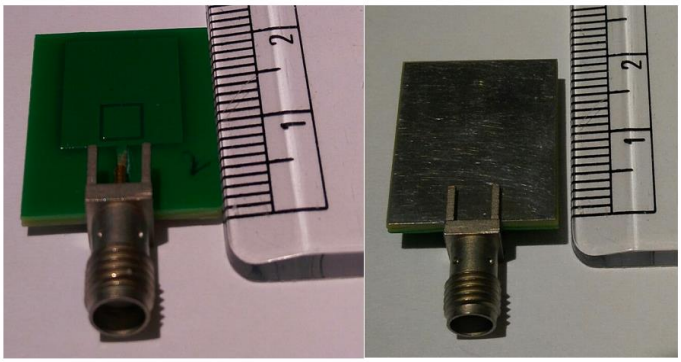
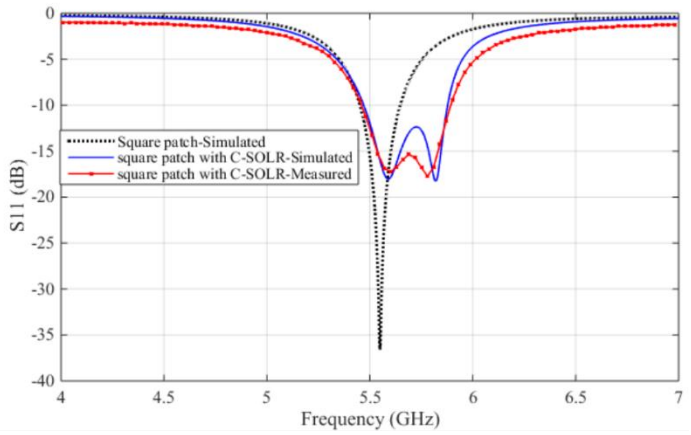


Figure 5: Influence of gap g on VSWR bandwidth.



(a) (b)



(c)

Figure 6: Fabricated antenna and corresponding simulated and measured reflection coefficient, a) antenna front view, b)antenna rear view, c) reflection coefficient.

Thus, the slotted SOLR antenna offers almost the same gain and directivity as a conventional square patch antenna with 4.25 dB gain and 6.1 dB directivity as shown in Figure 8. From simulations, the proposed antenna offers a uniform gain and directivity. In Figure 8, the gain tend to decrease above 5.87 GHz and indicates the appearance of cross polarized component as a consequence of using the C-SOLR structure.

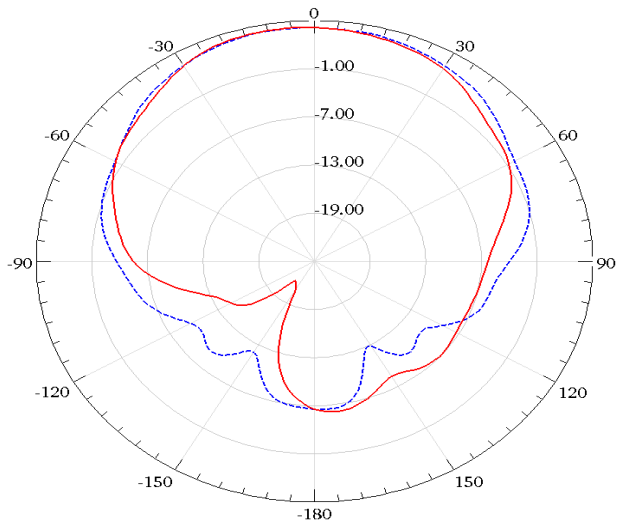


Figure. 7: Simulated 2-D radiation characteristics of C-SOLR patch antenna. E-plane (solid line) and H-plane (dashed line) pattern.

Thus, if we didn't chose the optimum frequency for the C-SOLR, this cross-polarizing component will appear inside the band of interest, causing a non-uniform gain. The uniform gain is an advantage of SOLR in comparison to a rectangular loop resonator in which, the longitudinal arm of the resonator produces the cross polarized component within the band of interest. The rectangular loop resonator causes a non-uniform gain but has comparable bandwidth with that of the SOLR.

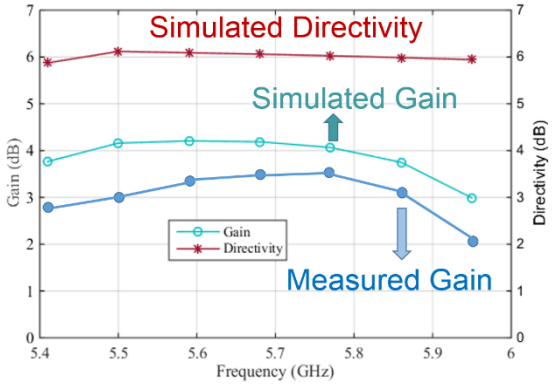
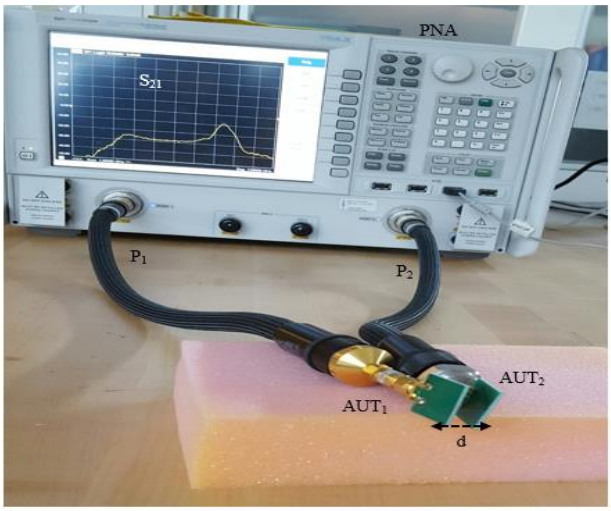


Figure.8: Simulated/measured gain and directivity of C-SOLR patch antenna.

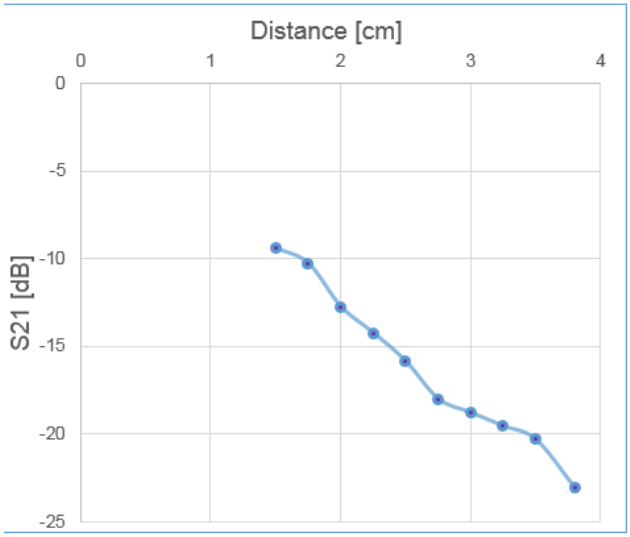
The power transfer measurement of the proposed antenna has also been conducted as depicted in Figure 9 (a). Two identical antennas separated with a distance *d* were connected to the two ports of the Network Analyzer and corresponding S21 was measured. The power receiving capability of the antenna was tested for different distances. As shown in Figure 9 (b), the magnitude of power level is decreasing with increasing the distance. The measurement was conducted for a range of *d* from 1.5 cm to 3.8 cm.

4. Conclusion

The presented article proposes a novel C-SOLR patch antenna with enhanced bandwidth with the support of measurement. A proximity coupled complementary square open loop resonator is placed on the patch. C-SOLR rejects this frequency and, as a result, the bandwidth is enhanced. Results proves that C-SOLR antenna



(a)



(b)

Figure 9: Power transfer measurement of the proposed antenna. a) Measurement setup, b) received power versus distance between the two antennas.

offers more than twice the bandwidth of a conventional square patch antenna, without disturbing the radiation characteristics. The proposed antenna offers the same uniform gain and directivity as that of the square patch antenna, as well. It can operate from 5.47 GHz- 5.87 GHz and could be used for WiFi applications. The antenna can be easily redesigned for another frequencies, e.g. to meet the extensive bandwidth requirement of 200 MHz for LTE standards applied to 4G systems.

References

[1] M.M. Honari, A. Abdipour, G. Moradi, "Bandwidth and Gain Enhancement of an Aperture Antenna with Modified Ring Patch," IEEE Antennas and Wireless Propagation Letters, vol. 10, pp. 1413-1416, 2011.

[2] A.G. Koutinos, G.A. Ioannspoulos, M.T. Chryssomallis, and G.A. Kyriacou, "A Dual Feed Rectangular Patch Antenna for Bandwidth Enhancement", 10th Loughborough Antennas and Propagation Conference(LAPC), UK, pp.698-702, November, 2014.

[3] K. Mandal, S. Sarkar, and P. P. Sarkar, "Bandwidth enhancement of microstrip antenna by staggering effect,"Microw. Opt. Technol. Lett., vol. 53, pp. 2446–2447, 2011.

[4] X. L. Bao and M. J. Ammann, "Small patch/slot antenna with 53% input impedance bandwidth," Electron. Lett. vol. 43, pp. 146–147, 2007.

[5] S.L.S.Yang, A.A.Kishk, and K.F. Lee, "Frequency reconfigurable U-slot microstrp patch antenna, "IEEE Antennas Wireless Propaga. Lette. Vol.7,pp. 127-129, 2008.

[6] S. Capora Dell Bario, M. Pelosi, and G.F. Pedersen, "On the effieciency of reconfigurable high-Q antennas for 4G standars,"Electronic Letters, vol.48, issue.16, pp. 982-983, Aug.2012

[7] P. Bahramzy, O.Jagielski, S. Svendsen, P. Olesen, G.F. Pedersen, "Aspects of high-Q tunable antennas and their deployment for 4-G mobile Mobile Communications", IEEE Antennas and Propagation Magazine, vol. 58, issue. 4, pp. 70-81, 2016.

[8] S.F. Roslan, M.R. Kamarudin, M. Khalily, and M. H. Jamaluddin, "An MIMO rectangular dielectric resonator antenna for 4G applications," IEEE antennas and wireless propagation letters, vol. 13, pp. 321-324, 2014.

[9] F.M. Caimi, "Antenna Design challenges for 4G," IEEE Wireless Communications, vol.18, issue.6, pp. 4-5, 2011.

[10] Nasimuddin and Z. N. Chen, "Wideband Microstrip Antennas with Sandwich substrate," Microw., Antennas Propag., vol. 2, pp. 538–546, 2008.

[11] C.-H. Lai, T.-Y. Han, and T.-R. Chen, "Broadband aperture- coupled microstrip antennas with low cross polarization and back radiation," Prog. Electromagn. Res. Lett., vol. 5, pp. 187–197, 2008.

[12] J. S. Chen and K. L. Wong, "A single-layer dual-frequency rectangular Microstrip patch antenna using a single probe feed", Microwave Opt. Technol. Lett. Vol.11, pp. 83–84, Feb. 5, 1996.

[13] M. T.Ali,N.Nordin, N. Ya'acob, andM.N. M. Tan, "Design of wideband microstrip patch antenna using L probe fed at 2.6 GHz," in Proc. ICCCE, 2012, pp. 961–965.

[14] A.G. Koutinos, G.A. Ioannspoulos, P.K. Gkonis, M.T. Chryssomallis, and G.A. Kyriacou, "A Novel Quadruple-Fed Rectangular Patch Antenna with Improved Bandwidth Performance", 11th Loughborough Antennas and Propagation Conference(LAPC), UK, November 2015.

[15] J. Hong and M. Lancaster, "Couplings of microstrip square open-loop resonators for cross-coupled planar micorwave filters," IEEE Transcations on Microwave Theory and Techniques, vol. 44, no.12, pp.2099-2109, December 1996.

[16] K. Chang, and L-H.Hsieh, "Microwave Ring Circuits and Related Structures,"Second Edition, Wiley Series in Microwave and optical Engineering, 2004.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1472-1477 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

A comparative study for using the LBC format for compressing static medical images

Ildikó-Angelica Szöke*,[1], Vasile Stoicu-Tivadar[1], Diana Lungeanu[2]

[1]*Automation and Computers, Politehnica Timisoara, 300223, Romania*

[2]*Victor Babes University of Medicine and Pharmacy Timisoara, 300041, Romania*

ARTICLE INFO

Article history:
Received: 30 May, 2017
Accepted: 21 July, 2017
Online: 10 August, 2017

Keywords :
Local Binary Compressed
Local Binary Pattern
Static medical images
Image compression

ABSTRACT

This paper is an extension of work originally presented at conference Applied Machine Intelligence and Informatics (SAMI), 2017 IEEE 13th International Symposium. The article performs a comparative study between the image compression technique Local Binary Compressed format (LBC) proposed by the authors and the standard image compression techniques used today (BMP, JPEG, PNG and GIF). The study is carried out for large medical images that need to be stored for longer periods of time. In the study, the most common types of medical images used are: X-rays and ultrasounds images. This study aimed to determine the distribution laws parameters used to reconstruct the image as well as to determine the influence of the most used medical image formats on the compression process. The effective representation of image format LBC is proven using an image quality index type, the Structural Similarity Index (SSIM).

1. Introduction

The progresses made in the computer technology field in the last decade, had as result the almost exponential increase of generated data. Thus, the amount of generated data has increased from five Exabyte's (EB) in 2003 to 2.8 zettabytes (ZB) in 2012. Compared to 2010, the amount of data [1, 2] increased more than 50 times. In 2020, the world will produce an amount of about 40 ZB data. According to the same study, only 1% of the amount of data produced in 2012 was analyzed. It is worth noting that in late 2012 there were 550 billion archived images comprising over 60% of medical type. For an average size of 20MB of images, amount of archived information [3] with an index of 0.15 compression is 550 billion * 20MB / .15 = 73 Exabyte's. Therefore, compressing images, especially those of medical nature, has become a topical issue, which will require finding new techniques and formats to be represented by reduced size. This article makes a comparative analysis of LBC compression technique proposed by the authors [1, 2] compared to standard techniques of representation images.

2. Related work

The image compression term refers to the process of reducing the amount of data required to represent the image quality subjective acceptable. This can be achieved by removing the redundancy present in an image [4]. The compression advantage is that it reduces the important resources consumption, such as hard disk space. Implementing an image compression involves storing the image in a bit stream as compact as possible and image decoding as accurately as possible. The needed items are an encoder and a decoder. The encoder receives and converts the image into a series of binary data, which then is transmitted to the decoder input. The decoder decodes picture as accurately as possible [5].

The most important properties of a compression algorithm are the compression rate and the reconstruction quality. The compression ratio is the ratio between the number of bits necessary for representing the data before and after compression.

The best-known image compression techniques are lossless and lossy compression. The first technique does not cause any loss on image quality and is used in cases where image accuracy is very important, like technical drawings. When using the lossy compression, the original image differs from the reconstructed one but to determine the algorithm's efficiency an estimation between the two versions is needed.

The following methods are some known image compression techniques [6, 7].

*Corresponding Author: Ildikó-Angelica Szöke, Automation and Computers, Politehnica Timisoara, 300223, Romania | Email: ildiko.szoke@yahoo.com

The Run Length Encoding (RLE) technique relies on image colors repetition. If the image is read from the top left, reading one row of pixels at a time, often rows of pixels of the same color will be obtained. If more than 3 pixels from the row have the same color, then the color and the color number are stored. Using this technique, no information is lost. RLE has been used in PCX format that no longer exists.

Standard Joint Photographic Experts Group (JPEG) specifies the algorithm used for compression and decompression (codec). This format involves loss of visual information. JPG format is good for photos but not equally suited for high-contrast images, such as screenshots or computer art. It is based on the fact that the eye is not sensitive to small color changes. Compression ratio may vary when the image is saved. This format is re-compressed every time a backup is performed; repeated compressions to an image can lead to loss of quality.

Graphics Interchange Format (GIF) is based on limiting the colors depth used in the image. Usually up to 256 colors are used to achieve the color palette - a table that assigns colors to 256 numbers from 0 to 255. The image pixels are then stored using an 8-bit number that represents the color position table. This format is suitable for cartoons and for produced computer art - supports transparency.

Portable Network Graphics (PNG) is using a lossless data compression. PNG is an open-source format, which was created to improve GIF's algorithm.

Current compression methods are not final. Interesting issues, such as obtaining accurate models of images, optimal representations of these models and rapid processing of those representations are the greatest challenges we face.

3. Proposed Method

The Local Binary Pattern (LBP) method is used to summarize the local structure of images. It labels the image pixels and creates a binary number used for classification in computer vision. This method "looks" at each pixel and tests it with its neighborhood. LBP is tolerant to monotonic illumination changes. The important property is the computational simplicity, which makes it possible to analyze images in challenging real-time settings.

Timo Ojala [5,8] introduced the Local Binary Pattern (LBP) as a fine scale-texture descriptor. LBP is the particular case of the Texture Spectrum model proposed in 1990 [9].

The LPD method divides the image into small 3x3-pixel matrix; each pixel is then compared to its eight neighbors. Then the LBP operator creates label's forms for the image pixels by looking at the surrounding points (of a central point) and tests whether the surrounding points are greater than the central point resulting in a binary result. It is applied to a 3x3 matrix neighborhood, which means that there are 8 points of comparison thus it produces an 8-bit binary number ($2^8 = 256$ different labels), which will be converted to decimal and used as a texture descriptor. This operator is effective in the computing plan because it works only with integer arithmetic (allows to reach real-time productivity in some problems), and also invariant to brightness changes on the image, it may be caused by shooting in various conditions of illumination.

As shown in [1, 2], Local Binary Pattern technique (LBP) allows to describe the image texture using a vector with various sizes, depending on the LBP version used. In the classic version, LBP describes a 3x3-texture image area through a single pixel. This pixel defines the border zone relative to the center pixel from the 3x3 area. The values obtained from the entire image texture analysis, lead to a vector with a 256 values size that uniquely defines the original image, which came from. Using this concept, the authors proposed a new image compression technique, which is called Local Binary Compressed format (LBC). The compression technique described in [1, 2] consists in defining the image through two complementary images (by the average and by the dispersion values) of much smaller size obtained by applying the Local Binary Pattern technique method. These two images obtained from analyzing the original image are used for image compression or image decompression [10].

3.1. Image compression

Based on the LBP value constructing mode in the base form [4] previously described, the description for a grayscale image is proposed using two smaller size images [1, 2]:

The first image represents the central values from the 3x3 size areas, which are used in building the LBP value. This image is seen as representing an average pixel defining image texture.

The second image represents the LBP values obtained using the basic version of this technique. This set of values is the ratio (the pixel has a higher or lower value) between the pixels from the 3x3 border area and the central pixel. For this reason, this image is seen as a pixels "dispersion" from the border relative to the center pixel area.

The two images sets represents the compressed image seen in Figure 1.

3.2. Image decompression

Based on the two images – the average and "dispersion" values – the original image can be reconstructed by generating pixels around the average value. The generation process uses existing information in the "dispersion" image. The reconstruction around the average pixels value can use two techniques. Both techniques assume that the distribution pixel values law around the mean value can be a uniform distribution law or a Gaussian distribution law. In both cases, it is necessary to define a dispersion factor [11].

The dispersion factor is used for building the time interval that pixels can take values. Dispersion factor is a numerical percentage applied to the average value type used for constructing the range of pixel dispersion around the mean value. After defining this range, based on the distribution law, pixels values can be generated, then can be added or removed from the central pixel value, depending on the operation direction specified by the attached "dispersion" image (Figure 1).

The factor's dispersion size and the distribution law are experimentally determined. The criterion in scope is to obtain a similarity index as big as possible between the original image and the resulted image after the compression / decompression. The study was conducted using the ultrasound medical image database containing 6264 JPEG images [12, 13]. All the images from this database were compressed and reconstructed using the [0.1 - 2] % interval dispersion around the LBP range, which was followed by a step of 0.1 in the Gaussian and uniform distribution case. In figures 2a and 2b, the similarity index variation can be seen when using Gaussian or uniform distribution that is used for calculating the dispersion interval of LBP values. It can be noted, for a value of one, for the dispersion range - percentage constant - around the LBP value was obtained a maximum value for the similarity index.

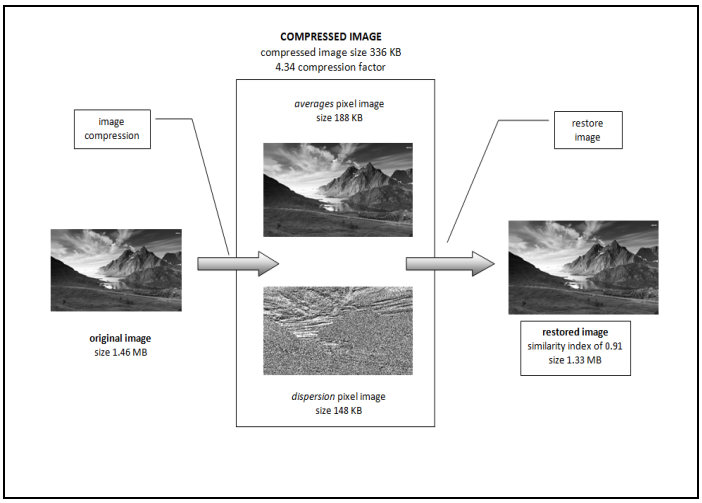


Figure 1. The principle of representing the LBC image format

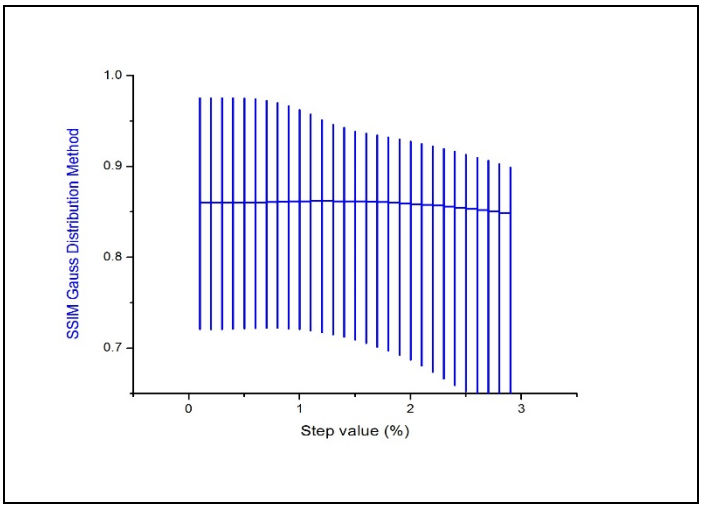


Figure 2a

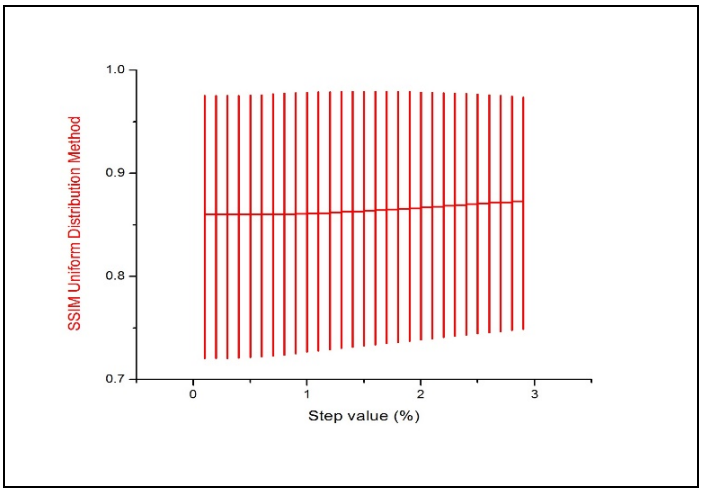


Figure 2b

Figure 2a, 2b Similarity index variation for Gaussian or uniform distribution, which is used for calculating the dispersion range of LBP values.

It has been shown that compression index value and the distribution law depend on the image nature and their formats (JPEG, PNG, BMP, JPEG.).

4. Experimental results

Medical images, regardless their source (radiographs, ultrasound, various types of CT scans) have a characteristics set which distinguish them from the photographic images taken with a modern photo equipment. The main feature refers to the fact that they are large images and for the same medical condition, there is a relatively large number of images taken from different angles, trying to capture that medical condition. In general, these are greyscale images. Color images origin from special histological analysis cases or from special ultrasound cases.

The study was carried out on two types of medical images, radiography and ultrasound images. The Structural Similarity Index Metric (SSIM) is used for measuring the similarity between two images and was used [12 - 17] to assess the similarity between the original image and the reconstructed one. For each image types like BMP, JPEG, GIF and PNG formats were considered. The analysis was carried with a dispersion factor of one for the uniform distribution as well as for the Gaussian distribution; a tool developed by the authors was used. A number of six medical images were taken to be analyzed - ultrasound and radiography images.

4.1. X-rays medical image compression

For the experiments, two types of images were taken into consideration: X-ray medical image in JPEG format, for the lossy image compression technique and X-ray medical image in BMP format, for the lossless image compression technique. The obtained results are shown in the following tables: 1a and b. The images taken into account for this study are presented in Figure 3.

Figure 3 X-ray medical image

Table 1a. X-ray image compression on BMP format

Image	OrigSize (KB)	Compression index (LBP method)	SSIM index Uniform distribution	SSIM index Gaussian distribution
a	550.223	4.038	0.902	0.862
b	357.528	4.091	0.958	0.911
c	414.877	4.167	0.950	0.909
d	603.895	4.318	0.868	0.828
e	536.009	4.020	0.873	0.832
f	365.305	3.678	0.948	0.913

Table 1b. X-ray image compression on JPEG format

Imagine	OrigSize (KB)	Compression index (LBP method)	SSIM index Uniform distribution	SSIM index Gaussian distribution
a	514.960	3.895	0.896	0.858
b	380.746	3.966	0.953	0.913
c	387.021	3.578	0.951	0.914
d	720.333	4.310	0.856	0.813
e	642.341	3.970	0.865	0.823
f	383.333	3.838	0.884	0.849

Table 1c. X-ray image compression on GIF format

Imagine	OrigSize (KB)	Compression index (LBP method)	SSIM index Uniform distribution	SSIM index Gaussian distribution
a	482.378	2.085	0.883	0.835
b	417.590	2.174	0.918	0.872
c	434.548	2.176	0.845	0.806
d	544.431	2.185	0.862	0.831
e	672.021	2.209	0.892	0.864
f	548.002	2.137	0.855	0.844

Table 1d. X-ray image compression on PNG format

Imagine	OrigSize (KB)	Compression index (LBP method)	SSIM index Uniform distribution	SSIM index Gaussian distribution
a	550.223	4.038	0.902	0.862
b	357.528	4.091	0.958	0.911
c	414.877	4.167	0.950	0.910
d	544.431	2.185	0.862	0.831
e	536.009	4.020	0.873	0.832
f	420.034	3.986	0.884	0.855

For the experiments were also used GIF and PNG images format.

Analyzing the results, the compression ratio is in the range [2.1, 4.3] for the radiography images types; the similarity index context for uniform distribution is in the range [0.8, 0.96] and the Gaussian distribution it belongs to the range [0.8, 0.92].

4.2. Ultrasound medical image compression

Images, taken into account in this study, are those from Figure 4. Analyzing the results, only JPEG images format were used.

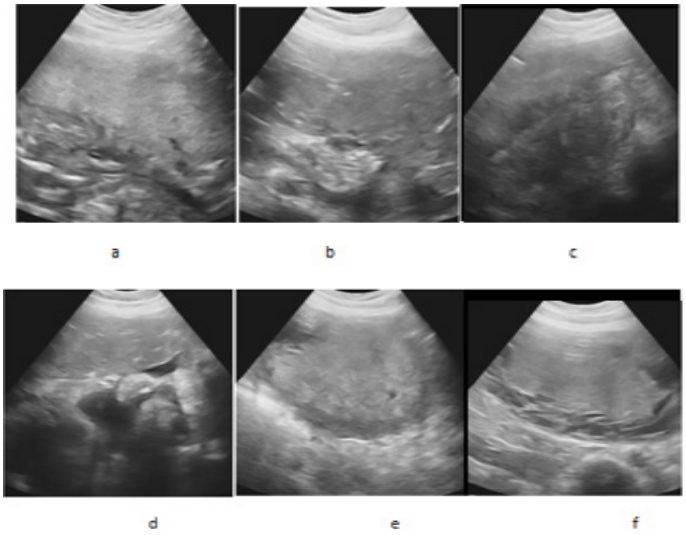


Figure 4. Ultrasound medical images

The results are shown in table 2:

Table 2. JPEG ultrasound medical images

Image	OrigSize (KB)	Compression index (LBP method)	SSIM index Uniform distribution	SSIM index Gaussian distribuion
a	642.341	3.970	0.865	0.823
b	720.333	4.310	0.856	0.813
c	471.541	3.885	0.948	0.907
d	387.021	3.578	0.951	0.914
e	514.960	3.895	0.896	0.858
f	320.896	3.851	0.960	0.913

The calculations made to determine the static medical image format influence on the compression process led to the results presented in Table 3. It can be observed that the index compression has the lowest value for the JPG image format (around 3.6), a little higher for BMP image format (around 3.99) and for PNG image format (around 4.01). Behavior is the opposite when we talk about the similarity index, so JPG images show the highest similarity index, and BMP, PNG images have a slightly lower similarity. We can assume that the image format slightly influences the compression index and related similarity regardless of the distribution law type used to reconstruct the compressed image (uniform or normal) [18].

Image format	Compression index	SSIM Uniform Distribution Method	SSIM Gauss Distribution Method
BMP	4.013	0.899	0.914
JPG	3.593	0.912	0.933
PNG	3.999	0.899	0.915

Table 3 Relationship between image format, compression factor and similarity index.

5. Conclusions and future work

Images have become necessary and useful nowadays. The image quality and clarity was developed and improved over time. The image formats performance has increasingly improved over time because our society has become digital. The most used image formats are still widely used because of Microsoft Company - JPEG, BMP, GIF and PNG.

Digital data almost exponential growth is due to the computer technology progress from the last decade and because of the continuous improvement or creation of new algorithms for processing digital data, particularly digital images.

Local Binary Pattern operator is one of the tools that led to improvement in the quality and performance o image compression / decompression, and it is a statistical method and structural analysis for improving the image texture.

One of the many LBP operator benefits is that the topologies can be changed and was changed, according to our app, to get the best results. The method presented so far, use the LBP operator to analysis the static images; there is a controller for dynamic images analysis named space-time. It is used to analyze moving textures, patterns along the proposed method (volume) on the three orthogonal planes.

Compression with the LBC image technique, which is defined by two components (an average value and a dispersion value calculated with the LBP technique type), had better result when for restoring the original image a dispersion factor of 1 was used, in the uniform distribution context used for generating the pixel values.

1. The proposed method allows obtain a compression factor in the range [2, 4.5] for medical images. The compression index depends on the image format (PNG, JPEG and BMP). The similarity index obtained belongs to an interval [0.81, 0.94]. From this point of view, the method is part of the lossy compression methods [18]. The study led to the following conclusions:

- The maximum indices performance (compression factor, SSIM similarity index) was obtained when the value 1 is used as the constant that determines the LBP dispersion range size.
- It was determined that the best law that can be used to reconstruct the pixel values is the uniform one. Between the use of the uniform distribution law and the Gaussian distribution law for the reconstruction of pixels, studies have shown that there are small differences between the two laws with a higher quality (compression factor, SSIM similarity index) for the uniform distribution law cases.
- Studies have shown that the original image type that is compressed (the most commonly used medical image formats - BMP, JPG and PNG) do not significantly affect the compression process. However, a conclusion is that BMP and PNG images lead to a higher compression factor with a decrease on the similarity index insignificant.

2. The LBC compression method lends itself to further developments; the similarity index can be increased by using new versions of the existing extraction technique – Local Binary Pattern feature

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

Politehnica University Timisoara, Romania.

References

[1] Ildikó-Angelica Szöke, Vasile Stoicu-Tivadar, Diana Lungeanu, "*A comparative study of using the LBC format for compression medical images*", Applied Machine Intelligence and Informatics (SAMI), 2017 IEEE 13th International Symposium; 10.1109/SAMI.2017.7880285.

[2] Ildikó-Angelica Szöke, Diana Lungeanu, Stefan Holban, "*Image Compression Techniques Using Local Binary Pattern*", Applied Machine

Intelligence and Informatics (SAMI), 2015 IEEE 13th International Symposium; 10.1109/SAMI.2015.7061863; pag. 139-143, 2015.
[3] R.Devakuncha, *Analysis on big data over the years*, International Journal of Scientific and Research Publications, Volume 4, Issue 1, January 2014.
[4] Sonal, Dinesh Kumar ,*" A study of various image compression techniques*", Guru Jhambheswar University of Science and Technology, Hisar , 2005.
[5] T. Ojala, M. Pietikäinen, and D. Harwood, "*A Comparative Study of Texture Measures with Classification Based on Feature Distributions*", Pattern Recognition, vol. 29, pp. 51-59, 1996.
[6] Matti Peitkinen, Abdenour Hadid, Guoying Zhao, Timo Ahonen- *Computer Vision Using Local Binary Patterns*, Computational Imaging and Vision, Volume 40 Springer-Verlag London Limited 2011.
[7] Loris Nannia, Alessandra Lumini, Sheryl Brahnam, *Local binary patterns variants as texture descriptors for medical image analysis*, Artificial Intelligence in Medicine, 49(2), pp. 117-125, 2010.
[8] T.Ojala, *Multiresolution gray-scale and rotation invariant texture classification with local binary patterns*, IEEE Transactions On Pattern Analysis And Machine Intelligence, Vol. 24, Issue. 7, July 2002.
[9] Wei - Yi Wei ,"*An Introduction to Image Compression*" National Taiwan University, Taipei, Taiwan, ROC, 2008.
[10] G.M.Padmaja, P.Nirupama, *Analysis of Various Image Compression Techniques*, ARPN Journal of Science and Technology, vol. 2, no. 4, pp. 371-376, may 2012.
[11] Zhang, Hong., Zhang, Xiaofei & Cao, Shun," *Analysis & Evaluation of Some Image Compression Techniques*", High Performance Computing in Asia Pacific Region, 2000 Proceedings, 4th Int. Conference, Vol. 2, pp 799-803, may 2000.
[12] Z. Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli, *Image quality assessment: From error visibility to structural similarity*, IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600-612, 2004.
[13] Wang, Zhou; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P.,*Image quality assessment: from error visibility to structural similarity*, IEEE Transactions on Image Processing, 13 (4), 2004, pp. 600–612.
[14] Dosselmann, Richard; Yang, Xue Dong , *A comprehensive assessment of the structural similarity index. Signa*l, Image and Video Processing 5 (1), 2009, pp. 81–91
[15] Fuxiang Lu, Jun Huang, *An improved local binary pattern operator for texture classification,* Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on, 20-25 March 2016.
[16] L Liu, P Fieguth, Y Guo, X Wang, M Pietikäinen, *Local binary features for texture classification: Taxonomy and experimental study*, Pattern Recognition Volume 62, February 2017, Pages 135–160.
[17] D Liu, S Wang, D Huang, G Deng, F Zeng, *Medical image classification using spatial adjacent histogram based on adaptive local binary patterns,* Computers in Biology and Medicine Volume 72, 1 May 2016, Pages 185–200.
[18] Ildikó-Angelica Szöke, "Contributions to the digital medical images compressions," Ph.D. Thesis, University Politehnica Timisoara, 2017.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1478-1486 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Image Navigation Using A Tracking-Based Approach

Stephen Craig Stubberud[1], Kathleen Ann Kramer[2,*], Allen Roger Stubberud[3]

[1]*Advanced Programs, Oakridge Technology, San Diego, 92121, USA*

[2]*Department of Electrical Engineering, University of San Diego, San Diego, 92110, USA*

[3]*Department of Electrical and Computer Engineering, University of California, Irvine, 92697, USA*

ARTICLE INFO

Article history:
Received: 31 May, 2017
Accepted: 20 July, 2017
Online: 10 August, 2017

Keywords:
Image correlation
Image navigation
Bearings-only tracking
Phase-only filter

ABSTRACT

Navigation is often performed using the Global Positioning Satellite system. However, the system is not always reliable for use at all times in all contexts. Reliance solely on such systems could endanger a vessel or get people lost. Another approach to provide navigation is to use known landmarks and triangulate one's position. A navigation solution can be automatically generated using image data. In this paper, a target-tracking based approach for navigation using imagery is developed. The technique is based on an extended Kalman filter that uses angle measurements only. The technique is developed to use landmarks known a priori and landmarks that are first tracked by the vessel navigating and then used as known landmarks. The issues of observability to the development are considered. Image tracking from video frame to video frame is a further element. The frame-to-frame image correlation technique, the phase-only filter, is used to provide a low uncertainty estimate of the position of elements based upon the image frame. This provides an estimate of bearing angle that can be combined with a bearing from other known landmarks to fix and track the platform to provide a navigation solution.

1. Introduction

Visual cues used in navigation have a very long tradition, as is evident from Homer's *The Odyssey* [1],

At night he (Odysseus) never closed his eyes in sleep, but watched the Pleiades, late-setting Bootes, and the Great Bear that men call the Wain, that circles in place opposite Orion, and that never bathes in the sea. Calypso, the lovely goddess, had told him to keep that constellation to larboard as he crossed the waters.

Besides the stars, landmarks such as Poseidon's Temple at Cape Sunion, the Lighthouse of Alexandria, and the Roman lighthouse of Dover Castle provide testament to ancient use of upon a relative point. Using visual cues for navigation has long fixed reference points from which to understand location based been a standard procedure to ensure that ships could know where they were and get to desired locations.

Today, the primary navigational resource is GPS for systems from cell phones to vehicle navigation systems. In some cases, GPS signals are used to direct autopilot systems. But, at times, total reliance on GPS can be a detriment to navigation. There are places where the loss GPS can play havoc with a commuter. GPS signals are not available in urban canyons [2], such as in the heart of New York City or in Boston's Ted Williams Tunnel near Logan airport.

GPS also can be jammed or spoofed. In [3], the weakness and danger of sole reliance on GPS was presented. In 2011, a United states spy drone was reported captured by Iran through jamming and spoofing of the GPS signal [4]. Even if untrue, the fact that it could be considered a plausible scenario, it shows the need for other navigation system in places when vessel or individual needs to navigate in precarious places.

*Corresponding Author: Kathleen Ann Kramer, Department of Electrical Engineering, University of San Diego, San Diego, 92110, USA
Email: kramer@sandiego.edu

For their submarine fleet, the uses "eyes-on" navigation and piloting for the ingress to or egress from a port. This kind of navigation of a submarine through the channel is both time and human-power intensive. This procedure uses a labor-intensive visual-based approach with a company of sailors stationed on the sail and hull of the vessel to point out channel markers, landmarks, and hazards. Information on what they see is relayed to the captain who can then act upon the information. Besides the potential dangers with relying on GPS, other navigation capabilities available to the submarine are not capable enough for navigation. Standard inertial navigation systems (INS are not currently accurate enough for such tight piloting [5]. Other types of sensors such as electromagnetic log and velocity sonar are unreliable in littoral conditions [5]-[7].

A number of techniques have been developed to overcome the issues that can arise relying solely on GPS navigation. Many incorporate inexpensive INS [2],[8]. It is well known, though, that these inertial systems have drifts and biases that without corrections will grow [8], [9]. For short periods of time and for operations where an human can control the motion of the navigating object easily, these systems are useful. To compensate for the drifts and biases, some of these non-GPS reliant navigators uses other sensors. In [10], the uses of imagery of the helmet that contains INS is used to compensate for the INS's biases and drifts. Some systems, such as unmanned autonomous vehicles or drones, may be equipped with vision-based systems (video capabilities) rather than INS elements. In [11], [12] are typical techniques that use imagery for navigation. They use the image to provide a map over time. Complex image processing is then used to determine the relative position of the navigating object. In [13], [14], the concept of navigation using a target tracking approach based on the imagery was developed.

Unlike imaging techniques, such as [15], [16], the proposed method does not rely on an artifically-generated range estimate but instead on the angle-only information. Angle-only navigation is a common littoral navigation technique [17]. It is can also be considered the dual angle-only tracking [18]. In the tracking problem, the sensor platform has a known location and velocity while the targets' kinematics are unknown. For angle-only navigation, landmarks with known fixed locatios are known. The sensor platform which wants to navigate uses its measured relationship to the landmarks to provide a crossfix of its positio. By repeating the crossfix estimate over time, a navigation track of the vessel can be computed. This method of navigation based upon tracking can handle obfuscation of individual landmarks, and can be used to locate new landmarks.

In [14], the concept of new landmarks being added to the existing set was discussed. That work is extended with the development of the approach and analysis of the performance.

In Section 2, the concept and effects of observability for angle-only navigation are presented. The use of the Phase-only Filter (POF) for the frame-to-frame image tracking and bearing generation is discussed in Section 3. Section 4 provides the improved navigation technique, including its functional flow. In Section 5, the bearings-only Kalman filter used within the technique is summarized along with the procedure to switch from tracking to navigation. Then this is then applied to frame-to-frame tracking in Section 6. An analysis of a simulation a maritime channel with multiple located landmarks some known and some unknown is provided.

2. Observability and Analysis Summary

2.1. Maintaining the Integrity of the Specifications (Heading 2)

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

Observability provides the ability to determine the state of system from the measurements. In [19], the standard observability Gramian for time-invariant linear system defined as

$$\mathbf{x}_{k+1} = \mathbf{F}\mathbf{x}_k + \mathbf{G}\mathbf{u}_k \tag{1}$$

$$\mathbf{y}_k = \mathbf{H}\mathbf{x}_k \tag{2}$$

was shown to be

$$\mathbf{G}_O = \begin{bmatrix} \mathbf{H} \\ \mathbf{HF} \\ \mathbf{HF}^2 \\ \vdots \\ \mathbf{HF}^{n-1} \end{bmatrix}, \tag{3}$$

if $\mathbf{G_o}$ is of maximal rank then the system is considered fully observable.

The image navigation problem defined for this problem is more complex. The measurement is defined as

$$\mathbf{z}_k = \beta_{i,k} = \text{atan}\left(\frac{y_{LMi} - y_{platform}}{x_{LMi} - x_{platform}}\right) \tag{4}$$

where the subscript *LMi* denotes the ith landmark that can be used by the sensor platform. As seen in Figure 1, the platform and two landmarks are present. The measurement to a single landmark is an angle, which is a nonlinear relationship from the landmark position and the platform position to the measurement space. As shown in [20], one way to obtain observability is to use multiple landmarks. This provides the well-known concept of a cross-fix as seen in Figure 1.

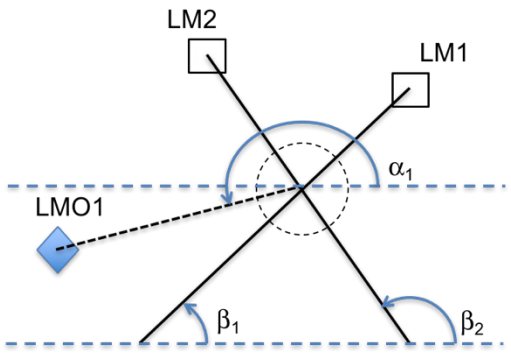


Figure 1. Example of creating a position of a nonsurveyed landmark to be used for crossfixing when one of the surveyed landmarks is no longer visible.

The navigation-solution initialization is created using bearing angles and knowledge of landmark locations:

$$x_{platform} = \left(\left(y_{LM} - y_{LM2}\right) - \left(\text{atan}\left(\beta_2\right)x_{LM2} - \text{atan}\left(\beta_1\right)x_{LM1}\right)\right) \cdot \left(\frac{1}{\text{atan}\left(\beta_2\right) - \text{atan}\left(\beta_1\right)}\right)$$

$$y_{platform} = y_{LM2} - \text{atan}\left(\beta_2\right)\left(x_{LM2} - x_{platform}\right)$$

Since the horizontal location of the platform contains the information of both landmarks, only a single landmark's information is necessary for the vertical component.

Figure 2 shows the old harbor entrance at La Rochelle, France. The two towers would be used as fixed landmarks on egress. The two channel markers in the front of the picture would make landmarks of opportunity as they are not as permanent structures as the towers would be.

For a landmark of opportunity, the structure being used must be fixed in position. The location of a fixed target can be determined by a moving platform. If the platform's location is known at two time points, then the landmark of opportunity can have its position generated with a crossfix:

$$x_{LMO1} = \left(\left(y_{platform}\left(t_2\right) - y_{platform}\left(t_1\right)\right) - \left(\text{atan}\left(\alpha_2\right)x_{platform}\left(t_2\right) - \text{atan}\left(\alpha_1\right)x_{platform}\left(t_1\right)\right)\right) \cdot \left(\frac{1}{\text{atan}\left(\alpha_2\right) - \text{atan}\left(\alpha_1\right)}\right)$$

$$y_{LMO1} = y_{platform}\left(t_2\right) - \text{atan}\left(\alpha_2\right)\left(x_{LMO1} - x_{platform}\right)$$

Crossfixing works very well for the case of small measurement uncertainties. However, since the platform is moving, a tracking algorithm, such as an extended Kalman filter (EKF) should be used to capture not only the position but the velocity dynamics. To capture the landmark of opportunity a tracking solution should also be used because of the uncertainty in the image sensor and the platform location. An EKF will smooth the error estimation noise than the single cross-fixes described in (5) - (8). To measure the observability of both the navigation solution and the landmark of opportunity position, the observability Gramian

$$\mathbf{G}_O = \sum_{k=k_0}^{k_f} \mathbf{F}^T\mathbf{H}^T\mathbf{H}\mathbf{F} \tag{9}$$

is used. Since the measurement is the bearing angle between the navigating platform and the ith landmark, β_i from (13) and (14), the measurement-coupling matrix **H** becomes the linearized Jacobian of bearing angle with respect to each position's coordinate direction of the platform position for navigation

$$\mathbf{H} = \begin{bmatrix} \frac{\partial \beta_i}{\partial x_{platform}} & 0 & \frac{\partial \beta_i}{\partial y_{platform}} & 0 \end{bmatrix}$$

where the partial derivatives of each angle to the fixed, known i^{th} landmark with respect to the platform coordinates is

$$\frac{\partial \beta_i}{\partial x_{platform}} = \frac{\left(y_{LMi} - y_{platform}\right)}{\left(1+\Delta^2\right)\left(x_{LMi} - x_{platform}\right)^2}$$

$$\frac{\partial \beta_i}{\partial y_{platform}} = \frac{-1}{\left(1+\Delta^2\right)\left(x_{LMi} - x_{platform}\right)^2}$$

$$\Delta = \frac{y_{LMi} - y_{platform}}{x_{LMi} - x_{platform}}$$

A similar Jacobian is generated with respect the landmark of opportunity's position for the tracking problem.

$$\frac{\partial \beta_i}{\partial x_{LMOi}} = \frac{-\left(y_{LMi} - y_{platform}\right)}{\left(1+\Delta^2\right)\left(x_{LMi} - x_{platform}\right)^2} \tag{13}$$

$$\frac{\partial \beta_i}{\partial y_{LMOi}} = \frac{1}{\left(1+\Delta^2\right)\left(x_{LMi} - x_{platform}\right)^2} \tag{14}$$

where Δ is defined as above.

The state-coupling matrix **F** is

$$\mathbf{F}_{Nav} = \begin{bmatrix} 1 & dt & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & dt \\ 0 & 0 & 0 & 1 \end{bmatrix}, \tag{15}$$

for the navigation solution, while it is defined as

$$\mathbf{F}_{LMOtrack} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{16}$$

for the landmark tracking problem. The uncertainty of the sensor and platform location is incorporated by modifying (9) to be

$$\mathbf{G}_O{}^T \left(\mathbf{R}+\mathbf{I}\right)^{-1} \mathbf{G}_O, \tag{17}$$

where

$$\mathbf{R}_k = E\left[\left(\mathbf{z}_k - a\tan\left(\frac{y_{LMi}-y_{platform}}{x_{LMi}-x_{platform}}\right)\right)\cdot \left(\mathbf{z}_k - a\tan\left(\frac{y_{LMi}-y_{platform}}{x_{LMi}-x_{platform}}\right)\right)^T\right]$$

$$= E\left[\mathbf{z}_k\mathbf{z}_k^T - \left(2\mathbf{z}_k a\tan\left(\frac{y_{LMi}-y_{platform}}{x_{LMi}-x_{platform}}\right)\right) + a\tan\left(\frac{y_{LMi}-y_{platform}}{x_{LMi}-x_{platform}}\right)^2\right]$$

$$= E\left[\mathbf{z}_k\mathbf{z}_k^T\right] - E\left[\left(2\mathbf{z}_k a\tan\left(\frac{y_{LMi}-y_{platform}}{x_{LMi}-x_{platform}}\right)\right)\right] + E\left[a\tan\left(\frac{y_{LMi}-y_{platform}}{x_{LMi}-x_{platform}}\right)^2\right]$$

$$= E\left[\mathbf{z}_k\mathbf{z}_k^T\right] + E\left[a\tan\left(\frac{y_{LMi}-y_{platform}}{x_{LMi}-x_{platform}}\right)^2\right]$$

$$\approx \mathbf{R} + f\left(\frac{\sigma(y_{LMi})}{r}, \frac{\sigma(x_{LMi})}{r}, \beta\right) \tag{18}$$

$$\approx \mathbf{R} + f\left(\frac{\sigma(y_{LMi})}{\hat{r}}, \frac{\sigma(x_{LMi})}{\hat{r}}, \hat{\beta}\right)$$

where **R** is the uncertainty covariance of the image measurement which models the sensor uncertainty. The added function f incorporates the uncertainty of the tion of the sensor platform (navigating vessel) for the Landmark of Opportunity tracking problem and, for the navigation problem, the Landmark of Opportunity's own position uncertainty. This function is based on the estimated range from the navigating platform to the landmark and relative bearing from the navigating platform to the landmark. Since these two values are not known, they must be estimated from the filter solutions.

Figures 2 - 5 show the bearing uncertainty for four different relative platform-to-landmark angles. These angles are 0º, 30º, 60º, and 90º. The x-coordinate depicts the estimated position error of the platform (for tracking) or the landmark (for navigation) in relation to the actual range. The x-position error is in percent of the range. Figures 2-5 each depict 41 different plots corresponding to the range error in the y-direction. The range of error starts at -50% of the range and, in 2.5% steps, goes to 50% error in range.

These graphs were created by generation of 10000 Monte Carlo runs to compute the bearing error for each x-y position error. The plots indicate the expected symmetry in the position errors from -50% position error to +50% position error in each direction. As the position error in one direction gets smaller and nearer to direction of dominance (0º for the x-direction and 90º for the y-direction), the bearing error diminishes.

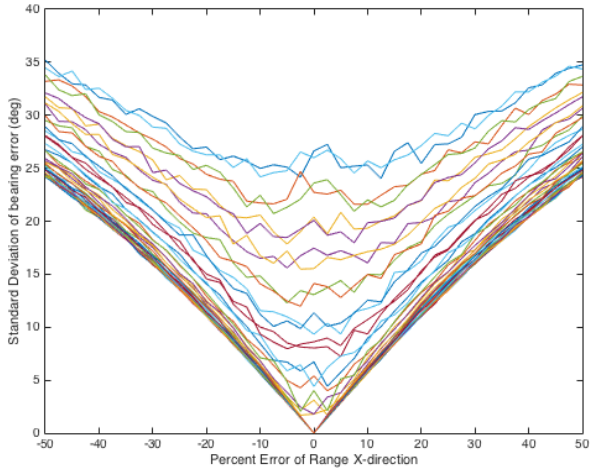


Figure 2. Bearing uncertainty (target is at 0º from the platform) based on the x-direction error in position.

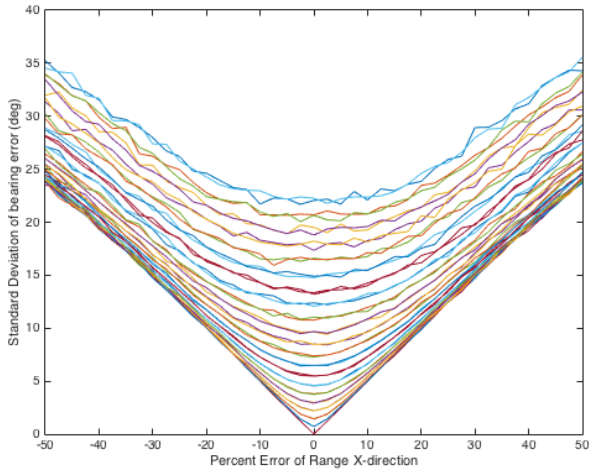


Figure 3. Bearing uncertainty (target is at 30º from the platform) based on the x-direction error in position.

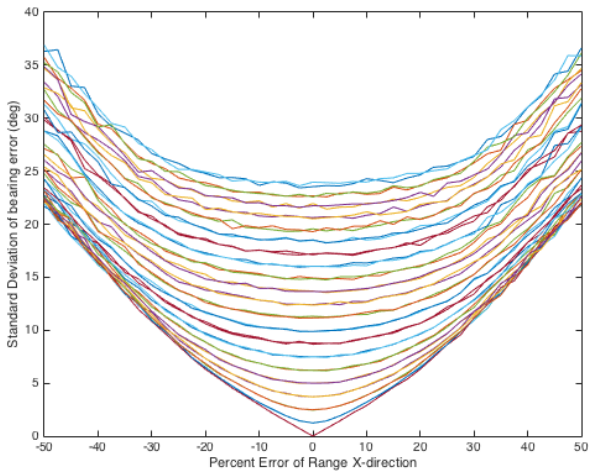


Figure 4. Bearing uncertainty (target is at 60º from the platform) based on the x-direction error in position.

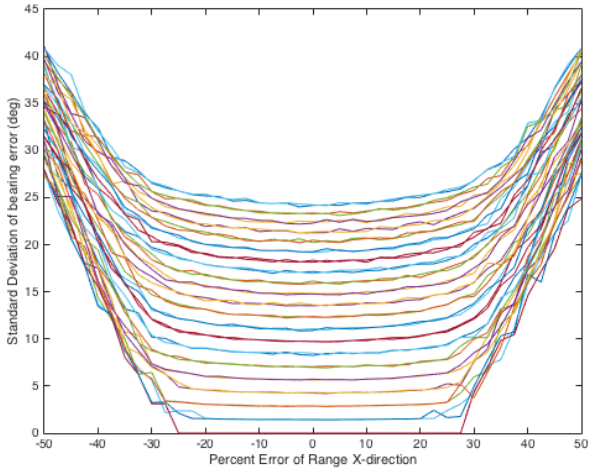


Figure 5. Bearing uncertainty (target is at 90º from the platform) based on the x-direction error in position.

3. The Phase-Only Filter

To generate the bearing measurements, the processor must know the orientation of the image sensor from the front of the navigating vessel. For a submarine this would be the rotation angle of the periscope from the bow of the submarine. This provides the center of the image frame. The landmarks in the image frame are then measured from the centroid to provide the angle adjustment from the pointing angle of the sensor.

The approach that has been chosen to generate the image angle uses a two-dimensional image correlation process referred to as the phase-only filter (POF). The POF is a variant the Fourier correlation method. Using a two-dimensional Fast Fourier Theorem (FFT) algorithm, two images can be correlated element-by-element with speed and efficiency. Despite the large number of operations necessary to do such correlation, the process is still applicable to real-time situations. The frequency-domain version of the correlation function [21] is defined as (19):

$$\begin{aligned}&\sum\sum g^{*}(x,y)*f(x+X,y+Y)\\&=\sum\sum G^{*}\left(k_x,k_y\right)F\left(k_x,k_y\right)e^{-j\left(k_xX+k_yY\right)}\\&=\mathrm{Re}\left\{\mathbf{F}^{-1}\left\{G^{*}\left(k_x,k_y\right)F\left(k_x,k_y\right)\right\}\right\}\end{aligned}$$

While the FFT approach may seem useful, the actual physics of an image correlation problem using an optical correlator show that primarily the comparable information of images lies in the phase of the frequency information [21]. Mathematically this causes the amplitude information of the input image to be normalized. Thus, the correlation function becomes the phase-only filtered image correlation:

$$\sum\sum G^{*}\left(k_x,k_y\right)\frac{F\left(k_x,k_y\right)}{\left|F\left(k_x,k_y\right)\right|}e^{-j\left(k_xX+k_yY\right)}$$

$$=\mathrm{Re}\left\{\mathbf{F}^{-1}\left\{G^{*}\left(k_x,k_y\right)\frac{F\left(k_x,k_y\right)}{\left|F\left(k_x,k_y\right)\right|}\right\}\right\}$$

The image correlation requires the template and the comparison image to be the same size, so padding the template image with zeroes to match the size of the scene image allows 2-D image correlation to indicate a match and also provides the location of the match within the scene.

For this application, the template would be the current image from the image sensor. While this seems backwards, the image being tracked is trying to be located in the image. As seen in Figure 6, Arabic letter pairs are being located on a license plate. The POF result of Figure 7 shows two peaks that align with the centroid of the letter pairs. The image angle will be computed as

$$\alpha=\left(num\_horz\_pixel-\frac{1}{2}\cdot total\_num\_pixels_{horiz}\right)\cdot angle\_ppixel \tag{21}$$

where the term *angle_ppixel*, the view angle per pixel, is computed from the optical view angle of the sensor and the number of pixels horizontally the sensor has:

$$angle\_ppixel=\frac{field\_of\_view\_angle}{total\_num\_of\_pixels_{horiz}} \tag{22}$$

The input image is from a database or from the previous frame. The POF image correlation has been demonstrated to be a reliable technique to match and localize an object within a scene in many efforts [22]. In [14], the improved localization of the image correlation centroid, because of a sharper correlation peak, was shown than that for the straight FFT correlation method. This reduction in centroid area reduces the uncertainty of the total image uncertainty.

Figure 6. Example of POF correlation. Two Arabic characters are looked for in a license plate. Two sets of the character pairs exist.

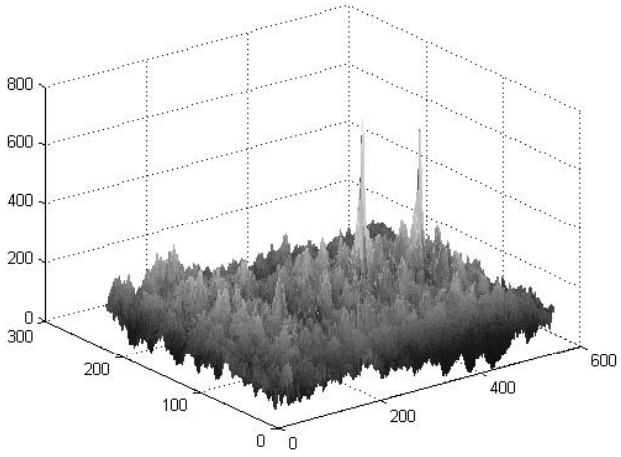


Figure 7. POF correlation result in two correlation peaks located at the center of each character pair.

4. Image-Based Navigation

Figure 8 outlines the functional flow of the overall image-based navigation.

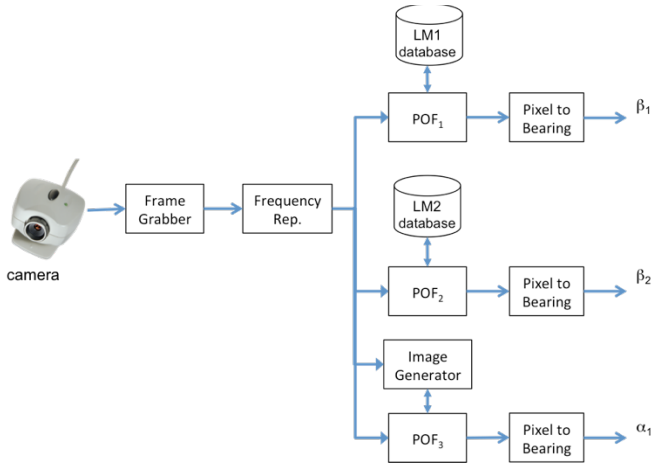


Figure 8. The functional flow of the image navigation system when unsurveyed landmarks are visible and can be used when surveyed landmarks are not available.

This section outlines the 10-step process, delving into the details of the components that this effort focuses on: image landmark tracking and image-based tracking. The steps are as follows:

Step 1: Initialize the navigation state of the platform. To determine which landmarks are surveyed in your database, a general knowledge of your location is required. This initialization need not be exact.

Step 2: Initialize the landmark database files. Once the known landmarks are defined, then the existing image databases for the local surveyed landmark positions are incorporated into the navigation algorithm.

Step 3: Grab the next video frame.

Step 4: Compute the angle of the electro-optical (EO) sensor with respect to the front of the vessel. This angle provides the relative angle of the center of the image with respect to the vessel.

Step 5: Correlate the known landmarks with the frame image. With the image data bases from Step 2 as the input to the POF algorithm and the video frame from Step 3 as the template, generate correlation peaks. If the peaks pass a predetermined threshold, the correlation peaks are considered a detection.

Step 5a: Calculate the angle of the landmark within the image. Using the computations from (21) and (22) for each correlation peak, the angle of the landmark with respect to the center of the image is computed.

Step 5b: Compute the landmark angle relative to the vessel. Add the angle from Stepb 5a to that of Step 4. This new angle is the measurement angle to the landmark.

Step 5c: Update or track image. As discussed in [13], the image is updated based on the previous frame to get a better aspect angle of the landmark for the next comparison. An image tracking algorithm can also be incorporated to estimate where the landmark should lie within the frame based on the estimated location of the vessel and the angle orientation of the sensor.

Step 6: Reinitialize the navigation solution with a crossfix. When two or more known landmarks are first available and their angles are computed, a crossfix initialization can be computed.

Step 7: Repeat Steps 5a – 5c. Start generating the angles for the navigation tracking algorithm.

Step 8a: Predict the estimation location of the vessel. Using the prediction equations of the EKF,

$$\mathbf{x}_{k+1|k} = \mathbf{F}_k \mathbf{x}_{k|k} \tag{24}$$

$$\mathbf{P}_{k+1|k} = \mathbf{F}_k \mathbf{P}_{k|k} \mathbf{F}_k^T + \mathbf{Q}_k \tag{25}$$

where **x** is the navigation state. The matrix **P** is the estimated error covariance of the state vector which indicates its quality. The state transition matrix **F** is defined in (15) while **Q,** the process noise is given as the integrated white noise process from [23],

$$\mathbf{Q}_k = q^2 \begin{bmatrix} \frac{dt^3}{3} & \frac{dt^2}{2} & 0 & 0 \\ \frac{dt^2}{2} & dt & 0 & 0 \\ 0 & 0 & \frac{dt^3}{3} & \frac{dt^2}{2} \\ 0 & 0 & \frac{dt^2}{2} & dt \end{bmatrix} \tag{26}$$

The location of where the vessel should be based on past measurements is the calculated result. This prediction takes the navigation estimate to the current time when a new measurement is available.

Step 8b: Update the estimation of the vessel's location. Using the bearing measurements generated in Step 5b and the estimated location from Step 8a for computing the Jacobian, the location of the vessel is updated. The equations for the EKF are defined as

$$\mathbf{K}_k = \mathbf{P}_{k|k-1}\mathbf{H}_k\Big|_{\hat{x}_{k|k-1}}^{T}\left(\mathbf{H}_k\Big|_{x_{k|k-1}}\mathbf{P}_{k|k-1}\mathbf{H}_k\Big|_{\hat{x}_{k|k-1}}^{T}+\mathbf{R}_k\right)^{-1} \quad (27)$$

$$\mathbf{x}_{k|k} = \mathbf{x}_{k|k-1} + \mathbf{K}_k\left(\mathbf{z}_k - \mathbf{h}\left(\mathbf{x}_{k|k-1}, \mathbf{x}^{LM}\right)\right) \quad (28)$$

$$\mathbf{P}_{k|k} = \left(\mathbf{I} - \mathbf{K}_k\mathbf{H}_k\Big|_{x_{k|k-1}}\right)\mathbf{P}_{k|k-1} \quad (29)$$

The matrix-vector **K** is referred to as the Kalman gain. The measurement uncertainty **R** is the composite measurement error defined in (18). The Landmark location is a fixed value. The superscript is used only in these equations to avoid confusion with the timing subscript.

Step 8c: If bearings of multiple landmarks are available, the update process of Step 8b is repeated. Since the landmarks are updated at the same time a zero-time-step prediction is used. This means that Step 8a is not repeated. Equations (28)-(29) are repeated but with the previously updated state values from the process of Step 8b.

Step 9: Define new landmarks of opportunity. As discussed with Figure 1, landmarks of opportunity are defined. Once identified, the image is placed in it own database.

Step 10a: Track each landmark of opportunity. An bearing-only stationary tracking is implemented. The EKF for this application is defined as

$$\mathbf{K}_k = \mathbf{P}_{k|k-1}\mathbf{H}_k\Big|_{\hat{x}_{k|k-1}}^{T}\left(\mathbf{H}_k\Big|_{x_{k|k-1}}\mathbf{P}_{k|k-1}\mathbf{H}_k\Big|_{\hat{x}_{k|k-1}}^{T}+\mathbf{R}_k\right)^{-1} \quad (29)$$

$$\mathbf{x}_{k|k} = \mathbf{x}_{k|k-1} + \mathbf{K}_k\left(\mathbf{z}_k - \mathbf{h}\left(\mathbf{x}_{k|k-1}, \mathbf{x}_k^{platform}\right)\right) \quad (30)$$

$$\mathbf{P}_{k|k} = \left(\mathbf{I} - \mathbf{K}_k\mathbf{H}_k\Big|_{x_{k|k-1}}\right)\mathbf{P}_{k|k-1} \quad (31)$$

$$\mathbf{x}_{k+1|k} = \mathbf{F}_k\mathbf{x}_{k|k} \quad (32)$$

$$\mathbf{P}_{k+1|k} = \mathbf{F}_k\mathbf{P}_{k|k}\mathbf{F}_k^T + \mathbf{Q}_k \quad (33)$$

The Jacobian **H** is defined by (10) while the target kinematic dynamics $\mathbf{F}_{LMOtrack}$ are defined by (16). The measurement uncertainty is defined not only by the sensor noise by also the location error of the navigation solution. The new measurement uncertainty **R** is calculated from (18).

Step 10b: Incorporate the new landmark into the available set. After the landmark has reach an acceptable level of accuracy, the tracking algorithm is stopped. The measurement noise is updated using (18). This measurement noise is used throughout the use of the new landmark. The new landmark and associated measurement noise is then used in Step 3 - 9. Note that a landmark of opportunity cannot be used in the initialization step as that would contain too many unknowns to solve the location calculations.

5. Analysis Scenario

In [14], a scenario using road data along United States Interstate 15 in the state of California was demonstrated to provide excellent performance when surveyed landmarks were available. In this effort, the analysis extends to a simulation of what occurs when landmarks of opportunity are incorporated into the scenario.

The example is a variant of the scenario developed in [20]. The scenario simulates a simple port entry and is illustrated in Figure 9. A platform wishing to use angle-only navigation heads north for 1000 s at 4.0 kts. The platform then heads east for 500 s at 4.0 kts. The platform starts at location 0, 0. There are four landmarks, Landmarks 1 – 4 in order are located at positions (1.5 nmi, 0.6 nmi), (0.375 nmi, -1.25 nmi), (0.5 nmi, 1.5nmi), and (2.2 nmi, -1.125 nmi), respectively. Landmarks 1 and 2 are defined as known landmarks with pre-surveyed locations. Landmarks 3 and 4 are defined as landmarks of opportunity. A 5.0° Gaussian random error was placed on each bearing measurement as the sensor component of the error.

After 60 seconds, the vessel starts tracking Landmarks 3 and 4 as landmarks of opportunity. This tracking continues to 850 seconds. After that point, the landmarks of opportunity are used with Landmarks 1 and 2. After 1000 seconds into the scenario, Landmarks 1 and 2 are no longer used. Only Landmarks of 3 and 4 are used through the rest of the scenario.

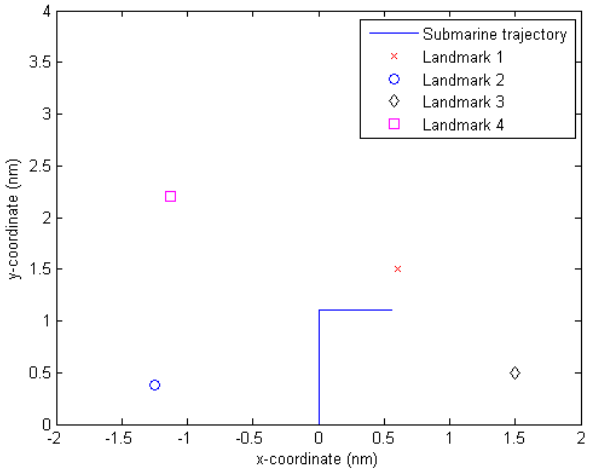


Figure 9. A Simple Scenario of a Submarine Navigating with Potentially Four Landmarks.

6. Analysis Results

The tracking error estimate provides approximately a 3.0% of error in the x-direction and 3.5% error in the y-direction for Landmark 3. For Landmark 4, a 2.0% error and 2.2% error in the

x and y directions, respectively are generated. Using the results in Figures 2 and 3, the added bearing error is approximately a 2.1° error. This makes the matrix **R** increase from 0.0076 radians-squared to 0.0090 radians-squared using (18). The tracking solution for Landmark 3 is tracked to be at position (0.492, 1.488). The tracking solution for Landmark 4 is tracked to be at position (2.296, -1.232).

The uncertainty for Landmark 3 navigation is defined as 23% for each direction. This results in a 12° added error to the measurement uncertainty. For Landmark 4 uncertainty in position is 4% and 4.5% in relation to the range in the x and y directions, relatively.

The results of the navigation accuracy are shown in Figure 10. As clearly seen, the navigation error is quite small when the surveyed landmarks are used. When only the landmarks of opportunity are used, error spikes appear. Deeper analysis indicates there are two reasons for this. The first is that Landmark 3 has too great of an uncertainty in its tracked location for it too be very effective. The second is that, at times near the end of the scenario, Landmarks 3 and 4 align reducing the observability.

The analysis indicates that the landmarks of opportunity require that the a figure of merit be used to determine if such a landmark is viable for navigation purposes. Also, the angular disparity between two landmarks is a necessary decision in selection.

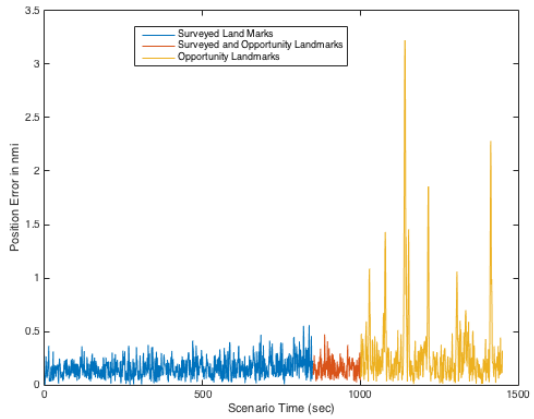


Figure 10. Position error of the navigation solution for three segments of the scenario: two surveyed landmarks, two surveyed landmarks and two landmarks of opportunity, and two landmarks of opportunity.

7. Conclusions

An image-based navigation system was developed using a bearings-only tracking system. The technique uses image correlation to measure the bearings by fixing on visible surveyed landmarks. The initial simulations and the test case have shown that this technique has merit as a navigation approach for ingress and egress operations where more standard forms of navigation are not available but surveyed locations in the region are. Tracking potential landmarks of opportunity can help if the tracking solutions are a quality level of accuracy.

In the next phase of this system's development, maritime data will be collected and utilized in testing. The results of this paper also indicate the need to develop figure of merit tool to determine if landmarks of opportunity can be incorporated into the approach.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] A. T. M. Homer, George E. Dimock, The Odyssey. Cambridge, MA: Harvard University Press, 1995.

[2] W. Soehren and W. Hawkinson, "Prototype personal navigation system," IEEE Aerospace and Electronic Systems Magazine, vol. 23, no. 4, pp. 10-18, April 2008. https://doi.org/10.1109/MAES.2008.4493437

[3] T. E. Humphreys, "Statement on the vulnerability of civil unmanned aerial vehicles and other systems to civil GPS spoofing," Subcommittee on Oversight, Investigations, and Management of the House Committee on Homeland Security, 2012.

[4] J. Leydon, "Iran spy drone GPS hijack boasts," The Register, 2013.

[5] H. Rice, S. Kelmenson, and L. Mendelsohn, "Geophysical navigation technologies and applications," in *Proceedings of the Position Location and Navigation Symposium - 2004.* IEEE, pp. 618-624. https://doi.org/10.1109/PLANS.2004.1309051

[6] I. B. Vaisgant, Y. A. Litvinenko, and V. A. Tupysev, "Verification of EM log data in marine inertial navigation system correction," Gyroscopy and Navigation, vol. 2, 2011, p. 34-38. https://doi.org/10.1134/S207510871101010X

[7] R. A. Smith, L. Russell, B. D. McKenzie, and D. T. Bird, "Synoptic data collection and products at the Naval Oceanographic Office," in OCEANS, 2001. MTS/IEEE Conference and Exhibition, vol. 2, 2001, pp. 1198-1203. https://doi.org/10.1109/OCEANS.2001.968283

[8] S. Moafipoor, D. A. Grejner-Brzezinska and C. K. Toth, "Multi-sensor personal navigator supported by adaptive knowledge based system: Performance assessment," 2008 IEEE/ION Position, Location and Navigation Symposium, Monterey, CA, 2008, pp. 129-140. https://doi.org/10.1109/PLANS.2008.4570049

[9] A. Kealy, N. Alam, C. Toth, T. Moore, et al, "Collaborative navigation with ground vehicles and personal navigators," in 2012 International Conference on Indoor Positioning and Indoor Navigation (IPIN), Sydney, NSW, 2012, pp. 1-8. https://doi.org/10.1109/IPIN.2012.6418893

[10] M. Harada, J. Miura and J. Satake, "A view-based wearable personal navigator with inertial speed estimation," in 2013 IEEE Workshop on Advanced Robotics and its Social Impacts, Tokyo, 2013, pp. 119-124. https://doi.org/10.1109/ARSO.2013.6705516

[11] D. Bradley, A. Brunton, M. Fiala and G. Roth, "Image-based navigation in real environments using panoramas," in IEEE International Workshop on Haptic Audio Visual Environments and their Applications, 2005, pp. 57-59. https://doi.org/10.1109/HAVE.2005.1545652

[12] P. Lukashevich, A. Belotserkovsky and A. Nedzved, "The new approach for reliable UAV navigation based on onboard camera image processing," in 2015 International Conference on Information and Digital Technologies, Zilina, 2015, pp. 230-234. https://doi.org/10.1109/DT.2015.7222975

[13] T. C. Glaser, N. Nino, K. A. Kramer and S. C. Stubberud, "Submarine port egress and ingress navigation using video processing," in 2015 IEEE 9th International Symposium on Intelligent Signal Processing (WISP) Proceedings, Siena, 2015, pp. 1-6. https://doi.org/10.1109/WISP.2015.7139153

[14] S. C. Stubberud, K. A. Kramer and A. R. Stubberud, "Navigation using image tracking," in 2016 IEEE International Conference on the Science of Electrical Engineering (ICSEE), Eilat, Israel, 2016, pp. 1-5. https://doi.org/10.1109/ICSEE.2016.7806050

[15] Z. P. Barbaric, B. P. Bondzulic, and S. T. Mitrovic, "Passive ranging using image intensity and contrast measurements," Electronics Letters, vol. 48, pp. 1122-1123, 2012. https://doi.org/10.1049/el.2012.0632

[16] M. Hirose, H. Furuhashi, and K. Araki, "Automatic registration of multi-view range images measured under free condition," in Proc. of the Seventh International Conference on in Virtual Systems and Multimedia, 2001, pp. 738-746. https://doi.org/10.1109/VSMM.2001.969738

[17] F. J. Larkin, Basic Coastal Navigation: An Introduction to Piloting. Dobbs Ferry, NY: Sheridan House, 1998.

[18] S. S. Blackman, Multiple-Target Tracking with Radar Applications. Dedham, MA: Artech House, 1986.

[19] M.S. Santina, A.R. Stubberud, G.H. Hostetter, Digital control system design, Saunders College Pub., 1994.

[20] S. C. Stubberud and K. A. Kramer, "Navigation using angle measurements," Mathematics in Engineering Science and Aerospace (MESA), **8**(2), 201-213, 2017.

[21] C. J. Stanek, B. Javidi, A. Skorokhod, and P. Yanni, "Filter construction for topological track association and sensor registration," in *Proc. of SPIE*, 2002, pp. 186-202. https://doi.org/10.1117/12.467794

[22] E. Ramos-Michel and V. Kober, "Correlation filters for detection and localization of objects in degraded images," in Progress in Pattern Recognition, Image Analysis and Applicatio*ns*. vol. 4225, J. Martínez-Trinidad, J. Carrasco Ochoa, and J. Kittler, Eds., ed: Springer Berlin Heidelberg, 2006, pp. 455-463. https://doi.org/10.1007/11892755_47

[23] Y. Bar-Shalom, X.-R.K.T. Li, Estimation with applications to tracking and navigation, Wiley, 2001.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1487-1490 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Analytical and Algorithmic Approaches to Determine the Number of Sensor Nodes for Minimum Power Consumption in LWSNs

Nukhet Sazak*, Ali Soner Kilinc

Sakarya University, Department of Electrical and Electronics Engineering, 54187, Turkey

ARTICLE INFO

Article history:
Received: 30 May, 2017
Accepted: 21 July, 2017
Online: 10 August, 2017

Keywords:
LWSNs
Power consumption
Energy efficiency

ABSTRACT

A Linear Wireless Sensor Network (LWSN) is a kind of wireless sensor network where the nodes are deployed in a line. Since the sensor nodes are energy restricted, energy efficiency becomes one of the most significant design issues for LWSNs as well as wireless sensor networks. With the proper deployment, the power consumption could be minimized by adjusting the distance between the sensor nodes which is known as hop length. In this paper, analytical and algorithmic approaches are presented to determine the number of hops and sensor nodes for minimum power consumption in a linear wireless sensor network including equidistantly placed sensor nodes.

1. Introduction

This paper is an extension of work originally presented in conference 2016 National Conference on Electrical, Electronics and Biomedical Engineering (ELECO 2016) "Eşit Mesafeli Çok-Sekmeli Kablosuz Algılayıcı Ağlarda Minimum Güç Tüketimi için Optimum Düğüm Sayısının Belirlenmesi"[1].

A Wireless Sensor Network (WSN) consists of a large number of low power, low cost, wirelessly communicating sensor nodes which can be deployed either inside the phenomenon or very close to it. The features of ease of installation, self-organization, scalability and low maintenance requirements make WSNs have a great potential in many different application areas from disaster relief to health applications, from military to environmental applications [2].

Sensing unit, processing unit, transceiver and power unit are the basic components of a wireless sensor node. According to the application requirements, location finding system, mobilizer and power generator could be added [2]. The main tasks of a sensor node are sensing the events, processing the data and transmitting it. Communication is the most energy consuming process among them [3]. Sensor nodes usually supply power from batteries in wireless sensor networks. Since changing or recharging these batteries is very hard or impossible in most of the applications, efficient use of these restricted energy sources is vital importance from the aspect of lifetime. There are several various definitions of sensor network lifetime found in the literature. WSN lifetime is defined as "the time span from the deployment to the instant when the network is considered as nonfunctional." in [4]. The most common lifetime definition is that the time from the instant the network starts operating to the first sensor node failure [5]. The sensor nodes with depleted energy source are assumed as "dead", they could not transmit data. Therefore the lifetime of the network depends on how the limited amount of energy is utilized. Many researches about energy efficiency in WSNs have been found in the literature. This paper presents both analytical and algorithmic approaches to determine the number of hops and sensor nodes for minimum power consumption in linear wireless sensor networks.

The remainder of the paper is organized as follows. In section 2, linear wireless sensor network (LWSN) definition, classes and applications of LWSNs are given. The details of the power consumption model used are in section 3. Algorithmic and analytical approaches are explained in section 4 and section 5, respectively. The conclusions of the study are presented in section 6.

2. Linear Wireless Sensor Networks (LWSNs)

Linear Wireless Sensor Networks have a large number of applications. Monitoring of railways, borders, AC power lines, oil,

*Corresponding Author: Nukhet Sazak, Sakarya University, Department of Electrical and Electronics Engineering, 54187, Turkey, nsazak@sakarya.edu.tr

gas and water pipelines are some application examples of LWSNs [6]. A subclass of wireless sensor networks satisfying some conditions is called as Linear Wireless Sensor Networks (LWSNs). If all nodes are aligned on a straight line or all of the nodes exist between two parallel lines that extend for a relatively long distance as compared to their transmitting range and the distance separating them constitute a semi-linear sensor network, this wireless sensor network is called as linear wireless sensor network [6].

LWSNs can be classified into some categories from topological and hierarchical perspectives. There are three types of sensor nodes according to their functions. Basic sensor nodes sense and transmit data to the relay nodes. Data relay nodes collect the data from their one-hop neighborhood. Data dissemination nodes deliver the collected data to the network control center. One-level, two-level and three-level are the categories of hiearchical classification. At one-level sub-category, there are only basic sensor nodes which sense, aggregate, compress and route information to the sink. All nodes consisting of basic sensor nodes and data relay nodes are lined up linearly at two-level structure. Basic sensor nodes sense and transmit data to the data relay nodes which have routing, aggregation and compression tasks. Basic sensor nodes sense and transmit data to the data relay nodes in three-level hierarchical structure. Routing, aggregation, compression are the duties of data relay nodes. Data dissemination nodes disseminate the collected data to the network control center.

There are three levels as thin, thick and very thick for topological classification. The LWSN where sensor nodes are lined up in a 1-dimensional linear form is called as thin LWSN. This kind of structure provides linearity of the nodes and the predictability of the structure to minimize energy consumption and increase the network reliability [6]. In the kind of thick structure, basic sensor nodes can have a 2- or 3-dimensional geographic distribution. Linearity only exists at the levels of data relay and data dissemination nodes. At all levels, nodes are placed randomly between two parallel lines that extend linearly for a long distance and this is called as very thick LWSN.

Multi-hop communication is the common method used to gather data for linear wireless sensor networks. In a multi-hop fashion, all nodes (sensor and relay) send their data to the sink. A multi-hop 1-dimension linear WSN model [7] consisting of equidistantly deployed sensor nodes is given in Figure 1. The aim of this network is to deliver the data generated at the source node (S) to the destination node (D). The relay nodes operate as routers and they are numbered as N_i, i=1,2,.n-1. According to the deployment type and sensor functions, this network structure can be called as thin-one level LWSN since each node acts as both a sensor and relay node.

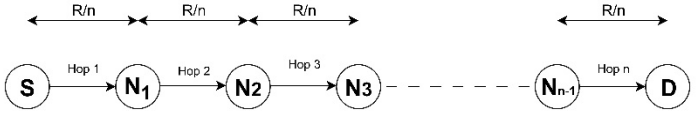


Figure 1. Multi-hop 1-D linear WSN model [7]

3. The Power Consumption Model Used

Wang et.al. [7] compared single-hop and multi-hop communication by using their own power consumption model and concluded that multi-hop communication is the optimal solution when sensor nodes are placed equidistantly between the source and destination nodes.

In this study, Mica2dot sensor nodes data and power consumption model proposed in [7] are used. The main difference of this model from the others in the literature is its dependence of the power amplifier performance which is drain efficiency (η).

There are some assumptions for this model. Communication bandwidth is so much low that interference and collisions could be avoided by using simple protocols without significant power consumption. Physical communication rate is assumed as constant.

The multi-hop power consumption over a given distance with equal hop distance between nodes is given in (1):

$$P(n) = (n-1)P_{R0}+nP_{T0}+[(nx\varepsilon x(R/n)^{\alpha})/\eta] \quad (1)$$

n: number of hops

P_{R0}: power consumption of a sensor node for receiving data

P_{T0}: power consumption of a sensor node for transmitting data

ε: constant

R: distance between source and destination nodes

R/n: hop length

α: path loss exponent

η: drain efficiency (the ratio of RF output power to DC input power)

Table 1 Typical values of path loss exponent [8]

Environment	α
Free-Space	2
Urban area LOS	2,7 to 3,5
Urban area no LOS	3 to 5
Indoor LOS	1,6 to 1,8
Factories no LOS	2 to 3
Buildings no LOS	4 to 6

The path loss exponent has different values depending on the propagation environment as given in Table 1 [8]. In this study, the value for free space (α=2) is used.

4. Algorithmic Approach

The aim of the algorithm is to determine the optimum hop number at which minimum power consumption is obtained. The distance between the source and destination nodes is the input, the optimum hop number is the output of the algorithm. Optimum number of nodes, power consumption value at the optimum hop number, the required minimum number of hops and nodes are also given as the output at the end of the algorithm.

The flow diagram of the algorithm is given in Figure 2. The algorithm gets the distance between source and destination nodes as input. Minimum number of hop is obtained by dividing this distance to the Mica2dot sensor node maximum transmission

distance. If the result is a non-integer number, it is rounded to the upper integer number and assigned to a variable. According to the power consumption model, the number of hop (n) is increased one by one starting from the minimum required hop number. The power consumption is calculated for each n. The number of hop and power consumption at this n are displayed on the screen every computation. The power consumption value for each n is compared to the one at the previous hop within the algorithm. If the previous power consumption is greater than current one, the computation is continued by increasing the n. When the computed power consumption value is firstly greater than the previous one, that critical point is determined as optimum hop number and the iteration is ended.

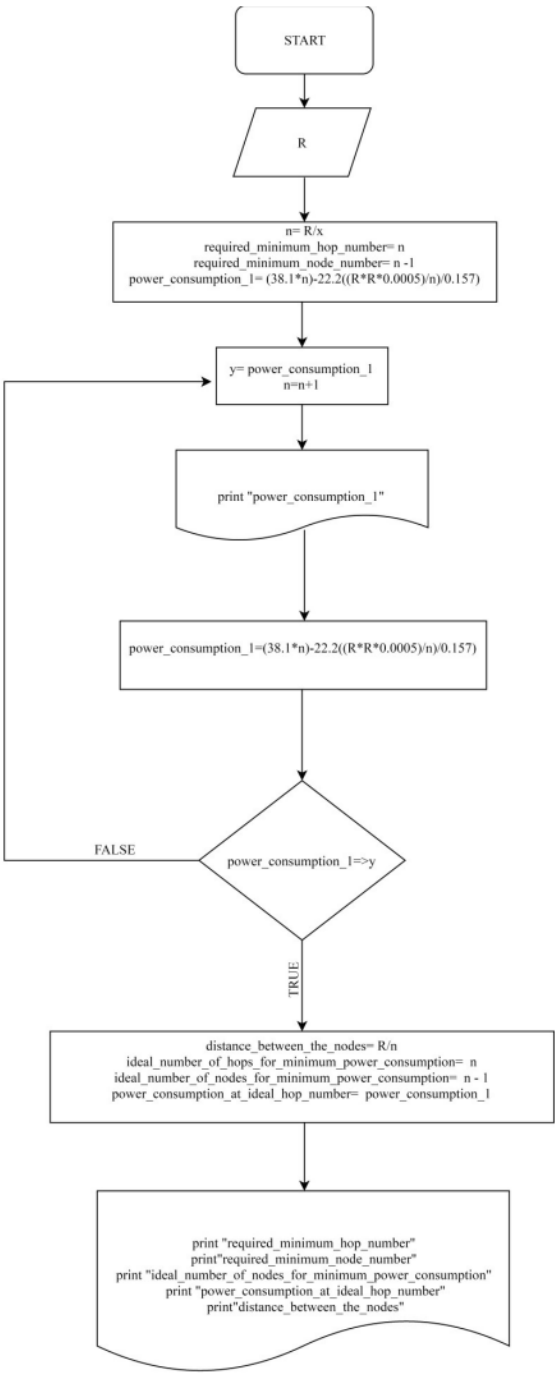


Figure 2. Flow diagram of the algorithm

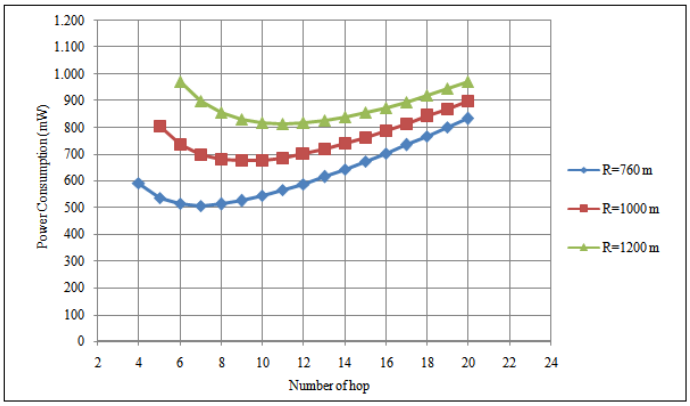


Figure 3. Power consumption versus number of hop

The distance between the source and destination nodes is asked by the algorithm. Then the beginning node number is computed by dividing this distance value to the node distance with maximum transmission power which is 230m for Mica2dot sensor node. The number of nodes is increased one by one from the starting value and iteration goes on until the minimum power consumption value is obtained. The power consumption decreases at first and it starts to increase after reaching its minimum value as expected from the results given in [8]. Since the minimum power consumption value is determined, no more calculation is required and the algorithm ends up.

During the development of this algorithm, Mica2dot sensor nodes are used [9] and Mica2dot sensor data is given in Table 2. However, this algorithm can be employed for any type of sensor node by using its hardware parameters or asking the required paramater values at the beginning as the input.

Table 2 Mica2 sensor node data

P_{R0}	22.2 mW
P_{T0}	15.9 mW
η	15.7 %
α	2
ε	0.0005

For the distance between source and destination nodes as 760m, 1000m and 1200m, power consumption versus number of hop graph is obtained and given in Figure 3. When the number of hops is increased, the power consumption decreases at the beginning and then increases after reaching the minimum value.

This algorithm provides not only the optimum hop length and as a result optimum hop number but also the number of sensor nodes required for minimum power consumption directly without computing power consumption for each hop number.

5. Analytical Approach

As it is seen from Figure 3, the power consumption decreases at the beginning and then increases after reaching the its minimum value. Therefore, this is considered as mathematical issue finding minimum point and analytical approach is developed [1].

The hop number at which the minimum power is consumed is called as optimum hop number. Taking the derivative of (1) with respect to hop number, optimum hop number (n_{opt}) is obtained as:

$$n_{opt} = R\{[\eta(P_{R0}+P_{T0})]/[\varepsilon(-1+\alpha)]\}^{-1/\alpha} \quad (2)$$

For the α=2 value, (2) can be rearranged as:

$$n_{opt} = R\{[\eta(P_{R0}+P_{T0})]/\varepsilon\}^{-0.5} \quad (3)$$

As it can be seen from (3), optimum hop number depends on sensor node receiving power (P_{R0}), transmitting power (P_{T0}), α, ε, η parameters and also the distance between source and destination nodes.

The optimum number of sensor nodes is defined as the sensor nodes required for minimum power consumption in a linear wireless sensor network with equidistantly placed sensor nodes. After optimum hop number is obtained, the optimum sensor nodes (d_{opt}) can also be calculated by subtracting 1 from (3):

$$d_{opt} = n_{opt} - 1 \quad (4)$$

The more generic expression for optimum number of sensor nodes is given in (5).

$$d_{opt} = R\{[\eta(P_{R0}+P_{T0})]/[\varepsilon(-1+\alpha)]\}^{-1/\alpha} - 1 \quad (5)$$

Thanks to these mathematical expressions, the number of sensor nodes which are placed with equal hop length can be calculated directly from P_{R0}, P_{T0}, ε, η, α for a linear wireless sensor network.

6. Conclusion

WSNs can be deployed over the locations where are difficult to install traditional systems and this makes them attractive. On the other hand, sensor nodes are left unattended for a long period of time, lifetime of the network depends on how the limited energy supply of the nodes is used. Therefore energy efficiency issue is required to be considered during the application design. From the engineering design perspective, the aim of this study to determine the number of hops and sensor nodes for minimum power consumption in LWSNs including equidistantly deployed sensor nodes. For this purpose, both algorithm and analytical approaches are presented in this paper. Thanks to these approaches, the number of hops and thus the number of sensor nodes required for a LWSN application are determined according to the hardware parameters of the sensor nodes during design phase.

References

[1] N. Sazak and A. S. Kilinc, "Eşit Mesafeli Çok-Sekmeli Kablosuz Algılayıcı Ağlarda Minimum Güç Tüketimi için Optimum Düğüm Sayısının Belirlenmesi" in 2016 National Conference on Electrical, Electronics and Biomedical Engineering (ELECO 2016), Bursa, Turkey, 2016.

[2] I.F. Akyildiz, W. Su, Y. Sankarasubramaniam, "Wireless sensor networks: A Survey" Computer Networks, vol. 38, 393-422, 2002.

[3] M. Ilyas, I. Mahgoub, Handbook of Sensor Networks: Compact Wireless and Wired Sensing Systems, CRC Press, 2005.

[4] Y. Chen, Q. Zhao, "On the Lifetime of Wireless Sensor Networks," IEEE Communication Letters, vol.9, no.11, 976-978, 2005.

[5] N. H. Mak, W. Seah, "How long is the Lifetime of a Wireless Sensor Network?" in International Conference on Advanced Information Networking and Applications, AINA'09, 2009.

[6] I. Jawhar, N. Mohamed, and D. P. Agrawal, "Linear wireless sensor networks: Classification and applications," Journal of Network and Computer Applications, 34, 1671-1682, 2011.

[7] Q. Wang, M. Hempstead, W. Yang, "A Realistic Power Consumption Model for Wireless Sensor Network Devices" in Proc. 3rd Annual IEEE Communications Society on Sensor and Ad Hoc Communications and Networks, Reston, 2006.

[8] M. Kheireddine, R. Abdellatif, "Analysis of hops length in wireless sensor networks," Wireless Sensor Network, 6(6), pp. 109–117, 2014.

[9] N. Sazak and A. S. Kilinc, "An Algorithm to Determine the Ideal Hop Length for Minimum Power Consumption in WSNs" in 2nd International Conference on Engineering and Natural Sciences (ICENS 2016), Sarajevo, Bosnia and Herzegovina, 2016.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1491-1501 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

System for the Automatic Estimation of the Tilt Angle of a Flat Solar Collector

Jorge Fonseca-Campos*, Paola N. Cortez-Herrera, Israel Reyes-Ramírez

Unidad Profesional Interdisciplinaria en Ingeniería y Tecnologías Avanzadas del Instituto Politécnico Nacional.

Av. IPN 2580, Col. Barrio La Laguna Ticomán, GAM, Ciudad de México, C.P. 07340. México.

ARTICLE INFO

Article history:
Received: 01 June, 2017
Accepted: 20 July, 2017
Online: 15 August, 2017

Keywords:
Tilt Angle
Geographical Coordinates
GPS
Python
Flat Solar Collector

ABSTRACT

In this work, a compact system for the automatic estimation of the tilt angle at any location of the world is presented. The system components are one computer, one GPS receiver and one Python program. The tilt angle is calculated through the maximization of the flux of direct radiation incident upon a flat solar collector. An estimation of the adjustments of this angle at different time periods are obtained. This angle is calculated in steps of six minutes during a whole year. Daily, monthly, biannually and yearly averages of this value are obtained. A comparison of the energetic gain when the tilt angle changes at the different time periods is made as well. Because, the algorithm doesn't receive as an input the solar radiation incident upon a surface at the location of the calculation, a comparison was made between the results obtained and the results reported for the monthly tilt angle of 22 different places. The root mean square error obtained with this comparison was between 1.5 and 9.5 degrees. The monthly tilt angle estimated deviated in average for less than 6.3° with respect to the values reported for the different locations. Finally, the application of a correction factor in the monthly estimated angles is proposed, which might increase the collected energy.

1. Introduction

Recently, we reported a system having as a purpose the calculation of the sun trajectory during one day in a specific location [1]. In this paper, another system presenting enhanced capabilities with respect to the system presented in the above work is reported.

The energy that can be extracted from a flat solar collector depends to a large extent on the direction of solar radiation in relation to the surface of the panel. This will be maximum, when the sun rays and the surface are perpendicular to each other. Through, a solar tracking system the above condition can be fulfilled the whole time. However, it is cheaper and simpler to keep a fix solar collector for a period of time than to make two rotations along two different axes to achieve tracking. In systems with dual axis tracking the panel is rotated along the line pointing to the vertical of the place (surface azimuth angle) and it is rotated along the line east-west (tilt angle or slope).

In general, for the fixed panels the optimal surface azimuth angle is zero degrees. Therefore, the tilt angle is the only one that is adjusted to maximize the collected energy. In this respect, several studies to obtain the optimum tilt angle β_{opt} at different locations of the world have been conducted. As for example: Athens [2], Izmir [3], four places in India [4], 5 regions in China [5], Sanliurfa [6], Tabass [7], two locations in India [8], eight provinces of Turkey [9], New Delhi [10], Helwan [11], Brunei [12], Syria [13], Surabaya [14], 35 sites of the Mediterranean region [15], Taiwan [16], Madinah [17] and the examples could continue.

In the literature, it is suggested that in the northern hemisphere the panels must be facing south. The optimum tilt angle depends of the geographical latitude of the place ϕ. Not all of the studies are focusing in the calculation of an annual β_{opt}. In some cases, this angle is reported in monthly or biannual (twice a year) time periods. Because, the selection of different time periods impacts the efficiency of the solar collector. For example, it was reported that in Madinah, a monthly adjustment of the tilt angle can enhance in 8% the captured energy than the one that can be obtained when this angle is kept fix during the whole year [17]. Numerous suggestions for β_{opt} are found in the literature. For example, in the

*Corresponding Author: Jorge Fonseca-Campos, Col. Barrio La Laguna Ticomán, GAM, Ciudad de México, C.P. 07340. México, Email: fonsecj@live.com

References [3, 11, 18] is proposed $\beta_{opt} = \phi \pm 15^o$, in the Reference [19] is suggested $\beta_{opt} = \phi \pm 10^o$, where the minus sign refers to the summer season and the plus sign is for the winter. In the Reference [10] is proposed $\beta_{opt} = \phi + 90^o$ for the winter and $\beta_{opt} = \phi - 60^o$ for the summer. In the Reference [20] is recommended $\beta_{opt} = 0.917\phi + 0.321^o$, in the Reference [21] is suggested $\beta_{opt} = \phi + (10^o \rightarrow 30^o)$, in the Reference [22] is proposed $\beta_{opt} = \phi - \delta$, where δ is the declination and the examples could continue. Unfortunately, the only way to obtain the optimum tilt angle is through radiation measurements at the place of interest for an extended period of time. This is due to the fact that solar radiation has different components, which are the direct, the diffuse and the albedo radiation. The optimum tilt angle, which is the ones that maximize the energy extraction is influenced in the way that these radiation components are distributed.

In this work, a system for estimation of the tilt angle considering exclusively the direct radiation is presented. This can be used in the locations that does not have a record of radiation measurements, or when the users or designers of the solar installations do not have a direct access to available radiation data. With this system in a very few minutes and with a minimum of interaction of the end user, several data used in the solar designs are obtained. Also, the tilt angle is calculated in time periods of six minutes, daily, monthly, biannual and yearly. A comparison of the energetic gain when the tilt angle changes at the different time periods is made as well. Finally, a comparison was made between the results provided with the proposed system and the ones reported for the monthly tilt angle of 22 different places. The root mean square error (*RMSE*) obtained through this comparison was between 1.5 and 9.5 degrees. The estimation of the monthly tilt angle deviated in average for less than 6.3°. Also, an adjustment for each of the estimated monthly slope angles is proposed, which might reduce the energy loss.

2. Direct Radiation Incident on a Tilted Solar Flat Panel

2.1. Sun Position

In a fixed point on the Earth, due to the Earth's rotation and its movement around the sun, the direction of the beam radiation is changing in time. It is customary to describe the trajectory that follows the sun by using two angles, when the representation of the Celestial sphere is used. These are: the hour angle ω and the declination angle for an equatorial coordinate system and the solar altitude angle α_s and the solar azimuth angle γ_s for a horizontal coordinate system. The angle between the two planes; the one containing the Earth's axis and the meridian plane of the point of observation, and the other containing the Earth's axis and the meridian plane of the sun is ω. At solar noon, i.e., when the sun is on the observer meridian $\omega = 0^o$. The angle between the equator and the sun position at solar noon is the declination. This angle in degrees is given by [23]

$$\delta = 23.45 \sin\left(360 \frac{284 + n}{365}\right), \tag{1}$$

where n is the integer day number, being January 1st $n = 1$. The declination is changing in time, but during one day is almost constant.

In the Figure 1 are shown the angles α_s, γ_s and θ_{sz}, where the latter angle is the solar zenith angle. α_s is the angle between the direct sun rays and the horizontal plane of the specific location, $\alpha_s = 0^o$ when the sun is on the horizontal plane, $\alpha_s = 90^o$ when the sun is at the zenith. During the sun hours this is an angle between 0° and 90°. Its value is negative before sunrise and after sunset. The relationship between the solar altitude angle and solar zenith angle is $\theta_{sz} = 90^o - \alpha_s$. γ_s is the angle between the south and the projection of the direct sun rays on the horizontal plane. This is an angle between -180° and 180°. Displacements east of the south are negative and west of the south are positive.

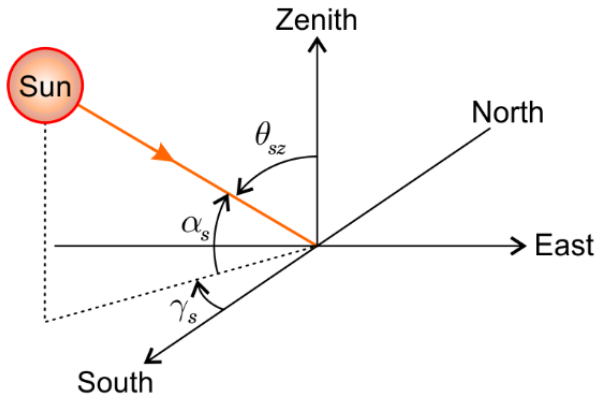


Figure 1. Sun position relative to one fixed point on Earth.

In the References [24-28] have been reported algorithms to calculate the sun position. The complexity and accuracy differ from each other. In this work, the fifth algorithm of the Reference [28] is employed, because it is easier to program and it is accurate from 2010 to 2110. The output of this algorithm are the angles ω, δ, α_s and γ_s. The inputs are the date (month, day, year), the time in decimal format, the latitude, the local longitude L_l, the atmospheric pressure P and the room temperature T. In this algorithm, the following conventions are applied: the latitude is an angle between -90° and 90°, in locations up north of the equator it is positive, otherwise is negative. The longitude is an angle between 0° and 360°, being 0° at the prime meridian. At the east of this meridian has a value between 0° and 180°. It is measured in a counter clockwise fashion to the east.

The angles of the equatorial and horizontal coordinate systems are related to each other by [18]

$$\cos\theta_{sz} = \sin\alpha_s = \cos\phi\cos\delta\cos\omega + \sin\phi\sin\delta \tag{2}$$

and

$$\gamma_s = \text{sign}(\omega)\left|\cos^{-1}\left(\frac{\cos\theta_{sz}\sin\phi - \sin\delta}{\sin\theta_{sz}\cos\phi}\right)\right|, \tag{3}$$

where the sign function is equal to $+1$, if ω is positive; otherwise is -1.

2.2. Angle of Incidence on a Tilted Flat Solar Collector

In a fixed location, the sun moves across the sky along the day. By this reason, the beam radiation vector will make an angle θ_i with respect to the normal of the collector surface. Most solar energy can be captured when $\theta_i = 0^o$. If a unit vector $\hat{\mathbf{n}}$ is pointing from a fixed point in the Earth to the direction of the beam radiation and another unit vector $\hat{\mathbf{m}}$ is normal to the collector surface, the angle between both vectors can be computed with the dot product, i.e., $\theta_i = \cos^{-1}\left(\hat{\mathbf{m}} \cdot \hat{\mathbf{n}}\right)$.

To minimize the angle θ_i during a specific time period, the fixed flat solar panels are subjected to two rotations. On one hand, the panel is rotated in a counterclockwise fashion an angle β along the line east-west (tilt angle or slope). On the other hand, the panel is rotated in a clockwise fashion an angle γ along the line pointing to the zenith of the place (surface azimuth angle). Both rotations are shown in the Figure 2.

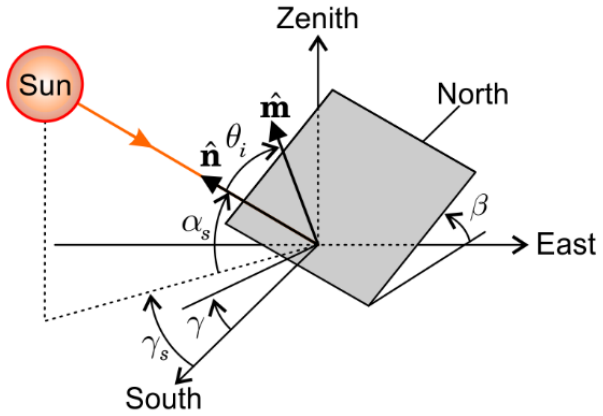


Figure 2. Flat surface rotated along the lines east-west and south-north.

According to the geometry shown in the Figure 2. The unit vectors $\hat{\mathbf{m}}$ and $\hat{\mathbf{n}}$ are given by

$$\hat{\mathbf{m}} = \sin\beta\cos\gamma\hat{\mathbf{e}}_x - \sin\beta\sin\gamma\hat{\mathbf{e}}_y + \cos\beta\hat{\mathbf{e}}_z \tag{4}$$

and

$$\hat{\mathbf{n}} = \cos\alpha_s\cos\gamma_s\hat{\mathbf{e}}_x - \cos\alpha_s\sin\gamma_s\hat{\mathbf{e}}_y + \sin\alpha_s\hat{\mathbf{e}}_z\,. \tag{5}$$

Thus, the angle of incidence between both vectors after a few simplifications is

$$\cos\theta_i = \cos\alpha_s\cos(\gamma - \gamma_s)\sin\beta + \sin\alpha_s\cos\beta\,. \tag{6}$$

The tilt angle is an angle between -90° and 90°, when $\beta > 0^o$ the collector is facing the south, which is the most used slope in the northern hemisphere. If the panel is horizontal $\beta = 0^o$ and $\theta_i = \theta_{sz}$. Finally, if $\beta < 0^o$ the surface of the panel is facing the north. In most situations, the collectors are adjusted to have an azimuth angle $\gamma = 0^o$. This angle is varied, when there are objects that produce shadows over the collector surface, or in some cases when a set of two fixed panels are used for extracting energy. In this configuration, one of them is used in the morning and the other is employed after the solar noon [14].

2.3. Maximization of the Flux of Energy in Terms of the Tilt Angle

Recently, Handoyo and co-workers have exposed a mathematical procedure to calculate the optimum tilt angle of a solar flat collector [14]. In that work, $\cos\theta_i$ is expressed in the equatorial coordinate system in terms of the following angles: δ, ω, β, γ and ϕ. This equation is reported elsewhere as in Reference [18]. They do the first derivative and the second derivative test for a single variable to maximize $\cos\theta_i$ in terms of β. At first, $\left(d\cos\theta_i/d\beta\right) = 0$ is calculated and later the condition $\left(d^2\cos\theta/d\beta^2\right) < 0$ is tested. These tests will be applied to (6), which is expressed in terms of a horizontal coordinate system. The value of β, which corresponds to the critical point of (6) is

$$\beta_{max} = \tan^{-1}\left(\frac{A}{B}\right), \tag{7}$$

where $A = \cos\alpha_s\cos\left(\gamma - \gamma_s\right)$ and $B = \sin\alpha_s$. The second derivative is

$$\frac{d^2\cos\theta_i}{d\beta^2} = -\left(A\sin\beta_{max} + B\cos\beta_{max}\right). \tag{8}$$

Computing β_{max} of (7) requires as inputs the date (month, day and year), the hour in decimal format, the geographical coordinates and the panel azimuth angle. If the magnitude of any of these variables changes, the value of β_{max} is modified as well.

2.4. Sunrise, Sunset, Day Length and Local Standard Time

In the equator, the day length is 12.0 h during the whole year. But at latitudes different to this geographical point, this quantity is changing in daily basis. Up north of this latitude, during the summer the sun light is shining for a longer time, on the contrary during the winter. Therefore, the net energy that can be captured by a collector will depend to some extent on the day length. Of course, if there is no sun light, no solar energy can be obtained. By this reason, it is interesting to calculate the evolution of this variable, which is related to the sunrise and the sunset time. From (2), the sunrise hour angle in degrees can be obtained by making $\alpha_s = 0^o$, i.e.,

$$\omega_{sr} = -\cos^{-1}\left(-\tan\phi\tan\delta\right). \tag{9}$$

The sunset hour angle is $\omega_{ss} = -\omega_{sr}$. The day length DL in decimal hour is

$$DL = 2\,\mathrm{abs}\left(\omega_{sr}/15^o\right). \tag{10}$$

The solar time ST in decimal hour is related to the hour angle by

$$ST = \left(\omega/15^o\right) + 12\,. \tag{11}$$

In terms of the local standard time LST and a time correction factor TC is

$$ST = LST + TC\,. \tag{12}$$

The time correction factor in minutes is given by

$$TC = 4\left(L_{st} - L_l\right) + EoT - C_{DST}\,, \tag{13}$$

where L_{st} is the standard longitude, L_l is the observer's longitude, C_{DST} is a time correction due to the daylight-saving time (DST) and EoT is the equation of time. In (13) the following conventions must be employed: L_l is an angle between -180° and 180°. It is considered negative at the west of the Greenwich meridian, and positive at the east of it. In this work, the DST changes are not considered, thus $C_{DST} = 0$. Also, $L_{st} = 15^{\circ} TZ$, where TZ is the time zone in decimal hour. TZ is positive at the west of the prime meridian and negative at the east of it.

The equation of time in minutes is

$$EoT = 9.87\sin\left(2B\right) - 7.53\cos B - 1.5\sin B\,, \tag{14}$$

where B in degrees is

$$B = \left(\frac{360}{365}\right)\left(n - 81\right). \tag{15}$$

By combining (11) and (12), it can be obtained the local standard time in decimal hour as

$$LST = \frac{\omega}{15^{\circ}} + 12 - \frac{TC}{60}. \tag{16}$$

With (9) and (16) the sunrise time LST_{sr} is obtained. Similarly, by substituting ω_{ss} in (16) the sunset time LST_{ss} is obtained.

The coordinated universal time UTC is employed to calculate the sun position. It is related to the LST by

$$UTC = LST + TZ\,. \tag{17}$$

This time doesn't observe the DST, and TZ follows the same convention applied to (13).

3. System Description

3.1. GPS Receiver

The global positioning system (GPS) receiver used in this work was the model ND-105C GPS Dongle from GlobalSat WorldCom Corporation. But, any GPS receiver that meets the National Marine Electronics Association (NMEA) requirements and that has the possibility of sending sentences through the serial port can be employed.

The above GPS receiver for the output data follows the protocol NMEA 0183. It returns four the sentences: GGA, GSA, GSV and RMC. The GGA sentence contains all the useful data introduced as an input in the program as the validity of satellite fix, the geographical latitude with the corresponding hemisphere and the local longitude with its relative position with respect to the prime meridian.

3.2. Python Program

The program used to perform the calculations was written in Python 3.6. It can be requested by email to the corresponding author. The flowchart of the program is shown in the Figure 3.

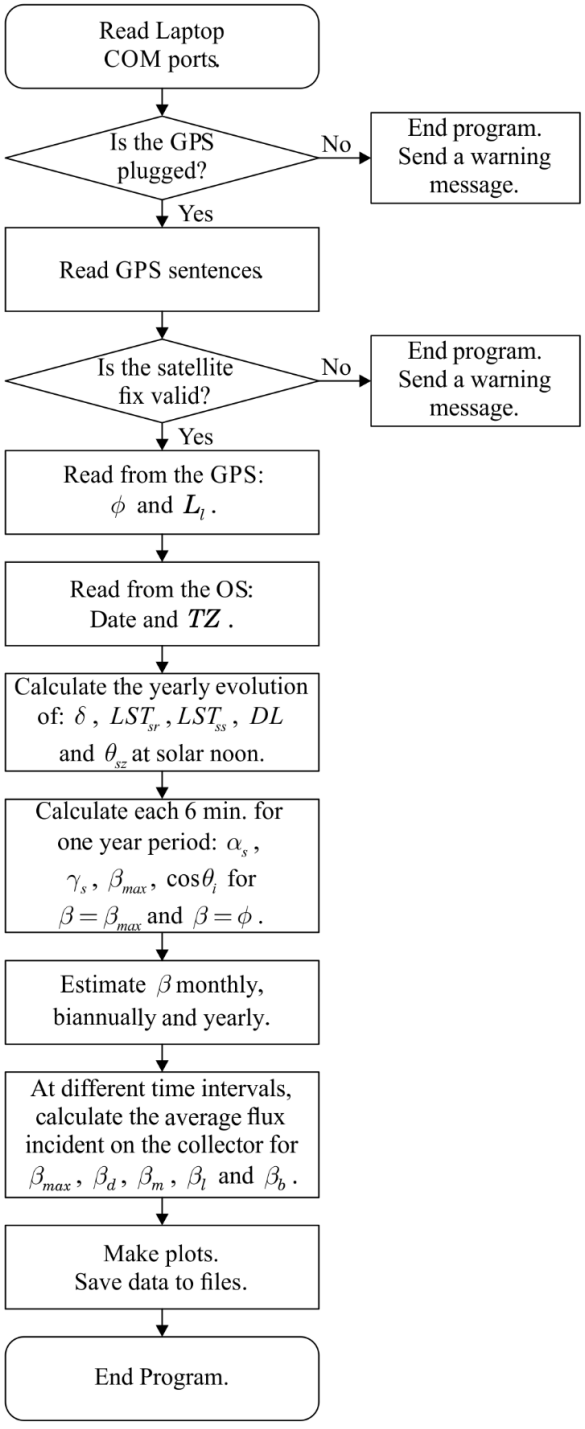


Figure 3. Flowchart of the Python program.

The time frame set in the calculations corresponds to one year. They will begin in the first day of the month at which the program

was run, and they will end in the same day and in the same month of the next year.

The Spyder integrated development environment (IDE) included in the Anaconda distribution was chosen to debug and to run the program. The communication between the GPS and the computer is realized through the serial port. Therefore, pyserial module must be installed, it is freely available at: (https://pypi.python.org/pypi/pyserial). In order to the program to work properly the end user must provide the communication (COM) port at which the GPS is connected to the computer and it must configure the time zone of the Windows operating system (OS). This must correspond to the region at which the calculation of the tilt angle is going to be done. The DST changes are not considered in the program, and they must be ignored in the above configuration.

As it shown in the Figure 3, if the GPS is plugged, the GPS sentences are read through the serial port. From the GGA sentence all the required inputs are extracted (valid fix, ϕ and L_l). When the fix is invalid the program will stop displaying a warning message. Otherwise, the program will continue doing computations. From the OS the date and the TZ are read by the program.

To obtain DL the program computes (9) and (10). The first equation requires the declination, which is calculated with (1). The latitude is provided by the GPS. LST_{sr} is obtained with (16), when ω is substituted by ω_{sr}, which is obtained with (9). LST_{ss} is get when ω is substituted by ω_{ss} in (16). The zenith angle at solar noon is calculated with the declination and the latitude. For places with greater latitudes than the tropics, this angle never will reach zero. For both tropics will be zero once a year. In places between the above latitudes will be reach zero twice a year.

Once the daily behavior of LST_{sr} and LST_{ss} are obtained a daily time interval is set for the calculation of the variables α_s, γ_s, β_{max} and $\cos\theta_i$. Initially, (6) is evaluated with two different tilt angles ($\beta=\beta_{max}$ and $\beta=\phi$). A few averages are realized once β_{max} is obtained. These are done for the following time periods: daily β_d, monthly β_m, from April to September and from October to March (biannual) β_b and yearly β_y.

However, the magnitude of the above angles depend of the daily time interval selected in the simulation. For example, their magnitude will have one value for the daily time interval $t\in\left[LST_{sr},LST_{ss}\right]$ h, which differs to the obtained if $t\in\left[10.0,14.0\right]$ h is considered. Therefore, a daily time interval for the simulation must be chosen. In this work, the user can select one of two possible options for the daily time interval. The first comprehends the daily time interval $t_{long}\in\left(t_i,t_f\right)$ in steps of 0.1 h (six minutes). The initial time of this interval is $t_i=\left(\max\left(LST_{sr}\right)-\min\left(LST_{sr}\right)\right)/2$, and the final time of this interval is $t_f=\left(\max\left(LST_{ss}\right)-\min\left(LST_{ss}\right)\right)/2$. This daily time interval will be called long time. The functions max and min are referring at the maximum and minimum values that reaches the variable over a one year period. The second time interval will be called short time. This is the default time considered in the program. It is given by $t_{short}\in\left(\max\left(LST_{sr}\right),\min\left(LST_{ss}\right)\right)$ in steps of 0.1 h. It is worth mentioning, that the user can input the magnitude of γ, which appears in the coefficient A of (7), but its default value is set to $\gamma=0^{o}$. If t_{long} is chosen in the simulations, for some days and in some specific times could result that $\alpha_s<0^{o}$ .This means that the sun is not in the local horizon. In these particular situations, the program assign NaN values to all variables.

To estimate the average flux of solar radiation incident in the solar collector, the different values of β (β_{max}, β_d, β_m, β_b and $\beta=\phi$) are evaluated in (6). Finally, the average of $\cos\theta_i$ for the each of the above angles in different time periods is calculated.

All the calculated variables are saved in files, which have a comma separated value (csv) format. The plots of the yearly evolution of LST_{sr}, LST_{ss}, DL, θ_{sz} at the solar noon, β_d, β_m and β_b are generated by the program. In addition, the plot of the annual variation of $\cos\theta_i$ is generated, when is evaluated with the following angles: β_{max}, β_d, β_m, β_b and $\beta=\phi$. After this step, the program ends.

3.3. System Applications

The applications that can have the system can be briefly summarized as follows:

- To have an insight of the working hours of the solar installations during a whole year, because these can operate just during the sun hours. The above variable is related to the day length, the sunrise and sunset times, which are provided graphically and in csv format files.
- To control a dual axis tracking system by the knowledge of the solar azimuth and the solar altitude angles. Both angles are provided by the system.
- To estimate the tilt angle that should follow one single axis tracking system in different periods of time. This can be done every six minutes, daily, monthly or biannual. To have an insight of the energetic gain that can have any of the above systems.
- To modify the program to extend its capabilities.

4. Results and Discussion

As stated above, the system can work at different locations. Just two conditions must be meet. On one hand, the correct COM port at which the GPS is detected by the computer must be set. On the other hand, the time zone of the location and the configured in the OS must match.

Several tests were realized to verify the proper operation of the system. However, they were realized in different places of same city, which was Mexico City.

As an example of the output data provided by the program, the results obtained during one test made in Mexico City will be presented.

4.1. Annual Evolution of LST_{sr}, LST_{ss}, DL and θ_{sz} at Solar Noon

In the Figure 4 the plot of LST_{sr} for 366 days is shown. The minimum value of the sunrise time is 6:09 h and it will be reached in June 6th, 2017. The maximum value is 7:17 h and it will be reached in January 19th, 2018. Due to the geographical location of Mexico City, there is no a significant daily variation associated to this variable.

In the Figure 5 is shown the plot of the yearly variation of LST_{ss}. The minimum value of the sunset time is 17:52 h and it will occur in November 23rd, 2017. The maximum value is 19:14 h and it will occur in July 6th, 2017.

In the figure 6 the yearly evolution of DL is presented. The shortest day length will be in December 21st, 2017 and it has a value of 10 h and 49 min. It is not surprising that this date corresponds to the winter solstice. The day length reaches a maximum value of 13 h and 10 min. This will happen in June 21th, 2017, which corresponds to the summer solstice.

In the Figure 7 the yearly evolution of the θ_{sz} at solar noon is presented. Due to the fact that Mexico City is located between the equator and the Tropic of Cancer, twice a year θ_{sz} is zero at solar noon. One of these dates corresponded to May 18th, 2017 and the other will occur in July 25th, 2017. The global maximum of θ_{sz} is $42.96°$ and it will be reached in the winter solstice. The local maximum, with a value of 3.94°, will happen in June 21st, 2017, which corresponds to the summer solstice.

Once LST_{sr} and LST_{ss} are known the daily time interval (t_{long} or t_{short}) is obtained.

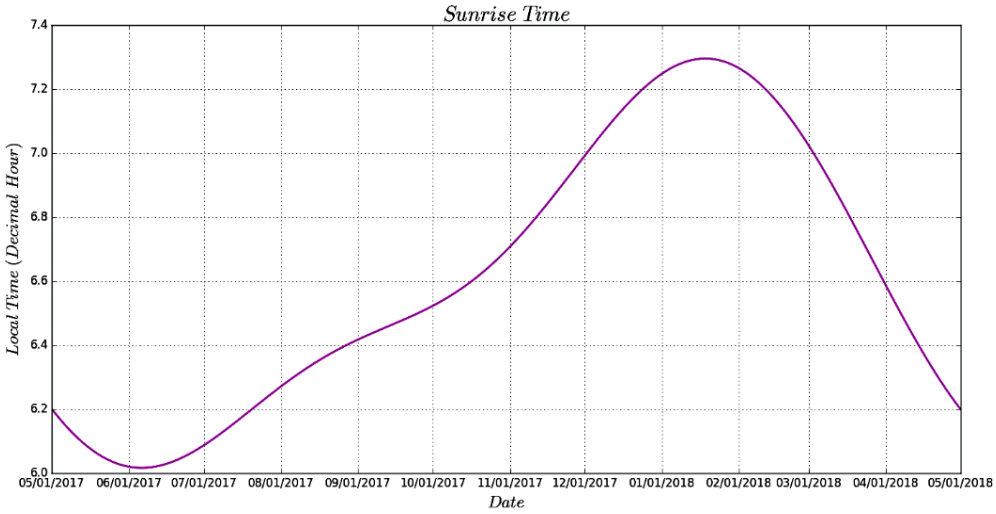


Figure 4. Yearly evolution of the sunrise time

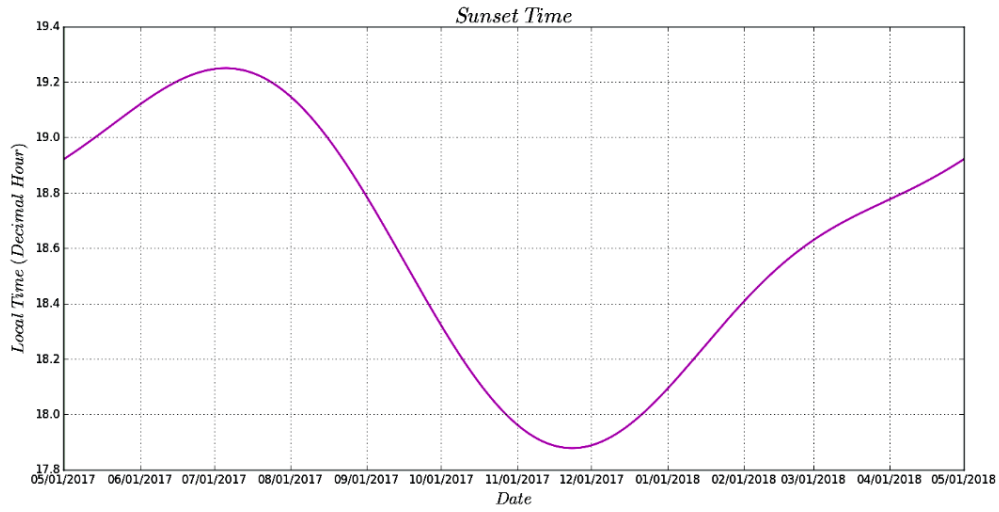


Figure 5. Yearly evolution of the sunset time.

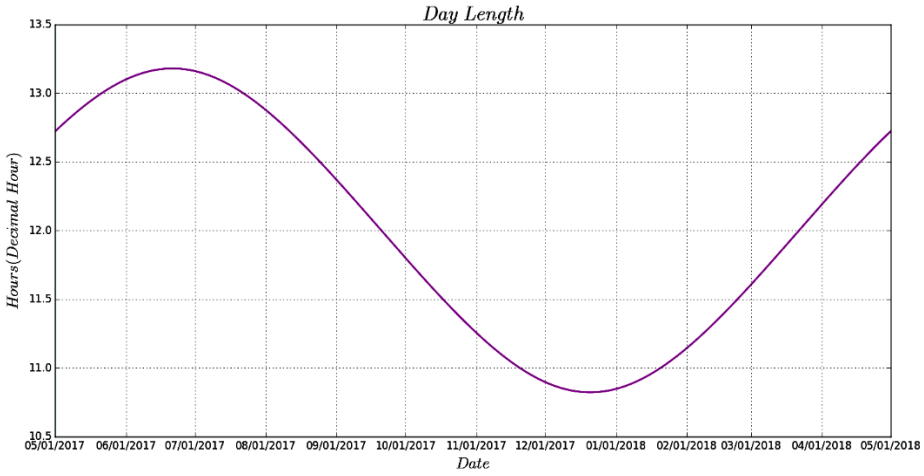


Figure 6. Yearly evolution of the day length.

Figure 7. Yearly evolution of the solar zenith angle at solar noon.

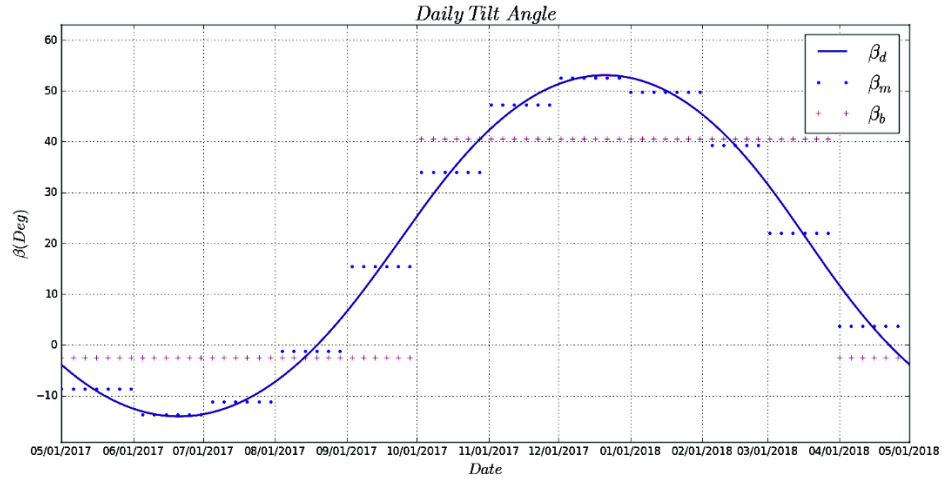


Figure 8. Yearly evolution of the daily, the monthly and the biannual tilt angles.

In the Figure 8 are shown the values of β_d, β_m and β_b, which were obtained averaging β_{max}. This angle was calculated in steps of 0.1 h. The simulation was made by choosing the daily time interval as t_{short}. As it can be seen in the figure 8 from April 23th to August 18th the daily tilt angle must be negative. It means that during this period the solar collector must face to the north. When a monthly adjustment of the tilt angle is realized, the panel should face the north from May to August. As well as $\beta_b < 0^{\circ}$ from April to September. In the Figure 9 it is shown how $\cos\theta_i$ is varying during an entire year in terms of the angles β_{max}, β_d, β_m, β_b and β_l, where $\beta_l = \phi$.

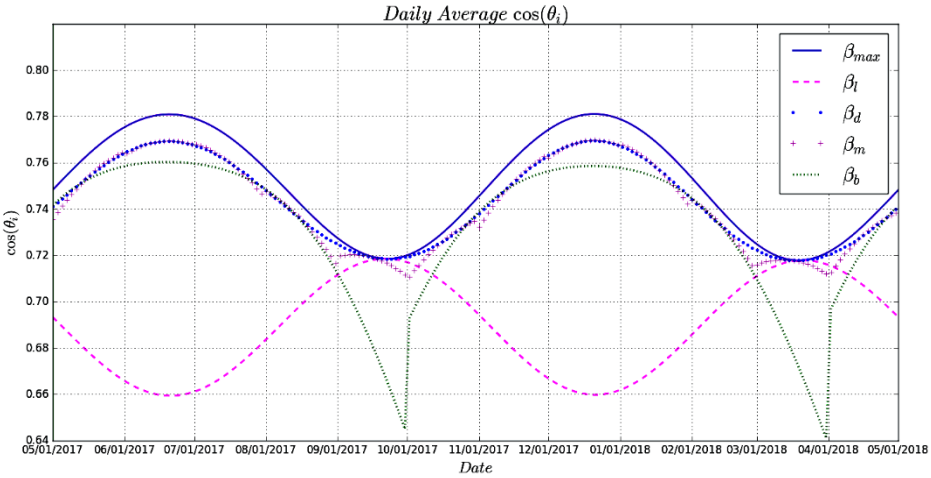


Figure 9. Yearly evolution of average radiation flux incident on the collector surface. When different tilt angles are evaluated in (6).

Time Period	β_m (Deg)	$\cos\theta_i(\beta_{max})$	$\cos\theta_i(\beta_d)$	$\cos\theta_i(\beta_m)$	$\cos\theta_i(\beta_b)$	$\cos\theta_i(\beta_l)$	DL (h)	LST_{sr} (h)	LST_{ss} (h)
May, 2017	−8.73	0.763	0.753	0.752	0.752	0.679	12.93	6.09	19.01
June, 2017	−13.7	0.779	0.768	0.768	0.76	0.661	13.16	6.04	19.19
July, 207	−11.22	0.77	0.76	0.759	0.756	0.67	13.05	6.17	19.22
August, 2017	−1.18	0.742	0.736	0.735	0.734	0.698	12.64	6.35	18.99
September, 2017	15.41	0.721	0.72	0.718	0.681	0.717	12.09	6.47	18.56
October, 2017	33.98	0.73	0.727	0.725	0.717	0.708	11.53	6.6	18.13
November, 2017	47.23	0.761	0.751	0.75	0.749	0.68	11.06	6.84	17.9
December, 2017	52.61	0.779	0.768	0.768	0.758	0.662	10.84	7.13	17.97
January, 2018	49.76	0.768	0.758	0.757	0.753	0.672	10.97	7.28	18.25
February, 2018	39.29	0.739	0.733	0.731	0.73	0.699	11.36	7.17	18.52
March, 2018	22.06	0.72	0.719	0.717	0.678	0.716	11.89	6.82	18.7
April, 2018	3.68	0.733	0.729	0.727	0.72	0.707	12.45	6.39	18.84
Yearly	19	0.751	0.743	0.742	0.732	0.689	12	6.61	18.61
April-September	−2.48	0.755	0.748	0.747	0.738	0.693	12.72	6.25	18.97
October-March	40.6	0.753	0.747	0.745	0.735	0.694	11.27	6.97	18.24

As expected, the greater flux incident on the collector surface is obtained when the panel is moved every six minutes. Surprisingly, $\cos\theta_i$ doesn't vary much when is evaluated with β_{max}, β_d or β_m. In September and in March a biannual movement of the panel can provide less energy than the energy extracted with the tilt angle equal to the latitude of the location. Most of the year, $\cos\theta_i(\beta=\beta_l)$ will have the smallest magnitude.

One file with the csv format summarizing the above results is generated, when the program is executed. This report is presented in the Table 1

The Table 1 must be read in the following way. The first twelve rows are referred to the monthly values. Taking as an example the month of May, β_m should be -8.73^{o}, the radiation flux incident on the collector surface in average will be 0.763, 0.753, 0.752, 0.752, 0.679, when β is evaluated in (6) with angles β_{max}, β_d, β_m, β_b and β_l, respectively. The day length will be 12.93 h in the decimal hour, and the sunrise and the sunset times will be 6.09 h and 19.01 h, respectively. Both values are as well given in decimal hour.

It is not very surprising to find that the yearly tilt angle is almost equal to the latitude as it is shown in the first column of the table 1, that is $\beta_y \approx \phi$. In the last two rows of the first column the results for β_b are shown. As expected yearly DL is 12.0 h. Finally, in the row related with the yearly values, and average of $\cos\theta_i$ for the different tilt angles is shown. These magnitudes are 0.751, 0.743, 0.742, 0.732 and 0.689, for β_{max}, β_d, β_m, β_b and β_l, respectively. It is interesting to note that an energetic gain of ~1% is obtained by adjusting the tilt angle every six minutes than doing it every month.

4.3. Evaluation of the Algorithm

The optimum tilt angle at a given location requires of the measurements of the solar radiation. Therefore, to evaluate the algorithm one comparison was made between the results obtained and the results reported for the monthly tilt angle of 22 different places. To do this comparison the geographical coordinates were

input manually to the algorithm. The monthly optimum tilt angle of this places is reported in the References [3-13]. As a measure of the deviation the $RMSE$ was calculated. This is given by

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}\left(x_{1,t} - x_{2,t}\right)^2}{n}}, \quad (18)$$

where $x_{1,t}$ and $x_{2,t}$ are two time series, and n in (18) is the number of data. The results of this comparison are shown in the Table 2. During these simulations the daily time interval was always t_{short}. In each of the rows corresponding to the twelve months of the year the difference between the reported tilt angle and the calculated with the algorithm is shown. A negative sign means that β_m was underestimated, on the contrary the angle was overestimated. In these simulations, the average β_m always was overestimated as it can be seen in the last row. Near the winter and summer solstices the variation was larger than in other months. The better estimations of β_m were obtained in March and in October. The greatest $RMSE$ was found for the city of Mohe (9.21°). The smallest discrepancy for this variable was found for the city of Aligarh (1.59°).

Table 2. Comparison between the results of this work and the results of 22 locations reported in the literature.

City	Jan. (Deg)	Feb. (Deg)	March (Deg)	April (Deg)	May (Deg)	June (Deg)	July (Deg)	Aug. (Deg)	Sept. (Deg)	Oct. (Deg)	Nov. (Deg)	Dec. (Deg)	RMSE (Deg)
Izmir	0	−1	−1.5	1.6	6.9	8.6	9.7	3.8	-0.2	−1.7	-0.5	0.6	4.46
Sanya	6.07	6.16	3.84	−0.24	-0.44	−0.57	-1.36	−0.22	1.79	4.75	6.13	8.31	4.32
Shanghai	10.54	6.52	6.6	5.77	5.69	5.54	6.17	4.86	4.39	6.53	10.26	9.13	7.10
Zhengzhou	9.42	6.58	4.95	5.36	6.94	8.02	7.73	6.14	6.21	6.87	6.42	8.18	7.00
Harbin	3.11	1.41	2.12	7.36	11.32	14.87	13.4	9.5	4.03	2.12	3.19	4.48	7.85
Mohe	0.57	−2.71	-0.64	6.9	15.53	17.47	16.56	11.05	4.09	1.09	0.65	1.62	9.21
Lhasa	−0.1	-0.02	0.1	1.37	4.87	6.61	5.45	3.35	0.4	−0.98	-0.9	0.59	3.06
Sanliurfa	5.31	3.47	−1.5	−4.31	−5.87	−5.81	−5.48	−4.05	−1.53	3.32	4.59	5.98	4.52
Tabass	−0.65	−1.63	−1.97	0.77	5.98	3	5.35	2.9	−2.23	−2.86	−1.34	0.07	2.94
Aligarh	1.09	−0.07	−0.87	1.08	1.12	−3.94	−1.6	2.15	−0.82	−1.1	0.25	1.26	1.59
Adana	5.94	4.05	2.95	3.99	7.36	6.41	8.75	5.23	2.87	1.88	4.23	6.62	5.39
Ankara	7.1	5.38	3.82	6.46	9.1	10.21	10.55	5.85	3	2.42	5.47	8.78	7.01
Diyarbakır	5.47	3.66	1.86	4.22	7.72	7.8	10.14	4.53	0.9	−0.39	2.81	6.16	5.44
Erzurum	2.25	1.49	1.83	6.34	8.9	10.01	10.35	5.69	2.94	1.51	2.64	3.96	5.79
Istanbul	8.91	7.23	5.8	6.58	8.28	11.42	10.75	7.01	3.05	4.34	7.32	10.6	7.99
Izmir	5.12	4.3	2.52	4.9	7.4	8.5	9.86	5.25	1.62	1.34	3.54	6.85	5.70
Samsun	7.52	6.85	6.41	7.17	9.87	12	11.33	6.58	4.63	4.92	5.89	8.19	7.93
Trabzon	8.34	7.67	8.24	9.01	9.71	11.84	11.16	9.42	7.45	6.75	7.73	9.02	8.97
New Delhi	1.4	2.28	−0.75	0.2	2.12	−2.94	−0.6	3.11	−0.91	1.69	1.55	2.06	1.85
Helwan	7.37	0.26	−0.62	−5.46	−0.45	−5.47	−6.11	−9.41	−6.06	−5.17	2.58	6.05	5.37
Darussalam	9.86	7.24	3.92	−5.4	−5.55	−5.63	−3.31	−5.6	−1.22	10.73	7.69	8.55	6.75
Damascus	0.79	−0.49	−0.86	2.17	6.54	8.76	7.7	3.65	−0.59	−1.5	−0.07	1.07	4.10
Average tilt angle deviation ($\Delta\beta$)	5.02	3.11	2.1	2.99	5.59	5.75	6.2	3.67	1.53	2.11	3.64	5.36	

In each of the rows corresponding to the twelve months of the year the difference between the reported tilt angle and the calculated with the algorithm is shown. A negative sign means that β_m was underestimated, on the contrary the angle was overestimated. In these simulations, the average β_m always was overestimated as it can be seen in the last row. Near the winter and summer solstices the variation was larger than in other months. The better estimations of β_m were obtained in March and in October. The greatest *RMSE* was found for the city of Mohe (9.21°). The smallest discrepancy for this variable was found for the city of Aligarh (1.59°).

4.4. Correction Factor

The results of the section 4.3 suggests that by considering exclusively the direct sun radiation, β_m is overestimated. Therefore, we propose an empirical expression, in which the monthly tilt angle might be estimated better by

$$\hat{\beta}_m = \beta_m - \Delta\beta \,, \tag{19}$$

where $\hat{\beta}_m$ is a new estimation of the monthly tilt angle and $\Delta\beta$ is the correction factor, which can be obtained from the last row of the table 2. This correction might help to improve the value of the monthly tilt angle calculated with the program presented in this work.

5. Conclusions

A system for the estimation of the tilt angle in different world locations has been implemented. The computation of this angle is realized for different time periods.

Data provided by a GPS receiver are automatically passed to an algorithm without the intervention of the user. This approach might be more effective and it may avoid errors.

The code is written in a high-level programming language, which makes the process of increasing its capabilities simpler.

The estimation of the yearly evolution of average radiation flux incident on the collector surface for different angles shows that the monthly variation of the tilt angle produces almost the same energetic gain than moving this angle every six minutes.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The authors want to thank to the Instituto Politécnico Nacional (IPN) and the Comisión de Operación y Fomento de Actividades Académicas COFAA-IPN for the financial support provided to do this work.

References

[1] J. Fonseca-Campos, L. Fonseca-Ruiz, P. N. Cortez-Herrera "Portable system for the calculation of the sun position based on a laptop, a GPS and Python" in 2016 IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC 2016), Ixtapa Mexico, 2016. https://doi.org/10.1109/ROPEC.2016.7830580

[2] E. D. Mehleri, P. L. Zervas, H. Sarimveis, J.A. Palyvos, N. C. Markatos, "Determination of the optimal tilt angle and orientation for solar photovoltaic arrays" Renew. Energ., **35**(11), 2468–2475, 2010. https://doi.org/10.1016/j.renene.2010.03.006

[3] H. Gunerhan, A. Hepbasli, "Determination of the optimum tilt angle of solar collectors for building applications", Build. Environ., **42**(2), 779–783, 2007. https://doi.org/10.1016/j.buildenv.2005.09.012

[4] A. Agarwal, V. Kumar Vashishtha, S.N. Mishra, "Comparative approach for the optimization of tilt angle to receive maximum radiation", Int. J. Engine. Res. & Tech. (IJERT), **1** (5), 1–9, 2012. http://www.ijert.org

[5] H. Zang, M. Guo, Z. Wei, G. Sun, "Determination of the optimal tilt angle of solar collectors for different climates of China" Sustain., **8** (654), 1–16, 2016. https://doi.org/10.3390/su8070654

[6] M. Kacira, M. Simsek, Y. Babur, S. Demirkol, "Determining optimum tilt angles and orientations of photovoltaic panels in Sanliurfa, Turkey Renew. Energ., **29** (8), 1265–1275, 2004. https://doi.org/10.1016/j.renene.2003.12.014

[7] H. Khorasanizadeh, K. Mohammadi, A. Mostafaeipour, "Establishing a diffuse solar radiation model for determining the optimum tilt angle of solar surfaces in Tabass, Iran" Energ. Convers. Manage., **78**, 805–814, 2014. https://doi.org/10.1016/j.enconman.2013.11.048

[8] B. Jamil, A. T. Siddiqui, N. Akhtar, "Estimation of solar radiation and optimum tilt angles for south-facing surfaces in humid subtropical climatic region of India" Engine. Sci. and tech. an Int. J., **19**(4), 1826–1835, 2016. https://doi.org/10.1016/j.jestch.2016.10.004

[9] K. Bakirci, "General models for optimum tilt angles of solar panels: Turkey case study" Renew. Sust. Energ. Rev., **16**(8), 6149–6159, 2012. https://doi.org/10.1016/j.rser.2012.07.009

[10] M. J. Ahmad, G. N. Tiwari, "Optimization of tilt angle for solar collector to receive maximum radiation" The Open Renew. Energ. J., **2**, 19-24, 2009. https://doi.org/10.2174/1876387100902010019

[11] H. K. Elminir, A. E. Ghitas, F. El-Hussainy, R. Hamid, M. M. Beheary, K. M. Abdel-Moneim, "Optimum solar flat-plate collector slope: Case study for Helwan, Egypt" Energ. Convers. Manage., **47**(5), 624–637, 2006. https://doi.org/10.1016/j.enconman.2005.05.015

[12] M. A. H. M. Yakup, A. Q. Malik, "Optimum tilt angle and orientation for solar collector in Brunei Darussalam" Renew. Energ., **24**(2), 223–234, 2001. https://doi.org/10.1016/S0960-1481(00)00168-3

[13] K. Skeiker, "Optimum tilt angle and orientation for solar collectors in Syria" Energ. Convers. Manage., **50**(9), 2439–2448, 2009. https://doi.org/10.1016/j.enconman.2009.05.031

[14] E. A. Handoyo, D. Ichsani, Prabowo, "The optimal tilt angle of a solar collector", Enrgy. Proced., **32**, 166–175, 2013. https://doi.org/10.1016/j.egypro.2013.05.022

[15] H. Darhmaoui, D. Lahjouji, "Latitude based model for tilt angle optimization for solar collectors in the Mediterranean region" Enrgy. Proced., **42**, 426–435, 2013. https://doi.org/10.1016/j.egypro.2013.11.043

[16] Ch. Ying-Pin, "Optimal the tilt angles for photovoltaic modules in Taiwan" Int. J. Elec. Power & Energy Syst., **32**(9), 956–964, 2010. https://doi.org/10.1016/j.ijepes.2010.02.010

[17] M. Benghanem, "Optimization of tilt angle for solar panel: Case study for Madinah, Saudi Arabia" Appl. Energ., **88**(4), 1427–1433, 2011. https://doi.org/10.1016/j.apenergy.2010.10.001

[18] J. A. Duffie, William A. Beckman, Solar engineering of thermal processes, John Wiley & Sons, Inc.; 2006.

[19] H. Heywood, "Operating experiences with solar water heating" J. Inst. Heat., Vent. & Energ., **39**, 63–69, 1971.

[20] H. Moghadam, F. F. Tabrizi, A. Z. Sharak, "Optimization of solar flat collector inclination" Desalination, **265** (1–3), 107–11, 2011. https://doi.org/10.1016/j.desal.2010.07.039

[21] G. O. G. Löf, R. A. Tybout, "Cost of house heating with solar energy" Sol. Energy, **14**(3), 253–78, 1973. https://doi.org/10.1016/0038-092X(73)90094-7

[22] C. Stanciu, D. Stanciu, "Optimum tilt angle for flat plate collectors all over the World–A declination dependence formula and comparisons of three solar radiation models" Energ. Convers. Manage., **81**, 133–43, 2014. https://doi.org/10.1016/j.enconman.2014.02.016

[23] P. I. Cooper, "The absorption of solar radiation in solar stills" Sol. Energy, **12**(3), 333–46, 1969. https://doi.org/10.1016/0038-092X(69)90047-4

[24] M. Blanco-Muriel, D. C. Alarcón-Padilla, T. López-Moratalla, M. Lara-Coira, "Computing the solar vector" Sol. Energy, **70**(5), 431–441, 2001. https://doi.org/10.1016/S0038-092X(00)00156-0

[25] R. Grena, "An algorithm for the computation of the solar position" Sol. Energy, **82**(5), 462–470, 2008. https://doi.org/10.1016/j.solener.2007.10.001

[26] J. J. Michalsky, "The astronomical Almanac's algorithm for approximate solar position (1950–2050)" Sol. Energy, **40**(3), 227–235, 1988. https://doi.org/10.1016/0038-092X(88)90045-X

[27] I. Reda, A. Andreas, "Solar position algorithm for solar radiation applications" Sol. Energy, **76**(5), 577–589, 2004. https://doi.org/10.1016/j.solener.2003.12.003

[28] R. Grena, "Five new algorithms for the computation of sun position from 2010 to 2110" Sol. Energy, **86**(5), 1323–1337. 2012. https://doi.org/10.1016/j.solener.2012.01.024

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1502-1507 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Source Code Vulnerabilities in IoT Software Systems

Saleh Mohamed Alnaeli[*,1], Melissa Sarnowski[2], Md Sayedul Aman[3], Ahmed Abdelgawad[3], Kumar Yelamarthi[3]

[1]*CSEPA, University of Wisconsin-Colleges, 53715, USA*

[2]*Computer Science, University of Wisconsin-Fox Valley, 54952, USA*

[3]*College of Science and Engineering, Central Michigan University, 48859, USA*

ARTICLE INFO

Article history:
Received: 02 June, 2017
Accepted: 21 July, 2017
Online: 15 August, 2017

Keywords:
unsafe commands
vulnerable software
scientific
security
static analysis
historical trends

ABSTRACT

An empirical study that examines the usage of known vulnerable statements in software systems developed in C/C++ and used for IoT is presented. The study is conducted on 18 open source systems comprised of millions of lines of code and containing thousands of files. Static analysis methods are applied to each system to determine the number of unsafe commands (e.g., strcpy, strcmp, and strlen) that are well-known among research communities to cause potential risks and security concerns, thereby decreasing a system's robustness and quality. These unsafe statements are banned by many companies (e.g., Microsoft). The use of these commands should be avoided from the start when writing code and should be removed from legacy code over time as recommended by new C/C++ language standards. Each system is analyzed and the distribution of the known unsafe commands is presented. Historical trends in the usage of the unsafe commands of 7 of the systems are presented to show how the studied systems evolved over time with respect to the vulnerable code. The results show that the most prevalent unsafe command used for most systems is memcpy, followed by strlen. These results can be used to help train software developers on secure coding practices so that they can write higher quality software systems.

1. Introduction

This paper was originally presented in 2016 IEEE 3rd World Forum on Internet of Things (WF-IoT) [1]. Additional research has been done to extend the research to examine the security of a much larger group of IoT software systems. While the IoT continues to grow to billions of devices running a huge variety of software systems, both open source and proprietary, security becomes a major concern for individuals and organizations who use the IoT in both academia and industry. The security of every software system becomes vital as each software system and device could be a target or an access point to hackers, or lawbreakers [2-4]. Most of the software and embedded systems used for IoT applications are currently available as open source systems and are therefore developed by programmers from varying disciplines and with different backgrounds. Many of those programmers have little to no background on security challenges imposed by the usage of vulnerable source code in some programming languages (e.g., C/C++) caused by commands that are well known to the research community as being unsafe and are banned by companies such as Microsoft.

According to [5, 6], most of the detected security threats are due to vulnerabilities in the code. Thus, minimizing usage of insecure commands can play an important role in protecting the software systems from any potential attacks. In order to minimize the use of unsafe commands at the source code level, developers need to be made aware of those unsafe commands and the security issues their usage can cause. Educating developers and ensuring that they follow good programming practices can minimize the time and effort spent on finding and fixing them in later stages, as well as lessen the expense of fixing a security issue if it's exploited

[*]Corresponding Author: Saleh Mohamed Alnaeli, 1478 Midway Rd, Menasha Wisconsin 54952, (920) 832-2615, USA | Email: saleh.alnaeli@uwc.edu

by an attacker. For example, C/C++ programming languages are preferred by developers because of the level of performance, flexibility, and efficiency that they offer. However, security vulnerabilities (e.g., integer vulnerability, buffer overflow, and string vulnerability) need to be addressed and avoided from the beginning when writing the source code to save the time and expense that comes with having to refactor the system at a later date to remove those security vulnerabilities. Additionally, programmers also need to make sure that they do not use commands that are known to cause security concerns (e.g., memcpy, strlen, and strcmp).

As an example, the standard C library includes a function called gets() that is used primarily for reading strings input by the user. This function accepts a pointer from data type char as a parameter and reads a string of characters from the standard input, placing the first character in the location specified by that pointer and subsequent data consecutively in memory. The gets() function will continue reading until a newline is detected, at which point the buffer is terminated with a null character. The issue is that the developer cannot determine the size or length of the buffer passed to gets() prior to runtime. Because of this, when the buffer is size bytes, an attacker trying to write size + extra bytes into the buffer will always succeed if the data excludes newlines [5]. Consequently, memory locations that are adjacent to the buffer in the memory may be overwritten, which could lead to sensitive data being modified if it is stored in those adjacent memory locations. Additionally, an attacker could even overflow the stack and lead the program to run into an arbitrary or unexpected status. A safe alternative to gets() is fgets(), which, unlike gets(), also accepts an integer number as a parameter that acts as the limit of characters copied into the string, including the null-character at the end.

The process of protecting software systems from vulnerability issues at the source code level is done by either completely removing known unsafe commands and functions, or by replacing them with safer replacements (e.g., strncpy, strncmp, strncat etc.). Here, we have extended a previous empirical examination of some of the open source software systems used for IoT to better understand how well IoT developers do when it comes to the usage of vulnerable code by studying a larger group of IoT systems and the history of some of those systems to analyze the change in the usage of unsafe commands over time. We are particularly interested in determining the most prevalent unsafe statements that occur in a wide variety of IoT software applications and if there are general trends, and are interested in seeing if the trends stay the same or change among a larger group of systems, as well as if there are historical trends for the group of systems whose history was also examined. While this work does not directly address the problem of removing unsafe source code from the systems, it does serve as a foundation for understanding the problem requirements in the context of a broad set of applications. Moreover, the focus of this research is on unsafe statement detection and identifying their distribution over time.

In this study, several software systems used for IoT and developed in C/C++ are analyzed and evaluated with respect to the usage of vulnerable commands that can affect the quality of IoT systems. Each of the studied systems are written in C/C++. For each system, the source code is analyzed a count is created for each of the unsafe functions as well as the safer replacement functions. These counts are compared against the counts of each other system to uncover trends and make observations about the usage of unsafe functions. Furthermore, the history of some of the systems, if it was available and of a sufficient length, is examined and the number of detected unsafe functions is calculated for each release.

This work focuses on addressing the following questions: how many unsafe functions are used in these systems, which of those are the most frequent, and what are their respective distributions? Are the numbers or distributions of unsafe functions changing over the history/versions of a software system? When a larger group of systems is studied, do the trends uncovered in the original study remain the same, or do new trends emerge?

This work contributes by extending one of the only studies on the usage of vulnerable functions in IoT software applications. We found that the functions memcpy and strlen represent most of the unsafe functions occurring in these systems, in both the original systems and the additional systems studied in this extension. Moreover, the findings show that, for the most part, the systems reviewed have become more vulnerable to attacks over time due to the increase in the number of the unsafe functions used over time. This knowledge will assist researchers in formulating and directing their work to efficiently address this problem when refactoring and designing new systems.

The remainder of this paper is organized as follows. Section 2 gives background information and related work on the topic of source code vulnerabilities and security of IoT and its challenges. Section 3 describes the methodology used in the study along with how we performed the analysis to create counts for each unsafe function. Section 4 presents the findings of our study of 18 open source software systems used for IoT. There is a discussion of results in the same section as well. The historical trends of 7 of the systems found are explained in section 5, followed by threats to validity and future enhancement in 6 and 7. Finally, conclusions are found in section 8.

2. Background and Related Work

Most of the previous research conducted on IoT security has focused on identifying security concerns related to the communication processes and authentication methods used with IoT. Some other studies have examined data privacy within various levels. To the best of our knowledge, no study has examined the usage or distribution of vulnerable code in IoT systems that are developed in C/C++.

A study has been conducted by VERACODE [7] to investigate a selection of always-on consumer IoT devices to understand the security posture of each product. They found that product manufacturers prioritized design instead of security and privacy, putting consumers at risk for an attack or physical intrusion. Their team performed a set of uniform tests across all devices and organized the findings into four different domains: user-facing cloud services, back-end cloud services, mobile application interface, and device debugging interfaces. The results showed that most of the tested devices exhibited vulnerabilities across most categories. The findings prove that there is a need to perform security reviews on device architecture and accompanying applications to minimize the risk to users.

Hui Suo, Jiafu Wan, Caifeng Zou, and Liu in [2] reviewed security in the IoT and analyzed security characteristics and requirements from four layers: perceptual layer, network layer, support layer, and application layer. They discussed the research status in this field from encryption mechanism, communication security, protecting sensor data, and encryption algorithm. While they confirmed the need to develop technologies and methodologies that meet the IoT needs for meeting the higher requirements of security and privacy, they did not discuss security concerns that can be caused by programming standards or by using vulnerable code.

A research group from George Mason University and National Institute of Standards and Technology in [3] presented a set of use cases that leverage commercial off-the-shelf services and products to raise awareness of security challenges in current practices and prove that there is a need for IoT security standards to be developed, as well as their possible implications. They recommended that experts begin formulating suitable guidance and identifying the right security and privacy primitives for more secure and reliable IoT products. However, the study has not discussed any security issues related to vulnerabilities at the source code level and possible security risks that might be caused by the usage of unsafe statements.

Although literature is rich with studies that focus on the methods and tools used for detecting vulnerable source code [5], no studies have been conducted specifically in the domain of IoT that evaluate the usage and the distribution of vulnerable source code in IoT software systems and applications prior to the study of which this is an extension.

This work extends a previous work on IoT security in which we conducted an empirical study of unsafe functions on the source code level. We empirically examined 18 systems to determine how many unsafe functions are used and their distributions to help software engineers develop better software systems for IoT and to show how these systems evolve over time in terms of secure programming standards based on the usage of unsafe versus safe replacement functions. We extended the work by examining a larger group of IoT systems, adding 15 more to the original group of software systems studied.

3. Methodology for detecting unsafe functions

A function is considered unsafe if it is one of the functions well-known to both the research community and industry to cause security concerns. Some of those unsafe functions and commands are already banned by compiler producers (e.g., Microsoft). Literature is abundant with lists of unsafe C/C++ commands. We used a tool, UnsafeFunsDetector, developed by one of the main authors, to analyze source code files and, if they contain any unsafe function calls, create a count for each unsafe function. First, we collected all files with C/C++ source-code extensions (i.e., c, cc, cpp, cxx, h, and hpp). For the systems whose history was analyzed, the last version of the system for each year was used. Then, we used the srcML (www.srcML.org) toolkit [1,11] to parse and analyze each file. The srcML format wraps the statements and structures of the source code syntax with XML elements, allowing tools, such as UnsafeFunsDetector, to use XML APIs to locate pieces of code, such as unsafe functions, and to analyze expressions in a quick and efficient manner. Once the system is converted into XML, UnsafeFunsDetector iteratively parses every source code unit to find each call of the unsafe functions and safe replacement functions and adjusts the counters. That is, a count of each unsafe function was recorded. Finally, all calls of unsafe functions were counted and their distributions determined.

Table 1. The 18 studied systems along with the total unsafe functions used in them and the most prevalent unsafe function.

System	Total Unsafe Functions	Most Prevalent Unsafe Function
ApacheMyNewtOS	1,524	memcpy
AtomThreads	62	strlen
Contiki	1,859	memcpy
DistortOS	3	memcpy
Embox	10,286	memcpy
FemtoOS	0	N/A
FreeOSEK	48	memcpy
Lepton	3,928	memcpy
nOS	4	memcpy
OpenTag	58	memcpy
openWSN	220	memcpy
PicoOS	12	free
POK	49	memcpy
TinyOS	772	memcpy
Tneo	1	memcpy
Trampoline	637	free
uOS-Embedded	15,556	puts
Zephyr	3,340	memcpy

The systems chosen in this study were carefully selected to represent a variety of open source systems developed in C/C++ and used for IoT and well-known to both academia and research IoT communities. Our findings are presented and discussed later in this paper, along with the limitations of our approach.

Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar.

4. Findings, results and discussion

We now study the usage of unsafe functions in the studied systems and with their distributions, along with the historical trends of their distributions of for 7 of the studied systems.

OpenWSN is an open source project that provides open-source implementations of a complete protocol stack based on Internet of Things standards, for a variety of software and hardware platforms. This implementation can help both academia and industry verify the applicability of these standards to the Internet of Things for those networks to become truly ubiquitous [8].

The TinyOS is an open source, operating system designed for low-power wireless devices, such as those used in sensor networks, ubiquitous computing, smart buildings, and smart meters [9]. Contiki is a lightweight and flexible operating system for tiny networked sensors [10]. These were the original 3 IoT systems studied in the original paper, but this paper has extended the work

to analyze 18 total systems, along with the most recent 5-year history of 7 of those systems.

4.1. Design of the Empirical Study

This study focuses on three aspects regarding the security of software systems used for IoT in terms of unsafe function and vulnerable code usage. First, we examine the number of calls to known unsafe functions. This gives an idea of how much of the system needs to be refactored to remove or replace the vulnerable code and, therefore, increase its security and quality. Next, we examine which unsafe functions are the most prevalent. This can give the developers of IoT software an idea about the most prevalent unsafe function to they should make a priority, if they plan on refactoring for the purpose of improving their system's security and quality, so that they can increase their system's security in the most efficient way possible. Finally, we examine how the presence of unsafe commands (as opposed to some safer replacements functions) changes over the lifetime of a software system.

The following is our study defined by a set of formal research:

RQ1: What is the total number of unsafe functions called in each system?

RQ2: Of those called, which unsafe functions are the most prevalent?

RQ3: Over the history of a system, is the presence of unsafe functions, and the use of safe replacement functions, increasing or decreasing?

We now examine our findings within the context of these research questions.

4.2. Number of Detected Unsafe Function and their distribution

In the original study, Contiki had the largest number of unsafe functions. In this extension, uOS-Embedded has the largest number of unsafe functions at 15556 total unsafe functions, as opposed to Contiki's 1859 unsafe functions. The smallest number of unsafe functions was 0 for FemtoOS. To address RQ1, Table 1 shows the total number of unsafe functions for each of the 18 systems studied. For the systems whose history was also studied, Table 1 shows the number of unsafe functions of the most recent version. Some of the systems had very few unsafe functions called, while others had thousands of calls to unsafe functions.

In this study, snprintf is considered an unsafe function. Some compliers consider it as a safe replacement function, but it was considered unsafe in this study as it is banned by Microsoft.

For most of the systems, memcpy was the most called unsafe function. In some of the systems, memcpy was the only unsafe function called (e.g., DistortOS, nOS, and Tneo). This trend matches the trend for the original study on only Contiki, TinyOS, and openWSN. Even when a much larger group of IoT systems is studied, mempy remained the most prevalent unsafe function for the majority of the systems, which could allow us to start making better generalizations about the IoT domain when it comes to security than we could when a smaller group was studied.

To address RQ2, Table 2 shows the top 3 most prevalent unsafe functions called across the 18 IoT systems studied. When the top 3 most prevalent for each system was analyzed, memcpy appeared in the top 3 most prevalent functions of 15 systems, strlen appeared in the top 3 most prevalent functions of 10 systems, and strcmp appeared in the top 3 most prevalent functions of 6 systems. The rest of the functions were in the top 3 most prevalent functions of 2 or 3 systems, or they were not present in the top 3 most prevalent functions of any system. The difficulty in generalizing the top 3 most prevalent unsafe functions was that many of the systems varied in the unsafe functions that they called the most, and for some systems, the only unsafe function called was memcpy. While memcpy was the most prevalent unsafe function for the majority of the systems, it was not the most prevalent for all of the systems. The most prevalent for uOS-Embedded was puts, which was not in the top 3 most prevalent unsafe functions for any of the other systems studied.

Table 2. The top three most prevalent unsafe functions across most of the systems

System	memcpy	strcmp	strlen
ApacheMyNewtOS	771	221	220
AtomThreads	12	0	24
Contiki	712	90	343
DistortOS	3	0	0
Embox	2,363	1,726	1,720
FemtoOS	0	0	0
FreeOSEK	17	10	9
Lepton	766	398	662
nOS	4	0	0
OpenTag	42	0	0
openWSN	211	0	2
PicoOS	0	0	0
POK	28	4	6
TinyOS	236	133	94
Tneo	1	0	0
Trampoline	78	40	135
uOS-Embedded	1,500	2,016	1,790
Zephyr	1,778	404	616

Clearly, there is still a noticeably presence of well-known unsafe functions in most of the studied systems, which complicates security concerns when it comes to systems used for IoT. But no matter how we present the data, it is apparent that unsafe functions present one of the most serious security issues that need to be addressed through refactoring systems to remove the unsafe functions. While literature is rich with studies focusing on addressing the problems of how to remove those unsafe functions from software systems, it appears that software developers are underestimating, or lack understanding of, the real threats imposed by the use of unsafe functions.

5. Historical Change of unsafe function frequency

In order to address RQ3, we looked at the most recent 5-year history for 4 additional systems to the original 3 studied. We examined the most recent version of each year and recorded the number of unsafe functions used, along with the number of safe replacement functions used. We are interested in knowing whether

the number of unsafe functions is increasing over time and the security risks are becoming more prevalent, or if the number is

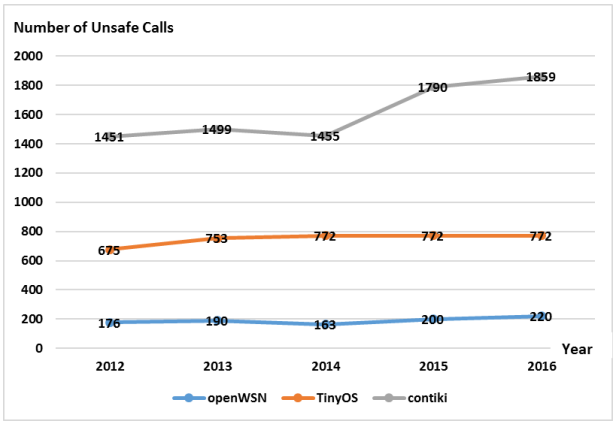


Figure. 1. The change in use of unsafe functions over time for the original three systems studied previously.

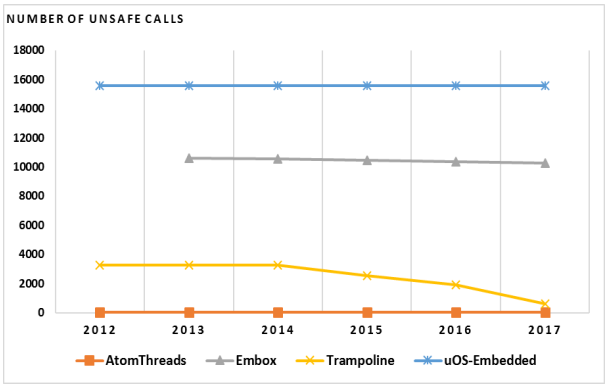


Figure. 2. The change in use of unsafe functions for the additional four systems studied.

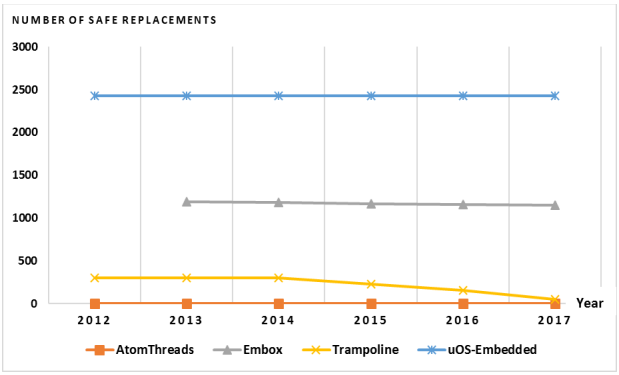


Figure. 3. The change in use of safe alternatives for the four additional systems studied.

lowering, possibly that the developers worked on removing the unsafe functions to increase the security and quality of their system. Figure. 1. shows the change in number of unsafe functions used for the original study [1]. Figure. 2. shows the change in number of unsafe functions for the 4 additional systems studied.

It can be seen that in the original 3 systems, the trends for two systems were relatively flat, while the trend for one system showed an increase in the use of unsafe functions over the 5-year period. We can also see in Figure. 2. that in the 4 additional systems, 3 of them showed very flat trends with 2 of those not having any change in the number of unsafe functions or safe replacements, shown in Figure. 3., beyond the first few years, over the 5-year period. The system, Trampoline, that did not show a flat trend instead showed a decreasing trend in the use of both unsafe functions and safe replacement functions. In 2012, Trampoline had 3265 unsafe functions. In 2017, that number had decreased to 637, although there had been a flat trend for the years 2012-2014.

While most of the systems don't show an increasing trend, as in they are not becoming more vulnerable to security risks, they are also not showing a decreasing trend, which can be seen as equally bad. The presence of unsafe functions remains the same, and the developers are not removing those functions in order to increase the security and quality of their systems. This could be that the developers of those systems are unaware of the risks imposed by the use of well-known unsafe functions, which leads to the need for better education on the topic of software security at the source code level to help those developers create higher quality systems.

6. Threats to Validity

The tools we developed and used for this study only work with languages supported by srcML (C/C++). Because of this limitation, we were unable to include systems written in languages such as Python and Java for this study.

Another limitation is that the tool we used is unable to differentiate between unsafe functions used in dead code, which means that some of the unsafe functions counted may never be called during the system's runtime. This might affect the accuracy of the results we present in terms of the systems' security and vulnerability. Additionally, the calls to unsafe functions that are included in wrappers are not excluded.

7. Future Enhancement

In the future, we are planning to improve the tool so that it only includes active code and exclude the calls to the unsafe functions that are protected with wrappers. We would also like to be able to include more systems written in different programming languages.

In this study, all calls to unsafe functions were counted including regardless the potential wrappers. We are planning to improve the tool so that it excludes the calls that are protected by proper wrappers.

8. Conclusion

This study empirically examined the usage of known unsafe functions and commands in eighteen open source software systems. The systems are all IoT applications written in C/C++ specifically for IoT architectures. There are no other studies of this type currently in the literature. The results show that usage of vulnerable functions is still common for most of the systems, although some systems had very few or even no calls to unsafe functions. Of the eighteen systems studied, memcpy was the most prevalent for the majority of the systems followed by strlen, free and strcmp. The historical trend, for selected systems, shows that developers are not working to improve a problem that still exists.

The vast majority of literature, concerning IoT security, focused on the security issues in the communication layer rather than vulnerabilities at the source code level. As such, more attention needs to be placed on dealing with source vulnerability,

reduce the usage of unsafe statements, especially the most prevalent statements to improve IoT platforms in terms of security, and educate developers on ways to both refactor their systems and to avoid the use of unsafe functions from the beginning when writing code, thus enhancing performance.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This work was supported in part by a grant from the US National Science Foundation (NSF) Grant no. 1542368.

References

[1] S. M. Alnaeli, M. Sarnowski, M. S. Aman, A. Abdelgawad, and K. Yelamarthi, "Vulnerable C/C++ code usage in IoT software systems," in 2016 IEEE 3rd World Forum on Internet of Things (WF-IoT), 2016, pp. 348-352.

[2] H. Suo, J. Wan, C. Zou, and J. Liu, "Security in the Internet of Things: A Review," in Computer Science and Electronics Engineering (ICCSEE), 2012 International Conference on, 2012, pp. 648-651.

[3] C. Kolias, A. Stavrou, J. Voas, I. Bojanova, and R. Kuhn, "Learning Internet-of-Things Security; 'Hands-On'," IEEE Security & Privacy, vol. 14, pp. 37-46, 2016.

[4] A. M. Gamundani, "An impact review on internet of things attacks," in Emerging Trends in Networks and Computer Communications (ETNCC), 2015 International Conference on, 2015, pp. 114-118.

[5] J. Viega, J. T. Bloch, Y. Kohno, and G. McGraw, "ITS4: a static vulnerability scanner for C and C++ code," in Computer Security Applications, 2000. ACSAC '00. 16th Annual Conference, 2000, pp. 257-267.

[6] R. K. McLean, "Comparing Static Security Analysis Tools Using Open Source Software," in Software Security and Reliability Companion (SERE-C), 2012 IEEE Sixth International Conference on, 2012, pp. 68-74.

[7] Veracode, "The Internet of Things: Security Research Study". (2015). https://www.veracode.com/sites/default/files/Resources/Whitepapers/internet-of-things-whitepaper.pdf

[8] The University of California. (2016). openWSN https://openwsn.atlassian.net/wiki/pages/viewpage.action?pageId=688187

[9] The TinyOS Working Group, (2013), TinyOS. http://www.tinyos.net/

[10] A. Dunkels, B. Gronvall, and T. Voigt, "Contiki - a lightweight and flexible operating system for tiny networked sensors" in 29th Annual IEEE International Conference on Local Computer Networks, 2004.

[11] S. M. Alnaeli, A. A. Taha, and T. Timm, "On the Prevalence of Function Side Effects in General Purpose Open Source Software Systems," in Software Engineering Research, Management and Applications, R. Lee, Ed., ed Cham: Springer International Publishing, 2016, pp. 115-131.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1508-1514 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Lookup Tables-based mean level detection of spatially distributed targets in non Gaussian clutter

Nabila Nouar*, Atef Farrouki

Laboratoire SISCOM, Département d'électronique, Faculté de Sciences et Technologies, Université des Frères Mentouri, Constantine 1, Algeria.

ARTICLE INFO

Article history:
Received: 30 May, 2017
Accepted: 18 July, 2017
Online: 15 August, 2017

Keywords :
Distributed targets
MDS
Lookup Tables
CFAR detection
Inverse Gamma distribution
Parameters estimation
Maximum Likelihood

ABSTRACT

In this paper, Constant False Alarm Rate (CFAR) detection of spatially distributed targets embedded in compound Gaussian clutter with Inverse Gamma texture is addressed. By taking into account the fact that clutter parameters are unknown in practical situations, we propose mean level based on Lookup Tables detectors, that operate as a two-step approach, which consists of computing threshold factors that maintain a Constant Probability of False Alarm (Pfa) using intensive Monte Carlo simulations, and storing these factors in Lookup Tables, this first step is done independently from the detection algorithm. Then, at the detection stage, the detectors structure is associated to the Maximum Likelihood (ML) estimation technique to estimate the shape and the scale parameters, and compare them to the closest integer and half integer values in the Lookup Tables, to select the suitable threshold factor. Under the High Resolution Radar (HRR) assumption, the target is spread over a number of cells according to the Multiple Dominant scattering centers (MDS) representation. The binary total binary hypothesis tests are derived using the expression of the overall target energy, which is computed as the sum of the energies reflected from each cell.

Performance analysis of the proposed detectors is carried out using Monte Carlo simulation for various couples of clutter parameters and MDS models, and are compared to those of the Cell Averaging Based on Lookup Tables detector (CA-LT). In order to assess the performances of the proposed detectors in terms of the radar resolution, their performances are compared to the case of point-like targets.

1. Introduction

It is well established in Radar detection systems that High Resolution Radar systems (HRR) can resolve a target into a number of scatterers, leading to a spatial distribution of the target energy on a number of cells, which is commonly referred to as "distributed targets". This paper is an extension of the work presented in [1] , in which the detection of spatially distributed targets in compound Gaussian clutter with Inverse Gamma texture has been addressed. In fact, Two detectors are proposed: the Greatest Of based on Lookup Tables (GO-LT) , and the Smallest Of based on Lookup Tables detector(SO-LT) . As for the CA-LT [2], and Both detectors are designed to operate without à priori information about the clutter parameters by first storing threshold factors that maintain a Constant Probability of False Alarm (Pfa) in Lookup Tables, and then estimating the actual clutter parameters using the Maximum Likelihood (ML) method to select the suitable threshold factor to be used in the hypothesis test.

The target energy is considered spread over a number of cells, according to the MDS concept, three MDS models are taken into account, and both the target energy profile and the scatterers locations are specified for each considered model. Moreover, expressions of binary hypothesis tests for both detectors are derived using the expression of the total target energy, formerly proposed in [2]. Detection Performances of (GO-LT) and (SO-LT) are analyzed in terms of the Probability of Detection (Pd) using intensive Monte Carlo simulations, and is compared to those of the Cell-Averaging based on Lookup Tables detector (CA-LT) [2] assuming different MDS models, number of primary cells and couples of clutter parameters.

Detection of range spread targets under the Gaussian assumption has been widely explored with reference to the MDS target model [4]. Initially, it has been analyzed under the Gaussian

*Correponding author: Nabila Nouar, Département d'électronique , Université des Frères Mentouri Constantine, Algeria | E-mail : nabila_nouar@yahoo.com

assumption (see [5,6 and references therein]) . However, with the support of experimental data, it has been demonstrated that the clutter samples from sea clutter returns are better described as a *compound Gaussian process.* The latter *is* the product of a temporally slowly varying texture component and a locally rapidly varying speckle component [4, 7, 8, and 9]. Consequently, various adaptive detection schemes of range spread targets embedded in compound Gaussian clutter have been proposed and investigated. For instance, the GLRT-based detector [8], the detectors with Rao and Wald tests [4], the GLRT linear quadratic (GLRT-LQ) [9]. Also, the Non-Scatterer density dependent GLRT (NSDD-GLRT) [10], the normalized adaptive matched filter (NAMF) [11] have been investigated under compound Gaussian assumptions.

In [2], the authors introduced the Cell averaging based on Lookup Tables detector (CA-LT), which detects spatially distributed targets embedded in K-distributed clutter with unknown parameters. An expression of the binary hypothesis test for range spread targets has also been derived using the expression of the total target energy, and the performances of the CA-LT have been compared to those of the Logarithmic Cell Averaging detector CAL [12]. Another based on Lookup Tables approach, with non-coherent integration of multiple pulses have been proposed. The latter is designed to detect range spread targets without *a priori* information about the clutter parameters. By taking into account the non-coherent integration technique, and the target energy profile, the authors derived a new expression of the range spread target energy, and an expression of the binary hypothesis test. The performances of the proposed detectors have been compared to those of OS-GLRT [13], and results indicated that even though both detectors present better detection performance when the target is uniformly distributed on the primary cells, the (M-pulse CA-LT) still outperforms the OS-GLRT [13].

2. Assumptions and problem formulation

2.1. Clutter model

Under the HRR assumption, sea clutter returns are better described using the compound Gaussian model. In fact, the clutter e_i is sampled from the ith range cell, and is expressed as the product of two components, namely, the texture τ_i and the speckle S_i, as follows:

$$e_i = \sqrt{\tau_i} S_i \tag{1}$$

Where $S_i = S_{iI} + jS_{iQ}$ is a stationary complex Gaussian process, and S_{iI} and S_{iQ} represent the in-phase (I) and quadrature (Q) components of S_i respectively [16]. In this work, we consider that the texture component follows the Inverse Gamma distribution [14], with the Probability Density Function (PDF) of τ_i [14, 15]:

$$f_{\tau_i}(\tau;\alpha,\beta) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} \tau^{-(\alpha+1)} e^{-1/\beta\tau} \tag{2}$$

Where α is the shape parameter, β represents the scale parameter and Γ (.) is the Gamma function.

Consequently, the magnitude of the clutter, denoted by X_i can be written as [14,15]

$$X_i = |e_i| = \sqrt{\tau_i}|S_i| \tag{3}$$

The PDF of X_i , namely $f_{X_i}(x)$, is expressed as [16]:

$$f_{X_i}(x) = \frac{2x\beta\Gamma(\alpha+1)}{(\beta x^2+1)^{\alpha+1}\Gamma(\alpha)} \tag{4}$$

2.2. Target model

According to the MDS representation, the target is considered spread over a finite number of cells , also referred to as "primary data, and its total energy , namely Δ, is expressed as the sum of the partial energies reflected by each range location as follows [2]:

$$\Delta = \sum_{k=1}^{Np} a_k {X_0}^k \tag{5}$$

Where a_k are multiplicative factors representing the energy proportion of the kth range cell, and ${X_0}^k, k = 1,2, \ldots Np$ refer to the group under test. Motivated by the work presented in [1,2].We consider in this paper three MDS models , with three and (Np=3) four primary cells (Np=4) . The total useful energy backscattered from each cell and corresponding range locations are given in Table 1.

Table 1: Discrete Scatterers Locations and the amount of the total reflected energy.

Model Number	Cell number			
	1	2	…	Np
1	1/Np	1/Np	…	1/Np
2	1/2	1/2	0	0
3	1	0	0	0

3. Greatest Of and Smallest Of based on Lookup Tables detectors

3.1. Working principle of GO-LT and SO-LT

The proposed detectors are based on the Greatest Of (GO) detector and the Smallest Of (SO) detector to detect spatially distributed targets embedded in compound Gaussian clutter with inverse Gamma texture. The (GO-LT) and (SO-LT) are designed to detect distributed targets using Lookup Tables (LT) and online clutter parameters estimation. As illustrated in Figure 1, the threshold factors $T(\alpha,\beta)$ that maintain a constant Pfa are computed for different couples of clutter parameters (α,β) using intensive Monte Carlo simulations and are stored in Lookup Tables (LT). At the detection stage, the clutter parameters the (α,β) are not *a priori*, and must be estimated. Therefore, the detectors structures are associated to the Maximum Likelihood estimation (ML) method, which estimates the shape and the scale parameters $(\hat{\alpha},\hat{\beta})$. These estimates are compared to the couples of clutter parameters in the Lookup Tables and are approximated to the closest integer or half integer pre-computed value of (α,β), once the closest couple of clutter parameters is selected, the suitable threshold factor $T(\alpha,\beta)$ is indexed accordingly .

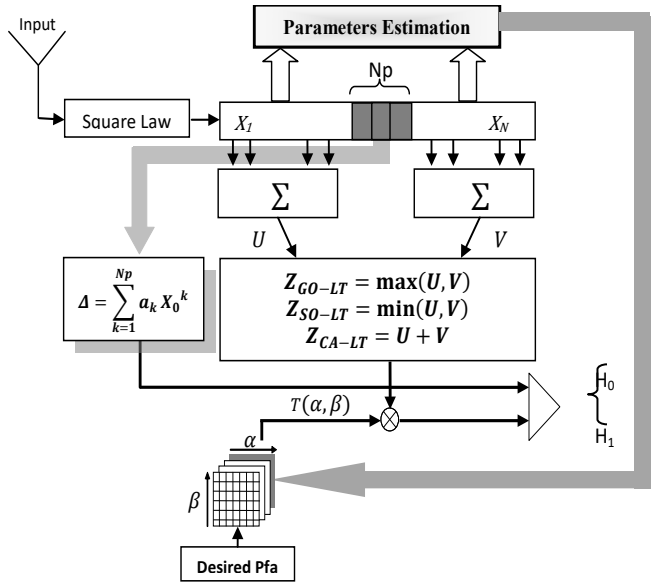


Figure 1: Working principle of GO-LT, SO-LT and CA-LT

The decision about the presence or the absence of a distributed target among the group under test ${X_0}^k, k = 1,2, \dots Np$ if taken by comparing the total reflected energy Δ to the adaptive threshold $T(\alpha,\beta)$ multiplied by the clutter level estimate Z. With respect to the proposed detectors, (GO-LT) and (SO-LT), we refer to the clutter level estimates as $Z_{GO\text{-}LT}$ and $Z_{SO\text{-}LT}$ respectively. The latter are estimated using the N reference cells by selecting the greatest and the smallest of the outputs *U* and *V*, as follows:

$$Z_{GO-LT} = \max(U,V) \tag{6}$$

$$Z_{SO-LT} = \min(U,V) \tag{7}$$

Here, the outputs U and *V* are computed as the average of the leading and the lagging reference windows respectively:

$$U = \frac{2}{N}\sum_{i=1}^{\frac{N}{2}} X_i \tag{8}$$

$$V = \frac{2}{N}\sum_{i=\frac{N}{2}}^{N} X_i \tag{9}$$

Where N/2 refers to the length of the lagging and the leading window.

Moreover, it is well known that in the CA-LT detector [2], the clutter level, namely, $Z_{CA\text{-}LT}$ is estimated by computing the mean of the reference cells surrounding the primary cells as follows:

$$Z_{CA-LT} = U + V \tag{10}$$

Consequently, the binary hypothesis tests of the GO-LT, SO-LT and CA-LT detectors can be respectively written as:

$$\sum_{k=1}^{Np} a_k {X_0}^k \underset{H_0}{\overset{H_1}{\gtrless}} T(\alpha,\beta) \begin{cases} \max(U,V) & GO-LT \\ \min(U,V) & SO-LT \\ \mathrm{U+V} & \mathrm{CA-LT} \end{cases} \tag{11}$$

It is important to note that the previous binary hypothesis test can be used to detect point like-targets by setting *Np* to 1, and Δ to X_0 in (11). Otherwise stated, when the number of primary cells reduces 1, we obtain a single cell under test X_0 which leads to:

$$X_0 \underset{H_0}{\overset{H_1}{\gtrless}} T(\alpha,\beta) \begin{cases} \max(U,V) & GO-LT \\ \min(U,V) & SO-LT \\ U+V & \mathrm{CA-LT} \end{cases} \tag{12}$$

3.2. Clutter parameters estimation using the Maximum Likelihood (ML) approach

Since the clutter parameters are not known in real radar application, many estimation methods, such as the (ML) estimation method [16] have been explored. In this study, we consider two mean level based on Lookup Tables detectors that are each associated to the (ML) method. The latter is associated to the detectors structure to estimate the shape and the scale parameters in order to automatically switch to the suitable threshold factor $T(\alpha,\beta)$ among the pre-computed values in the Lookup Tables. This method is used to estimate the clutter parameters (α,β) of the compound Gaussian clutter with inverse Gamma texture. As presented in [16], the ML estimate of the shape parameter, namely $\hat{\alpha}$, is expressed as a function of the clutter samples as follows:

$$\widehat{\alpha_{ML}} = \frac{N}{\beta \sum_{i=1}^{N} \frac{X(i)^2}{\beta X(i)^2 + 1}} - 1 \tag{13}$$

The scale parameter is estimated using the following concentrated Log Likelihood function of *β*, $\mathcal{L}_c(\beta)$ [16]:

$$\mathcal{L}_c(\beta) = \frac{N\beta \sum_{i=1}^{N} \frac{X(i)^2}{\beta X(i)^2 + 1}}{N - \beta \sum_{i=1}^{N} \frac{X(i)^2}{\beta X(i)^2 + 1}} - \sum_{i=1}^{N} log\,(\beta X(i)^2 + 1) \tag{14}$$

The ML estimate of *β*, namely, $\hat{\beta}$, is a zero point of this function [16]. Therefore, we use the following approach:

First, we compute the function $\mathcal{L}_c(\beta)$ for different values of the scale parameter β. Then, the estimate of β corresponds to the value of β that leads to a zero point of the function.

4. Results and discussion

Performance analysis of the proposed detectors is carried out using Monte Carlo simulations based on 100/Pfa independent trials and considering a nominal Pfa=10^{-3} and Pfa=10^{-2}.

We analyze the effect of different clutter parameters *(α,β)*= (1.5,1), (2,1),(2.5,1), and (3,1) on the detection performances of the proposed detectors. For the target, it is considered distributed over Np=3 and Np=4 primary cells respectively, and three MDS models that describe the target energy profile are investigated. Moreover, by taking into account the clutter parameters estimation approach, we set N=64 reference cells

We also consider the following Signal to Clutter Ratio (SCR):

$$SCR = 10log_{10}\left(\frac{2\sigma^2}{\frac{1}{\beta}\frac{\Gamma(\alpha-1)}{\Gamma(\alpha)}}\right) \quad (15)$$

Here , The parameter σ is used to generate the primary cells ${X_0}^k, k = 1,2, \dots Np$ for different SCR. The term $2\sigma^2$ refers to the target power. As for the CA-LT detector [2], the cells under test ${X_0}^k$ are scaled by a_k and are summed according to (5). The target's envelope echoes are described as independent random variables distributed according to the Chi-square law with mean square value ${\sigma^2}/{Np}$ [1].

4.1. Threshold factors T(α,β)

The threshold factors $T(\alpha, \beta)$ maintaining a Constant Pfa of 10^{-3} of the detectors GO-LT, SO-LT and CA-LT , assuming Np=4 primary cells are given in Table 2, Table 3 and Table 4 respectively. From Table 2 , that refers to the (GO-LT) detector, we observe that for the couple (α,β)=(2,1) , the threshold factor T=0.1514, where as for the couple (α,β)=(1.5,1) the threshold factor T=0.1920 . In fact, the threshold factor values continue decreasing as the shape parameter α increases, which is the case of α=3, where the highest values of T (T=0.1262) is required to maintain a Constant Pfa. Moreover, for couples of parameters with the same values of the scale parameter β, we have the same value of T. For instance, if we consider different couples (α,β) with α=1.5 : (α,β)=(1.5,1) , (1.5,1.5) and (1.5,2) , the required value of T is 0.1920.

From Table 3 and Table 4, we observe that the (SO-LT) and the (CA-LT) detectors exhibit the same behavior with regards to the threshold factors. Lower values of the threshold factor *T(α, β)* are required to maintain a Constant Pfa for higher values of the shape parameter α, and the same values of T are required to maintain a Constant Pfa when different values of β are considered.Hence, the scale parameter β does not affect the Pfa of all detectors. Hence, it can be scaled to 1 *(β=1)*. We conclude that the proposed detectors guarantee the CFAR property with regards to the scale parameter. Also, by comparing the threshold factors of all detectors, we observe that the SO-LT exhibits the lowest threshold factors for any couple of parameters *(α, β)*.

Table 2: Threshold factors of the detector GO-LT Np=4 , Pfa=10^{-3}

α / β	1.5	2	2.5	3
1	0.1920	0.1514	0.1352	0.1262
1.5	0.1920	0.1514	0.1352	0.1262
2	0.1920	0.1514	0.1352	0.1262

Table 3: Threshold factors of the detector SO-LT Np=4 , Pfa=10^{-3}

α / β	1.5	2	2.5	3
1	0.2573	0.1950	0.1715	0.1605
1.5	0.2573	0.1950	0.1715	0.1605
2	0.2575	0.1950	0.1715	0.1605

Table 4: Threshold factors of the detector CA-LT Np=4 , Pfa=10-3

α / β	1.5	2	2.5	3
1	0.1090	0.0842	0.0745	0.0681
1.5	0.1090	0.0842	0.0745	0.0681
2	0.1090	0.0842	0.0745	0.0681

Table 5: Threshold factors of the detector GO-LT Np=4, Pfa=10-2

α / β	1.5	2	2.5	3
1	0.1080	0.0915	0.0900	0.0848
1.5	0.1080	0.0915	0.0900	0.0848
2	0.1080	0.0915	0.0900	0.0848

Table 6: Threshold factors of the detector SO-LT Np=4, Pfa=10-2

α / β	1.5	2	2.5	3
1	0.1320	0.1202	0.1109	0.1100
1.5	0.1320	0.1202	0.1109	0.1100
2	0.1320	0.1202	0.1109	0.1100

For comparison purposes, we also present threshold factors $T(\alpha, \beta)$ maintaining a Constant Pfa of 10^{-2} of the detectors GO-LT, SO-LT and CA-LT , assuming Np=4 in Table 5, Table 6 and Table 7 respectively. We observe that the threshold factors exhibit the same behavior with regards to the clutter parameters, and that the scale parameter β still does not affect the threshold factors values. We also observe that among the three detectors, the SO-LT still requires the lowest value of the threshold factor. On the other hand, by comparing the obtained values of the threshold factors with Pfa=10^{-2} (Table 5, Table 6, Table 7) to those obtained with Pfa=10^{-3},(Table 2, Table 3, Table 4), it is clear that the threshold factor is affected by the value of the prescribed Pfa. In other terms, if we consider the same couple of parameters (α,β), the value of the threshold factor that required in the case of Pfa (10^{-2}) is lower than the the threshold factor required in the case of (10^{-3}).

Table 7: Threshold factors of the detector CA-LT Np=4, Pfa=10^{-2}

α / β	1.5	2	2.5	3
1	0.0581	0.0520	0.0488	0.0475
1.5	0.0581	0.0520	0.0488	0.0475
2	0.0581	0.0520	0.0488	0.0475

At this point , it is also important to stress that the previous results have been obtained considering four primary cells (Np=4). Thus , for comparison purposes , and in order to analyze the proposed detectors for different scenarios, we also considered the case of three primary cells (Np=3). The threshold factors *T(α, β)* that are required to maintain a Constant Pfa of 10^{-3} in case of three primary cells (Np=3) for the GO-LT, SO-LT and CA-LT are given in Table 8, Table 9 and Table 10 respectively. An interesting result is that higher values of threshold factor T are required to maintain a Constant Pfa for a lower number of primary cells. Otherwise stated , by comparison to the case of (Np=4) , if we decrease the number of primary cells (Np=3), the threshold factor that is required to maintain a Constant Pfa increases .

Table 8: Threshold factors of the detector GO-LT Np=3, Pfa=10^{-3}

α / β	1.5	2	2.5	3
1	0.2210	0.1750	0.1520	0.1450
1.5	0.2210	0.1750	0.1520	0.1450
2	0.2210	0.1750	0.1520	0.1450

Table 9: Threshold factors of the detector SO-LT Np=3, Pfa=10-3

α / β	1.5	2	2.5	3
1	0.2785	0.2240	0.1950	0.1800
1.5	0.2785	0.2240	0.1950	0.1800
2	0.2785	0.2240	0.1950	0.1800

Table 10: Threshold factors of the detector CA-LT Np=3 , Pfa=10-3

α / β	1.5	2	2.5	3
1	0.1244	0.0954	0.1520	0.0778
1.5	0.1244	0.0954	0.1520	0.0778
2	0.1244	0.0954	0.1520	0.0778

4.2. Detection performance

In this section, we examine the detection performances of the GO-LT, SO-LT and CA-LT detectors considering different couples of clutter parameters , MDS models and numbers of primary cells Np , including the conventional case of single range cell , i.e, point like target.Performances are plotted using the Pd versus SCR plots. Detection performances of GO-LT, SO-LT and CA-LT for Np=4 (MDS-1) and Np=3 (MDS-1) considering different couples of clutter parameters (*α,β)* are presented in Figure 2 and Figure 3, respectively

We first observe that all detectors exhibit better detection performance for higher values of the shape parameter α, which is the case of the couple (α,β)=(3,1). However , all detectors behave the same for any couple of parameters, i,e, all detectors present the same behavior with the couples (α,β)=(1.5,1), and (α,β)=(3,1). This result remains true for any number of primary cells (Np=3 or Np=4).

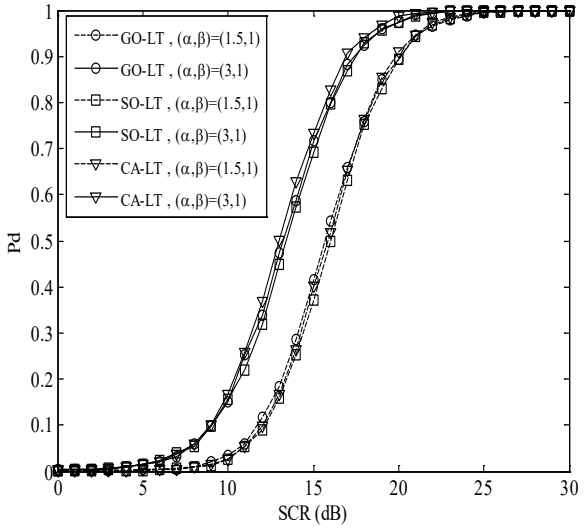


Figure 2: Pd of GO-LT, SO-LT and CA-LT for different clutter parameters (α, β) for Np=4 , MDS-1 ,Pfa=10^{-3}

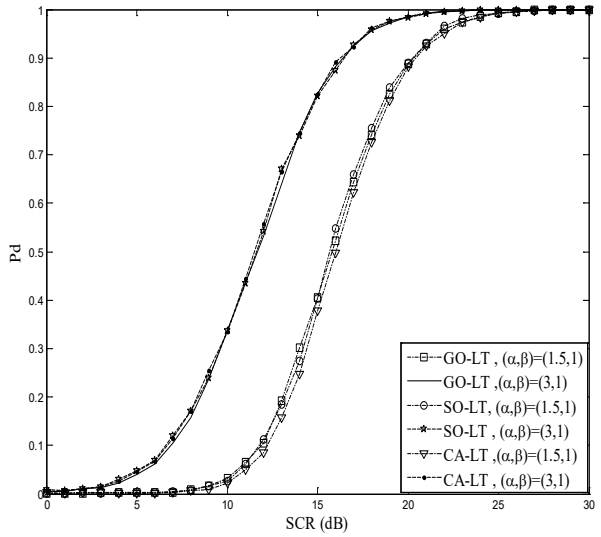


Figure 3: Pd of GO-LT, SO-LT and CA-LT for different clutter parameters (α, β) for Np=3 , MDS-1 ,Pfa=10^{-3}

Moreover, it is important to analyze the effect of the target energy profile, i,e, the MDS model on the detection performances of the proposed detectors. Hence, the Pds of GO-LT, SO-LT and CA-LT against SCR are illustrated in Figure 4, Figure 5 and Figure 6 respectively. We consider that the target is spread over four primary cells (Np=4) according to three different MDS models (MDS-1, MDS-2 and MDS(3). It is shown that all detectors present the best performance with MDS-1 model, and it is degraded with MDS-2 and MDS-3. This result can be interpreted by the fact that all detectors perform best when the target energy is uniformly distributed on the primary cells $X_0{}^k, k = 1,2, \dots Np$, and it is degraded as the energy becomes more concentrated in a specific cell as it is the case of MDS-2, and MDS-3 models. Therefore, the MDS-1 model is adopted for the next simulations.

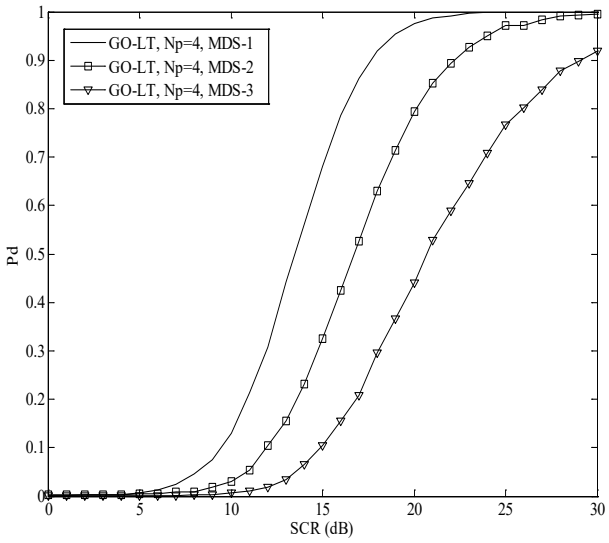


Figure 4: Pd of GO-LT for different MDS models for Np=4

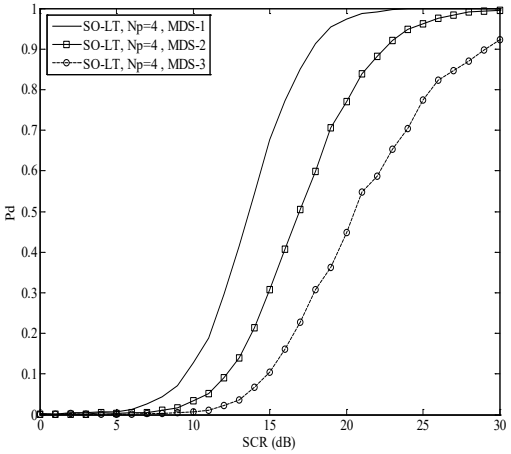


Figure 5: Pd of SO-LT for different MDS models for Np=4

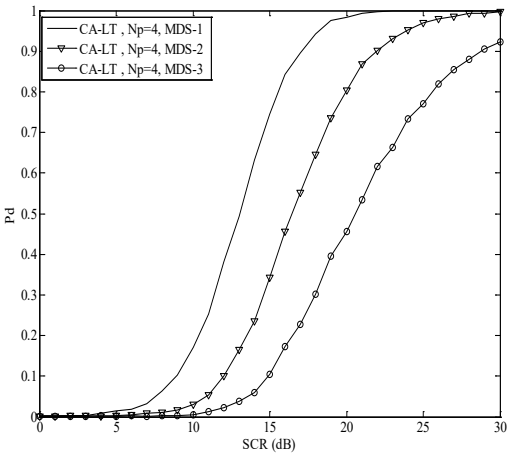


Figure 6: ,Pfa= Pd of C-LT for different MDS models for Np=4

On the other hand, it is important to compare the performances of the proposed detectors considering different numbers of primary cells Np. In Figure 7, we analyze the performances of the GO-LT, SO-LT and CA-LT detectors considering spatially distributed targets with Np=4 and Np=3 , and compare them to the case of point-like targets (Np=1). The plots indicate that the best detection performance is achieved with Np=4 , and it is degraded as the number of primary cells Np decreases (Np=3). We also observe that the proposed detectors exhibit the worst performance with point like targets (Np=1), which is expected since the detection performance is enhanced by increasing the radar resolution, i,e, the number of primary cells.

Another interesting aspect of the proposed detection approach is the online clutter parameters estimation. Therefore, it is important to check the dependency of the Pfa of the proposed detectors on the clutter parameters. In fact, we present the Pfa of the GO-LT , SO-LT and CA-LT , considering a nominal Pfa of 10^{-3} and 10^{-2} in Figure 8 and Figure 9 respectively. It is clear that the Pfa of the three detectors does not depend on the clutter parameter α, and it is maintained almost constant for all the values of α. As stated above, the Pfa of the proposed detectors does not depend on the values of the scale parameter β, which indicates that the proposed detection approach ensures the CFAR property with respect to both the shape and the scale parameters.

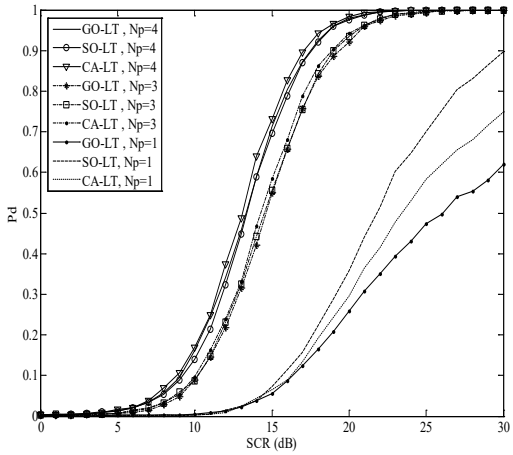


Figure 7: Pd of GO-LT , SO-LT and CA-LT for distributed targets (Np=4) and point-like targets (Np=1)

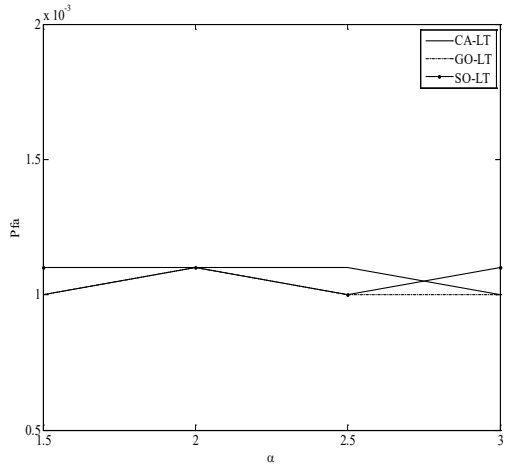


Figure 8: Pfa of CA-LT, GO-LT and SO-LT as a function of the shape parameter α for a nominal Pfa=10^{-3}

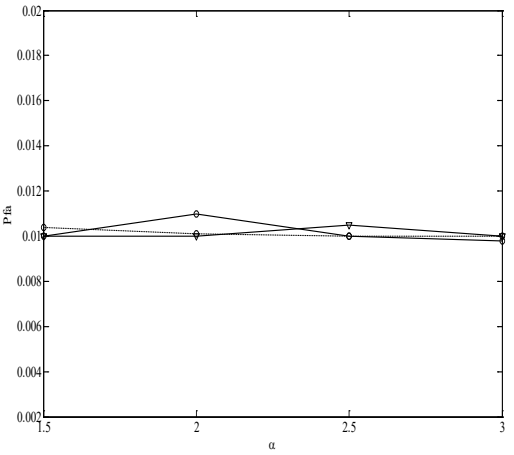


Figure 9: Pfa of CA-LT, GO-LT and SO-LT as a function of the shape parameter α for a nominal Pfa=10^{-2}

Finally, the effect of the scale parameter β on the performances of the CA-LT detector is presented in Figure 10. We observe that the same performance is achieved for different values of β.

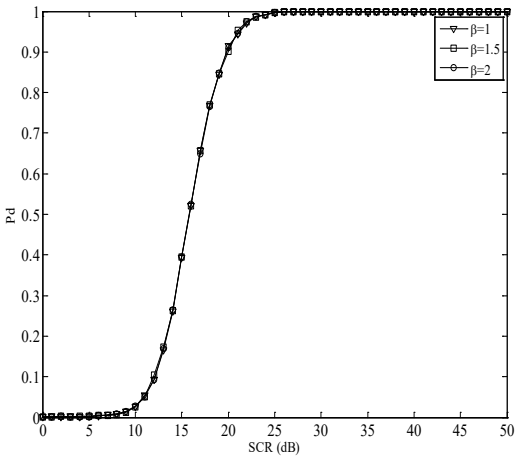


Figure 10: Effect of the scale parameter β on the detection performances of the CA-LT

5. Conclusion

In this paper, the detection of distributed targets in compound Gaussian clutter, with Inverse Gamma texture have been examined by designing a new Look-up Tables based detection approach . In fact, the proposed detectors are based on Lookup Tables that contain threshold factors, maintaining a constant Pfa, for different clutter parameters. Then, the ML technique is associated to the detectors structure in order to automatically switch to the most suitable threshold factor. To decide the presence of the target among the group of cells under test, the selected threshold factor is multiplied by the clutter level estimate and is compared to the total target energy. The performances of the proposed detectors have been investigated assuming three MDS models and compared to the point like targets. Simulation results indicated that all detectors present the best performance, when the target energy is uniformly distributed over the primary cells. However, they still present similar performance when the same MDS model is considered.

Another interesting result with regards to the proposed target energy profile is that all detectors perform better when the number of primary cells increases. Indeed, increasing the number of primary cells in the proposed detection scheme is equivalent to increasing the radar resolution, which can enhance the detection performances.

Finally, we conclude that the proposed detectors are applicable in practical situations since they can operate in more realistic situations such as: target energy profile when HRR systems are considered, and non-availability of priori information about the clutter parameters.

References

[1] N.Nouar, A.caFarrouki, “Detection of distributed targets embedded in homogeneous compound Gaussian clutter with Inverse Gamma texture”. In 2017 Seminar on Detection systems Architectures and Technologies (DAT), Alger, Algeria, 2017

[2] N.Nouar, A.Farrouki, “CFAR detection of spatially distributed targets in K-distributed clutter with unknown parameters”. In Proc.22nd European Signal Processing conference (EUSIPCO ’14), Lisbon, Portugal, 2014.

[3] N.Nouar, A.Farrouki “Lookup tables-based detection of range spread targets in compound Gaussian environment with multiple-pulse non-coherent integration”. Signal, Image Video Process. 11(61), 1–8 (2017).

[4] E. Conte, A.De. Maio, “Distributed target detection in compound- Gaussian noise with Rao and Wald tests,” *IEEE Transactions on Aerospace and Electronic Systems.,* **39**(2), 568-582, . 2003.

[5] E. Conte, A.De. Maio, “GLRT-based adaptive detection algorithms for range-spread targets,” *IEEE Transactions on Signal Processing.,***49**(7), 1336-1348,2001.

[6] H.V. Trees, *Detection, Estimation, and Modulation Theory*, Part III Wiley, New York, 2001.

[7] G. Alfano, A.De. Maio, A. Farina, “Model-based adaptive detection of range-spread targets,” *IEE Proceeding of Radar, Sonar and Navigation.,* **151**(1), 2-10, 2004.

[8] E. Conte, A.De. Maio, G.Ricci, “CFAR detection of distributed targets in non-Gaussian disturbance,” *IEEE Transactions on Aerospace and Electronic Systems.,* **38**(2), 612-621, 2002.

[9] F. Gini, M. Greco, “Suboptimum approach for adaptive coherent radar detection in compound-Gaussian clutter,” *IEEE Transactions on Aerospace and Electronic Systems.,***35**(3), 1095-1104,1999.

[10] K. Gerlach, “Suboptimum approach for adaptive coherent radar detection in compound-Gaussian clutter,” *IEEE Transactions on Aerospace and Electronic Systems.,***35**(3), pp. 1095-1104, July. 1999.

[11] E. Conte, M. Lops, G. Ricci, “Adaptive matched filter detection in spherically invariant noise,” *IEEE Signal Processing Letters,***3**(8), 248-250, 1996.

[12] T. Bucciarelli, P.Lombardo , and S. Tamburrini, “Optimum CFAR detection against compound Gaussian clutter with partially correlated texture ,” *IEE proceedings-Radar, Sonar and Navigation*, **148**(4), 95-104, 2001.

[13] Y. He, T. Jian, F. Su, C. Qu, X. Gu, “Novel Range Spread target detectors in Non-Gaussian clutter”. *IEEE Transactions on Aerospace and Electronic Systems.,**46**(3), 1312-1328, 2010.*

[14] A. Dogandzic, A. Nehorai, J. Wang, “Maximum likelihood estimation of compound-Gaussian clutter and target parameters,” in *Proc. 12th Annual Int. Workshop Adaptive Sensor Array Processing (ASAP ’04),* Lincoln Laboratory, Lexington, March 2004.

[15] K.L. Lange, R.J.A. Little, and J.M.G. Taylor, “Robust statistical modeling using the t distribution.” *Journal of the American Statistical Association,* **84**, 881-896, 1989.

[16] A.Balleri, A.Nehorai, and J. Wang, “Maximum Likelihood estimation for compound-Gaussian clutter with inverse Gamma texture,” *IEEE Transactions on Aerospace and Electronic Systems,* **43** (2), 775-779, 2007.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1515-1526 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

A Machine Learning based Framework for Parameter based Multi-Objective Optimisation of Video CODECs

Maryam Al-Barwani*, Eran A. Edirisinghe

Department of Computer Science, Loughborough University, Loughborough, UK

ARTICLE INFO

Article history:
Received: 03 June, 2017
Accepted: 15 July, 2017
Online: 15 August, 2017

Keywords:
Video Codec
Encoder/Decoder
H.264/MPEG4-AVC
High Efficiency Video Coding
Multi-objective Optimisation

ABSTRACT

All multimedia devices now incorporate video CODECs that comply with international video coding standards such as H.264 / MPEG4-AVC and the new High Efficiency Video Coding Standard (HEVC), otherwise known as H.265. Although the standard CODECs have been designed to include algorithms with optimal efficiency, a large number of coding parameters can be used to fine-tune their operation, within known constraints of for example, available computational power, bandwidth, energy consumption, etc. With the large number of such parameters involved, determining which parameters will play a significant role in providing optimal quality of service within given constraints is a further challenge that needs to be met. We propose a framework that uses machine learning algorithms to model the performance of a video CODEC based on the significant coding parameters. We define objective functions that can be used to model the video quality as Peak Signal-to-Noise Ratio (PSNR), CPU time utilization and Bit-Rate. We show that these objective functions can be practically utilised in video Encoder designs, in particular in their performance optimisation within given constraints. A Multi-objective Optimisation framework based on Genetic Algorithms is thus proposed to optimise the performance of a video codec. The framework is designed to jointly minimize the complexity, Bit-rate and to maximize the quality of the compressed video stream.

1. Introduction

This paper is an extension of work originally published in Future Technologies Conference FTC 2016 San Francisco, United States[1].

Applications that benefit from accurate video capture, efficient representation and coding, error-free transmission and subjectively optimised display, have been growing over the years due to the availability of higher network bandwidth, faster processor speed and advanced capture and display technologies. Recent studies have shown that coded video data is contributing a the major part of consumer internet traffic with a predicted share of 90% by 2019.

Some of the most extensively used applications include real-time video conferencing, video streaming over broadband networks and digital TV broadcasting. Most current mobile hand-held devices come equipped with a video camera that is able to capture and encode a video stream in a standard format. These devices also include video players, which can decode and play back video. All the above developments continuously demand more efficient video coding algorithms that are able to reduce the bitrate without sacrificing video quality or to enable the increase of video resolution, without increasing the bitrate. High Efficiency Video Coding (HEVC) also known as H.265 is the most recent answer to this consumer demand, demand which supersedes the more widely used video coding standards such as MPEG-2 and H.264.

All advanced video CODECs have many parameters that can be used to control their operational characteristics, both at the encoder and decoder ends, enabling the possibility of fine tuning their operation for maximum efficiency within environments and application scenarios that are bound by various constraints. For example the available bandwidth will have an upper limit, the network will be subjected to delays and the decoder/display unit may have limitations in processing and display capabilities. Yet the encoder, transmission and decoder have many parameters that can be adjusted for them to be efficiently operational under the above mentioned constraints. Identifying the values of these parameters that results in the CODECs optimal performance under given constraints remains an open research problem of vital importance.

*Correponding author: Maryam Al-Barwani, Department of Computer Science, Loughborough University, UK | Email: m.al-barwani@lboro.ac.uk

The first step of parameter based optimisation of a video CODEC is the identification of the coding parameters that have a significant impact on its key properties, such as, bandwidth, image/video quality, and CPU cycles. Although an experienced user of a video CODEC can guess these parameters with some accuracy when the content of the video is known, a formal scientific approach is needed to accurately decide the parameter set, with minimum subjective error. Having obtained these parameters, it is then possible to model the key properties of the video CODEC described above based on the significant parameters. These models can then be used to optimise the performance of the video codec when operated under practical constraints, thus making the parameter based characterisation and modelling practically useful.

In this paper we propose a framework that is based on the fundamentals of machine learning that can be used to scientifically determine the significant coding parameters of a video CODEC. These parameters are then used to model the operational behaviour of the video CODEC for which machine learning algorithms are further utilised. We also show that this model can be used to establish the foundations of a multi-objective optimisation framework. Optimisations algorithms are widely used to solve many difficult optimisation problems in other research areas. Although the experiments conducted are limited to H,.264 and H.265 standards, the proposed framework can be used in relation to any video coding standard.

For clarity of presentation, the remainder of this paper is structured as follows: Related work and the background of H.264 & H.265 video coding is introduced in section 2. Section 3 presents the proposed framework for performance modelling and the experiments conducted for establishing the framework. Section 4 presents a comprehensive analysis of the results of the performance modelling of H.264 along with analysis of Optimisation. Section 5 presents the results and analysis of H.265 video codec and Optimisation stages carried out using a Matlab based implementation. Finally, section 6 concludes the paper.

2. Related Work

A significant amount of research has been conducted in the past and presented in literature on the optimization of video coding/compression algorithms. Parameter-based optimization focuses on the selection of the optimal set of coding parameters that will influence the optimal overall performance of the video CODEC, given operational constraints such as bandwidth, distortion, and CPU.

In [2] , a joint power-rate-distortion (P-R-D) framework, for the analysis, control and the optimization of the behaviour of rate-distortion (R-D) algorithms of a wireless video communication system, under constraints of energy consumption, was proposed.

In [3] it was shown that the approach presented in [4] cannot be easily extended to other video encoders. It presented a novel power-rate-distortion (P-R-D) optimization algorithm that can be used to minimize energy consumption of delay tolerant applications in portable video communication.

2.1. H.264 video codec.

A joint power-distortion model was presented in [5] and was analysed under two constraints, namely, power consumption and video quality. The approach jointly considered the power consumption and video quality and analysed the two problems within a uniform optimization framework. The work presented in [6], proposed a power-rate-distortion (P-R-D) model of a video encoding system to maximize its operational lifetime. In [7] a novel equation for the prediction of distortion was proposed that was used for the optimization of quantization parameter (QP) selection. An improvement of image quality as compared to the standard rate control algorithm of the H.264 reference software JM15.01 was recorded.

A large number of coding parameters poses a practical operational problem of how to best select the set of parameters that will ensure the CODECs optimal performance, especially under multiple operational constraints. In [8], the challenge to determine H.264 parameter settings that have low complexity but still offer high video quality was investigated.

An improvement to the work presented in [8] was proposed by the same authors in [4], in which two further algorithms for finding additional parameter settings for the GBFOS-basic algorithm were presented. However, a significant constraint of the work of [7,3] is its limited use only within two constraints namely, complexity and video quality.

The author in [9] presented a detailed study of the importance of a multi-objective optimisation framework and the approach presented in [10] as a solution. The presented framework focused on the development of a joint complexity-rate distortion (C-M-R-D) optimization framework for a H.264 video CODEC, which could be extended to cover any number of constraints and to be used within any type of video CODEC.

2.2. H.265 video codec.

High Efficiency Video Coding (HEVC) is the next generation video coding standard being developed, the newest video coding standard of the ITU-T Video Coding Experts Group and the ISO/IEC Moving Picture Experts Group [11].

In [12] it was shown that HEVC provides significantly improved compression performance, i.e. an approximately 50% reduced bit rate as compared to the best existing video coding standards, under the same visual quality. The paper proposed a hardware-friendly method for RDO of HEVC intra coding. The results of the study showed that the proposed RD cost function provides 85.8% area reduction and 1260% throughput improvement in hardware design, with slight loss of bitrate and PSNR, which is suitable for real-time encoder applications.

A performance comparison of H.265, VP9 and H.264 encoders was presented in [13]. According to the experimental results, obtained for a whole test set of video sequences, by using similar encoding configurations for all three examined representative encoders, H.265/MPEG-HEVC was shown to provide significant average bit-rate savings of 43.3% and 39.3% relative to VP9 and H.264/MPEG-AVC, respectively.

In [14] it was shown that for resolutions of up to HD (1920x1080), code optimizations including heavy use of single instruction multiple-data (SIMD) instructions are sufficient to achieve HEVC real-time software decoding. It was further shown that when it came to decoding UHD video (3840x2160), single threaded execution with code optimization was not enough.

To improve the compression performance of current video coding standards by 50%, especially when it comes to transmitting high resolution video like 4K over the internet or in broadcast, the

50% bitrate reduction is essential. [15] shows that real-time decoding of 4K video with a frame- level parallel decoding approach using four desktop CPU cores is feasible.

Emerging video compression standard H.265/HEVC provides up to 2 times better compression efficiency compared to H.264/AVC standard. Iterative intra prediction search in [16] was proposed for the H.265/HEVC encoder to reduce the number of prediction modes for estimation: about 40% encoding time reduction for HM 10.1 intra-only coding with negligible bitrate increase and PSNR quality degradation. Additional speed-up techniques, including fast prediction error estimation, were offered.

3. Proposed Framework for Performance Modelling

The proposed framework for a MOO Multi-Objective Optimisation [2] is developed to determine the optimum coding parameters for a H.265 video CODEC, when working under multiple constraints as shown in Figure 1. The MOO framework is intended to minimize the complexity (CPU utilisation), bit-rate and to maximize the quality of the compressed video stream. The MOO framework proposed is accomplished by following the steps below.

1) Profiling experiments on the encoder and decoder were carried out to determine the coding parameters that have a significant impact on each of the objectives/constraints related to rate, distortion and CPU utilization. This was achieved by measuring the impact of each parameter (while being varied) on each of the above aspects.

2) Developing the objective function for each objective/ constraint, based on the above significant parameters, by using a suitable regression procedure.

3) These objective functions can then be used within a genetic algorithm (GA) based multi-objective optimization framework to determine optimal parameter values.

In a practical multimedia application scenario a device captures a video, encodes it and transmits it via a network to another device that decodes and displays the content to a viewer. Assuming that the network has bandwidth constraints and the device in which the encoder is placed has compute power constraints and the potential viewers of content may demand at least meeting minimal quality levels, a situation occurs in which the proposed MOO framework can be used may arise.

The significant number of encoder parameters that control the encoder's bit rate, quality and computational power requirements can be selected, to ensure the encoder performance is optimal, under the given multiple constraints. However this requires the modelling of the encoder's bit-rate, quality and CPU utilisation, based on the large number of selectable encoder parameters. If mathematical objective functions can be derived for each of the above, a standard approach to optimisation can be used. Deriving objective functions, for example using mathematical regression, will need the determination of the significant coding parameters, the key focus of the research presented below. The same explanation can be applied to the selection of decoder parameters that results in optimal decoder performance. Within the research context of this paper, we assume that the data transmission network is assumed to be perfect, i.e. no delays, no bit losses, no errors etc. Therefore the bit stream generated by the encoder is transmitted without any loss or alteration to the decoder, in real-time. The following section proposes the experimental process adopted to determine the significant coding parameters for both the encoder and decoder.

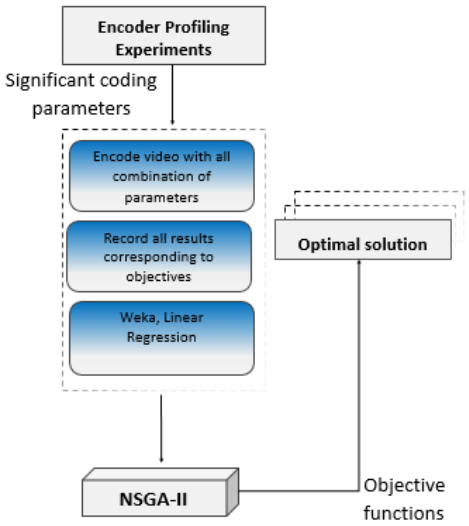


Figure 1: Proposed Multi-objective Optimisation framework.

4. Machine Learning based Framework for the analysis of significant coding parameters of H.264.

4.1. Profiling Experiments/ Determining the Significant Coding Parameters of H.264.

This experiment was carried out using a configuration file of the encoder software named JM (Joint Model) Reference software version 18.6 [17]. Six video sequences were encoded and decoded in the H.264 encoder by using a configuration file to set parameters of the mentioned video sequences. In this analysis, six QCIF video sequences were chosen for the experiment with a QCIF Quarter Common Intermediate Format, a videoconferencing format that specifies data rates of 30 frames per second (fps), with each frame containing 144 lines and 176 pixels per line. This is one fourth the resolution of Full CIF.

The first 30 frames were encoded for each video. Starting with Claire, a video sequences have a simple motion foreground and a non-moving area in the background, with a news presenter is talking by moving her head and eyes and mouth slowly. Coastguard video sequences have fast movement; on both regions foreground and background simultaneously, a boat is moving fast, waving water and a second boat will come into the scene within the final frames. A football video sequence has complicated fast motion with many players moving very quickly at the same time. Foreman video has a slight movement in the background; the Foreman is talking and his head is moving quite rapidly.

Mobile video sequences has a fast background and foreground movement; simultaneously, with a calendar moving upward, a ball and train are moving towards the left side with background. Tennis video sequences have a slow motion foreground and gentle movement in the background: the player's hand is moving and bouncing the ball and slightly zooming out of motion in the last few frames. The above six videos have different properties and motions as shown in Table 1.

Table 2 also tabulates the sample values used in our experiments for each parameter from within their corresponding value ranges. The control variable Intra-Period, can take values: 0 (meaning that the first frame is coded as an I-frame and subsequent frames are coded as P-frames), 5 and 8. The search window size is assumed to take either of the two values 16 or 32. The control variable Quantization Parameter (QP) is assumed to take two possible values 17 or 49. The number of reference frames (NRF) can take values 2, 5 and 8.

Table 1: H.264 Selected Video Sequences.

Claire 494 frames	**Coastguard 300 frames**
Football 260 frames	**Foreman 300 frames**
Mobile 300 frames	**Tennis 150 frames**

Table 2: Significant parameters and value used

Variables	Parameters	Values Range	Variable Type
IP= x(1)	Intra-Period	(0,5,8)	Numeric
SR= x(2)	Search-Range	(16, 32)	Numeric
QP= x(3)	Quantization Parameter	(17,49)	Numeric
NRF= x(4)	Number-Reference- Frames	(2,5,8)	Numeric

The computer chosen for the experiment has the following specification: HP Intel, Microsoft Windows 8.1 (64-bit), Intel® Core™ i5 CPU 4200Y @ 1.40 GHz, 4.00GB RAM.

The total encoding CPU time for each video sequence was recorded using Intel VTune Amplifier XE [18]. The following are the parameters used, with each video containing 3*2*2*3 = 36 Total Number of Instances. Table 3 presents 12 Instances out of the 36 Instances.

In each experiment one parameter will be changed while the rest of the parameters are fixed. This will help to observe the effect all parameters have on each Objective as shown in Table 3. The selected values of each coding parameter, the distortion (as PSNR is measured in decibels (dB)), bit-rate in kbps and the encoding time in seconds were recorded. Subsequently, the results were fed into the Linear Regression Analysis tool of WEKA [19] to generate the linear regression function for each objective.

Table 3: selected set of parameters for foreman sequence.

x_1	x_2	x_3	x_4	Bitrate	PSNR	Encoding Total Time
0	16	17	2	44.606	547.62	61.627
0	16	17	5	44.635	473.74	75.058
0	16	17	8	44.636	471.7	84.624
0	32	49	2	22.871	9.98	41.24
0	32	49	5	23.449	9.21	57.3
0	32	49	8	23.405	9.38	71.303
5	16	17	2	45.751	819.97	55.727
5	16	17	5	45.819	712.66	69.253
5	16	17	8	45.822	714.2	75.729
8	16	17	2	44.818	618.74	57.825
8	16	17	5	44.906	544.33	69.269
8	16	17	8	44.899	540.87	77.812

Based on the output of the linear regression, only the functions of three objectives of the Foreman video sequences are given in equation (1), (2) and (3). Following are the obtained models for each video sequence, with *f(1)* representing PSNR, *f(2)* rate and *f(3)* encoding-time. x_5, representing PSNR as parameter and x_6, representing Bit-rate.

Foreman Linear Regression Model

$$f(1)_{PSNR} = -0.5926 * x(3) + 0.0798 * x(4) + 0.0046\ x(6) + 52.0225 \quad (1)$$

$$f(2)_{Bitrate} = 6.8133 * x(1) - 18.5023 * x(3) + 890.0319 \quad (2)$$

$$f(3)_{Enc\_Time} = -0.5347 * x(1) + 0.1297 * x(2) - 0.5646 * x(3) + 4.0201 * x(4) + 60.362 \quad (3)$$

4.2. H.264 Encoder Analysis

Experimental analysis was conducted separately for the encoder and decoder. Table 4 tabulates the correlation coefficients of the objective functions. They range between 0-1. In analysing the objective functions above, higher positive coefficients of coding parameters indicate higher positive dependency and higher negative coefficients represent higher negative dependency. If a certain parameter is not present in the objective function that means that the objective is independent of that parameter.

Table 4: Encoder Correlation Coefficient.

Video	PSNR	Bit-rate	CPU
Claire	1	0.9424	0.9460
Coastguard	0.9865	0.9865	0.8917
Football	0.9997	0.9967	0.9849
Foreman	0.9998	0.9678	0.9746
Mobile	0.9999	0.9809	0.9588
Tennis	0.9998	0.9757	0.9883

A careful analysis of the coding parameters that have non-zero weighting factors in the objective functions obtained and a comparison of relative magnitudes of the coefficients can lead to a direct correspondence with the properties of the video. For example, the analysis of the linear regression equations obtained for the Foreman video sequence identifies all four parameters to have significant impact on CPU utilisation (Encoding time) as in equation (3), namely:

- IntraPeriod
- Searchrange
- Quantization parameter
- NumberReferenceFrames

The most significant impact on CPU utilisation is the number of reference frames. This is expected due to the need to repeat the motion estimation process when NRF increases. The next significant impact is from the Quantization parameter. The impact from Search Range (SR) and Intra Period (IP) is relatively insignificant. For most videos with fast movement of objects (i.e. Football and Mobile) there is no impact from the Search Range. This is true given the fact that for videos with fast moving objects, best matches will not be found quickly, i.e. without having to scan the entire video.

For the same video, the following parameters were identified to have a significant impact on Bit-rate as in equation (2).

- IntraPeriod
- Quantization

In equation (1) the parameters that are identified to have a significant impact on PSNR are:

- Quantization parameter
- NumberReferenceFrames
- Bit-rate

The parameter that has the most significant impact on PSNR is bit-rate. It is noted that these two parameters are highly dependent

4.3. H.264 Decoder Analysis

An H.264 decoder takes a .264 file as input and outputs a raw YUV video stream. **Error! Reference source not found.** shows the output video artifact of frame 30 with quantization parameter (QP) of 49 that gives very low quality with PSNR of 24.189 db and 5.83 Bitrate compared to QP 17 that has 43.418 db and 979.02 Bitrate.

(a) (b)

Figure 2: Sample image of frame at (a) QP= 17; (b) at QP= 49

The more QP is increased during the encoding of the video, the more the video lost information and the bitrate reduced.

Note that the Decoder parameters have no impact on Bit-rate and PSNR as these are determined by the encoder. In the proposed framework, the quality and the bit-rate received by the decoder are the same as the encoder output.

The computational complexity of the decoder is analysed using the same method used at the encoder end. For the six given video sequences, experiments were performed in order to find out those coding parameters that can significantly influence CPU utilisation. The objective functions thus obtained are listed within the equations below.

Claire Linear Regression Model

$$f(1)_{Dec\_Time} = 0.004 * x(1) + 0.0008 * x(2) - 0.0039 * x(3) + 0.373$$

Coastguard Linear Regression Model

$$f(1)_{Dec\_Time} = -0.0042 * x(2) - 0.0118 * x(3) + 0.8945$$

Football Linear Regression Model

$$f(1)_{Dec\_Time} = 0.0016 * x(1) - 0.022 * x(3) - 0.0033 * x(4) + 1.3455$$

Foreman Linear Regression Model

$$f(1)_{Dec\_Time} = -0.0146 * x(3) + 0.9353$$

Mobile Linear Regression Model

$$f(1)_{Dec\_Time} = 0.0071 * x(1) - 0.022 * x(3) - 0.0059 * x(4) + 1.3396$$

Tennis Linear Regression Model

$$f(1)_{Dec\_Time} = -0.0158 * x(3) - 0.0072 * x(4) + 1.0304 \quad (4)$$

Table 5 tabulates the correlation coefficients of the decoder objective functions. The Football video sequence has the highest correlation coefficient closely followed by mobile. From the analysis of the linear regression equations obtained to identify parameters that have significant impact on CPU utilisation, Equation 4 reveals that the quantization parameter has the most significant impact. QP has an impact in all the video sequences as evidenced by its presence in all objective functions and being the parameter having the highest magnitude coefficient.

Table 5: Decoder correlation coefficient

Video Sequences	Decoder Time
Claire	0.9593
Coastguard	0.9217
Football	0.9984
Foreman	0.9786
Mobile	0.9958
Tennis	0.9873

The Encoder and Decoder analysis indicates that the objective functions obtained as a result of using the proposed framework is able to accurately define the significant coding parameters and further detail the level of significance of each parameter. They can also be related to the motion and content information of the videos.

4.4. Multiobjective Optimisation of H.264.

This section presents a framework for multi-objective optimisation of video CODECs. Specifically, an optimization scheme is proposed to determine the optimum coding parameters for a H.264 AVC video codec in a bandwidth constrained environment, which minimises codec complexity and video distortion. Solutions to the optimization problem are reached through a Non-dominated Sorting Genetic Algorithm (NSGA-II). NSGA-II is implemented in the genetic algorithm gamultobj, available in the MATLAB optimisation tool-box and the settings are fixed as shown in Figure 3.

4.4.1. Optimising the Encoder of h.264

The objective functions given in section 4.1 were used to optimise the encoder. These functions are then provided to the NSGA-II optimization tool along with the fitness function and number of variables. The fitness function corresponding to the objective function of the encoded videos is computed. Then populations are generated by applying the crossover and mutation operators described with the settings shown in Figure 3.

The NSGA-II provides all sets of optimal results that jointly minimize complexity, bit-rate and maximize quality. Since a single 3D graph is complex to visualize the optimality of the results, pairs of graphs were plotted.

Since optimization implemented in MATLAB minimizes the objective or fitness function, it solves problems of the form

$$\min_{x} f(x).$$

If you want to maximize *f(x)*, *–f(x) should be minimised,* because the point at which the minimum of *–f(x)* occurs is the same as the point at which the maximum of *f(x)* occurs.

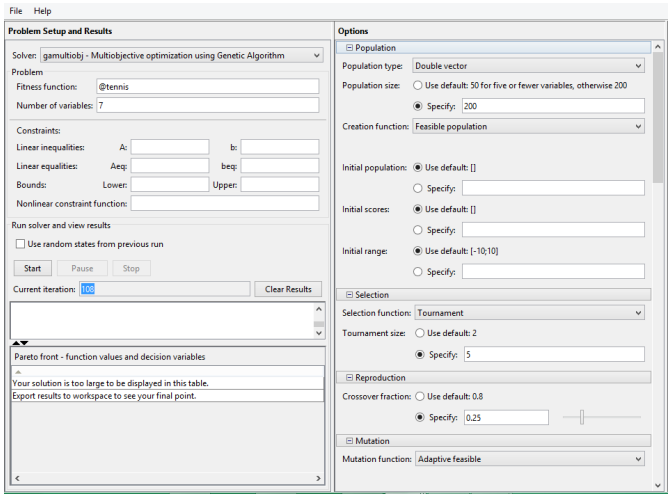


Figure 3: Set options for the problem

To obtain a Pareto front for two objective functions, the optimization is implemented using the above equations in section 4.1. A Pareto plots will appear as shown in Figure 4 – 6.

Figure 4 shows the Pareto front or set of non-dominated solutions for Bit-Rate and PSNR at the final stage of generations being used in the optimization process. One example optimised point is defined as:

IntraPeriod is -14.1153, SearchRange is 4.953125, QP is -9.39187, NRFrames is 21.66449, PSNR is 8.107256 and Bit-rate is 8.700779. Whereas the optimal values for PSNR is -59.357, Bit-Rate is 967.6315 as shown in Figure 4. Similarly, the results showing the Pareto front of non-dominated solutions for PSNR vs. CPU in Figure 5 and Bit-rate Vs. CPU time are presented in Figure 6.

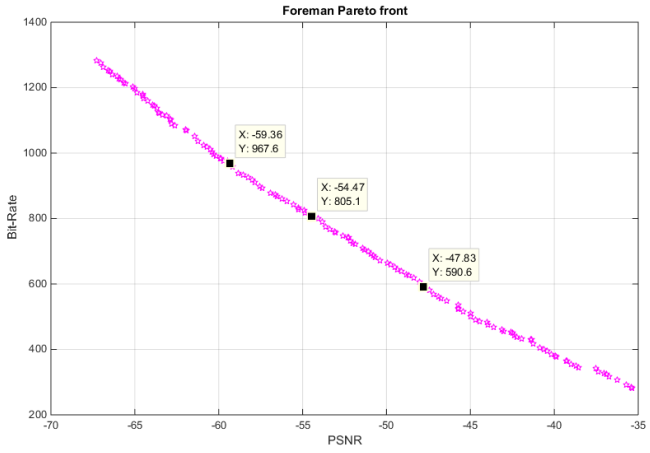


Figure 4: Pareto points for foreman PSNR vs. Bit-Rate.

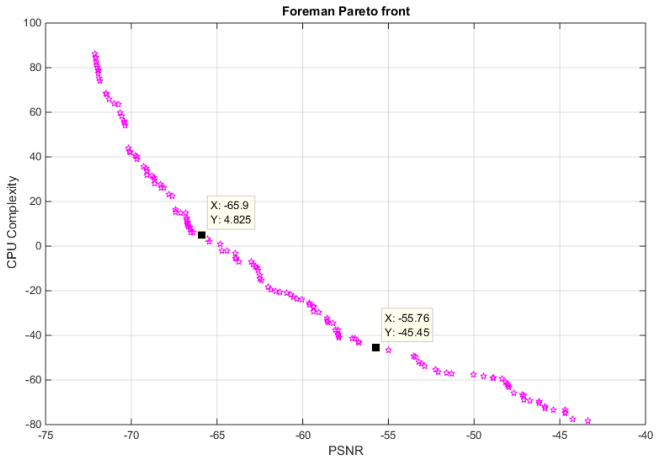


Figure 5: Pareto points for foreman PSNR vs. CPU.

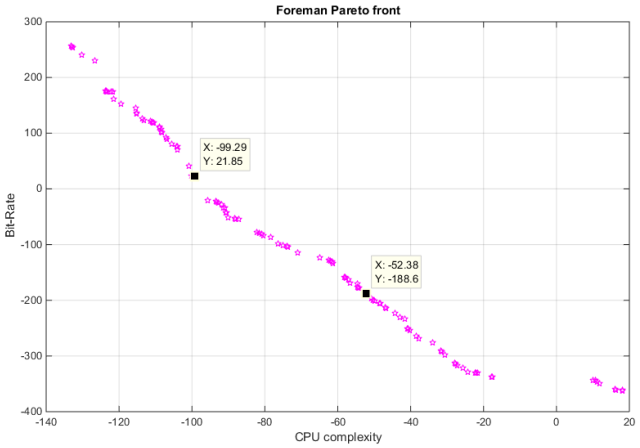


Figure 6: Pareto points for foreman CPU vs. Bit-Rate.

It is noted that the optimisation procedure described above results in a number of optimal solutions.

Table 6 Output data describing the results of MOO with GA for Figure 4 Figure 6.

Problem	Number of generations	Population Size	Pareto fraction	Size of non-dominated set	Function count	Average distance	Spread
PSNR vs. Bit-rate	107	200	.7	140	21601	0.0043	0.3793
PSNR vs. CPU	127	200	.7	140	25601	0.0196	0.3164
CPU vs. Bit-rate	246	200	.7	137	49401	0.0442	0.5231

According to Table 6, population size and Pareto fraction for the GA are set at 200 and 0.7, respectively, which are considered sufficient to generate a search for optimal solutions. The solver will try to limit the number of individuals in the current population that are on the Pareto front to 70 percent of the population size since the Pareto fraction is set to 0.7.

As the MOO algorithm is implemented in MATLAB, and stops at 107 generations and 21601 function counts, the GA selected 140 of the best individuals that are considered as non-dominated solutions out of 200 individuals in the population. Average distance between individuals is 0.0043, which indicates good convergence of the MOO solution, since it has a distance of less than 0.05 from the nearest point in the Pareto set.

4.4.2. Optimising the Decoder.

The analysis of the decoder is limited to decoder parameters that have a significant effect on only the decoder's computational complexity. In the proposed framework, the quality and the bit-rate received by the decoder are the same as the encoder output. This means the decoder receives all data transmitted by the encoder at the same rate. In such cases, the decoder totally depends on encoder coding parameters.

Figure 7 illustrate graphs between Bit-Rate vs. CPU complexity and PSNR vs. CPU complexity on the way to final generation.

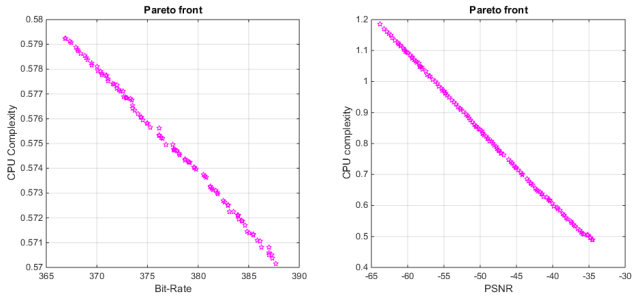


Figure 7: CPU a) Bit-Rate vs. CPU complexity and b) PSNR vs. CPU complexity

5. A Machine Learning based Framework for Parameter based Multi-Objective Optimisation of a H.265 Video CODEC.

5.1. H.265 Profiling Experiments.

This experiment was carried out using the Random Access (RA) configuration file of the Reference software for ITU-T H.265 high efficiency video coding named the HEVC test model (HM) version 16.8. Different resolutions can be used in each profiling experiment: 1080p which is representative for (Full HD) high definition systems with resolution of 1920x1080 pixels in a 16:9 aspect ratio, 2K Video a display resolution of 2560x1600 pixels with a 16:10 aspect ratio and 2160p (Ultra HD) which is representative for the next generation of high quality video. Each video sequence was encoded using a selected combination of possible parameter values of the initial set of encoder parameters.

In other words, each encoding instance corresponds to a combination of coding parameter values, selected from the possible exhaustive set that can be determined by varying each parameter within its entire range. For example, instead of using quantization parameter variations between 1-51 (that is the exhaustive set), only three sample values, 27, 37 and 45, were used (for further examples see Table 7). The table also tabulates the sample values used in our experiments for each parameter from within their corresponding value ranges.

Table 7: Settings for the Encoder in HM.

Parameter	Meaning	Values Range
SourceWidth SourceHeight	Specifies the width and height of the input video.	1920x1080 2560x1600
FrameRate	Specifies the frame rate of the input video.	Depends on video
Internal Bit Depth	Specifies the bit depth used for coding. When 0, the setting defaults to the value of the MSBExtendedBitDepth.	8
Coding Unit Size/Depth	Maximum coding unit width in pixel Maximum coding unit height in pixel	64/4 64/4
IntraPeriod	Period of I-frames. Specifies the intra frame period. A value of -1 implies an infinite period.	(16,32,48)
GOPSize	Specifies the size of the cyclic GOP structure.	8
FastSearch	The use of a fast motion search.	1:TZ search
SearchRange	Sets allowable search range for motion estimation.	(64,128)
Fast Encoding	Fast encoder decision	(0 or 1)
Quantization Parameter	Specifies the base value of the quantization parameter. If it is non-integer, the QP is switched once during encoding.	(27,37,45)
Asymmetric Motion Partitioning (AMP)	Enables or disables the use of asymmetric motion partitions.	1
Sample adaptive offset (SAO)	Enables or disables the sample adaptive offset (SAO) filter.	1
Rate Control	Rate control: enables rate control or not.	0

Table 8 shows selected sample frames of a set of six video sequences with different resolutions. Note that typical resolutions used in conjunction with H.265 video coding standard, i.e., 1080p and 2K resolution videos are used in all experiments, to carry out the analysis and make the relevant conclusions of this research. However, without any restrictions the proposed framework can be used in relation to a video sequence of any resolution, in particular

HD and full-HD 2K, 4k and beyond. The six selected video sequences have different properties of object motion, both in the foreground and background. Further differences exist in the scene content.

The experiments were initially conducted on a HP computer, running Microsoft Windows 8.1 (64-bit), having an Intel Core i5 CPU 4200Y @ 1.40 GHz and 4.00GB RAM. However it was found that coding HD resolution video is an intensive task that required for example, if encoded in the computer with the above specification, 10 hours to encode 50 frames of a 1920x1080 video at QP 37, intra period 48 and search range 64. Consequently, a decision was made to make use of a High Performance Computing (HPC) facility. Thus for all the experiments a HPC system using Redhat Enterprise Linux v6, with 20 cores of Intel Ivy Bridge Xeon E5-2670 containing 64GB RAM was used, significantly reducing the execution time per experiment.

Table 8: HEVC Tested Video Sequences.

YachtRide_1920x1080	**Traffic_2560x1600**
BasketballDrive_1920x1080	**Jockey_3840x2160**
Cactus_1920x1080	**YachtRide_3840x2160**

A sample of 36 data instances of the Cactus video sequence is presented in Table 9. These were used in the final stage of modelling the PSNR, Bit-rate and CPU utilisation. These are the inputs to the [19] linear regression based modelling process that result in the three objective functions that include the significant parameters, Intra Period as x_1, Search Range x_2, Quantization Parameter x_3 and Fast Encoding x_4.

The resulting objective functions for Bit-rate, PSNR and CPU time are the final outcomes of the performance modelling of the CODEC. Separate experiments are performed for each of the sample test videos.

Based on the output of the linear regression algorithms applied as explained above, the objective functions for the three objectives (for the Cactus video) are found as presented in equation (5). These functions provide the means to discuss in detail the significance of each parameter and how they affect the PSNR, Bit-rate and CPU encoding time. The following section provides an analysis of the experimental results. In particular the analysis considers the test videos separately and discusses the impact of each coding parameter given the known properties of the contents of each video. [Note that for each video, a different model is generated based on the video's inherent properties.]

$$f(1)_{Bitrate} = -22.4664 * x(1) - 386.2482 * x(3) + 18066.616$$

$$f(2)_{PSNR} = -0.0039 * x(1) - 0.4404 * x(3) + 48.873$$

$$f(3)_{Enc\_Time} = 3.9537 * x(1) + 2.5501 * x(2) - 36.2174 * x(3) + 545.1239 * x(4) + 2768.025 \quad (5)$$

Table 9: Selected Set of Parameters for Cactus video Sequence.

x_1	x_2	x_3	x_4	Bitrate	PSNR	Encoding Total Time
16	64	27	1	8422.096	36.8076	2000.97
16	64	37	1	2195.592	32.7418	1612.08
16	64	45	1	736.12	28.9058	1474.57
16	128	27	1	8424.696	36.8054	2140.34
16	128	37	1	2197.152	32.7447	1722.19
16	128	45	1	735.304	28.9092	1556.66
16	64	27	0	8414.944	36.8149	2559.02
16	64	37	0	2196.192	32.7507	2106.38
16	64	45	0	735.864	28.9111	1925.5
16	128	27	0	8414.864	36.8149	2778.27
16	128	37	0	2195.184	32.7516	2287.14
16	128	45	0	736.728	28.9173	2067.34
32	64	27	1	6993.976	36.7337	2115.22
32	64	37	1	1726.648	32.6277	1698.1
32	64	45	1	561.512	28.7824	1553.63
32	128	27	1	6991.08	36.7323	2271.37
32	128	37	1	1725.24	32.6279	1819.83
32	128	45	1	561.896	28.7877	1634.66
32	64	27	0	6994.6	36.7419	2684.76
32	64	37	0	1725.12	32.6325	2198.14
32	64	45	0	563.04	28.7949	2002.86
32	128	27	0	6991.312	36.7421	2923.43
32	128	37	0	1725.536	32.6336	2411.49
32	128	45	0	561.952	28.7952	2164.01
48	64	27	1	6914.36	36.7438	2126.42
48	64	37	1	1722.568	32.6039	1719.68
48	64	45	1	560.656	28.7335	1556.47
48	128	27	1	6911.576	36.7428	2295.48
48	128	37	1	1720.96	32.5996	1867.74
48	128	45	1	559.032	28.7422	1656.8
48	64	27	0	6911.616	36.7488	2690.7
48	64	37	0	1720.944	32.6059	2216.83
48	64	45	0	562.424	28.742	2018.13
48	128	27	0	6912.472	36.7515	2962.86
48	128	37	0	1720.176	32.6066	2437.26
48	128	45	0	560.872	28.7534	2200.32

5.2. H.265 Encoder Analysis.

The Encoder objective functions obtained as a result of the experimental procedure presented in section 5.1 enables one to discuss the significance of each of the coding parameters. Following are the obtained models for each video sequence, with $f(1)$ representing PSNR, $f(2)$ rate and $f(3)$ CPU encoding time.

Otabulates the correlation coefficients of the objective functions. They range between 0-1. A value closer to 1 represents the fact that the dependant variable (in this case Bit-Rate, PSNR or CPU utilisation) can be predicted very accurately from the coding parameters that play a role and have been included within the objective functions.

Table 10: Encoder Correlation Coefficient.

Video	PSNR	Bit-rate	Enc_Time
Cactus	0.9989	0.9551	0.988
YachtRide	0.9981	0.9532	0.9837

In analysing the objective functions (5), higher positive coefficients of coding parameters indicate higher positive dependency and higher negative coefficients represent higher negative dependency. If a certain parameter is not present in the objective function, that means that the objective is independent of that parameter. A careful analysis of the coding parameters that have non-zero weighting factors in the objective functions obtained and a comparison of relative magnitudes of the coefficients can lead to a direct correspondence with the properties of the video; for example, the presence of motion in the foreground and background, the speed of movement of objects, sudden scene changes, camera pan/tilt/zoom effects and the general characteristics of the content of the video as well.

For example, the analysis of the linear regression equations obtained for the cactus video sequence identifies all four parameters to have significant impact on CPU utilisation, namely:

- IntraPeriod
- Searchrange
- Quantization parameter
- Fast Encoding

For the same video, the following parameters were identified to have a significant impact on Bit-rate.

- IntraPeriod
- Quantization parameter

The parameters that are identified to have a significant impact on PSNR are:

- IntraPeriod
- Quantization parameter

A more detailed and video sequence specific analysis can be presented as follows.

- ***Analysis of the CPU Utilisation Experiment:***

The objective functions obtained for all tested video sequences for CPU encoding time indicates that the parameter that has the most significant impact on CPU is Fast encoder decision. Further, in the selection of the Intra-Period, more I frames (smaller intra period) results in a higher processing time. The next significant impact is from the Quantization parameter. The impact from search range (SR) and Intra Period (IP) is relatively insignificant.

When search range increases, encoding time will slightly increase. These tests have no major impact on the quality of the video. Disabling FEN will also slightly increase encoding time. However it has no major impact on quality.

- ***Analysis of the PSNR Experiment:***

The parameter that has the most significant impact on PSNR is QP. The PSNR results tabulated in 0 indicate that the two videos with the least amount of movement/changes, namely Cactus and YachtRide have the best correlation coefficients. This is expected due to the stability of the CODEC during the encoding of the individual frames of the coded sequence.

- ***Analysis of the Bit-Rate Experiment:***

The parameter with the most significant impact is the QP. Lower quantisers result in higher bitrate and correspondingly higher visual quality as illustrated in Figure 8. (QP) has a very important impact on the compression rate of H.265.

In cactus, both PSNR and Bit-Rate have no impact from the Search Range. This is true given the fact that for videos with fast moving objects, best matches will not be found quickly, i.e. without having to scan the entire video. All objective functions include a similar constant term indicating that a fixed computational cost for encoding is present, which is independent of the selection of coding parameters. This is expected given the processes that exist, which are independent of the coding parameters.

5.3. H.265 Decoder Analysis

The computational complexity of the decoder is analysed using the same method used at the encoder end. Experiments were performed in order to find out those coding parameters that can significantly influence CPU utilisation. The objective functions thus obtained are listed within equation (6).

$$f_{Dec\_Time} = 0.0023 * x(2) - 0.0819 * x(3) + 0.1178 * x(4) + 8.2928 \quad (6)$$

Table 11 Decoder Correlation Coefficient

Video	Dec_Time
Cactus	0.9509
YachtRide	0.9263

Table 11 presents the correlation coefficients of the objective functions. The cactus video sequence has the highest correlation coefficient. The analysis of the linear regression equations is carried out to identify parameters that have significant impact on CPU utilisation. Equation (6) reveals that the Fast Encoding has the most significant impact, being the highest magnitude coefficient.

The Encoder and Decoder analyses indicate that the objective functions obtained as a result of using the proposed framework are able to accurately define the significant coding parameters and further detail the level of significance of each parameter. They can also be related to the motion and content information of the videos. More importantly, these objective functions model the behaviour/properties of the encoder and decoder thus allowing them to be used in multi-objective optimisation as described in the next section.

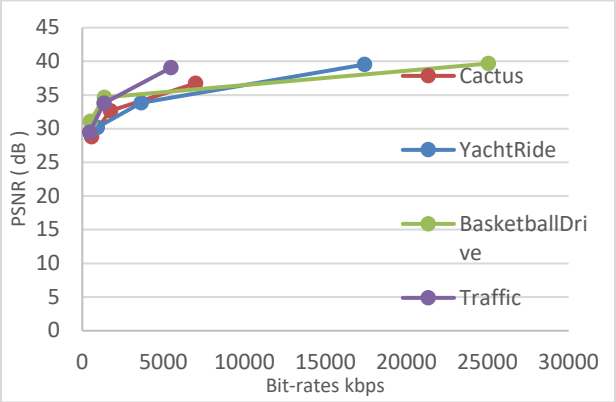


Figure 8: PSNR versus Bit-rate at QP 27, 37,45.

The HM encoder/ decoder configuration file takes a .265 file as input and outputs a raw YUV video stream as ReconFile. The output shown in Figure 9 indicates that the output video artifact of yacht video at frame 30 with quantization parameter (QP) of 45 gives very low quality with PSNR of 28.7877 dB compared to QP 27, which has 36.7323.

The more the QP is increased during the encoding of the video, the more the video loses information and bitrate reduces.

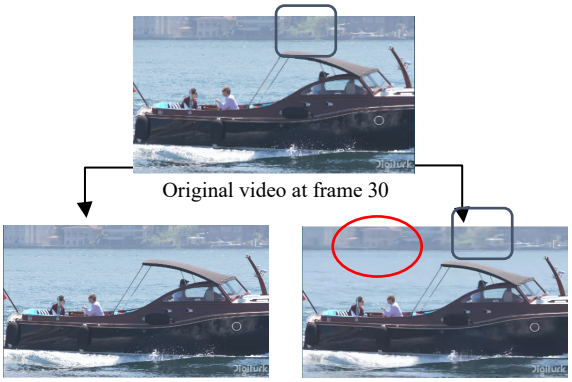


Figure 9: Shows the visual artifact with different QP

5.4. Optimising the Encoder of H.265

In this section, the objective functions given in equation (5) were used to optimise the encoder performance under multiple constraints. The experiment was conducted, and the following setting of parameters for the GA were chosen (see Table 12)

Table 12: Settings for Multiobjective Genetic Algorithm Solver.

gamultobj settings	
Fitness function:	@function
Number of variables:	4
Bounds Constraints:	lb = [16,64,27,0] ub = [48,128,45,1]
Creation Function:	Constraint dependent
Population Size:	60
Initial Population:	Default
Crossover Fraction:	0.8
Mutation Function:	Constraint dependent
Crossover Function:	Intermediate
Crossover Ratio:	0.8
Pareto Front Population Fraction:	0.35
Maximum Generations:	300
Plot functions:	Pareto front

The results of Multi-Objective Optimisation (MOO) Pareto set analysis are presented in Figure 10 and Figure 11. Figure 10 shows the Pareto front or set of non-dominated solutions for Bit-Rate and PSNR. The Pareto front illustrated is within a limited range and hence shows that the points lie on a straight line. In practice, when the range of testing is increased the shape of the curve would represent a typical shape of a Pareto curve. The Pareto curve allows one to select optimum performance points and hence select the corresponding coding parameters that resulted in the objective function optimal values for coding the videos.

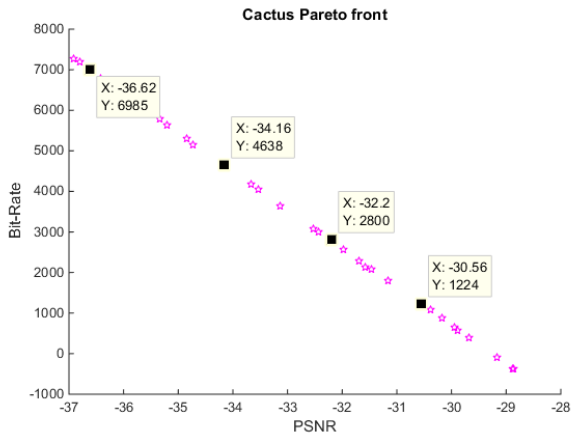


Figure 10: Pareto front for cactus video sequences PSNR vs. Bit-Rate.

For one optinal point of operations, the IntraPeriod is 27.49851, SearchRange is 83.34469, QP is 33.16743, Fast Encoding is 0.264558, PSNR. Whereas the optimal value for PSNR (X) is -34.1591, Bit-Rate (Y) is 4637.964 as shown in Figure 10. Similarly, the results showing the Pareto front of non-dominated solutions for Bit-rate Vs. CPU complexity are presented in Figure 11.

Table 13: Output data describing the results of MOO with GA for cactus.

Problem	Number of Generations	Size of population	Pareto fraction	Size of non-dominated set	Function count	Average distance	Spread
PSNRvs.Bit-rate	259	60	0.35	35	13001	0.0149	0.1487
CPU vs. Bit-rate	220	60	0.35	21	13261	0.0308	0.1050

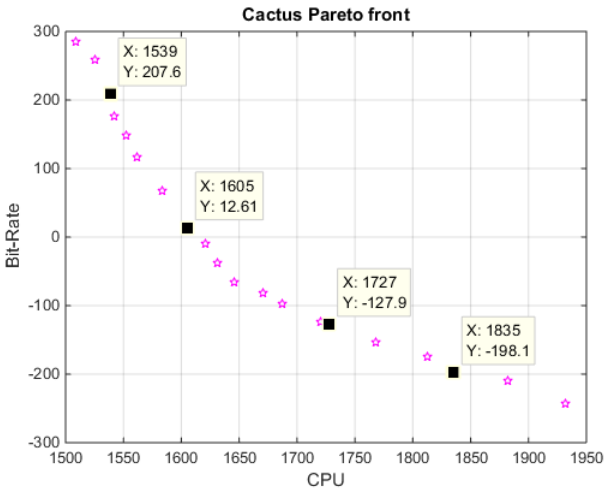


Figure 11: Pareto front for cactus video sequences CPU vs. Bit-Rate final Generation.

In Table 13, population size and Pareto fraction for the GA are set at 60 and 0.35, respectively, which are considered sufficient to generate search for optimal solutions.

At 259 generations and 13001 function counts, the GA selected 35 best individuals considered as non-dominated solutions out of 60 individuals in the population. Average distance between individuals is 0.0149, which indicates good convergence of the MOO solution.

Conclusion

In this paper we have proposed a machine learning based approach for the determination of significant coding parameters of a H265 video CODEC. In particular, we have used multivariate regression analysis in defining objective functions for CPU utilisation, PSNR and the bit-rate of a video CODEC when a given video is being Encoded/Decoded. We have been able to use known information about the content and the motion present in the test videos to justify the formation of the objective functions. We have shown that these regression equations provide the means for modelling the performance of a typical H.265 video CODEC. Finally, we have used these models to optimise the performance of a video CODEC under multiple constraints. For this purpose we demonstrated the effective use of a Genetic Algorithm based approach.

The proposed framework for the performance analysis, modelling and multi-objective optimisation of a H.265 video CODEC can be applied to any video CODEC and provides a useful contribution to the video coding community who are often faced with the dilemma of selecting values for a large number of coding parameters with the intention of obtaining optimal performance of the CODEC under multiple performance constraints.

6. Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgment

This research is supported by the Ministry of Man Power (MOMP) Muscat, Oman.

References

[1] M. Al-Barwani and E. A. Edirisinghe, "A machine learning based framework for parameter based multi-objective optimisation of a H.265 video CODEC," *2016 Future Technologies Conference (FTC)*. pp. 553–559, 2016.

[2] H. Zhihai, L. Yongfang, C. Lulin, I. Ahmad, and W. Dapeng, "Power-rate-distortion analysis for wireless video communication under energy constraints," *Circuits Syst. Video Technol. IEEE Trans.*, vol. 15, no. 5, pp. 645–658, 2005.

[3] H. Zhihai, C. Wenye, and C. Xi, "Energy Minimization of Portable Video Communication Devices Based on Power-Rate-Distortion Optimization," *Circuits Syst. Video Technol. IEEE Trans.*, vol. 18, no. 5, pp. 596–608, 2008.

[4] R. Vanam, E. A. Riskin, and R. E. Ladner, "H.264/MPEG-4 AVC Encoder Parameter Selection Algorithms for Complexity Distortion Tradeoff," in *Data Compression Conference, 2009. DCC '09.*, 2009, pp. 372–381.

[5] P. Wei, L. Yan, and W. Feng, "Joint Power-Distortion Optimization on Devices with MPEG-4 AVC/H.264 Codec," in *Communications, 2006. ICC '06. IEEE International Conference on*, 2006, vol. 1, pp. 441–446.

[6] K. Jaemoon, K. Jungsoo, K. Giwon, and K. Chong-Min, "Power-rate-distortion modeling for energy minimization of portable video encoding devices," in *Circuits and Systems (MWSCAS), 2011 IEEE 54th International Midwest Symposium on*, 2011, pp. 1–4.

[7] C. Fu-Chuang and H. Yi-Pin, "Rate-distortion optimization of H.264/AVC rate control with novel distortion prediction equation," *Consum. Electron. IEEE Trans.*, vol. 57, no. 3, pp. 1264–1270, 2011.

[8] R. Vanam, E. A. Riskin, S. S. Hemami, and R. E. Ladner, "Distortion-Complexity Optimization of the H.264/MPEG-4 AVC Encoder using the GBFOS Algorithm," in *Data Compression Conference, 2007. DCC '07*, 2007, pp. 303–312.

[9] F. Al-Abri and D. Eran, "Multi-Objective Optimization Of Video Coding and Transcoding," Loughborough University, 2010.

[10] F. Al-Abri, X. Li, E. A. Edirisinghe, and C. Grecos, "A Novel Framework for Multi-objective Optimization of Video CODECs," in *CyberWorlds, 2009. CW '09. International Conference on*, 2009, pp. 195–202.

[11] G. J. Sullivan, J. Ohm, W. Han, and T. Wiegand, "Overview of the High Efficiency Video Coding," *Ieee Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1649–1668, 2012.

[12] S. Weiwei, F. Yibo, H. Leilei, L. Jiali, and Z. Xiaoyang, "A hardware-friendly method for rate-distortion optimization of HEVC intra coding," in *VLSI Design, Automation and Test (VLSI-DAT), 2014 International Symposium on*, 2014, pp. 1–4.

[13] D. Grois, D. Marpe, A. Mulayoff, B. Itzhaky, and O. Hadar, "Performance comparison of H.265/MPEG-HEVC, VP9, and H.264/MPEG-AVC encoders," *2013 Picture Coding Symposium (PCS)*. pp. 394–397, 2013.

[14] B. Bross *et al.*, "HEVC real-time decoding," in *SPIE Optical Engineering+ Applications*, 2013, p. 88561R–88561R.

[15] B. Bross *et al.*, "HEVC performance and complexity for 4K video," *Proc. 2013 IEEE 3rd Int. Conf. Consum. Electron. - Berlin, ICCE-Berlin 2013*, pp. 44–47, 2013.

[16] M. P. Sharabayko and N. G. Markov, "Iterative intra prediction search for H.265/HEVC," *2013 Int. Sib. Conf. Control Commun. SIBCON 2013 - Proc.*, 2013.

[17] F.-I. H. H. I. Dolby Laboratories Inc. Microsoft Corporation, "H.264/AVC Reference Software." Joint Video Team, Germany, 2009.

[18] Intel, "Intel® VTune™ Amplifier XE," 2013. .

[19] E. F. Mark Hall Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witte, "The WEKA Data Mining Software," 2009. [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka/.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 2, No. 3, 1527-1539 (2017)
www.astesj.com
Special Issue on Recent Advances in Engineering Systems

ASTES Journal
ISSN: 2415-6698

2-D and 3-D Visualization of Many-to-Many Relationships

SeungJin Lim*

Computer Science, Merrimack College, 01845, U.S.A.

ARTICLEINFO

Article history:
Received: 22 May, 2017
Accepted: 24 July, 2017
Online: 19 August, 2017

Keywords:
Visual data mining
Visualization
3-D primitive
Many-to-many relationship

ABSTRACT

With the unprecedented wave of Big Data, the importance of information visualization is catching greater momentum. Understanding the underlying relationships between constituent objects is becoming a common task in every branch of science, and visualization of such relationships is a critical part of data analysis. While the techniques for the visualization of binary relationships are widespread, visualization techniques for ternary or higher relationships are lacking. In this paper, we propose a 3-D visualization primitive which is suitable for such relationships. The design goals of the primitive are discussed, and the effectiveness of the proposed visual primitive with respect to information communication is demonstrated in a 3-D visualization environment.

1 Introduction

Understanding the relationships between the entities is an important matter in any type of data analysis project whether scientific or non-scientific. Visualization is generally viewed as complementary to traditional text-based analysis. Visualization also plays an indispensable role in data analysis. At times, visualization comes with an unexpected surprise such that it reveals auspicious patterns that may not be intuitive otherwise. In this paper, we address the design issues of 2-D and 3-D visualization techniques for many-to-many relationships. This paper is an extension of work originally presented in the proceedings of the 19th IEEE International Symposium of Multimedia (ISM 2016).

Augustus De Morgan insightfully stated, "When two objects, qualities, classes, or attributes, viewed together by the mind, are seen under some connexion, that connexion is called a relation" [1]. It is critical to understand the relationships between the objects in order to analyze the full drama played by the objects. For example, given the information that each of the two pairs of persons A and B, and A and C holds a friendly relationship in a social network, it may be important to know if the friendly relationship "clique" comprising A, B, and C together exists at a statistically significant level in psychology or cyber security. The authors of [2] identified two types of friendship from their study on Facebook. One type is the sympathy group which consist of approximately 15 close friends. The other type is the support clique of "friends on whom you would depend for emotional/social support in times of crisis." The size of the second group is approximately 5.

Relationships may be characterized by attributes such as type (e.g., "co-authorship" [3] and "trust" in social networks [4], "presidential election", etc.), weight (boolean or continuous), direction (directional or undirectional) and etc. The set of attributes for relationships is determined by the nature of the application domain. In our work, relationships are attributed only by the degree of participation of an object in its relationships to other objects, defined below, and the weight of the relationship.

Definition 1.1 (Degree of participation in relationship) *Given a set of objects $O = \{o_1, o_2, \ldots, o_n\}$, the relationship of an arbitrary type t between the members of O is many-to-many when each o_i $(1 \leq i \leq n)$ holds a relationship of the same type t with zero or more other objects in O simultaneously. Furthermore, we use the term k-relationship to denote a relationship whose degree of participation is k.*

Among the various degrees of relationships in terms of participation, binary relationship is the most common in and the primary target of many data analysis projects. Examples of binary relationship include the gravitation between two masses in physics, symbiotic organisms in biology, term co-occurrence in information retrieval, co-citation [5] in bibliographic coupling studies, spatial co-location [6] in spatial data

*SeungJin Lim, North Andover, MA, U.S.A., 978-837-3404, lims@merrimack.edu

https://dx.doi.org/10.25046/aj0203191

mining, frequent 2-itemsets in association mining, and friend relationship between two persons in social networks.

At the same time, we recognize the rise of the relationships of a higher degree where three or more objects participate in the relationship simultaneously. Co-authorship [3] is such an example, because a paper can be published by more than two authors. The "support clique" of size 5 mentioned above is an example of relationships of a higher degree. As another example, the 34.7% of the papers are coauthored by three or more authors out of the 5.5 million journal articles, conference papers, and other publications on computer science listed at the DBLP Computer Science Bibliography website as of February in 2017. And the 0.5% of them are published by 10 or more co-authors as shown in Figure 1.

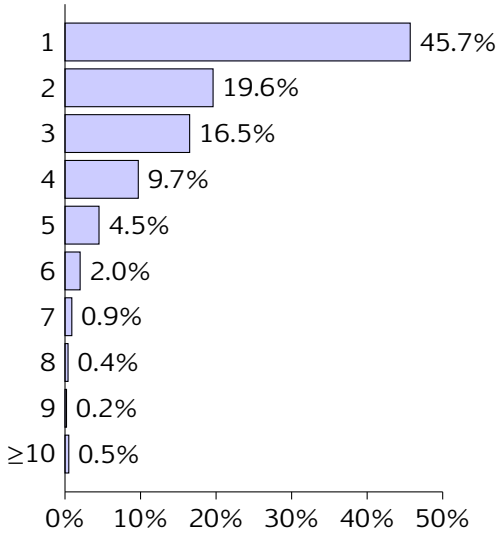


Figure 1: The fraction of publications listed in the DBLP Bibliography by number of coauthors.

Object relationships are commonly represented using a textual format in accordance with a data structure specifically designed for the data. A classic example is a tuple notation which is widely used in the relational database model. Relationships are also represented graphically. A graph $G = (V, E)$ where V is a set of vertices denoting objects and E is a set of edges representing relationships between the objects is a common example of this kind of representation, which has a dominant appearance in data/information visualization such as in social network analysis, web graph, and so forth.

When object relationships are analyzed over a large volume of data, overall descriptive patterns manifested in the underlying data is often the primary objective of the analysis. Individual or local patterns may or may not be of interest. Visualization helps us understand both overall and local patterns and offers a unique added value to data analysis of a large scale such as big data.

The visualization of binary relationships is conveniently and effectively done by drawing a line between the two related objects. In Figure 2, for example, the binary coauthorship relationship between Barkerma and Moursseau is apparent. We can also easily observe other binary relationships involving Barkerma, such as (Barkerma, Leeuw) and (Leeuw, Moursseau). A drawback of this visualization is that the readers are prone to mistakingly derive the ternary relationship (Barkerma, Leeuw, Moursseau) from (Barkerma, Leeuw), (Barkerma, Moursseau) and (Leeuw, Moursseau), unless noted otherwise.

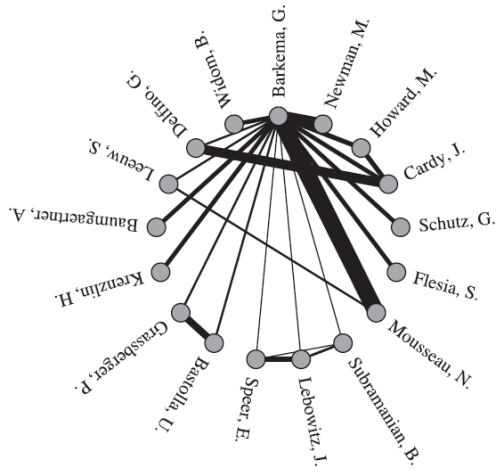


Figure 2: A visualization of G. Barkerma and collaborators in scientific publication [3].

The purpose of this paper is two-fold.

- We compare the merits of the selected 2-D visualization techniques and present the merits and challenges they impose for the purpose of data analysis when the degree of participation or the number of relationships grows large.
- We propose a geometric primitive which is suitable for visualizing ternary or higher-degree many-to-many relationships in 3-D. A set of design goals are also discussed. The effectiveness of the proposed primitive is demonstrated with various real-world datasets as well as artificial datasets.

The remainder of this paper is organized as follows: In Section 2, some selected related works are presented. We first discuss on the selected popular 2-D visualization layouts in Section 3. We then present the proposed visualization primitive along with design goals and technical details in Section 4. A demonstration of the primitive is presented in Section 5. We conclude our presentation with what we plan to do in the future in Section 6.

2 Related Works

The topic of our work is generally related to scientific visualization in which a global summary of data is often the primary interest. On the other hand, local

details are also important in visual data exploration and mining to which our work is closely related. A review on visual data exploration and mining can be found in [7]. Undirected relationships between entities are largely visualized as 2-D or 3-D matrices [8, 9] or undirected graphs [10, 3, 11, 12]. A number of different approaches exist. Determination of which approach is better than others is largely domain-specific as observed in [13]. We conveniently categorize most common approaches in the literature with respect to the layout of objects and visual cues representing relationships.

Difference in layout. The decision for the layout of objects is commonly influenced by the number of domains underlying the data objects or the number of attributes describing the data object. When the objects are from different domains, their membership to a different domain should be visually encoded so that the user may discern their membership easily. Intuitively, n orthogonal axes are used to make n domains distinguishable.

In case of bipartite relationships, for example, one approach is to arrange the objects in accordance to two axes that are visually perpendicular or parallel to each other. Bipartite graphs are examples of this kind. An alternative approach found in the literature is to arrange them anywhere in the visual space with two different object shapes to make them distinguishable. This approach can be thought of "zero" axis.

Though the objects belong to a single domain, we may still need multiple axes when these objects are associated with multiple attributes. A technique called parallel coordinates [15] is an example of this kind where objects are arranged vertically and each object is attributed by a number of parallel axes which are spaced out horizontally, each of which represents an attribute.

In our work, objects belong to a single domain. Two types of object layouts are commonly found in the literature and in practice.

1. Single set of objects: Only one set of objects is rendered in the visualization. The objects are arranged along the circumference of a circle (Figure 3a), on a single linear axis (Figure 3b), or anywhere in the given visual space (Figure 3c). Notice that the three visualizations in Figure 3 are *isomorphic*. In other words, they carry the exact same set of relationships between the same set of objects though they may produce different visual impression. Graph is the basis of these visualization techniques.

2. Two sets of objects: This layout can be thought of as an Euclidean space where one axis carries the full set of objects, and the second axis perpendicular to the first axis carries the same set of objects in the same ordering.

   In Figure 4a, a set of 19 performance statistics of baseball players is arranged along both the horizontal and the vertical axes to show the correlation (a particular type of relationship) between them for the 1986 season. The statistics include season and career batting averages and so forth, calculated from "Hits" and "At bats".

   This layout is extended often to capture the correlation between two variables, as illustrated in the H-R diagram of stars showing star categories with respect to temperature and luminosity (Figure 4b). The layout is also extended to a 3-D space to amplify the visual effect of the significance of relationships as illustrated in Figure 5b.

Difference in the representation of relationships.

1. Presence or absence of relationship: When the presence or absence of a relationship between two entities is what matters, connecting the two entities by an edge or a marker on a 2-D plot is sufficient. The three visualizations in Figure 3 and the one in Figure 4b are examples of this category. Unweighted graphs and scatter plot-like visualizations belong to this category.

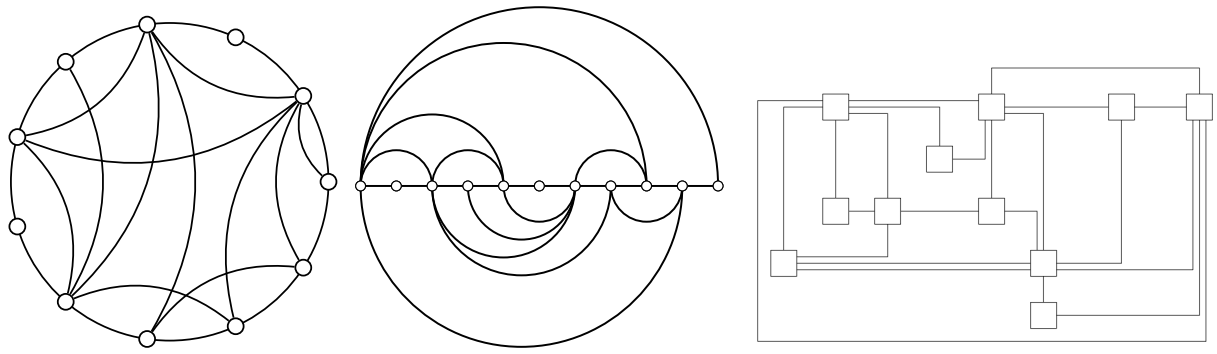
2. Significance of relationship: In contrast, Figures 2, 4a, and 5a show the significance of a relationship using another visual cue such as thickness of edges as in Figures 2 and 5a, color or density of color (Figure 4), or height of object in a 3-D space (Figure 5b).

As for the visualization techniques designed specifically for data mining, the authors of [9] chose to show cluster analysis results in a 3-D matrix such that the square grid of horizontal axes represent data objects and the vertical axis represents the similarity between the two corresponding objects. Visualization of geospatial point sets presented in [18] is inherently a similar 3-D matrix visualization since longitude and latitude values constitute X- and Y-axes. [10] uses multiple visual cues in a 2-D square matrix to enhance the visual feedback from a large dataset.

There also exist tools which employs a circle rather than a rectangle or square as the layout of the data objects. In [3], coauthorship is represented in a circular image such that the authors are arranged on a non-interactive 2-D canvas whose visual feedback may lose effectiveness as the data size grows large. As an extension of a circular grid, a concentric 2-D ring view of a coauthorship network is presented in [12] in which each ring represents the authors' contribution in a particular year, which is useful to show temporal trends in the underlying information.

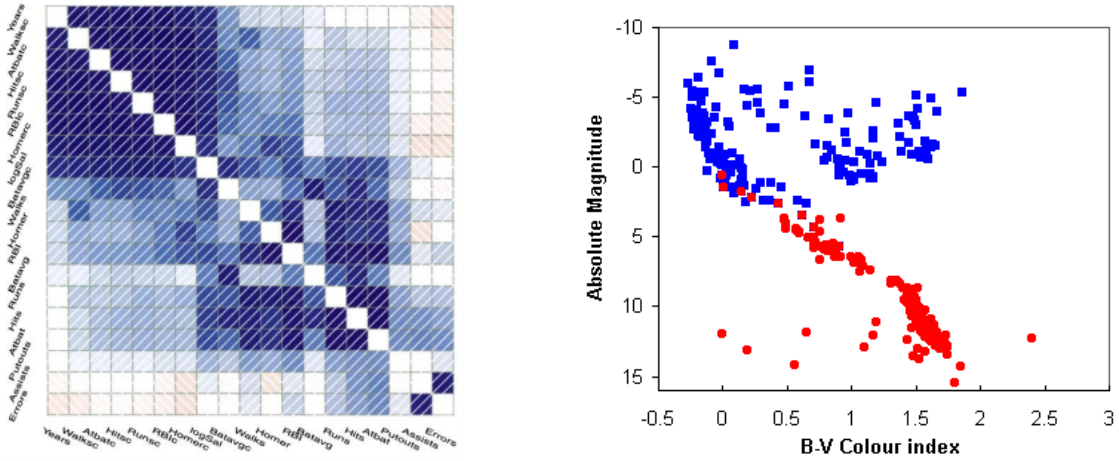
We notice that these visualization techniques discussed above are designed for binary relationships and we are not aware of a technique suitable for relationships whose degree of participation is higher than binary at the time of this writing.

Evaluation of the effectiveness of visualization techniques. Authors of [13] propose evaluation strategies for the effectiveness of visualization techniques. One approach is to solicit feedback from users in the form



(a) Circular layout (b) Kandinski layout (c) Podevsef layout

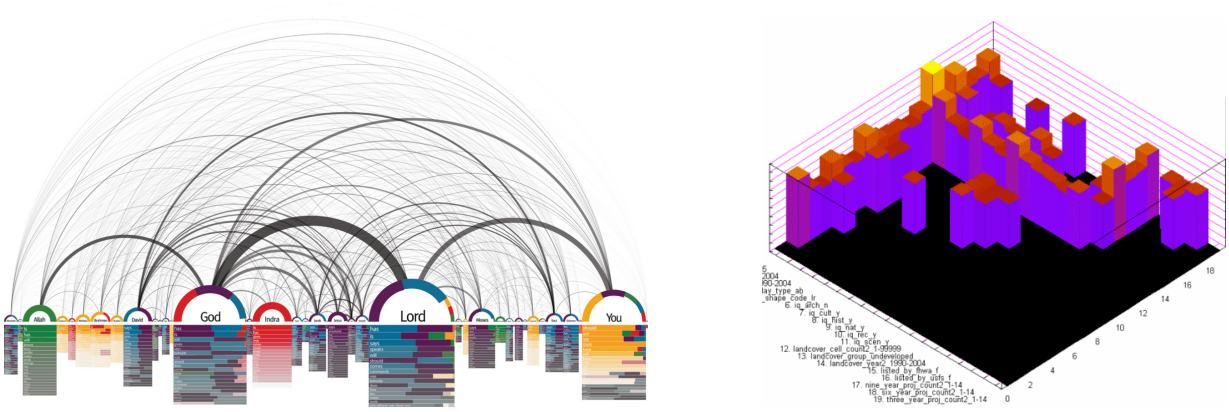
Figure 3: Various approaches to visualization of relationships using a single set of objects [14]. Objects are arranged along the circumference of the base circle (a) or along a single linear axis (b), or they can appear anywhere in the given visual area (c).



(a) 2-D visualization of the performance correlation between baseball players. The strength of a correlation is expressed by the intensity of color shade. [16]

(b) Hertzsprung–Russell diagram is a scatter plot of stars showing their relationships in terms of temperature and luminosity. (http://www.jb.man.ac.uk/distance/life/sample/stars/)

Figure 4: Matrix-like plots for visualization of relationships using two sets of entities.



(a) "A visualization of the similarities and differences between the holy books of five world religions: Christianity, Islam, Hinduism, Buddhism, and Judaism." [17]

(b) The significance of a relationship is represented by the height of the relationship as well as the color of the top surface as a 3-D bar graph.

Figure 5: Line thickness, color and object height are commonly used visual cues for the magnitude of interval and ratio values.

of qualitative interviews and quantitative usage statistics collected from the access log of the tool. Another approach is a more formal user study in which they found that visualization-based instructions help the

workers complete the assembly task 35% faster with fewer errors than traditional instructions.

3 2-D Visualization of Many-to-Many Relationships

In this section, we first compare the three 2-D strategies presented in [14]. The three approaches use a single set of objects. Our motivation is to see their suitability for the visualization of many-to-many relationships.

Data set. For an experiment of the visualization of relationships, we use a social network analysis data set that we generated from the book of Genesis in the Bible. The analysis was aimed at analyzing the interpersonal interactions in the book. There are 402 persons appearing in 1,064 interaction events in the book. The result of the analysis comprises 1,230 binary relationships and 225 maximal interactions. A maximal interaction in this study is defined as an interaction which is not contained in another interaction.

In this study, all the relationships belong to the same domain, namely *interpersonal interactions*. They have two attributes: degree of participation, ranging from 2 to 11, and significance of appearance, ranging between 0.0 and 1.0. The relationship in this data set is undirectional.

3.1 2-D Kandinski Layout

The objects are arranged on the single linear axis in the Kandinski layout illustrated in Figure 3b. The ordering of the objects is domain-specific. They can be arranged in the lexicographical order of the object labels. When the objects are associated with an ordinal quantity, such as rank or weight, they can be ordered according to the quantity. We can easily see that there are n-th factorial number of different arrangements of n objects and that a different ordering of objects will yield a different visual impression from the same set of relationships.

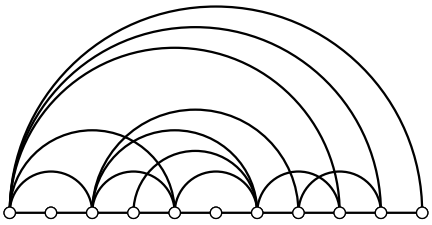


Figure 6: A variation of Kandinski shown in Fig. 3b where all the arcs are rendered only on the northern hemisphere of the axis.

The arrangement of arc between objects in the Kandinski figure is arguably arbitrary. Some arcs are arranged on the northern hemisphere while some others are on the southern hemisphere primarily for the purpose of the most comfortable visual effect. This approach may be prone to misinterpretation of the arcs. Viewers may mistakingly interpret the northern arcs may be superior to the ones on the south or vice versa.

Arcs can be rendered on a particular side of the axis for a specific purpose if relationships are associated with a bipolar attribute such as male or female and friendly or antagonistic.

As noted in Section 2, the three common layouts for a single set of objects shown in Figure 3 are *isomorphic*. In fact, the variation of Kandinski shown in Figure 6 is also isomorphic to the layout shown in Figure 3b. The benefit of this variation is to avoid the probability of misinterpretation of the arcs by the fact that the objects appear on a particular side of the axis.

Based on this observation, we generated Figure 7 using the Kandinski layout for the 1,230 binary relationships from the 1,064 interaction events in the book of Genesis. The objects (persons) are arranged in the order of name. The statistical significance of an interaction is encoded using the red-orange-yellow-green-blue-gray color scheme with red being the most significant level. With this configuration, it is relatively clear that Abraham has higher interaction with God, Isaac, Sarah and Lot than other associates. God has distinctively higher interaction with Abraham than with anyone else, followed by Joseph. Most other relationships rendered in gray are equally insignificant in terms of occurrence frequency.

One drawback of Kandinski is that viewers may interpret incorrectly that a relationship of a larger arc may be more significant than the arcs of smaller radius. However, the significance of a relationship cannot be expressed by means of the radius or height of the arc. Hence, if the relationship is not boolean, a visual cue, such as thickness or color, must be carefully designed to represent different significance. Note that Figure 5a uses thickness for the significance of an arc. We opt out this option for color because thick arcs tend to mask thin arcs especially as the number of arcs grows large.

The figure is generated using the TikZ library from a script produced by a tool developed in-house.

3.2 2-D Podevsef Layout

We experiment the same interaction relationships using the Podevsef layout. In this layout, objects are laid out freely anywhere in the visual space. However, the common objective is to spread the objects in a "desirable" order to minimize the link distance and node collision. One practical implementation of this layout is found in the Sankey plugin (`https://bost.ocks.org/mike/sankey/`) of the d3.js library.

Figure 8 shows the interaction relationships as a Sankey diagram. The algorithm built in the library has chosen to place 5 objects as middle layer nodes between the node on the left and the nodes on the right. We can see the interactions between a pair of persons by the link between them and the significance of the interaction by the width of the 'cable ribbon'-shaped

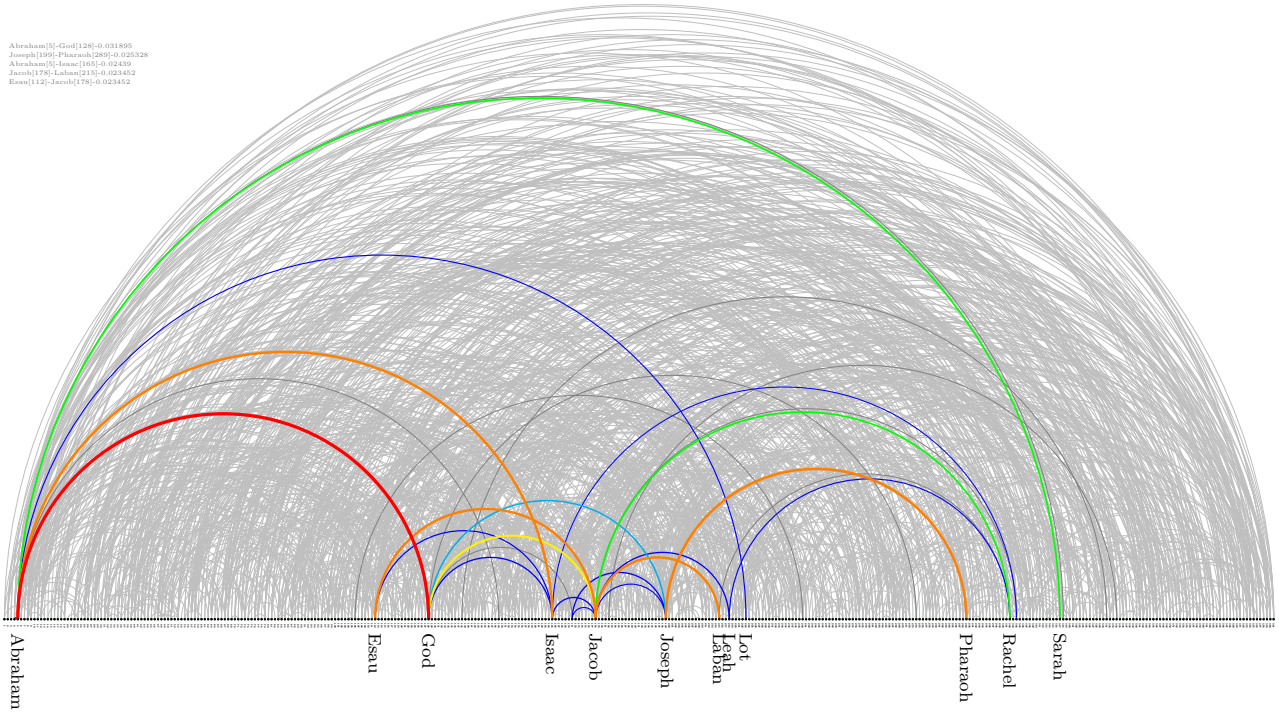


Figure 7: Interpersonal interactions among the 402 persons found in the book of Genesis, shown in Kandinski layout

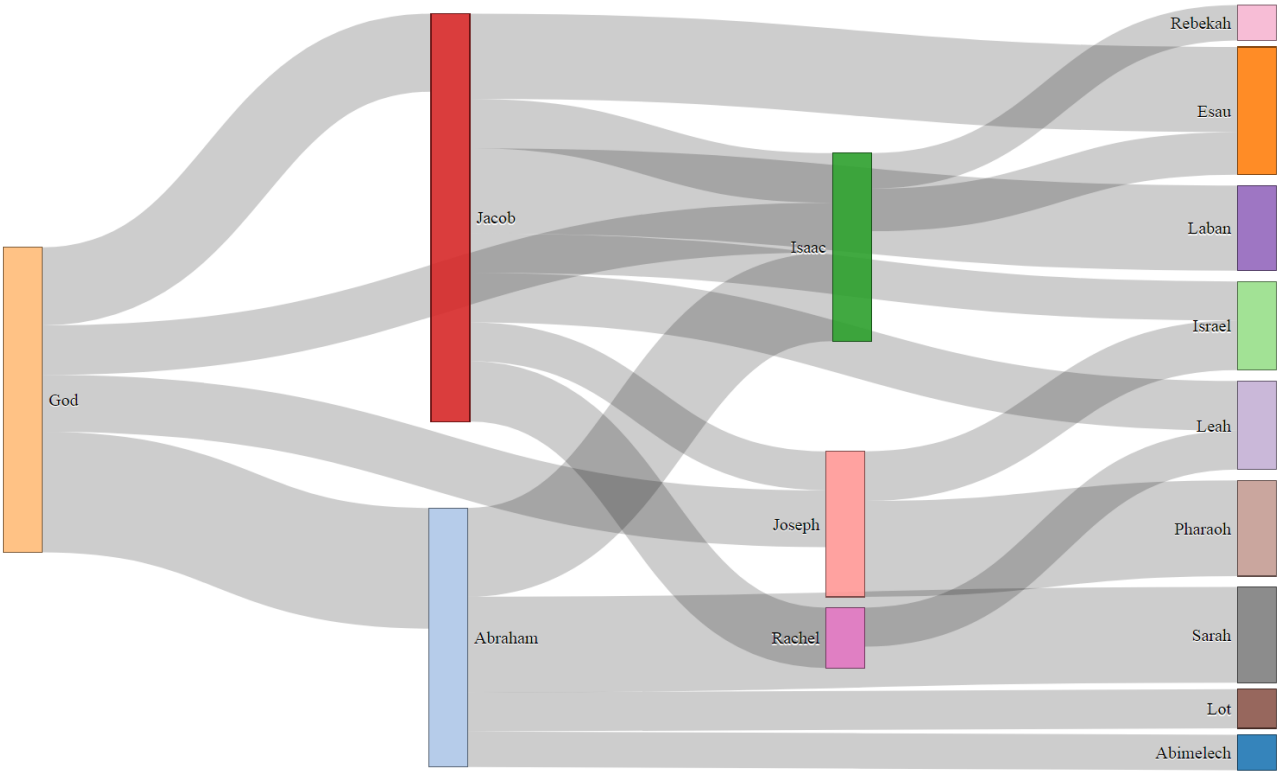


Figure 8: Interpersonal interactions among the 402 persons found in the book of Genesis, shown as a Sankey graph

link.

Note that Sankey graphs are generally used to visualize the flow of communication between the nodes, usually from left to right. Its application as a Podevsef layout should be used with care. Undirectional interactions are rendered on both sides of an internal node. Therefore, the cumulative total interactions of a person in the figure should be obtained by adding the width of every ribbon on both sides of a node. With this constraint, the Sankey visualization shows the dominating person-to-person interactions and the dominating persons in the interaction events.

We have observed that the Sankey cannot handle over 1,230 binary relationships of 406 objects effectively. The result was nearly unreadable, defeating the primary motivation of visualization, "a picture is worth a thousand words." Because of this scalability issue, it is mainly used with a small number of selected objects.

3.3 2-D Circular Layout

One way to avoid the potential visual bias caused by the orientation of the edges observed in Kandinski and Podevsef layouts is to place all the objects on the equal ground without the orientation of left and right and all the edges on the same side of the visual space. This idea is implemented using the circular layout (Fig. 3a).

In circular layout, the nodes are nominal, but not ordinal. There is no notion of left-end or right-end. There is no beginning nor end. Nodes may be placed in a random order or in the lexicographical order for the merit of object search visually. Edges are rendered within the disc.

In our approach, objects are arranged by alphabetical order of the object label. The significance of a relationship is represented by the color of the corresponding edge. The length of an edge is irrelevant. Using this configuration, we generated Figure 9 from the same interaction events found in the book of Genesis.

Circular layout exhibits the similar advantages and disadvantages of the Kandinski layout. However, a distinctive advantage of the circular layout over Kandinski is that it uses a twice larger area than Kandinski to present the same number of edges when both layouts use the diameter of the same length. Another advantage is that the circumference of a disc is 3.14 times longer than the diameter. Hence, the objects on a circular layout are 3.14 times more spaced out than the objects on the Kandinski axis. These differences generally make the circular layout more preferred.

We have presented thus far the visualization of binary relationships only using the three 2-D layouts that are commonly used. For the visualization of k-relationships where k is 3 or larger, we first propose a geometric shape which is suitable for such a purpose since we are not aware of any one in the literature. Recall that the ternary relationship (A, B, C) is not a valid logical consequence of three binary relationships $(A, B), (B, C)$, and (A, C). Suppose, for example, that $\{(A, B), (B, C), (A, C)\}$ is a set of co-authorship relationships. It does not necessarily mean A, B, and C publish an article together. For this reason, we opt out the traditional approach which connects a pair of related objects by a line.

4 A Geometric Shape for Many-to-Many Relationships

We now present a discussion on our strategy for the visualization of ternary or higher-degree relationships. The heart of the visualization of higher-degree relationships is to design a visualization primitive suitable for the purpose. The primitive should not only be capable of rendering k-relationships ($k \geq 3$) accurately, but it must also guarantee to prevent the visualized relationship from being interpreted incorrectly for any degrees less than k. In our presentation of the geometric shape, we assume that objects are arranged along the circumference of a circle using the circular layout for its advantages discussed in the previous section.

4.1 k-Relationships as Hyperedges

In developing an effective visual metaphor for k-relationships where k can be a reasonably large integer value, we investigated several geometric shapes as options. In particular, we have considered three approaches to the design of hyperedges.

The first approach is to connect the constituent objects to the center point of the objects with k lines for k-relationships. An example of this approach is shown in Figure 10a. This approach is computationally efficient. Also, each hyperedge is thin and occupies less pixels on the screen than the other two approaches in the same figure. Hence, it offers more visual information on the screen than the others. However, it presents a challenge to identify the correct connecting center point of a relationship. When the number of relationships is large and the degrees of the relationships are diverse, the problem will become unbearable.

The next approach is to add a geometric surface to the hyperdge at the expense of incurring more computation and leaving less visual information space. In Figure 10b, we add to the edges a polygon which has k sides for k-relationships. The advantages of this approach is that higher-degree relationships can be visually recognized quicker and easier with the aid of the polygon surface and that it is relatively easy to implement. However, it has a considerable drawback: each hyperedge occupies a much larger area than the previous edge-only approach, leaving less amount of unoccupied pixels on the screen. Hypothetically, when the degree of a relationship is close to the maximum degree of participation, it is possible that the single hyperedge would block nearly the entire visualiza-

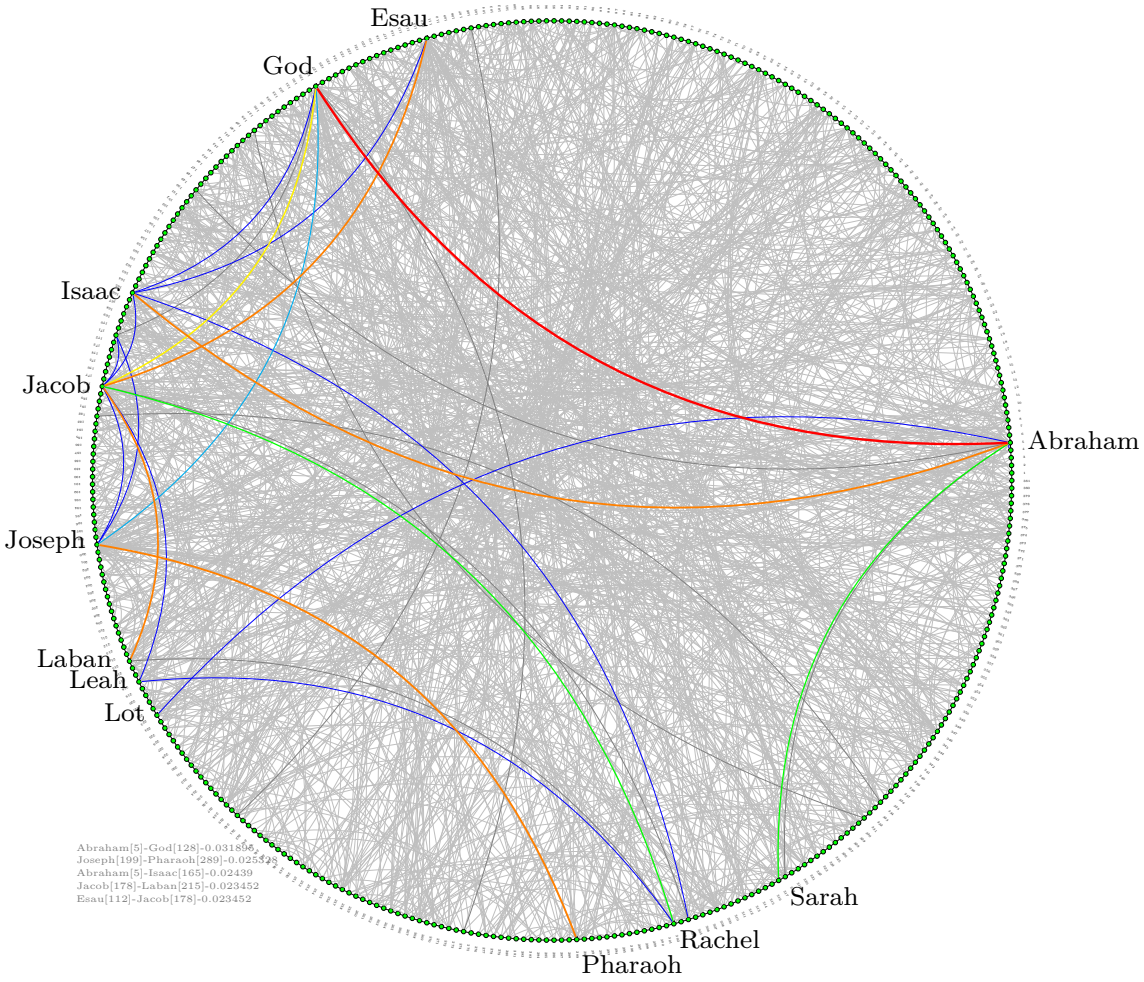
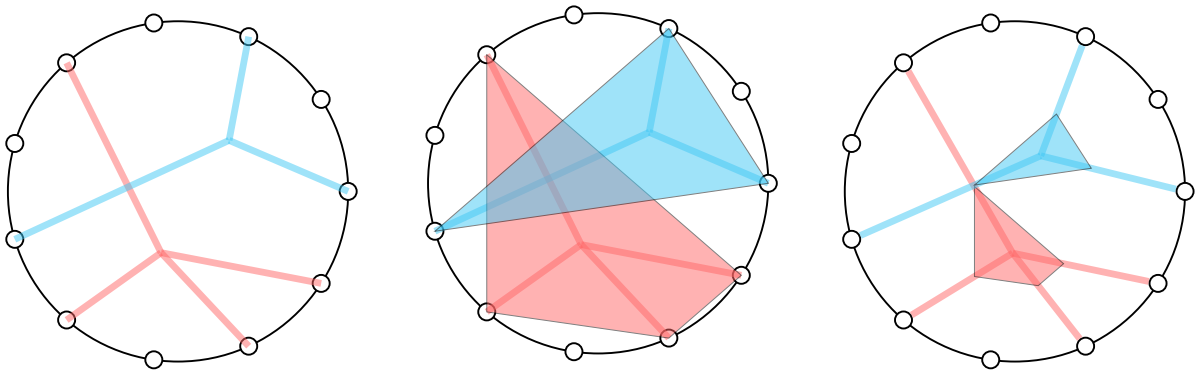


Figure 9: Interpersonal interactions among the 402 persons found in the book of Genesis, shown in circular layout.



(a) A geometric shape with k legs only. (b) A geometric shape with k legs and a covering surface. (c) A geometric shape with k legs and a smaller covering surface.

Figure 10: A conceptual design of the hyperedge for the visualization of k-relationships where $k = 3$ and $k = 4$.

tion space (although such high-degree relationships are rare in real-world datasets).

Between these two approaches, we take the second approach, with a modification, mainly because of its strength in distinguishing relationships and their degrees.

The improvement that we made to the second hyperedge approach is three-fold: we use smaller hyperedge covers as illustrated in Figure 10c, use concave sides for the polygon surface to further reduce the amount of pixels to use, and spherical edges are used so that the proposed hyperedges may be used in a 3-D visualization as well.

For an example of a hyperedge connecting 3 objects, we can employ a *hyperbolic triangle* that can be found in hyperbolic geometry. The characteris-

tics of a hyperbolic triangles are well studied in a 2-D space. However, extending the theory of hyperbolic triangles to a higher-dimensional geometric shape for the purpose of implementation requires substantially more effort. The novelty of our work in implementing spherical edges and accompanying concave-sided polygon surfaces is to use Bezier curves as discussed below.

4.2 Geometry of hyperedges

As discussed earlier, the proposed hyperedge primitive has edges (we call them "legs") and a cover that is well suited in a 3-D metaphor. A k-ary hyperedge has k legs plus a 3-D surface covering the midpoint of the legs to make it easier to identify the degree of the underlying relationship and to augment the 3-D look of the hyperedge. In designing and implementing such hyperedges, we ought to maximize the separation of hyperedges from each other for effective data analysis. We begin with the construction of legs.

Each k-relationship is first mapped to a k-ary hyperedge of k legs, each of which is a *second-order (quadratic) Bezier curve* standing vertically. A quadratic Bezier curve requires three control points to define, i.e., two line segments between the first and second points and between the second and the third points. Given the three control points p_0, p_1 and p_2, the quadratic Bezier curve is defined as a function of the parameter $t \in [0,1]$:

Definition 4.1 (Second-order Bezier Curve)

$$B(t) = (1-t)^2 \cdot p_0 + 2t(1-t) \cdot p_1 + t^2 \cdot p_2$$

where t denotes a curve segment.

The polynomial expressed in second-order Bezier curves can be approximated by the following repeated steps:

1. Start with a line segment L_1 connecting p_0 and p_1 and another line segment L_2 between p_1 and p_2.
2. Place a marker M_1 along L_1 at distance t from p_0 and another marker M_2 along L_2 at the same distance from p_1.
3. Draw a line L between M_1 and M_2, and place a marker at distance t from M_1. Emit the marker as a point on the Bezier curve.
4. Repeat the steps with respect to the next value of t.

Notice that the finer the line segments (controlled by the division of t), the smoother the resulting curve will be as illustrated in Figure 11. Since the calculation of the emitting point take a constant amount of time, the time complexity of drawing a second-order Bezier curve is proportional to the number of the curve segments s, i.e., $O(s)$.

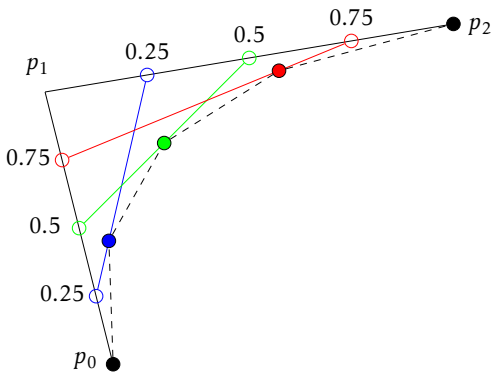


Figure 11: The effect of the number of curve segments on the Bezier curve generated from p_0 and p_2 with a control point p_1 in between. With 4 line segments at the interval of 0.25, the Bezier curve is approximated by the three interpolated points (in blue, green and red) plus p_0 and p_2, resulting in the coarse dashed curve.

Construction of hyperedge. Adapting the quadratic Bezier curve, we construct hyperedges as follows: Let I be a k-relationship comprising k objects, i_1, ..., i_k, which are located on the circumference of a circle in the circular layout, and let p_m be the center point of the k objects on the x-z plane. Then, we can construct the hyperedge for the k-relationship as follows: First, take i_1 as p_0 of the corresponding quadratic Bezier curve. We can compute the location of p_2 such that p_2 is the opposite point of p_0 on the circle about p_m. Furthermore, set the y-axis value (i.e., the height) of p_m to the statistical significance value of I. Now we can create a quadratic Bezier curve using the three control points p_0, p_m and p_2. Take the half of the curve on the p_0 side as the component curve for i_1. Repeat the steps to construct the other $k-1$ component curves from the other $k-1$ objects of I. The resulting hyperedge is generated by joining the k component curves at their peak point (located at the same x–y coordinate with p_m). Figure 12 shows how six half-length Bezier curves join at the crest in a 3-D space. The resulting hyperedge is parabolic-shaped, and its height is scaled by the statistical significance of the corresponding k-relationship automatically. The hyperedge is further decorated with a concave-sided polygon cover over the peak point for visual clarity.

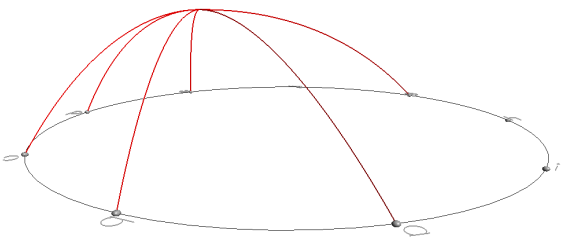


Figure 12: An hyperedge of 6 legs, each of which is a half-length second-order Bezier curve.

Figure 13a shows examples of hyperedge imple-

mentations for 3-, 4-, 5-, 6- and 7-relationships. Hyperedges for 3-relationships are constructed from three Bezier curve components with a cover at the joining point of the component curves as shown in the figure. The figure also illustrates that relationships of higher degree can also be constructed seamlessly using the same algorithm. Notice also in Figure 13c that each side of the polygon cover of a hyperedge is also a Bezier curve, the exact shape of which is determined by the locations of the control points.

The length of the portion of a leg which is not covered by the corresponding cover is about 50% of the leg length (measured from the bottom to the peak) by default. If the coverage ratio is larger, it tends to be easier to inspect the corresponding relationship visually, and visual ambiguity is reduced. However, it can easily mask other hyperedges at the same time. This ratio can be dynamically adjusted for maximum visual effect in our implementation. We found that the coverage ratio of 20% yields good visual feedback, while not dominantly masking other hyperedges. Once the location of each vertex of the hyperedge cover is determined on a leg, every pair of adjacent vertices (on two neighboring legs) and the peak point of the hyperedge are used as the three control points to generate the final hyperedge cover between the two legs.

Time complexity of hyperedge. As discussed earlier in this section, the time complexity of the construction of a second-order Bezier curve is $O(s)$, where s denotes the number of curve segments. Hence, the time complexity of the construction of a hyperedge with k legs without the cover is $O(ks)$. Each hyperedge with k legs has $k-1$ cover segments between a pair of legs. Each cover segment costs additional $O(s)$ operation when we use the same level of smoothness for each Bezier curve. Hence, adding the cover to the hyperedge will cost additional $O(ks)$, and the total time complexity of a hyperedge with k legs and a cover is $O(ks)$. From this analysis, we conclude that the overall complexity of the visualization of n relationships is $O(ksn)$.

Notice that the average value of k is relatively small in the real-world data when relationships obey the anti-monotone principle. In our implementation, n is the dominating factor for the scalability of the proposed approach. On the other hand, the capability of rendering a large number of relationships may bring an adversary effect, especially when the relationships are not evenly spaced out; that is, individual objects become less discernible. This issue is a system development issue and beyond the scope of this paper.

4.3 Visual cues

Visualization of relationships of arbitrary degrees would be challenging mainly due to the fact that the degrees of relationships can vary widely and the number of relationships, at least in theory, that are dealt with in data warehouses or in big data can be very large. Also, users may respond to the visual stimuli coming from a large number of visual primitives differently. For example, one user may be sensitive to variations in color but not in length, and vice versa. For these reasons, in designing a visual metaphor for relationships, we need to utilize visual cues to reinforce the meaning of each relationship in the context and to minimize any user-level distortion of the visual information.

For 3-D visualization of relationships, using multiple visual cues is generally more effective than using a single visual cue as long as they are used with consistency since viewers' responses to a particular visual cue may vary. In the proposed visualization primitive for k-relationships, multiple visual cues are used to maximize the discriminative power of the metaphors: shape, color, illumination, height, and thickness. These visual cues are demonstrated in the figures throughout this paper.

Relative location Since the peak point of each component Bezier curve of a hyperedge is a function of its associated control points, hyperedges tend to be spatially clustered when hyperedges are rendered on the same screen. For example, the four 6- and 7-relationships in Figure 13c form a clump. This tendency will increase toward the center of the screen as the degree of relationships increases (in other words, the distribution of the objects becomes more even on the base circle).

Color The hue of the legs representing a k-relationship r is determined by the statistical significance of the relationship using the following formula:

$$hue(r) = \frac{\max - weight(r)}{\max}$$

where max is the maximum significance in the dataset, and $weight(r)$ is the statistical significance of the corresponding relationship r. With this color scheme, the most significant legs are rendered in red and the least significant ones in blue, with yellow and green edges in between.

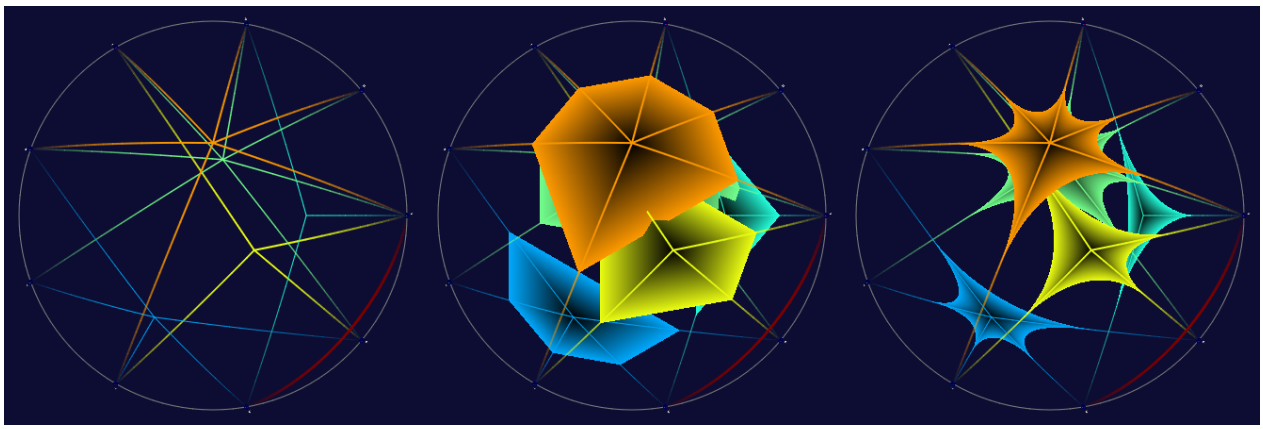
Illumination Shading is a natural phenomenon to real 3-D objects under illumination. We add a pseudo-shading effect to hyperedges to augment their 3-D look by applying varying illumination rather than uniform illumination.

The leg illumination is defined as a function of $t \in [0,1]$

$$illumination(t) = -\frac{(t-0.5)^2}{0.5^2} + 1$$

where t denotes a curve segment.

The above formula guarantees a leg to be the brightest at the vertex point (i.e., peak point) and gradually darker toward both end points, as illustrated in Figure 13.

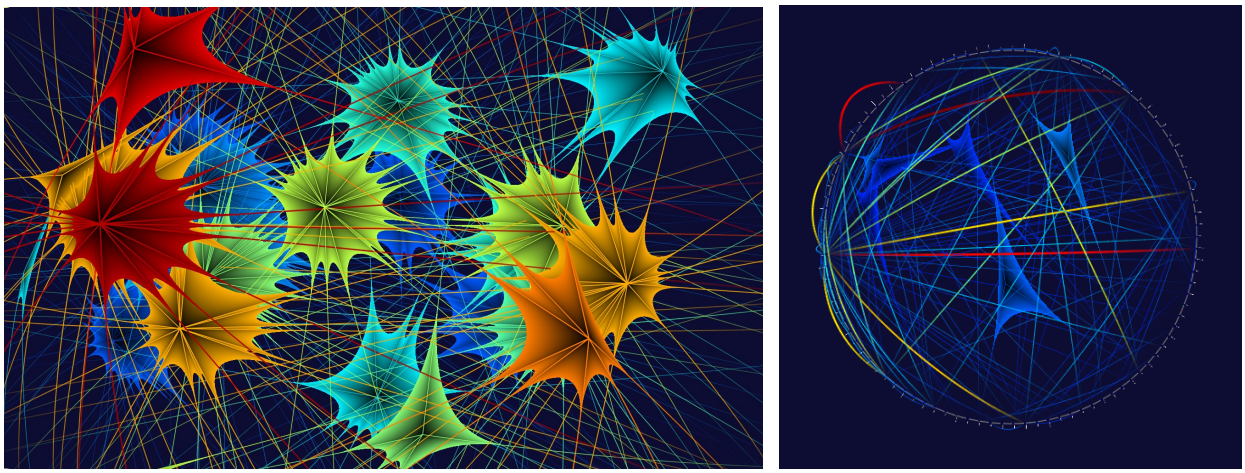


(a) Hyperedges with spherical edges only.

(b) Hyperedges with spherical edges and polygons with straight sides.

(c) Hyperedges with spherical edges and polygons with convex sides.

Figure 13: Implementation of hyperedges for 3-, 4-, 5-, 6- and 7-relationships.



(a) Hyperedges of various degrees and statistical significance from an artificial dataset. Each distinctive geometric shape with multiple visual cues help viewers visually inspect the overall and individual trend in the underlying relationships.

(b) Frequently co-occurring words in the first four books of the New Testament. The three outstanding ternary relationships are ("elders," "priests," "chief"), ("priests," "scribes," "chief"), and ("son," "man," "jesus").

Figure 14: Hyperedges rendered for artificial and real datasets.

Thickness The thickness of a leg is also proportional to the statistical significance of the corresponding relationship in our approach.

5 Implementation and Demonstration in Action

Implementation of the proposed visualization primitive. The proposed primitive is implemented in two different 3-D graphics programming environments, one in Java using the Java 3d library and the other in C using the C OpenGL library. Both libraries provide comparable 3-D graphics programming features. In our development, it took much longer time and effort to arrive at the current design from the idea of hyperbolic geometry than actual implementation.

Hyperedges in action. Figure 14a shows that highly significant relationships are in red and tall (over other hyperedges), whereas less significant ones are in blue and short. Note that the figure is generated from an artificial dataset to demonstrate the capability of the proposed visual primitive of various visual cues employed in our approach.

Figures 14b and 15 are generated using real-world datasets. Figure 14b shows frequently co-occurring words in the first four books of the New Testament. Overall, the visualization tells us that there are only a few word sets in which 3 or more words occur together at a statistically significant level. This is a natural phenomenon because frequently occurring words obey the principle of anti-monotone. In fact, among the 165 sets of co-occurring words with a minimum support of 0.35%, only ("elders," "priests," "chief"), ("priests," "scribes," "chief"), and ("son," "man," "je-

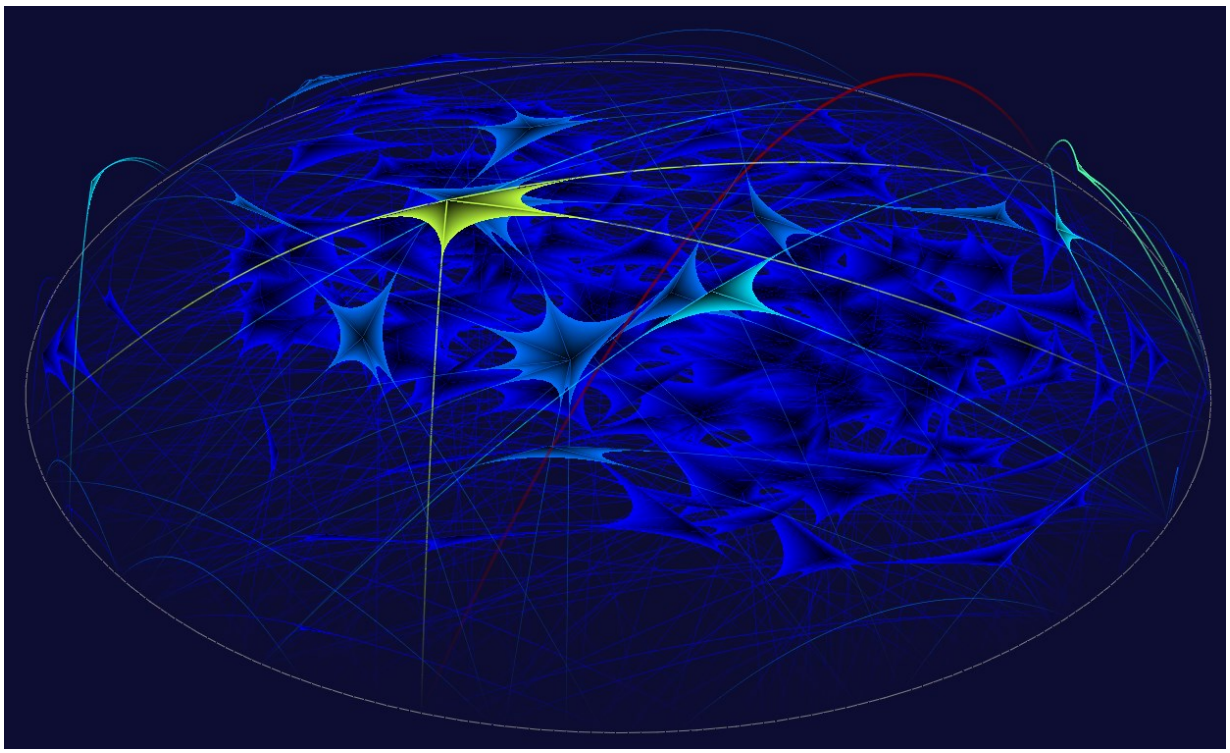


Figure 15: Interpersonal interactions found in the book of Genesis. The visualization shows that most interactions of more than 2 persons occur at a relatively low frequency. The binary interaction (God, Noah) shown in red is the most significant. The interactions between Noah and his three sons shown in yellow is one of the few outstanding interactions involving more than 2 persons.

sus") occurs more frequently than the threshold.

Figure 15 shows 402 persons along the circumference of the disc and the maximal interactions discovered from the book of Genesis. As these maximal interactions also obey the principle of anti-monotone, only a handful of relationships of higher degrees stand out in the visualization. The most significant relationship is the binary interaction between God and Noah (the red and tall arc). The most significant 4-relationship in yellow in the figure is between Noah and his three sons.

6 Conclusion

A 3-D visual metaphor for many-to-many relationships is presented in this paper. The proposed visualization primitive accurately conveys the degree of participation of a relationship with its statistical significance using multiple visual cues. Each relationship of ternary or higher degree of participation is modeled as a novel hyperedge having as many legs as the degree of the relationship and a cover at the junction point of the legs. The metaphor is implemented as visual primitives of high-quality graphical objects and shows a good separation of different degrees of participation.

The effectiveness of our model with respect to information communication is generally demonstrated through a handful of experiments in this paper. At the same time, we forewarn the reader that efficiency in terms of CPU cycles could be a concern in comparison to 2-D multivariate visualization techniques since we render information as high-quality 3-D objects. In the future, we plan to evaluate the effectiveness of the proposed visual primitive using quantitative usage statistics and a user study to assess how well our visualizations improve information processing, communication, and decision making as suggested in [13].

Conflict of Interest The authors declare no conflict of interest.

References

[1] Augustus De Morgan. On the syllogism, no. iii, and on logic in general. In *Transactions of the Cambridge Philosophical Society*, chapter 10, pages 173–230. Cambridge University Press, Cambridge, 1864.

[2] A. Sutcliffe, R. Dunbar, J. Binder, and H. Arrow. Relationships and the social brain: integrating psychological and evolutionary perspective. *British J. of Psychology*, 103(2):149–168, 2012. doi: 10.1111/j.2044-8295.2011.02061.x.

[3] M. E. J. Newman. Coauthorship networks and patterns of scientific collaboration. *Proc. of the National Academy of Sciences of USA*, 101(1):5200–5205, April 6 2004.

[4] Wanita Sherchan, Surya Nepal, and Cecile Paris. A survey of trust in social networks. *ACM Comput. Surv.*, 45(4):47:1–47:33, August 2013. doi: 10.1145/2501654.2501661.

[5] M. Kessler. Bibliographic coupling between scientific papers. *American Documentation*, 14(1):10–25, 1962.

[6] Zhongshan Lin and SeungJin Lim. Optimal candidate generation in spatial co-location mining. In *Proc. of the 24th Annual ACM Symposium on Applied Computing*, pages 1441–1445, Mar 2009. doi: 10.1145/1529282.1529604.

[7] Maria Cristina Ferreira de Oliveira and Haim Levkowitz. From visual data exploration to visual data mining: a survey. *IEEE TVCG*, 9(3):378–394, Jul-Sep 2003. doi: 10.1109/TVCG.2003.1207445.

[8] Y. Matsuo and M. Ishizuka. Keyword extraction from a single document using word co-occurrence statistical information. *Int'l J. on AI Tools*, 13:157–169, 2004. doi: 10.1142/S0218213004001466.

[9] Eduardo Tejada and Rosane Minghim. Improved visual clustering of large multi-dimensional data sets. In *Proc. of the 9th Int'l Conf. on Info. Vis.*, pages 818–825, July 2005. doi: 10.1109/IV.2005.61.

[10] James Abello and Jeffrey Korn. MGV: A system for visualizing massive multidigraphs. *IEEE TVCG*, 8(1):21–38, Jan-Mar 2002. doi: 10.1109/2945.981849.

[11] Joshua O'Madadhain, Danyel Fisher, Padhraic Smyth, Scott White, and Yan-Biao Boey. Analysis and visualization of network data using JUNG. *J. of Statistical Software*, VV, 2005.

[12] Tze-Haw Huang and Mao Lin Huang. Analysis and visualization of co-authorship networks for understanding academic collaboration and knowledge domain of individual researchers. In *Proc. of the 3rd CGIV*, pages 18–23, 2006.

[13] Maneesh Agrawala, Wilmot Li, and Floraine Berthouzoz. Design principles for visual communication. *Commun. ACM*, 54(4):60–69, April 2011. doi: 10.1145/1924421.1924439.

[14] Michael A. Bekos, Michael Kaufmann, Stephen G. Kobourov, and Antonios Symvonis. Smooth Orthogonal Layouts. 17(5):575–595, 2013.

[15] Julian Heinrich and Daniel Weiskopf. State of the art of parallel coordinates. In M. Sbert and L. Szirmay-Kalos, editors, *Eurographics 2013 - State of the Art Reports*. The Eurographics Association, 2013. doi: 10.2312/conf/EG2013/stars/095-116.

[16] Michael Friendly. Corrgrams: Exploratory displays for correlation matrices. *The American Statistician*, 56(4):316–324, Nov 2002.

[17] Philipp Steinweber and Andreas Koller. *Similar diversity*. in Visual Complexity: mapping patterns of information. 2011.

[18] Daniel A. Keim, Christian Panse, Mike Sips, and Stephen C. North. Visual data mining in large geospatial point sets. *IEEE CG&A*, 24(5):36–44, Sep-Oct 2004. doi: 10.1109/MCG.2004.41.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 2, No. 3, 1540-1548 (2017)
www.astesj.com
Special Issue on Recent Advances in Engineering Systems

ASTES Journal
ISSN: 2415-6698

On the Performance of a Clustering-based Task Scheduling in a Heterogeneous System

Hidehiro Kanemitsu[1], Masaki Hanada[2*], Emilia Ndilokelwa Weyulu[2], Moo Wan Kim[2]

[1]*Global Education Center, Waseda University, 169-8050, Japan*

[2]*Department of Information Systems, Tokyo University of Information Sciences, 265-8501, Japan*

ARTICLEINFO

Article history:
Received: 31 May, 2017
Accepted: 18 July, 2017
Online: 19 August, 2017

Keywords :
Task Scheduling
Task Clustering
Real Environment
List Scheduling

ABSTRACT

Recent task scheduling algorithms for a generalized workflow job in heterogeneous system adopt list-based scheduling. In those algorithms, the response time cannot be effectively reduced if the given workflow job is data-intensive. If the workflow job is computationally intensive, an attempt is made to assign tasks to many processors, which can lead to resource starvation. To this end, a task scheduling algorithm that is based on clustering tasks, called CMWSL (Clustering for Minimizing the Worst Schedule Length) has been proposed. In CMWSL, the lower bound of the assignment unit size for each processor is derived in order to suppress the total number of executing processors for effective use of processors. After the lower bound is derived, the processor as a assignment target is determined and then the assignment unit as a task cluster is generated. As a final phase of CMWSL, task ordering is performed for every assigned task. In this paper, we compare several task ordering methods in CMWSL in a real environment to find the best task ordering policy.

1 Introduction

Recent household computers have high-performance processing schemes, e.g., multi-core, hyper-threading technologies and so on. Such parallel processing schemes enable the expansion of conventional applications to scientific and large amount of data processing applications in household computers. This transition increases the needs for parallel and distributed processing schemes, e.g., grid computing[1], utility computing and cloud computing[2]. These computing models require an appropriate execution method to acquire the optimal value in terms of the response time or the schedule length, energy consumption, the economic cost and throughput. However, to obtain the optimal schedule length has been an *NP*-complete problem. In a heterogeneous system where each processing speed and communication bandwidth is arbitrary, a task assignment method as well as a task scheduling method are needed because the actual processing time and communication time are determined after each task is assigned to a processor.

In various application models, especially for work-flow type application, called Directed Acyclic Graph (DAG) is one of currently studied task scheduling fields, also known as a DAG scheduling problem. DAG scheduling is classified into two approaches, i.e., list-based scheduling[3, 4, 5, 6] and clustering-based task scheduling[7, 8, 9, 10, 11]. In HEFT (Heterogeneous Earliest Finish Time)[3], which is the most famous list scheduling for DAG applications in heterogeneous systems, a scheduling priority is assigned to each task by using the average processing time and the average communication time. Then each task is assigned to a processor by which its completion time is minimized. Though the critical part of a list scheduling is how to define the priority for ready tasks, list-based scheduling models such as HEFT, PEFT[4] and CEFT[5] apply averaged processing time and communication time for deriving a priority value for each task. As a result, a good schedule length cannot be obtained if the job is data-intensive or has a large dispersion among task sizes and data sizes. On the other hand, in task clustering for heterogeneous systems, each task belongs to a task cluster. Each task cluster is assigned to a processor and each task is scheduled in order to minimize the schedule length. Though a task clustering method is known as one of the promising methods in homogeneous systems, a good schedule length cannot be obtained in heterogeneous systems because it assumes a temporal homogeneous system for deriving each clustering priority.

We proposed a task clustering method for heterogeneous systems for effective use of processors in [12], the

*Corresponding Author Masaki Hanada, Department of Information Systems, Tokyo University of Information Sciences, 4-1, Onaridai, Wakaba-ku, Chiba 265-8501, Japan & mhanada@rsch.tuis.ac.jp

https://dx.doi.org/10.25046/aj0203192

objective of which was to minimize the schedule length using a small number of processors. The method proposed in [12] firstly derives the lower bound of the total execution time of tasks in a cluster to regulate the number of processors. Then each task is clustered in order to minimize the schedule length by reducing the Worst Schedule Length (WSL). According to the literature in [13], minimizing WSL leads to the reduction of both the lower bound and the upper bound of the schedule length. Then each task is ordered by Ready Critical Path (RCP) scheduling[14] on each assigned processor. However, there is no criteria to minimize WSL in the task scheduling. In this paper, we propose a task scheduling to minimize WSL after the task clustering proposed in [12] has been finished. Experimental results in a real environment show that the proposed task scheduling outperforms the method in [12] in terms of the schedule length. This paper is an extension of work originally presented in the 19th International Conference on Advanced Communications Technology(ICACT2017)[15].

2 Assumed Model

2.1 Job Model

An assumed job is expressed as a Directed Acyclic Graph (DAG), which is known as workflow type job. Let $G = (V, E)$, where V is the set of tasks and E is the set of edges (data communications among tasks). An i-th task is denoted as n_i. Let $w(n_i)$ be the workload of n_i, i.e., $w(n_i)$ is the sum of unit times taken for being processed by the reference processor. We define the data dependency and direction of data transfer from n_i to n_j as $e_{i,j}$. And $c(e_{i,j})$ is the sum of unit times taken for transferring data from n_i to n_j over the reference communication link.

One constraint imposed by a DAG is that a task cannot be executed until all data from its predecessor tasks arrive. For instance, $e_{i,j}$ means that n_j cannot be started until the data from n_i arrives at the processor that executes n_j. And let $pred(n_i)$ be the set of immediate predecessors of n_i, and $suc(n_i)$ be the set of immediate successors of n_i. If $pred(n_i) = \emptyset$, n_i is called START task, and if $suc(n_i) = \emptyset$, n_i is called END task. If there are one or more paths from n_i to n_j, we denote such a relation as $n_i \prec n_j$.

2.2 System Model

We assume that each computer is completely connected to others, with non-identical processing speeds and non-identical communication bandwidths. The set of processors is expressed as $P = \{p_1, p_2, \ldots, p_n\}$ and let the set of processing speeds in p_i be α_i. As for data communication among computers, let the communication bandwidth of p_i be β_i. Since the actual communication speed depends on the bottleneck in the network path, suppose the communication speed is the minimum value in the two bandwidths among p_i and p_j, i.e., $L_{i,j} = min\{\beta_i, \beta_j\}$, where $L_{i,j}$ is the communication speed among p_i and p_j.

The processing time in the case that n_k is processed on p_i is defined as

$$t_p(n_k, \alpha_i) = w(n_k)/\alpha_i. \tag{1}$$

The data transfer time of $e_{k,l}$ over $\beta_{i,j}$ is defined as

$$t_c(e_{k,l}, L_{i,j}) = c(e_{k,l})/L_{i,j}. \tag{2}$$

This means that both processing time and data transfer time are not changed with time, and suppose that data transfer time within one computer is negligible. In general, the communication setup time for the preparation, denoted by O_k occurs before p_k sends the data. If $c(e_{i,j})$ is sent from p_k to p_l, the communication time is defined by O_k and the communication speed $L_{k,l}$; in particular, $L_{k,l} = \min\{\beta_k, \beta_l\}$ and the communication time is given by

$$t_c(e_{i,j}, L_{k,l}) = O_k + c(e_{i,j})/L_{k,l}. \tag{3}$$

We assume the communication setup time is negligible, i.e., $O_k = 0$ for $\forall p_k \in P$.

3 Related Work

In this section, we discuss existing methods for Directed Acyclic Graph (DAG) scheduling in heterogeneous systems. We also clarify their disadvantages when they are applied to systems with a large number of processors and when not all processors are used for execution. DAG scheduling for heterogeneous systems is categorized into two classes in terms of the task assignment policy: list-based heuristics [3, 4, 5, 16, 17] and task clustering [7, 8, 9, 10, 11].

For list-based heuristics, both HEFT[3] and Predict Earliest Finish Time (PEFT)[4] use the average processing time and average communication time to derive the scheduling priority for each free task whose predecessor tasks have been scheduled. According to the HEFT algorithm, the scheduling priority of task n_k is based on the path length from n_k to the END task, which is denoted by $rank_u^*(n_k)$ as follows:

$$\begin{aligned} rank_u^*(n_k) &= t_p(n_k, \alpha_i) + t_c(e_{k,l}, L_{i,j}) \\ &+ \max_{n_l \in suc(n_k)} \left\{ rank_u^*(n_l, \alpha_j) \right\}. \end{aligned} \tag{4}$$

In the set of free tasks, the task with the maximum $rank_u$ value is chosen and assigned to the processor to minimize the completion time of the task using an insertion-based policy. PEFT[4] uses an Optimistic Cost Table (OCT); the OCT value of task n_k assigned to p_p is the longest path from n_k to the END task in the set of paths starting from n_k, provided that each path length assumes the minimum length in the processor assignment combinations when n_k is assigned to p_p. The scheduling priority of n_k in PEFT is the average OCT value of n_k through all of the processors in P. One difference between HEFT and PEFT is whether the actual processor assignment is considered during the prioritizing phase. Another difference is that PEFT uses the average value. In Constrained Earliest Finish Time (CEFT)[5], the pruning phase is performed first. In the pruning phase, the longest path, called the Constrained Critical Path (CCP), which is derived from the average processing time and average communication time, is pruned from the DAG. Then, the pruned DAG is traversed, and the new CCP is pruned. Pruning continues until all of the tasks are pruned. Tasks in a CCP are assigned to the same processor

where their completion times are minimized. The Minimizing Schedule Length (MSL) algorithm[16] uses Mean Execution Time (MET) and Total Communication Time (TCT) for the scheduling priority derived by MET+TCT. They are calculated based on the average processing time and the average communication time. The Heterogeneous Selection Value (HSV) algorithm[17] derives $rank_u^*$ values for all processors for each task. Then the average $rank_u^*$ value is calculated as $hrank_u$. The scheduling priority in HSV is derived by the out degree times $hrank_u$, which reflects the idea that a task having larger out degree can have great effect on the schedule length. Since HEFT, PEFT, CEFT, MSL, and HSV use the average value calculated from all processor information, each scheduling priority is not accurate for minimizing the schedule length. Longest Dynamic Critical Path (LDCP)[6] uses the actual processing time and communication time for each scheduling priority and outperforms HEFT in terms of the schedule length. However, LDCP requires a set of DAGs for all processor assignment patterns, and thus, a high space complexity is required if the system has a large number of processors.

4 Overview of CMWSL Algorithm

In this section, we present an overview of the clustering-based task scheduling, called CMWSL (Clustering for Minimizing Worst Schedule Length)[12]. CMWSL consists of: (i) the lower bound of the total processing time is derived to the processor by which the maximum worst schedule length (defined as WSL (Worst Schedule Length) in Section 4.1) is obtained. (ii) several tasks on WSL path are clustered until the total processing time exceeds the lower bound in order to minimize WSL. (iii) the execution order for each task on a processor is decided according to the scheduling priority. In the following sections, the definition of WSL, the lower bound, the task clustering procedures, and the task ordering procedures are presented. For more details, see the literature in [12].

4.1 WSL (Worst Schedule Length)

Since CMWSL assumes WSL, we present the basics of WSL. Intuitively, WSL is the maximum execution path length when each task is executed as late as possible in a node, provided that there is no data waiting time for each task once the node starts executing; that is, WSL is the upper bound of the schedule length if no data waiting time occurs for each processor.

Fig. 1 shows an example of WSL derivation. In this figure, (a) is the state after four tasks have been clustered, (b) is the meaning for each items in (a), and (c) is the given system information, where three processors exist, i.e., p_1, p_2 and p_3. At (a), a virtual processor having the maximum processing speed (=4) and the maximum communication bandwidth (=4) is assumed to be assigned to a task. From this state, WSL is derived by $A \to C \to F \to E \to G \to H$, i.e., the longest execution path length is obtained when E is scheduled as late as possible on p_2.

If a processor assignment can make WSL smaller, both the upper bound and the lower bound of the actual schedule length can be made smaller[13]. The schedule length cannot be obtained until every task execution order is determined. Thus, at the task clustering phase as a task assignment, the objective is to minimize WSL, not to minimize the actual schedule length.

4.2 Lower Bound of Total Execution Time for each Processor

In general, the more tasks are assigned to a processor, the more data communications can be localized on one processor. However, to increase the number of assigned tasks on a processor can lead to the reduction of the degree of parallelism. The balance between localization of data communication and parallelism should be considered for minimizing the schedule length. Furthermore, imposing the lower bound can reduce the number of assigned processors, thereby effective use of computational resources can be achieved. In CMWSL, the lower bound of the total processing time on a processor p_p is defined as follows (for more details, see the literature in [12]):

$$\delta_{opt}(\alpha_p,\beta_p) = \sqrt{\frac{1}{\alpha_p}\left\{\varepsilon\lambda + \frac{w_{min}}{\alpha_p}\sum_{i=1}^{N}\sum_{n_k\in seq_{s-1}^{<}(i)} w(n_k)\right\}}, \quad (5)$$

where

$$\varepsilon = \sum_{n_k\in seq_{s-1}^{<}} w(n_k) - \sum_{i=1}^{N}\sum_{n_k\in seq_{s-1}^{<}(i)} w(n_k),$$
$$\lambda = \left(\frac{c_{max}}{\beta_p} - \frac{c_{max}}{\beta_{max}} + \frac{w_{min}}{\alpha_{max}} + \frac{c_{min}}{\beta_{max}}\right),$$

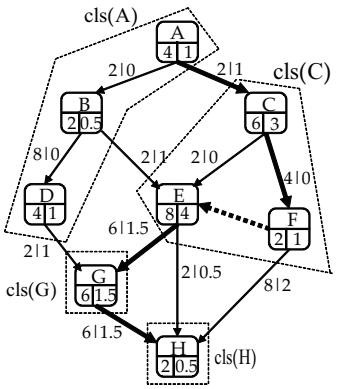
where $seq_{s-1}^{<}$ be the set of tasks in a path from a START task to the END task in the WSL sequence at the workflow after $(s-1)$ tasks have been clustered. In $seq_{s-1}^{<}$, let $seq_{s-1}^{<}(i)$ be the set of tasks that belong to the i-th cluster. In addition, c_{min} and c_{max} are the minimum and maximum data sizes in $seq_{s-1}^{<}$, and w_{min} is the minimum task size. Finally, α_{max} and β_{max} are the maximum processing speed and maximum communication bandwidth, respectively.

$\delta_{opt}(\alpha_p,\beta_p)$ means the value when the upper bound of WSL is minimized. That is, both the lower bound and the upper bound of the schedule length is minimized by applying $\delta_{opt}(\alpha_p,\beta_p)$ for each cluster.

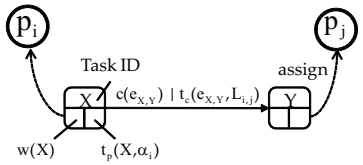
Next, we present how the processor to which an assignment unit (we call it as "cluster") is determined. CMWSL tries to find WSL path for each clustering procedure. On WSL path, there are two sub-paths, i.e., clustered path and unclustered path. Since tasks on a clustered path cannot be clustered anymore, the only set of tasks by which WSL can be varied is the unclustered path. Thus, on the unclustered path, the number of clusters whose total execution time exceeds δ (the variable for the lower bound), is defined as

$$N(\delta) = \frac{1}{\alpha_p\delta}\left(\sum_{n_k\in seq_{s-1}^{<}} w(n_k) - \sum_{i=1}^{N}\sum_{\substack{n_k\in seq_{s-1}^{<}(i),\\ cls_{s-1}(i)\notin UEX_{s-1}}} w(n_k)\right), \quad (6)$$

where $\delta = \delta(\alpha_p,\beta_p,G^s)$, G_s is the DAG after s tasks have been clustered, and $\alpha_p\delta$ is the task cluster size of p_p. $N(\delta)$



(a) State after 4 tasks have been clustered.



(b) Meanings of assigned values.

Task Cluster	Processor (p_i,α_i, β_i)
cls(A)	(p_1,4,2)
cls(C)	(p_2,2,4)
	(p_3,2,3)

(c) Processor Assignment at (a).

Figure 1: Example of WSL Derivation

is derived by taking the total workload of tasks on the unclustered path. Then, the increase of WSL by generating clusters by more r clustering steps on the unclustered path, is defined by

$$\begin{aligned} & \Delta WSL^*(M(G^{s+r})) \\ \leq & \sum_{i=1}^{N(\delta^*)-y} \Delta LV^{up}_{lnr}(cls_{s+r}(i)) + \sum_{i=N(\delta)-y+1}^{N(\delta)} \Delta LV^{up}_{nlnr}(cls_{s+r}(i)) \\ & + \sum_{\substack{e_{k,l} \in seq^{<}_{s-1}, \\ e_{k,l} \notin seq^{<}_{s+r}(i)}} \left(\frac{c(e_{k,l})}{L_{p,q}} - \frac{c(e_{k,l})}{\beta_{max}} \right) \\ \leq & \Delta WSL^*_{up}(M(G^{s+r})), \end{aligned} \tag{7}$$

where $\Delta LV^{up}_{lnr}(cls_{s+r}(i))$ is the increase of WSL by generating one linear cluster, i.e., all tasks in the cluster $cls_{s+r}(i)$ has dependencies each other. $\Delta LV^{up}_{nlnr}(cls_{s+r}(i))$ is the increase of WSL by generating one non-linear cluster, i.e., one or more tasks in the cluster $cls_{s+r}(i)$ has no dependency with others. $WSL^*_{up}(M(G^{s+r}))$ is the upper bound of the increase of WSL, i.e., $\Delta WSL^*(M(G^{s+r}))$. δ_{opt} in (5), it is obtained by differentiating $\Delta WSL^*_{up}(M(G^{s+r}))$ with respect to δ. Then $\Delta WSL^*_{up}(M(G^{s+r}))$ takes the local minimum value when δ equals to δ_{opt} defined at (5).

From the set of unassigned processors, the processor having the minimum value of $\Delta WSL^*_{up}(M(G^{s+r}))$ is selected as the next assigned target. In this phase, both the lower bound and the processor to be assigned to a cluster is determined.

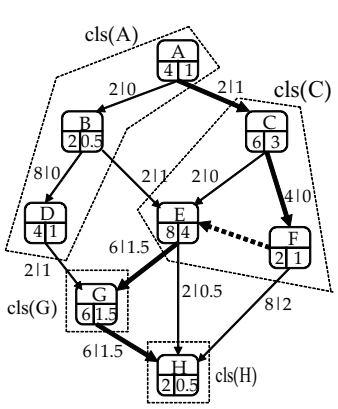
4.3 Clustering Tasks

In the clustering phase, let the selected processor by the previous section be p_p. Then the actual lower bound, i.e., $\delta_{opt}(\alpha_p, \beta_p)$ is derived. The unclustered task on WSL path becomes the start point for clustering. From such a task, the successor tasks are included in the same cluster until the total processing time exceeds $\delta_{opt}(\alpha_p, \beta_p)$ in order to minimize WSL.

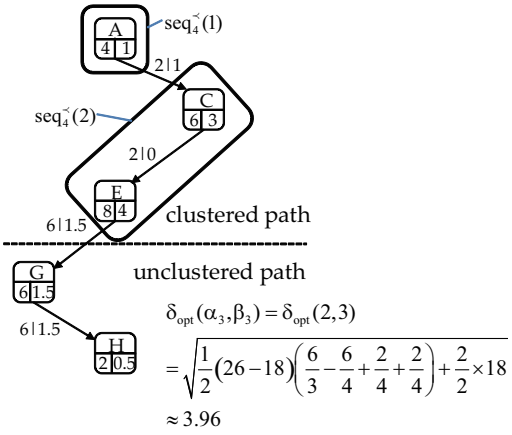
Fig. 2 shows an example of the task clustering phase in CMWSL. In this figure, (a) is the state after four tasks have been clustered, which is the same state as (a) at fig. 1. (b) in fig. 2 is the path on WSL path, i.e., A, C, E, G, H. In this path, the clustered path is $A \rightarrow C \rightarrow E$, and the unclustered path is $G \rightarrow H$. Thus, G becomes the start point for the next task clustering and p_3 is selected as the next assignment target, because only p_3 remained as the unassigned processor. From this state, the lower bound, i.e., $\delta_{opt}(\alpha_3, \beta_3) \approx 3.96$. Then at (c), G and H are clustered as a cluster $cls(G)$. Since the total processing time is 3+1 > 3.96, no more task are included at $cls(G)$. The clustering phase is finished if all clusters are generated with exceeding each lower bound.

5 Task Ordering

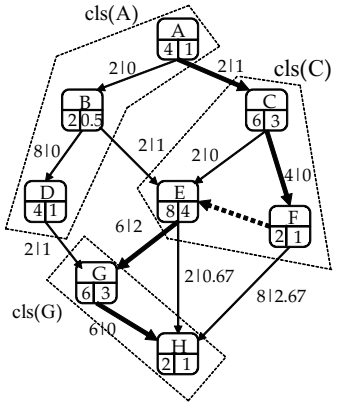
In this section, we present two list-based task ordering criteria in CMWSL. One of the objectives of this paper is to determine which one is better for minimizing the schedule length. Thus, the task ordering phase is performed after the task clustering phase presented in Section 4.3. In this phase, determining the actual execution order for each task is performed. Since the assigned processor for each task is known, the objective in this phase is to minimize the schedule length, not WSL. According to the literature in [18], a list-based task ordering method was proposed, where the task in the current WSL sequence is selected from the free list. Here, "free list" is the set of tasks, whose all immediate predecessor tasks have already been scheduled. We shall call the task ordering method as "List_maxLV". The fundamental idea behind List_maxLV is that the task in WSL



(a) State after 4 tasks have been clustered.



(b)$\delta_{opt}(\alpha_3, \beta_3)$ derivation with the path (A-C-E-G-H) being a part of WSL sequecnce, i.e.,{A,C,F,E,G,H}



(c) State after 5 tasks have been clustered.

Task Cluster	Processor (p_i,α_i, β_i)
cls(A)	(p_1,4,2)
cls(C)	(p_2,2,4)
cls(G)	(p_3,2,3)

(d) Processor Assignment at (c).

Figure 2: Example of Task Clustering in CMWSL

sequence can have an effect on the worst schedule length, i.e., such a task can prolong the actual schedule length depending on the execution ordering. Thus, such a task must be scheduled as early as possible. The task ordering method in CMWSL is quite different. The task ordering in CMWSL selects the task having the minimum DRT (data ready time) from the free list. DRT of task n_j on p_q is defined as

$$DRT(n_j, p_q) = \max_{\substack{n_i \in pred(n_j),\\ p_p \in P}} \left\{ t_f(n_i, p_p) + t_c(c(e_{i,j}), L_{p,q}) \right\}, \quad (8)$$

where $t_f(n_i, p_p)$ is the finish time of n_i on p_p, and n_j on p_q is supposed to reveive the data from n_i on p_p. We shall call the ordering method as "List_minDRT".

Fig. 3 demonstrates behaviors of both List_minDRT and List_maxLV. In this figure, A, B, and C are supposed to have been already scheduled. Thus, the content of the free list is $\{E, F, G\}$. In List_minDRT, DRTs of E, F, G are supposed to be 14, 16, and 6, respectively. Then G, having the minimum DRT is selected for assigning an idle time slot. In List_maxLV, the priority is defined as "level", which is the longest execution path length when the task is scheduled as late as possible. In this case, level values of E, F, and G are supposed to be 24, 12, and 23, respectively. Thus, E is selected because its level value is the maximum.

6 Experiment

In this section, we present the comparison results in terms of the response time as the speed-up ratio in a realistic environment. The objective of this paper is to confirm the performance of CMWSL in the real environment for practical use of CMWSL, as well as which list-based task ordering policy is better in List_minDRT and List_maxLV. We adopt Gaussian Elimination DAG and FFT (First Fourier Transform) DAG as target DAGs.

6.1 Comparison Target

As for task ordering methods in CMWSL, we adopt List_minDRT and List_maxLV. We name List_minDRT in CMWSL as "CMWSL_DRT", and List_maxLV in CMWSL as "CMWSL_LV". As other comparison targets, HEFT[3], CEFT[5], HSV[17], and PEFT[4] are adopted, because these heuristic algorithms are known to derive a good schedule length with low time complexity, i.e., they are applicable in a real situation.

6.2 Gaussian Elimination DAG

In this comparison, we prepared 20 nodes that were connected in the local network as a heterogeneous cluster. Ta-

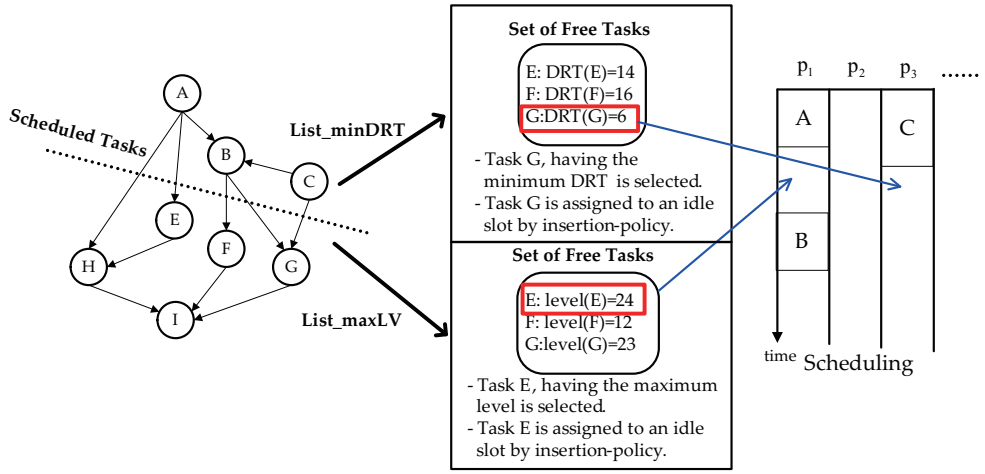


Figure 3: Example of Task Ordering Methods

Table 1: Specification of the Experimental Environment for Gaussian Elimination DAGs.

Node ♯	CPU Freq.	t_{proc} (μs)	α_p	Bandwidth	t_{comm} (μs)	β_p
1	300MHz	0.157	1.00	100Mbps	2.27	1.00
2	533MHz	0.136	1.16	1Gbps	0.45	5.04
3	800MHz	0.091	1.74	100Mbps	2.15	1.06
4	1.0GHz	0.057	2.74	1Gbps	0.62	3.66
5	1.2GHz	0.051	3.06	1Gbps	0.33	6.88
6	1.5Ghz	0.038	4.14	100Mbps	2.03	1.12
7	1.5Ghz	0.035	4.45	1Gbps	0.55	4.13
8	2.0GHz	0.032	4.92	1Gbps	0.34	6.68
9	2.4GHz	0.029	5.49	100Mbps	0.38	5.97
10	2.4GHz	0.027	5.76	1Gbps	0.27	8.41
11	2.6GHz	0.025	6.21	1Gbps	0.42	5.40
12	2.6GHz	0.024	6.56	100Mbps	2.2	1.03
13	2.8GHz	0.022	7.15	1Gbps	0.22	10.32
14	2.8GHz	0.021	7.61	1Gbps	0.37	6.14
15	2.8GHz	0.021	7.61	100Mbps	1.89	1.20
16	2.8GHz	0.019	8.43	1Gbps	0.28	8.11
17	3.0GHz	0.016	9.83	1Gbps	0.29	7.83
18	3.0GHz	0.015	10.26	1Gbps	0.39	5.82
19	3.0GHz	0.015	10.26	100Mbps	2.11	1.08
20	3.2GHz	0.013	12.42	1Gbps	0.21	10.81

ble 1 presents the specifications of the environment in terms of processing and communication performance for Gaussian elimination DAGs. In this table, t_{proc} represents the time taken to process one arithmetic operation, and t_{comm} is the time taken to send one byte from the node to the router in the local network. In particular, t_{comm} was derived by $RTT/(\sharp\, of\ bytes * 2)$ of pinging.

As for the DAG structure, we consider one task as a second inner loop, i.e., **for**$(j = k + 1; j \leq N; j++)$ at the kji Gaussian elimination without pivoting. That is, the same as the one in Section 5.5.1. Fig. 4 presents one example of Gaussian elimination DAG at $N = 6$. According to the study in [19], the processing time of a task $n_{k,j}$ and the communication time of a data point $c(e_{(k,j),(k+1,m)})$ are defined as follows:

$$\begin{aligned} w(n_{k,j}) &= (2(N-k)+1)t_{proc}, \\ c(e_{(k,j),(k+1,m)}) &= O_p + (N-k+1)t_{com}, \end{aligned} \tag{9}$$

where O_p is the setup time of p_p performed before sending data, t_{proc} is the processing time for one arithmetic operation, and t_{com} is the communication time per byte. This model was proved to be asymptotic to a real environment[19]. Then, we can rewrite (9) using $t_c(n_{k,j}, \alpha_p)$ and $t_c(c(e_{(k,j),(k+1,m)}), L_{p,q}))$, as follows:

$$\begin{aligned} t_p(n_{k,j}, \alpha_p) &= \frac{(2(N-k)+1)t_{proc}}{\alpha_p}, \\ t_c(c(e_{(k,j),(k+1,m)}), L_{p,q}) &= O_p + (N-k+1)\frac{t_{com}}{L_{p,q}}, \end{aligned} \tag{10}$$

where t_{proc} and t_{comm} are the values obtained by the reference processor. α_p is the processing speed ratio when that of the reference processor is equal to one, and $L_{p,q}$ is the communication bandwidth ratio when the reference communication bandwidth is equal to one.

First, we executed the simulation using the information listed in Table 1.[1] We obtain the mapping of a task and a processor through the results of the simulation. Then, we implemented the parallelized MPI (MPICH2[20]) program with non-blocking communication in the C language by hand-coding each algorithm. For each algorithm, the program was executed 100 times, and we averaged the speed-

[1] O_p depends on many factors, including the program coding style and the processor status at the execution. This makes it difficult to specify O_p for each processor. We fixed the setup time O_p=0 in the simulation.

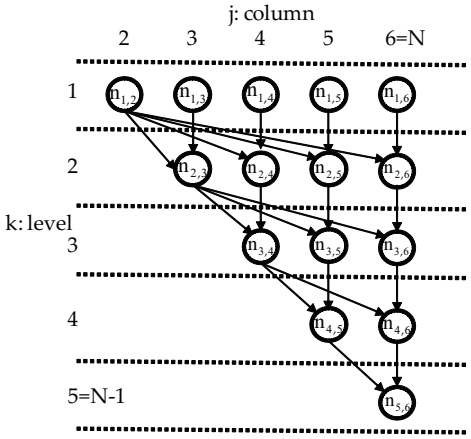


Figure 4: Gaussian Elimination DAG at $N = 6$.

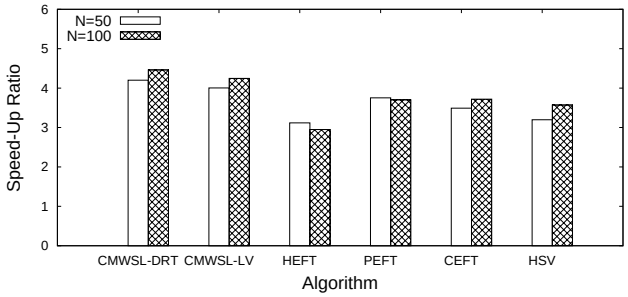


Figure 5: Comparison of Speed-Up Ratio for Gaussian Elimination DAGs in a Real Environment.

up ratio defined by the schedule length of "node 20" in Table 1, divided by that of each algorithm.

Fig. 5 presents the results in cases of $N = 50$ and $N = 100$. CMWSL-DRT and CMWSL-LV outperform the other algorithms in both cases ($N = 50$ and $N = 100$). We can see that PEFT, CEFT, and HSV output a smaller schedule length than HEFT. In the real environment, the setup time is nonzero for each processor, and therefore the communication time can have a greater effect on the schedule length than in the simulation. Even in such a case, both CMWSL-DRT and CMWSL-LV have a smaller schedule length than the other algorithms. As for the comparison between CMWSL-DRT and CMWSL-LV, it is observed that CMWSL-DRT outputs better speed-up ratio. In CMWSL-LV, a task, e.g., n_k by which WSL is obtained if it is scheduled as late as possible, is selected from the free list. However, even if n_k is selected, the actual schedule length cannot be made smaller if n_k is not a critical task, i.e., n_k does not belong to the actually dominating execution path in terms of the schedule length. In CMWSL-DRT, selecting a task having the minimum DRT can lead to the actual schedule length; that is, scheduling a task, e.g., n_k having the minimum possible start time can reduce the data waiting time from n_k. As a result, the total idle time for each processor can be suppressed.

6.3 FFT DAG

In FFT DAG, every task size is equal and every data size is equal. The FFT DAG used in this comparison contains $m \log m$ butterfly operation tasks, where m is the matrix size, as in Section 5.5.2. Each task contains the following operations:

$$\begin{aligned} y(k) &= y_e(k) + \omega^k * y_o(k), 0 \le k \le m/2 - 1, \\ y(k + m/2) &= y_e(k) - \omega^k * y_o(k), 0 \le k \le m/2 - 1, \end{aligned} \tag{11}$$

where y_e is the FFT of even points and y_o is the FFT of odd points. That is, every task involves the same number of operations. Although the data size among all tasks is the same, it contains both the real and imaginary parts as double type variables. Thus, we define each data size as 16 bytes.

Table 2 presents the specification of the environment in terms of processing and the communication performance for FFT DAGs. In this table, the "Task Size" is measured by executing (11) for each processor, while the "Data Size" is defined by multiplying t_{comm} from Table 1 by 16. From this information, α_p and β_p are derived for the simulation. Similar to the case of Gaussian elimination DAGs, we obtain the mapping of each task and each processor from the results of the algorithms in the simulation. Then, we implemented the parallelized MPI program with non-blocking communication in the C language.

Fig. 6 presents the comparison results in cases of $m = 128$ and $m = 256$, in terms of the speed-up ratio. In both cases, both CMWSL-DRT and CMWSL-LV outperform the other algorithms. However, the differences of the speed-up ratio between CMWSL (CMWSL-DRT and CMWSL-LV) and the other algorithms are not large compared with the case of Gaussian elimination, as seen in Fig. 5. As for the comparison between CMWSL-DRT and

Table 2: Specification of the Experimental Environment for FFT DAGs.

Node ♯	CPU Freq.	Task Size (μs)	α_p	Bandwidth	Data Size(μs)	β_p
1	300MHz	37.540	1.00	100Mbps	36.32	1.00
2	533MHz	17.114	2.19	1Gbps	7.2	5.04
3	800MHz	13.736	2.73	100Mbps	34.4	1.06
4	1.0GHz	8.438	4.45	1Gbps	9.92	3.66
5	1.2GHz	7.757	4.84	1Gbps	5.28	6.88
6	1.5Ghz	6.672	5.63	100Mbps	32.48	1.12
7	1.5Ghz	6.231	6.02	1Gbps	8.8	4.13
8	2.0GHz	4.003	9.38	1Gbps	5.44	6.68
9	2.4GHz	3.822	9.82	100Mbps	6.08	5.97
10	2.4GHz	3.539	10.61	1Gbps	4.32	8.41
11	2.6GHz	3.221	11.65	1Gbps	6.72	5.40
12	2.6GHz	3.183	11.79	100Mbps	35.2	1.03
13	2.8GHz	3.105	12.09	1Gbps	3.52	10.32
14	2.8GHz	2.994	12.54	1Gbps	5.92	6.14
15	2.8GHz	2.944	12.75	100Mbps	30.24	1.20
16	2.8GHz	2.919	12.86	1Gbps	4.48	8.11
17	3.0GHz	2.896	12.96	1Gbps	4.64	7.83
18	3.0GHz	2.863	13.11	1Gbps	6.24	5.82
19	3.0GHz	2.732	13.74	100Mbps	33.76	1.08
20	3.2GHz	2.556	14.69	1Gbps	3.36	10.81

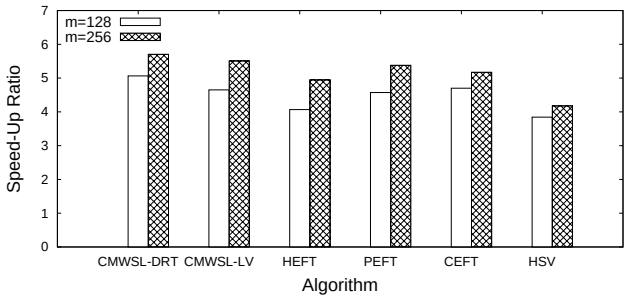


Figure 6: Comparison of Speed-Up Ratio for FFT DAGs in a Real Environment.

CMWSL-LV, CMWSL-DRT outputs better results in both cases ($m = 128, 256$). Similar to the result obtained in Section 6.2, selecting a task having the minimum possible start time can contribute to reduce the data waiting time for each processor. In general, the data size among tasks in FFT DAG is smaller than that in Gaussian Elimination DAG. This means that the effect of the communication delay is smaller in case of FFT DAG. Thus, the difference of schedule length among algorithms is also small.

From these results, we can infer that CMWSL can be applied to the FFT in a real environment.

7 Conclusion and Future Works

In this paper, we confirmed the performance and behaviors of CMWSL in a real environment using Gaussian Elimination and FFT program. According to the obtained results, CMWSL outperforms other list-based task scheduling algorithms in terms of the speed-up ratio, i.e., the schedule length. We also presented the comparison results among two task ordering methods, i.e., List_minDRT and List_maxLV for CMWSL. In CMWSL, List_minDRT was proved to output a better result than List_maxLV. That is, we can say that the task having the minimum possible start time should be selected for scheduling.

References

[1] I. Foster, "The Grid: A new infrastructure for 21st century science," *Physics Today*, Vol.55(2), pp. 42–47, 2002.

[2] S. Di and C.L. Wang, "Dynamic Optimization of Multiattribute Resource Allocation in Self-Organizing Clouds," *IEEE Trans. Parallel Distrib. Syst.,* Vol. 24, No. 3, pp. 464-478, 2013.

[3] H. Topcuoglu, S. Hariri, M. Y. Wu, "Performance-Effective and Low-Complexity Task Scheduling for Heterogeneous Computing," *IEEE Trans. Parallel Distrib. Syst.*, Vol. 13, No. 3 pp. 260-274, 2002.

[4] H. Arabnejad and J. G. Barbosa, "List Scheduling Algorithm for Heterogeneous Systems by an Optimistic Cost Table,", *IEEE Trans. Parallel Distrib. Syst.*, Vol. 25, No. 3, pp. 682-694, 2014.

[5] M. A. Khan, "Scheduling for heterogeneous systems using constrained critical paths, " *Parallel Comput.*, Vol. 38, No. 4-5, pp. 175-193, 2012.

[6] M. I. Daoud and N. Kharma, "A high performance algorithm for static task scheduling in heterogeneous distributed computing systems," *J. Parallel Distrib. Comput.*, Vol. 68, No. 4, pp. 399–409, 2008.

[7] A. Gerasoulis and T. Yang, "A Comparison of Clustering Heuristics for Scheduling Directed Acyclic Graphs on Multiprocessors," *J. Parallel Distrib. Comput.*, Vol. 16, pp. 276-291, 1992.

[8] B. Jedari and M. Dehghan, "Efficient DAG Scheduling with Resource-Aware Clustering for Heterogeneous Systems," *Comp. Inf. Sci.*, Vol. 208, pp. 249–261, 2009.

[9] S. Chingchit et al., "A Flexible Clustering and Scheduling Scheme for Efficient Parallel Computation," *in Proc. of the 13th Int. Parallel Distrib. Process. Symp.*, pp. 500–505, 1999.

[10] C. Boeres et al., "A Cluster-based Strategy for Scheduling Task on Heterogeneous Processors," *Proc. of the 16th Symp. on Comp. Arch. High Perform. Comput. (SBAC-PAD'04)*, pp. 214–221, 2004.

[11] B. Cirou and E. Jeannot, "Triplet: a Clustering Scheduling Algorithm for Heterogeneous Systems," *in Proc. Int. Conf. on Parallel Process. Workshops (ICPPW'01)*, pp. 231–236, 2001.

[12] H. Kanemitsu , M. Hanada, and H. Nakazato, "Clustering-based task scheduling in a large number of heterogeneous processors," *IEEE Trans. Parallel Distrib. Syst.*, Vol. 27, No. 11, pp. 3144–3157, 2016.

[13] Hidehiro Kanemitsu, Gilhyon Lee, Hidenori Nakazato, Takashige Hoshiai, and Yoshiyori Urano, "On the Effect of Applying the Task Clustering for Identical Processor Utilization to Heterogeneous Systems," *Grid Comput.*, pp. 29 - 46, 2012.

[14] T. Yang and A. Gerasoulis., "List scheduling with and without communication delays," *Parallel Comput.,* pp. 1321 – 1344, 1993.

[15] S. Hashimoto, E. H. Weyulu, K. Hajikano, H. Kanemitsu, and M. W. Kim, "Evaluation of Task Clustering Algorithm by FFT for Heterogeneous Distributed System", *in Proc. the 19th Int. Conf. Adv. Comm. Tech.(ICACT2017)*, pp. 94-97, 2017.

[16] D. Sirisha and G. V. Kumari, "A new heuristic for minimizing schedule length in heterogeneous computing systems," *in Proc. IEEE Int. Conf. on Elect. Comput. and Comm. Tech. (ICECCT)*, pp. 1–7, 2015.

[17] G. Xie et al., "Heterogeneity-driven end-to-end synchronized scheduling for precedence constrained tasks and messages on networked embedded systems, " *J. Parallel Distrib. Comput.*, Vo. 83, pp. 1–12, 2015.

[18] H. Kanemitsu and M. Hanada, "A Task Scheduling with Response Time Estimation for Efficient Processing in Heterogeneous Systems," *in Proc. of 2015 RISP Int. Work. Nonlinear Circuits, Comm. Sig. Process. (NCSP 2015)*, 2015.

[19] A. K. Amoura et al., "Scheduling Algorithms for Parallel Gaussian Elimination with Communication Costs," *IEEE Trans. Parallel Distrib. Syst.*, Vol. 9, No. 7, pp. 679–686, July 1998.

[20] MPICH web site: https://www.mpich.org/ .

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1549-1555 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Modeling and Control of Collaborative Robot System using Haptic Feedback

Pawar Mansi Shailendrasingh[*,1], Vivekananda Shanmuganatha[1], Lad Pranav Pratap[2], Pawar Mansi Shailendrasingh[3]

[1] *Vellore Institute of Technology, School of Mechanical Engineering, VIT University, 632014, India*

[2] *Vellore Institute of Technology, Production and Industrial Engineering, VIT University, 632014, India*

[3] *Vellore Institute of Technology, Electronics and Instrumentation Engineering, VIT University, 632014, India*

ARTICLE INFO

Article history:
Received: 06 April, 2017
Accepted: 03 May, 2017
Online: 21 August, 2017

Keywords:
Multi Robot Systems
Collaborative Robot Systems
Kinematics
Dexterous manipulation
Haptics
Telerobotics
Human-Machine Interaction

ABSTRACT

When two robot systems can share understanding using any agreed knowledge, within the constraints of the system's communication protocol, the approach may lead to a common improvement. This has persuaded numerous new research inquiries in human-robot collaboration. We have built up a framework prepared to do independent following and performing table-best protest object manipulation with humans and we have actualized two different activity models to trigger robot activities. The idea here is to explore collaborative systems and to build up a plan for them to work in a collaborative environment which has many benefits to a single more complex system. In the paper, two robots that cooperate among themselves are constructed. The participation linking the two robotic arms, the torque required and parameters are analyzed. Thus the purpose of this paper is to demonstrate a modular robot system which can serve as a base on aspects of robotics in collaborative robots using haptics.

1. Introduction

The Perception of robot motions is really important. In recent years, there has been in increasing interest in distributed robotic systems. In such a system, a task is not completed by a single robot but instead by a team of collaborating robots. Team members may exchange sensor information, may help each other to scale obstacles, or may collaborate to manipulate heavy objects. A team of robots has distinct advantages over single robots with respect to actuation as well as sensing. When manipulating or carrying large objects, the load can be distributed over several robots so that each robot can be built much smaller, lighter, and less expensive. As for sensing, a team of robots can perceive its environment from multiple disparate viewpoints. A single robot, on the other hand, can only sense its environment from a single viewpoint, even when it is equipped with a large array of different sensing modalities. There are many tasks for which distributed viewpoints are advantageous: surveillance, monitoring, demining, plume detection, etc.

The word "cobot". "Co" stands for "collaborative". We can find it in "co-working", "co-design", "co-development", and the list goes on. Work together, find better solutions together, and produce faster together. Here's the idea. Together. With new technologies coming up fast and better ways of communication, people and companies start to handle resources and skills in a better way. Make people work in open spaces, encourage them to interact with different background and culture individuals, try new flexible organization processes, are parts of this new collaborative mind. The idea is simple: work smarter together for better results. Cobots can have many roles — from autonomous robots capable of working together with humans in an office environment that can ask you for help, to industrial robots having their protective guards removed as they can react to a human presence under EN ISO 10218 or RSA BSR/T15.1.

Comparing the motions of a robot is very censorious activity. Elaborating some motions may be arduous as the operator should manage the overall motion; for example, these questions include: " The robots which are designed are with respect to the environment. Will the robot work with the same efficiency?", "Is

[*]Correponding author: Pawar Mansi Shailendrasingh, C-613, VIT University, Vellore,Tamil Nadu, India-632014, +91 9629773352, mansipawar29@gmail.com

the operator getting the same outcome?", "Are the motions of the robot constrained to specified trajectories?", "Will both the geometric specifications of the path as well as the velocity at which the robot traverses the path be followed?, "What are the control issues in the robotic control?" or "What are the fundamental problems in mobile robotics?"

When two robot systems can share understanding using any agreed knowledge, within the constraints of the system's communication protocol, the approach may lead to a common improvement. This has persuaded numerous new research inquiries in human-robot collaboration. We have build up a framework prepared to do independently following and performing table-best protest object manipulation with humans and we actualize two different activity models to trigger robot activities. The idea here is to explore collaborative systems and to build up a plan for them to work in a collaborative environment which has many benefits to a single more complex system. In the paper, two robots that cooperate among themselves are constructed. The participation linking the two robotic arms, the torque required and parameters are analyzed. Thus the purpose of this paper is to demonstrate a modular robot system which can serve as a base on aspects of robotics in collaborative robots using haptics. The primary aim is to design an actively articulated robotic arm and dynamically control its orientations i.e. yaw, pitch and roll using negative feedback to assure stability or the desired orientation at all times.

The collaborative robots seen so far are based on hard coding and are run with respect to each other. This barrier may create some issue during the interaction of the robot like if the time gives to robotic arm is different than the other after several cycles the robot will stop working in a synchronize mode and could create a problem. We are planning to an interactive platform that could create not only human-robot interaction as well as robot-robot interaction. This could reduce the barrier and could work in synchronized mode. Even we are able to make the system open source that it could be used by the college student and work on the background rather working on the fabrication part. We are using detailed simulation software only for the simulation, so the student could work on the developed CAD model to test the respective operation without direct using the provided hardware. This could reduce the risk of damage to the fabricated hardware.

2. Literature Review

In the previous paper [1], we had introduced how autonomous robots would benefit from improvements in haptic intelligence and overview of telerobotics systems. In the real world, we want robots to help to perform tasks in the remote environment. We have demonstrated the design and manufacture of a touch robotic arm that is inspired by a motive to increase efficiency. A multi-robot system can investigate an unknown domain speedier than a single robot system, even with the imperative of performing meeting to permit correspondence. In [2], the author has demonstrated the issue of how two heterogeneous robots can mastermind to meet in an obscure domain from obscure beginning areas. In [3], the author has ideated another approach for two-sided control in delta robot which has been proposed so as to lessen gravity impact [4]. In [5], [6], the author portrays the advancement of automated innovations that empower another vision of shared versatile assembling by groups of robots. In [7], the author. concentrates on the improvement and combination of the sensor arm used to convey a spectrometer from a multi-meter scale robot to centimeter scale rocks. The sensor arm joins off the rack equipment for movement control, activation, and detection. In [8], the author has tended to the inquiries of whether and when a robot ought to step up with regards to joint human-robot undertaking execution by contrasting three activity models with trigger robot activities: human-started help, robot-started receptive help, and robot started proactive offer assistance. In [8], [9], the author has clarified a review on participation between two master-slave modular robots in the paper. Kinematics of the system is contemplated. A tri-dimensional estimation arrangement of Optotrak 3020 is utilized to input the end position of the two robots. He tended to [9] the undertaking of organizing the movements of different controllers when either their directions or their ways are given. By exploiting a principal time scaling law for controllers in light of their elements, we distinguish adequate conditions for crash free coordination of the robots when the velocity profiles can be consistently time-scaled and the robot begins times can differ. [10] The demonstration of the transportation of numerous things to stations that make ask. The collaborative collection technique can outperform the individual accumulation strategy with up to 15.4% diminishment in maximum traveling cost. At that point [11] proposed scientific models for these surmisings in view of streamlining cost, attracting the similarity to the guideline of objective activity. In [12] and [13], the respective authors presented a robot control system which has been particularly intended for a subjective robot which offers space and undertaking with a human. The approach is productive in light of successful individual and collective abilities. In [14], the author looked into tactile sensing applications incapable robot hand control. The fundamental sorts of material sensors and their reconciliation with robot hands are discussed. Tactile data processing techniques and its applications are displayed.

3. Design

3.1. Specifications

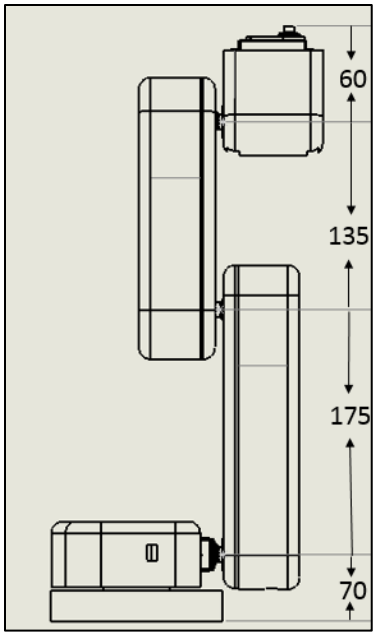


Figure 1: Robotic Arm Structure (dimensions in mm)

Table 1: Specifications of Robotic Arm

Sr. No.	*Parameters*	*Specification*
1	DOF	5 axes
2	L1	175mm
3	L2	135mm
4	L3	60mm
5	Height	440mm
6	Joint-1	180 degrees
7	Joint-2	180 degrees
8	Joint-3	180 degrees
9	Joint-4	180 degrees
10	Joint-5	180 degrees

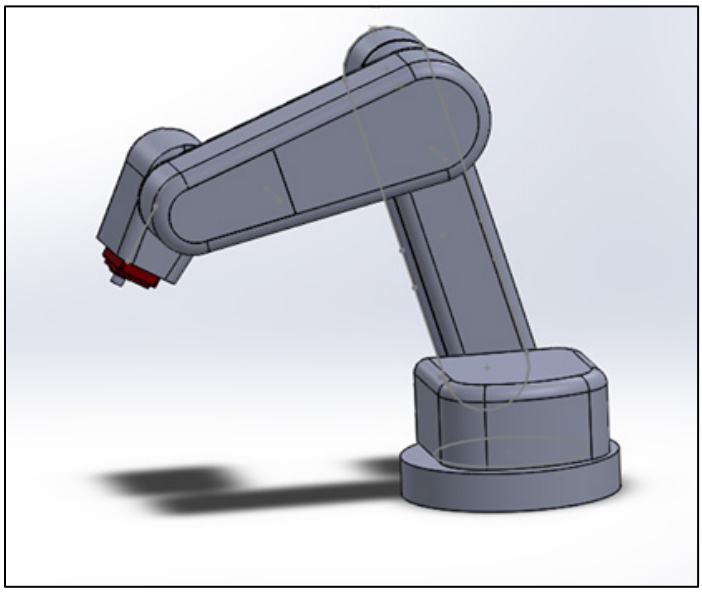


Figure 2: The redefined robot imported from the 3 D model into Solidworks model.

The aim was to build a collaborative robot system that could work with a single control unit. The primary aim was to develop a single system that could help the robot to communicate with each other and work together. There are five degrees of freedom (DOF) robotic arm. The entire robotic system was designed in Solidworks and manufactured using 3D printing process. The goal was to make the control system which could be easily operated and understood by the operator. A robot with the same dimension that is going to be manufactured is imported in the V-REP software. In this method, the robot is made to follow defined paths that are plotted using path planning function. The robot is programmed to follow the plotted path. Inverse kinematics helps the robot to follow the specified path. Individual joints calibrate their own angle and the end effector tries to reach the defined path. While following the path the robot joint angle values differ from joint to joint. Absolute joint values are obtained in V-REP during operation and even the XYZ values of the end effectors can be generated.

Figure 3: 3D printed Hardware model

4. Math

4.1. DH Parameters

The 5-DOF manipulator kinematic parameters are derived using Denavit Hartemberg formulation:

```
                % theta   d     a    alpha
L (1) = Link ([θ   70  -22.2  -pi/2     0]);
L (2) = Link ([θ  -70.0  175    pi      0]);
L (3) = Link ([θ  -4.0   135    0       0]);
L (4) = Link ([θ   0.1    60    pi/2    0]);
L (5) = Link ([θ   0.80    0    pi      0]);
L (6) = Link ([gripper    0]);
```

All measurements are in centimeters.

```
q0 =[0 -pi/3 -pi/3 pi/2 0 0 ];
```

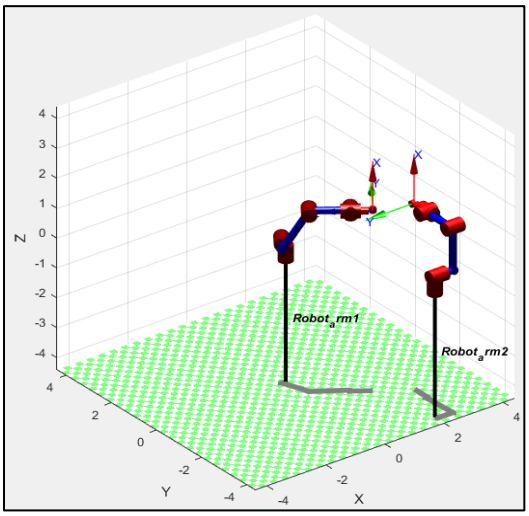


Figure 4: Sequential Robot movements of the Collaborative robotic arms for the required path.

The workspace of the end-effector relies on upon the physical furthest reaches of the five joints point, i.e. on the off chance that the point scope of the robotic arm joints is changed the workspace of the robot changes. So it is critical to consider the precision in deciding the edge scope of the robot joints to get the required workspace which covers the working range.

The general form of the Homogeneous Transformation Matrix is:

[cos(θ) -cos(a)*sin(θ) sin(α)*sin(θ) A*cos(θ)

sin(θ) cos(α)*cos(θ) -sin(α)*cos(θ) A*sin(θ)

0 sin(α) cos(α) d

0 0 0 1]

A1 = [1.0000 0 0 -22.2000

0 0.0000 1.0000 0

0 -1.0000 0.0000 70.0000

0 0 0 1.0000]

A2 = [0.5000 -0.8660 -0.0000 87.5000

-0.8660 -0.5000 -0.0000 -151.5544

0 0.0000 -1.0000 -70.0000

0 0 0 1.0000]

A3 = [0.5000 0.8660 0 67.5000

-0.8660 0.5000 0 -116.9134

0 0 1.0000 -4.0000

0 0 0 1.0000]

A4 = [0.0000 -0.0000 1.0000 0.0000

1.0000 0.0000 -0.0000 60.0000

0 1.0000 0.0000 0.1000

0 0 0 1.0000]

A5 =[1.0000 0 0 0

0 -1.0000 -0.0000 0

0 0.0000 -1.0000 0.8000

0 0 0 1.0000]

The link transformations matrix can be given as:

$^{0}T_{5}$ = A1*A2*A3*A4*A5

$^{0}T_{5}$= [0.0000 0.0000 -1.0000 201.0970

0 1.0000 0.0000 -66.1000

1.0000 -0.0000 0.0000 281.5501

0 0 0 1.0000]

4.2. Working

The flowchart representing the sequence of generating the CODE of the link transformations matrix is shown in Figure 5.

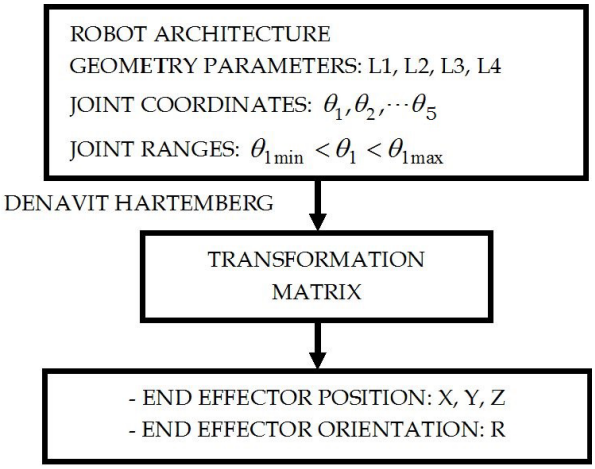


Figure 5: The flow chart of link transformation matrix.

There are four distinct strategies which are derived from figuring the direction arranging of the robot utilizing the code. The direction planning in the four techniques requires two parameters, specifically the joint point run and the last time required to finish the procedure. The first was chosen from the workspace expected to finish the procedure. For any situation, this subject to the sort of surgical applications and can without much of a stretch be distinguished. The final time needed is a vital parameter to determine the trajectory of the robotic arm. As every one of the outcomes that have been construed from the introduction, acceleration speed and torque of the robotic arm depend on this parameter. In the event that the time expands the speed needed declines and increasing speed i.e. the inertia of the arm connect therefore diminishes unintentionally regardless of the techniques utilized. The introduction must be seen precisely inspected and the conduct ought to increment step by step with the time from the underlying edge to the final point. The final time needed to finish the procedure for every technique ought to be improved and this is referred to as optimal planning. In this study, the trajectory planning has been derived and after that, correlation of the outcomes a particular joint is made. The best possible time which gives tasteful outcomes is obtained.

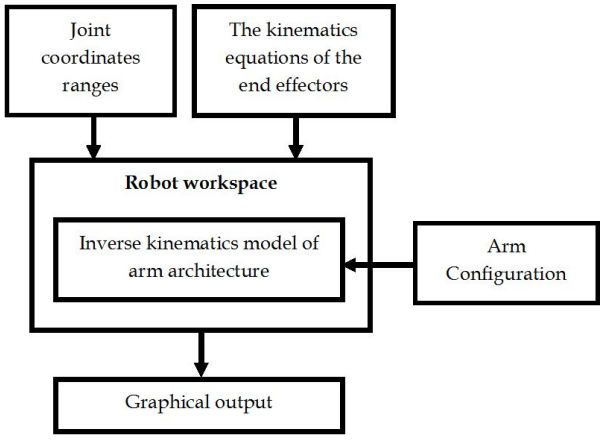


Figure 6: The flowchart for Robot workspace.

4.3. Electronics

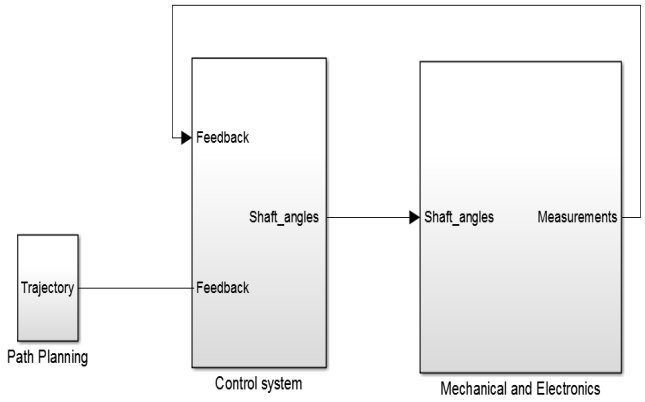


Figure 7: Simulink model of the Robot System.

These Simulink models are created in MATLAB using the robotics toolbox. The mechanical and electronic system contains the hardware as well as the electronics part. In this, the electronic components are the motors. The amount of rotation is given by the control system with the means of shaft angle. This shaft angle is provided by the robotics toolbox blocks. The motor tries to achieve the given shaft angle and provides the feedback to the control system. Then the control system provides the proper angle again reducing the error. The path planning is done and as shown in Figure 7 and the robot tries to move with respect to it. These types of feedback make the robot to easily achieve the respective position.

Figure 8: Coordination of Robotic arm in V-REP obatined from importing simulink model.

There are two lines appeared in the Figure 9. The blue one informs the movement of the end effector joint for the Robot arm 1 and the other line notifies the movement of Robot arm 2 regarding the z-axis. The movement of both the end effectors as for z-axis and time. The simulation done was the sample study of pick and place for corporative manipulators. As shown in the graph the robot one end effector moves with respect to z-axis move down practically to ground level. The robot tries to pick the object and remains there for 20 seconds. The robot picks the object and move to the upper level and move linearly in the XY plane. The robot achieves a specific position and the other robot arm which was very still going to a same position of the Robot arm1 yet the other robot is at a lower level to get the object and proceed with the cycle. As the robot arm1 drops the question from the gripper the haptic feedback recognizes the other robot and makes the other robot to get the object as the Robot arm 2 end effector is at a similar z-axis. The Robot arm 2 completes the movement. In this system, one robot arm communicates with the other robot arm with the sensing of the gripped object in the end effector as the object from the gripper tries to slip the other robot gripper tries to hold the object.

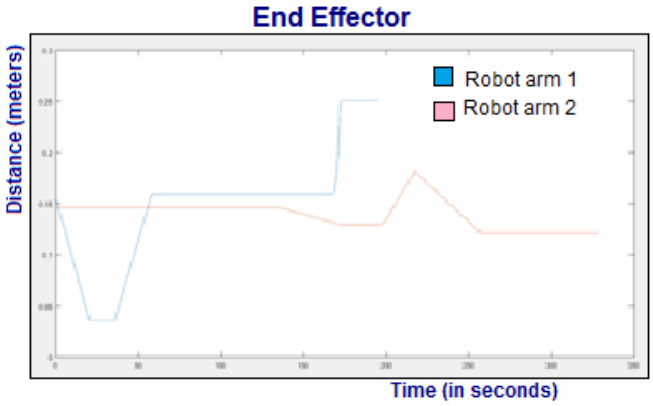


Figure 9: Position control of end effector with respect to z-axis. X axis is time in seconds and Y axis is distance in meters.

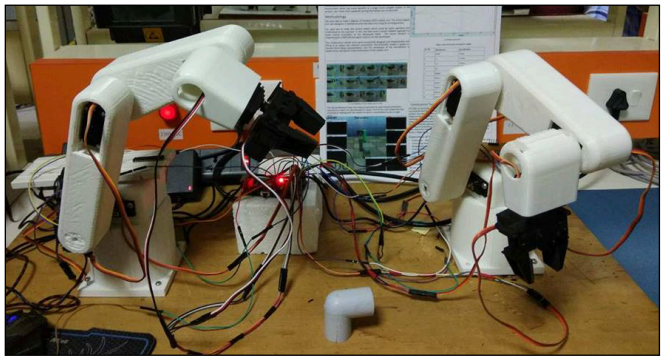


Figure 10: Hardware model of Robot System.

In order to conclude about the project execution, the project was carried out successfully and the desired output was obtained. Talking specifically about the project, the development of two robotic arms was successfully done and our main objective about making a single brain that could command both the arm was developed. Both the arms receive a command decently and behave as a single system.

4.4. Force Sensitive Resistor for Haptic Feedback

A force-sensing resistor is a material whose resistance changes when a force or pressure is applied. Figure 10 shows a demonstration of how to construct an Arduino-based pressure sensor that displays the measured pressure on a small LCD screen. With a minor adjustment to the code, one can just read the FSR values in the Arduino serial window. In that case, none of the LCD/potentiometer stuff is needed. Force Sensing Resistors (FSR) is a polymer thick film (PTF) device which exhibits a decrease in resistance with an increase in the force applied to the active surface. Its force sensitivity is optimized for use in human touch control of electronic devices. With a minor adjustment to the code, one can just read the FSR values in the Arduino serial window. In

that case, none of the LCD/potentiometers is needed. The Force Sensitive Resistor (FSR) with a sensing area of 0.5" diameter is attached to the gripper. The sensor uses a voltage divider. The FSRs get a 10k resistor, and the flex resistor gets a 22k resistor. The voltage divider works with the changing resistance of the sensors. When the resistance is higher for the variable resistor, there is a larger voltage drop across the variable resistor. The voltage dividers are supplied with +5V, and that drop in 5V is split between the two resistors. The Arduino measures the voltage between the resistors and returns value between 0 and 1023 based on the voltage reading (0 being 0V and 1023 being 5V).

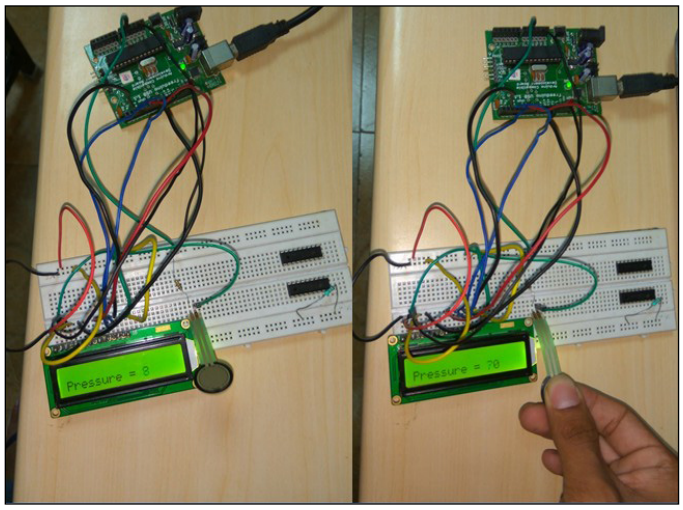


Figure 11: Display variance in pressure.

Figure 11 shows the interface of FSR with MATLAB. This figure demonstrates the calibration of a Force Sensitive Resistor (FSR) using the Arduino UNO board and MATLAB. Additionally, the UNO board logs real-time FSR data. The input and output data is as follows:

Force: [133g 175g 185g 250g 490g]

Voltage 1= [0.3955 1.0693 0.69 1.2793 1.330]

Voltage 2= [0.48 0.87 1.82 1.29 1.48]

Voltage 3= [0.65 0.94 1.015 1.15 1.41]

Average voltage= [0.5085 0.9597 1.048 1.406]

Therefore, plot between Force and Voltage:

Force_Xaxis= [133,175,183,250,490]

Voltage_Yaxis= [0.5085,0.9597,1.048,1.2397,1.406]

The Linear model Polynomial equation:

J(x) =Ax3+Bx2+Cx+D where, A=1.562x 10-7

B=-0.000152

C=0.04677

D=-3.391

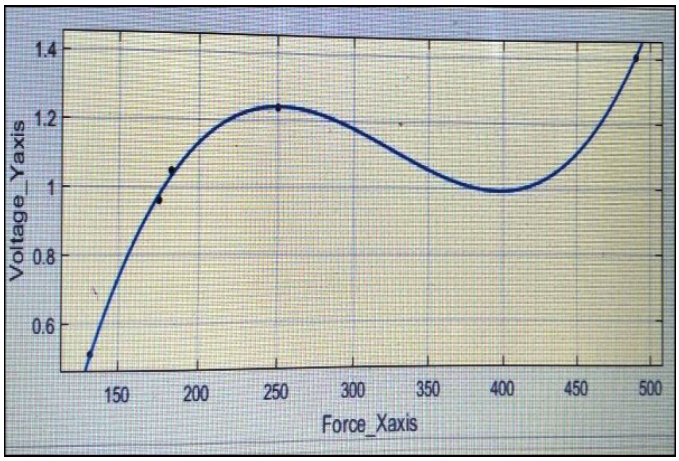


Figure 12: Interface of FSR with MATLAB

5. Results and Discussion

Through the processes of conceptualization, design, assembly, and final coding, we were required to complete tasks ranging from needs assessment, static and dynamic load analyses, design for manufacturing, machining, troubleshooting, and finally microcontroller coding, to name but a few. Therefore following the earlier paper we studied human-machine interaction and works in a similar way as the body goes in specific direction or manner and shows sensation at the same time. The collaborative robotic arms were successfully designed and implemented and lifting of an object was attained successfully. The kinematic model is based on Denavit-Hartenberg representation and the workspace of the end-effector is explained by solving the inverse kinematics of the robotic arm. It plainly appeared in this work the best technique for trajectory planning that gives the smooth set trajectory planning and best execution of the robotic arms under examination was the delicate movement trajectory planning on the grounds that the most vital explanation behind this choice was the torque history that has the least number ever of shooting and the shooting was dispersed frequently over the timeframe, unlike alternate strategies which have a long number of shootings and were circulated arbitrarily.

The collaborative robotic arms are capable of handling various varieties of applications in mining, agriculture, and defense. It is a platform provision for various autonomous assignments in agriculture comprising cultivation, chemical spraying, and mowing. Because of its robustness, in the mining sector, the vehicle is operated to survey and map mines. As for military use, it can transport autonomously materials to support the soldiers. This paper also accentuates the synergy between the human operator and a robot and has various applications in healthcare, medicines, entertainment, education, graphic arts, the industry as well as space. A high level of precision was observed in the movement of the robotic arm due to the generation of accurate control signals for the servo motors. In collaboration with the visual display, haptics can be exploited by people for tasks involving hand-eye coordination, essentially Space-ship maneuvering, and Robotic Tele- Surgery. It can also be used for games in which you feel and at the same time see your Interactions with images. The information regarding the different touch forces supports the approximation of the contact accomplished by the

sensor mounted on the gripper and permit the experiencing the slipping of an object.

6. Future Work

The machine learning calculations enhance and different advances, for example, neuromorphic chips, empower more brilliant robots, we'll be seeing them do tasks that no one but people can do until now, with the exception of, by and large, they will have the capacity to improve. They will have the capacity to get to incomprehensible databases, perceive individuals, machines and parts, have the capacity to determine bits of knowledge and follow up on them. That implies that robots will soon have the capacity to do short work, as well as analyze issues and devise arrangements. They'll have the capacity to do schedule, highly monotonous work like the quality affirmation, assessing parts and completed items. Likewise, the cost of all that capacity will probably descend significantly as innovation advances and the cost of parts decreases.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The authors would like to thank Dr. Vivekananda Shanmuganathan for his long-standing support of this work. We would also like to thank the Robotics Laboratory, of School of Mechanical Engineering and Vellore Institute of Technology University for providing scientific guidance, field expertise and excellent logistical support.

References

[1] Mansi Pawar, Pranav Lad,"Development of robotic arm using haptic feedback", IEEE International Symposium on Robotics and Manufacturing Automation(ROMA), Year 2016, Pages 1-7. DOI: 10.1109/ROMA.2016.7847815.

[2] Nicholas Roy, Gregory Dudek, "Collaborative Robot Exploration and Rendezvous: Algorithms, Performance Bounds and Observations", Autonomous Robots, September 2001, Volume 11, Issue 2, pp 117–136.

[3] Chowarait Mitsantisuk, Kiyoshi Ohishi, "Haptic human-robot collaboration system based on delta robot with gravity compensation", IECON 2016 - 42nd Annual Conference of the IEEE Industrial Electronics Society, Year: 2016, Pages: 5796 – 5801.

[4] Glenn Wagner , Minsu Kang, and Howie Choset, "Probabilistic path planning for multiple robots with subdimensional expansion", IEEE International Conference on Robotics and Automation (ICRA), , May, 2012, pp. 2886--2892.

[5] David Bourne, Howie Choset, Humphrey Hu, George A. Kantor, Christopher Niessl, Zack Rubinstein, Reid Simmons, and Stephen Smith, "Mobile Manufacturing of Large Structures", Proceedings 2015 IEEE International Conference on Robotics and Automation, May, 2015.

[6] Christopher Urmson, Benjamin Shamah, James Teza, Michael D. Wagner , Dimitrios (Dimi) Apostolopoulos , and William (Red) L. Whittaker," A Sensor Arm for Robotic Antarctic Meteorite Search", Proceedings of the 3rd International Conference on Field and Service Robotics, July, 2001.

[7] Jimmy Baraglia; Maya Cakmak; Yukie Nagai; Rajesh Rao; Minoru Asada,"Initiative in robot assistance during collaborative task execution", 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Year: 2016, Pages: 67 – 74.

[8] Liying Su, Lei Shi, Yueqing Yu, Qixiao Xia, "Bolt and Screw Assemblage through Collaborative Kinematics Operation of Two Modular Robots Based on the Position Feedback", Proceedings of the 2009 IEEE International Conference on Information and Automation, Pages: 1574 – 1579.

[9] S. Akella; S. Hutchinson, "Coordinating the motions of multiple robots with specified trajectories". Proceedings 2002 IEEE International Conference on Robotics and Automation (Cat. No.02CH37292), Year: 2002, Volume: 1, Pages: 624 - 631 vol.1.

[10] S. Akella; Jufeng Peng, "Time-scaled coordination of multiple manipulation", Robotics and Automation, 2004. Proceedings. ICRA '04. 2004 IEEE International Conference on Year: 2004, Volume: 4, Pages: 3337 – 3344.

[11] Nantawat Pinkam; François Bonnet; Nak Young Chong, "Robot collaboration in warehouse", 2016 16th International Conference on Control, Automation and Systems (ICCAS), Year: 2016, Pages: 269 – 272.

[12] Anca D. Dragan; Kenton C. T. Lee; Siddhartha S. Srinivasa, "Legibility and predictability of robot motion", 2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Year: 2013, Pages: 301 – 308.

[13] Felix Jimenez; Masayoshi Kanoh; Tomohiro Yoshikawa; Takeshi Furuhashi, "Effect of collaborative learning with robot that prompts constructive interaction", 2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Year: 2014, Pages: 2983 – 2988.

[14] Rachid Alami, "On human models for collaborative robots",2013 International Conference on Collaboration Technologies and Systems (CTS), Year: 2013, Pages: 191 – 194.

[15] Zhanat Kappassova, Juan-Antonio Corralesb, Véronique Perdereaua, "Tactile sensing in dexterous robot hands — Review", Volume 74, Part A, December 2015, Pages 195–220.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1556-1564 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Lifelong Learning Application: Mobile Application to Promote Lifelong Learning and Introduce Educational Institutions

Wongpanya Nuankaew[*,1], Pratya Nuankaew[2], Kanakarn Phanniphong[3]

[1]*Faculty of Information Technology, Rajabhat Mahasarakham University, 44000, Thailand*

[2]*School of Information and Communication Technology, University of Phayao, 56000, Thailand*

[3]*Faculty of Business Administration and Information Technology, Rajamangala University of Technology Tawan-Ok, 20110, Thailand*

ARTICLE INFO

Article history:
Received: 28 May, 2017
Accepted: 29 July, 2017
Online: 21 August, 2017

Keywords:
Lifelong Learning Application,
Datamining in Education,
Students' Academic Achievement,
Education Engineering Systems

ABSTRACT

Education is the raising of the thinking process and developing the mind in a broader perspective. However, education that is comprehensive and appropriate to the individual according to its potential, ability and skill is too hardest to describe. Thus, matching and referral institutions, including education model that suits the person's behavior has been receiving attention. The purpose of this research is to study and develop the applications that support the development of the learner's education process with the highest degree of achievement. The material and method are consisting of five main processes, including 1) requirement definition, 2) system and software design, 3) implementation and unit testing, 4) integration and system testing, and 5) operation and maintenance. The data collection were collected from 431 samples, which divided into 4 targets; 186 high school students at Phadungnaree School, Maha Sarakham province, 191 university students at Rajabhat Mahasarakham University, 12 head of institutions at Phadungnaree School and Rajabhat Mahasarakham University, and 42 secondary teachers and university lecturers at Phadungnaree School and Rajabhat Mahasarakham University. According to the results of research, it is indicated that the total satisfaction toward the application are highest, which is equal to 4.32, and the groups with the highest levels of satisfaction are university students, which is equal to 4.43. It can be concluded that the application is appropriate and it is advisable for students who wish to study at university.

1. Introduction

Nowadays, scientific and technological progress than in the past, it seems that everything around them with the facilities and the living is easy. Technology can provide support and resolve problems in every field of knowledge. Transferring of technology and know-how is fast and comprehensive for various purposes and reasons.

Even so, justice and equality in accessing development and knowledge of the individual remain questions. According to the Sustainable Development of the United Nations (UN), it mentioned that the need for "A world with equitable and universal access to quality education at all levels, to health care and social protection, where physical, mental and social well-being is assured [1]". Many questions arise from this mission, such as how to achieve the goals, what are the main hurdles, where to solve the problem, who to deal with, and many more.

One of the ways to achieve this, the international community has focused on health, education and social reforms involving individuals, systems and information enormously. The 2015 World Education Forum held at Incheon, South Korea, been shown that the future education vision is fully taken up with an offer of equal and fair quality to educational institutions as well as the promotion of lifelong learning opportunities for all learners [2].

[*]Correponding author: Wongpanya Nuankaew, Rajabhat Mahasarakham University, Thailand | Email: wongpanya.nu@rmu.ac.th

These two ideas discuss the relationship with Goal 4 of the United Nations Sustainable Development Goal, which is quality education and ensures that quality in education is equal and promotes equality. Everybody's lifelong learning opportunity [3, 4]. Based on the importance of the above, the research team studied the components of the recommendations for selecting the right university for the right student, and developed it as an application on the device, which is the intention of the international community, as shown in Figure 1.

Figure 1. The importance of the research

Figure 1 illustrates the importance and relevance of research development to the goals of the international community, which is a complex problem when attempting to match the most appropriate student to the institution. Moreover, this research is an extension of Nuankaew's research [5], which reflects the research problem as shown in Figure 2 and Figure 3.

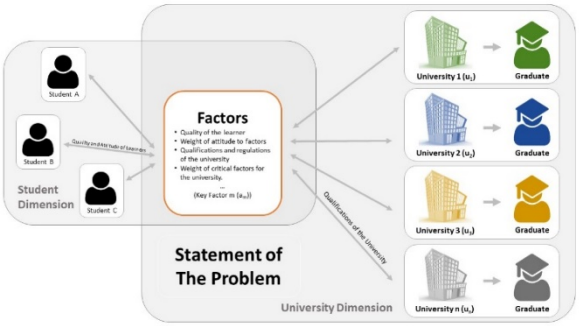


Figure 2. The Research Problem [5]

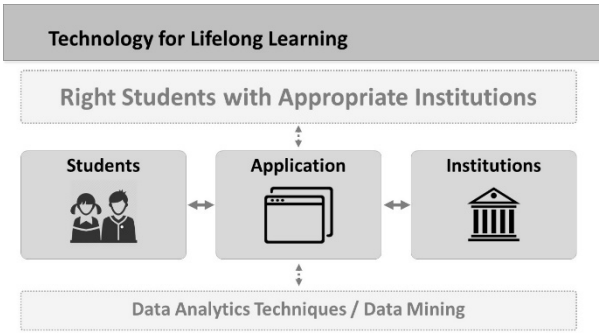


Figure 3. Using Modern Technology for Lifelong Learning

Once the research is done, the final step before publishing the application to the public is known and applied, the team has evaluated the performance and satisfaction of the application, which consists of five parts: 1) requirement definition, 2) system and software design, 3) implementation and unit testing, 4) integration and system testing, and 5) operation and maintenance as demonstrated in Figure 4. The details of each topic are clearly described in the material and methods sections.

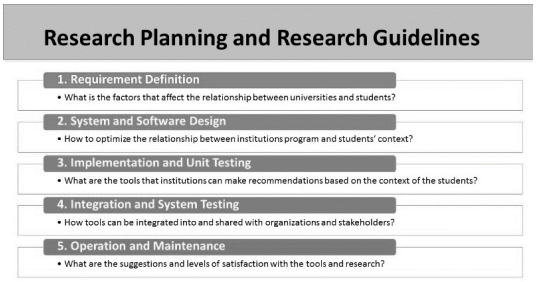


Figure 4. Research Planning and Research Guidelines

To complete the research work, the structure of the document is as follows: Literary review is compiled and presented in Section II. Section III defines and describes the variables, including a summary of all operational procedures. Section IV presents the results of the operation and discusses. Finally, Section V provides conclusions and future works, respectively.

2. Literature Review

Research and development tools for educational application consist of a combination of disciplines, including science, social sciences, education, engineering, and technology that can develop potential and success. From the research problem needed to solve the problem of matching learners to appropriate institutions. The research team summarized and collected relevant literature to present the ideas in the following topics.

2.1. Students' Academic Achievement

Measurement of student success can be done in several ways and factors. According to the research of Dougherty and Sharkey [6], it was found that factors contributing to the improvement of student achievement were three parts: grade point averages (GPAs), emotional competence, and social support. It offers a model of path analysis for mediation models with social support and emotional ability that reflects students' impact on GPA and achievement.

While the invisible academic relationship between students and institutions is contributing to academic achievement [7]. It clearly demonstrates that attitudes toward academia, along with their academic achievements, contribute to institutional decision-making and academic success. Another example of research related to academic achievement are dimension of learning styles and students' academic achievement [8]. It is suggested that an appropriate learning style could help students to achieve a good academic record in any subject they learned.

Therefore, it is concluded that the development of students to attain academic achievement must be in every dimension including the students, the learning environment, the education system and the stakeholders [9-11].

2.2. Educational Technology and Students' Performance

Students' performance refers to the academic performance of a student, which encompasses the processes that promote student success in education. While technology enables life and education is easy. Various tools and methods are used consists in teaching and learning in the classroom, including using of a mobile learning management system with online students [12], an applications on

the mobile to introduce the appropriate institution [5], prediction of students' academic failure [13], and using an educational computer program to enhance student performance [14].

No matter what tools or technologies are used to support education. Educational measurement and educational standards are still needed as a basis for the student's developmental and educational success. Therefore, the choice of tools must take into account the learning goals and achievement that will occur with the students.

2.3. Education System and Limitations

It cannot be denied that education is not only in the classroom, although education is attempting to develop a structure or form that is appropriate for the learner's age and environment. But, the out-of-class study and self-learning are surprisingly remarkable. Meanwhile, getting the knowledge and answers from the Internet can be easy and convenient. But there are few percent verified and accepted. Therefore, the education system must integrate modern technology and limit the scope of finding the correct and appropriate knowledge for the learners.

Some researches offer a dimension of education related to the education system in Thailand. Based on the research report of Sitthipon Art-in [15], it was found that the current situation and need in learning management for developing the analytical thinking of teachers in basic education of Thailand as overall in the high level. The problem with the lowest level of performance is that teachers have the ability to design and create learning plans to develop students' analytical thinking.

While research on patterns and methods of education affecting students has been discussed [16]. It is concluded that the students should be allowed to construct the body of knowledge by themselves through the Open Approach, which they are interested in and motivated to face of the problem situation. Therefore, in order to understand Thailand's education system and education structure, the following topics are divided into 2 main themes: 1) Thailand's formal education system, and 2) lifelong learning style.

2.3.1. Thailand's Formal Education System

In the public domain of Thailand, the education system is free and takes 12 years to complete. The study structure follows the "6-3-3" project, with 6 years of primary education, 3 years of lower secondary level and 3 years of upper secondary level. As displayed in Figure 5, elementary education is classified as Prathom 1-6, which correlates to Grades 1-6. While, secondary education is classified as Matthayom 1-6 and correlates to Grades 7-12.

<table>
<tr><th colspan="2">Stage</th><th colspan="2">Level / Grade</th><th>Typical Age</th><th>Notes</th></tr>
<tr><td rowspan="18">Basic Education</td><td rowspan="3">Early Childhood (Kindergarten)</td><td colspan="2" rowspan="3">Variable (Typically Anuban 1-3)</td><td>4</td><td rowspan="3"></td></tr>
<tr><td>5</td></tr>
<tr><td>6</td></tr>
<tr><td rowspan="12">Elementary</td><td colspan="2">Prathom 1</td><td>7</td><td rowspan="15">Compulsory Education</td></tr>
<tr><td colspan="2">Prathom 2</td><td>8</td></tr>
<tr><td colspan="2">Prathom 3</td><td>9</td></tr>
<tr><td colspan="2">Prathom 4</td><td>10</td></tr>
<tr><td colspan="2">Prathom 5</td><td>11</td></tr>
<tr><td colspan="2">Prathom 6</td><td>12</td></tr>
<tr><td rowspan="3">Lower-Secondary</td><td colspan="2">Matthayom 1</td><td>13</td></tr>
<tr><td colspan="2">Matthayom 2</td><td>14</td></tr>
<tr><td colspan="2">Matthayom 3</td><td>15</td></tr>
<tr><td rowspan="4">Upper-Secondary</td><td>General</td><td>Vocational</td><td rowspan="2">16</td><td rowspan="4"></td></tr>
<tr><td>Matthayom 4</td><td rowspan="3">Vocational Certificate (3 years)</td></tr>
<tr><td>Matthayom 5</td><td>17</td></tr>
<tr><td>Matthayom 6</td><td>18</td></tr>
<tr><td></td><td>Higher Education</td><td colspan="2">Variable</td><td>...</td></tr>
</table>

Figure 5. Thailand's Formal Education System

2.3.2. Lifelong Learning Style

In view of its success, to learn what love is one of the factors. It can be supported by Cabi's research [17]. It is recommended that students find out their own learning styles and choose the most effective strategies for their learning. It is important to learn for lifelong learning. In the definition of learning style, it is a set of techniques that make it easy for individuals to learn about yourself and learn to make the learning process information in a manner that is stable.

Theories of learning styles are continually evolving from the past. One popular theory is that Kolb's learning model, in which individual learning styles are cyclical [18]. There are four learning style in this cycle, including Abstract Conceptualization (AC), Active Experience (AE), Concrete Experience (CE), and Reflective Observation (RO). The student learning model is represented by four basic types of learning, and the individual's score indicates the most appropriate learning style as shown in Table I.

Table I. Four basic types of learning styles

Types	Elements	Meaning
Accommodating	AE and CE	Engage in real experiences and enjoy learning from exploration and application and search. They are open for strange thoughts.
Assimilating	AC and RO	They like structured and systematic knowledge. The knowledge presented must be detailed, structured and include visual and audio components.
Converging	AC and AE	The details are very important for those individuals and they are trying to understand by studying each section first. They learn to new knowledge first, then use it and not be afraid to make mistakes while learning.
Diverging	CE and RO	They like learning by observing real experiences instead of joining them. Need systematic presentation of knowledge on various topics. These individuals will be successful from the perspective of real cases.

Learning styles have a significant relationship to the nature of teaching and learning [18]. Therefore, teaching and learning management has an impact on the educational system. Laal, et al. [19] summarizes a group of lifelong learning and education systems, consisting of three main groups: formal learning, non-formal learning, and informal learning as details are shown in Table II.

Table II. Education systems

Categories	Definitions and Examples
Formal Learning (Formal Education)	It is learning that takes place in an organized and structured context. **Examples:** • Study in school or academic institutions • Training in a workplace or through an expert • Learning with curriculum or structured courses
Non-Formal Learning (Non-Formal Education)	It is driven by activities that are not clearly defined. **Examples:** • Short course training • Training in communities • Groups of adjustment according to students' needs.
Informal Learning (Informal Education)	It relates to learning from the daily activities of family and work. It is often called experiential learning. **Examples:** • Family activities • Education to improve individuals

2.4. Technology Support and Solutions

This section provides information and literature about applying technology related to education in three topics: 1) Data mining in education, which refers to the use of data mining technology for educational research. 2) Mobile web application technology, which offers technology to support and develop research tools. 3) Software development methodology, which is a process and step to develop an effective and complete application.

2.4.1. Data Mining in Education

Data mining technology is searching for patterns and processes that have values from a large number of data. In the education system, vast amounts of information being analyzed and synthesized for the main purpose is to develop the capacity of students and learners' potential, to achieve the academic achievement of the learner.

Examples of research such as Costa's research, which presents the results of evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure [13]. Natek's research demonstrates the application of data mining technologies to the analysis and development of decision models for planning and management in higher education institutions [20]. And the important example is an institution recommender system based on student context and educational institution in a mobile environment [5] that extends to this research. This work has been carried out to complete the process of the System Development Life Cycle (SDLC), which consists of seven steps: problem recognition, feasibility study, analysis, design, construction, conversion, and maintenance.

2.4.2. Mobile Web Application Technology

Nowadays, access to information and the internet is easy with the communication device everyone has. It is a mobile phone. Many applications have been developed and used to support educational systems.

For example, Briz-Ponce's research [21], which presents research findings on different drivers and factors that may affect student needs with mobile technology for learning. Yoo and Lee's research [22] uses a mobile application technology to develop knowledge and teaching for nursing students. Simonova and Poulova's research [23] applied mobile technology for learners' preferences in mobile-assisted higher education.

From the examples that bring mobile computing technology into use in various fields. As well as, the research team was interested in developing tools to support education, it developed a mobile for introducing appropriate institutions to students [5]. The research was extremely interesting, and the research team further developed the research that was proposed to assess application satisfaction and efficiency, which is presented in the next section.

2.4.3. Software Development Methodology

In software engineering, software development methodology is software development. It is a different procedure which consists of activities that are intentionally planned and managed better. It is often perceived as a subset of the development life cycle. Such methods may include predefined definitions of deliverables and inventions defined by project teams to develop or maintain programs. There are several software development methodology such as waterfall, prototyping, iterative and incremental development, spiral development, rapid application development, extreme programming and various types of agile methodology.

In this research, the waterfall model is used to assemble processes to develop applications and software for quality and performance, which consisting of five components: requirements engineering, design and implementation, testing, release, and maintenance as shown in Figure 6.

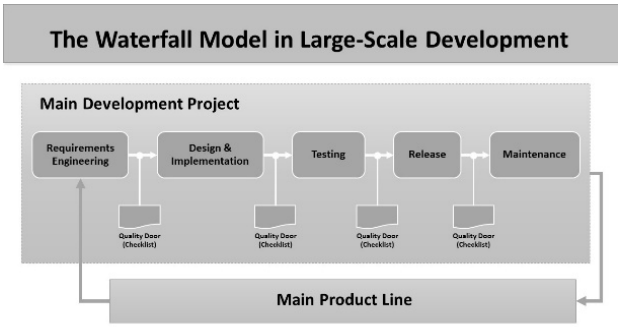


Figure 6. The Waterfall Development

2.4.3.1. Requirements Engineering

At this stage, the needs and significance of the problem are identified and documented at the abstract level. After that the specifications will be refined so that they can be used for design and operation. Specifications are stored in the required repository, and the requirements to be applied are selected from the repository. A number of requirements, depending on the resources available for the project. The quality will check that all terms are understood and agreed upon and documented. It also requires monitoring relevant stakeholders and solutions to support strategies and objectives.

2.4.3.2. Design and Implementation

At the stage of architectural design was created documented. After that, system development started. Developers also perform basic unit tests before code development during the testing phase. The quality checklist verifies that the architecture has been evaluated for deviation from the specifications compared to predetermined quality decisions.

2.4.3.3. Testing

In this period, integration has been tested on quality and functionality. In order to decide whether the system is usable, it will measure performance in the lab such as throughput. Because the goal is to solve the problem perfectly, then the test will be performed on a variety of different configurations. Phase results will be reviewed according to a checklist to track whether the system has been validated.

2.4.3.4. Release

In the release phase the product is returned into a shippable state. The release documentation is the final step, which contains documentation for installation and usage for the user. In addition, creating instructions for the system must be programmed. The build guide can be used to enable and disable features of the main line product to customize the system to meet the specific needs of the target. Quality must ensure that the response meets the needs of the target. The final result is delivered for the intended purpose and meets the quality requirements.

2.4.3.5. Maintenance

After the product has been released, it must be maintained. That is, if the user encounters a problem in the product, the report is developed and encouraged to solve the problem. If the problem is due to an error in the product package, the system update will be sent to the user.

3. Material and Methods

After gathering the literature and scope of the research area, this step details the material and the method. It consists of five main stages: 1) requirement definition, 2) system and software design, 3) implementation and unit testing, 4) integration and system testing, and 5) operation and maintenance as shown in Figure 7.

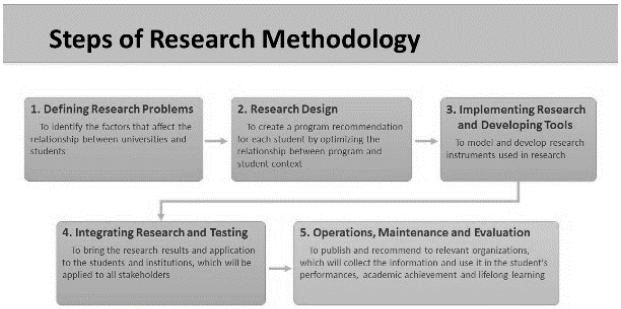


Figure 7. Steps of Research Methodology

3.1. Defining Research Problems

The objective of this step is to identify the factors that affect the relationship between universities and students [24]. This process was conducted by designing a questionnaire asking for opinions and attitudes from 256 students at two universities: University of Phayao and Rajabhat Maha Sarakham University. After receiving information from the respondents, the text mining process was analyzed to model and identify factors that are important to students and institutions.

The results show that the performance of the model employs two important factors; education system and interested in study, which are used to predict student's interest in higher education. While, the decision tree model with two depth levels was most appropriate with 81.46% accuracy as shown in Figure 8.

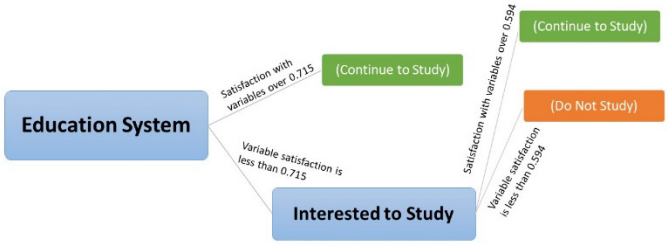


Figure 8. Decision Tree Model to Predict Student's Interest [24]

3.2. Research Design

After discovering that the education system and student interest have a bearing on the decision to study in higher education, the research team studied the characteristics of attitudes that resulted in the selection of educational programs, as well as educational institutions. Based on the data collected from 885 students of three higher educations, including Rajabhat Mahasarakham University, Mahasarakham University, and University of Phayao and the results of model analysis, the accuracy points out 19 significant factors; 15 factors of student context and 4 factors of program context, which is appropriate used to introduce institutions for the style and attitude of the students as show in Table III.

Table III. Summary of significant factors

Characteristics	Features
Student's Interests	• Interest in studying • Career path • Identity of the University • Obtained knowledge • Education system
Student's Characteristics	• Ability • Skills • Knowledge • Learning style
Student's Environment	• Social capital and Economic • Trust in institutions • Personal decision • Family income • Community • Purpose of study
Programs' Context	• The popularity of the University • Curriculum • Education standard • Quality assurance

In addition, the model has 76.50 % accuracy, showing that the generated model represents the appropriate relationship between student's context and programs' context for the proposed program recommendation, with 53 rules as shown in Figure 9.

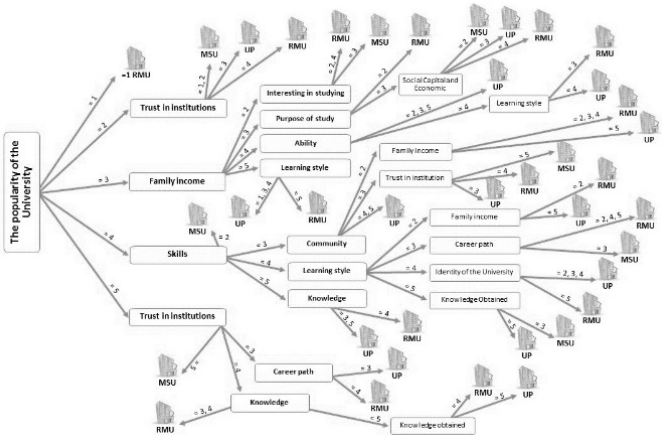
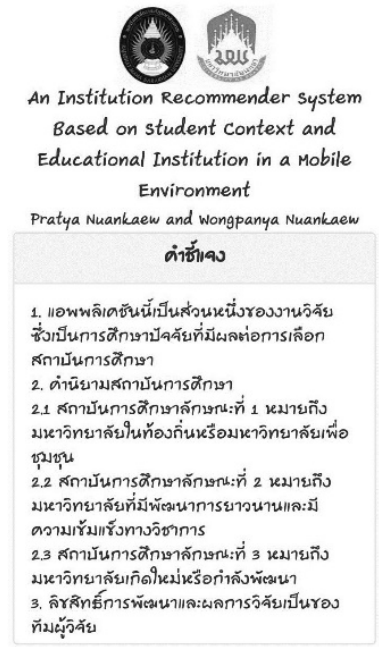


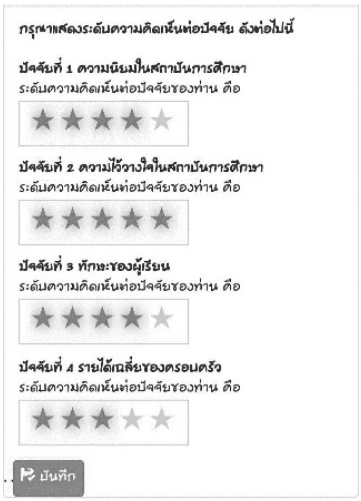
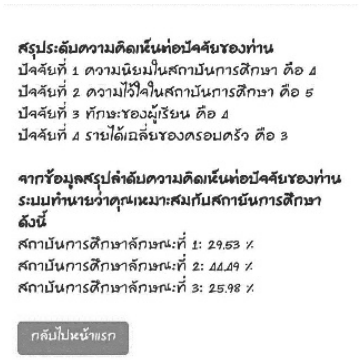
Figure 9. Decision Tree Model to Introduce Institutions

3.3. Implementing Research and Developing Tools

After knowing and discovering the factors that can be used to predict an interest in further education and behavioral characteristics that are consistent with, and appropriate educational programs and institutions. The application for introducing educational institutions based on student's context and educational institution in a mobile environment has been developed, which using the mobile web apps function [5]. It was applied to 1,109 students from 3 campuses; Rajabhat Mahasarakham University, Mahasarakham University, and University of Phayao to conduct and test the application, which the interface of the application is shown in Table IV.

Table IV. Application interface

Interface	Function
	This page aims to clarify the purpose and explain the details of the research.
	This page store user information and ask for feedback on the educational system.
	This page requests the level of satisfaction to study at university.
	This page shows the results of interest analysis of users who wish to study at university.

Interface	Function
	This page aims to gather attitudes towards additional factors in predicting the appropriate university for respondents.
	This page shows results and universities that fit the respondents.
	Thank you page

3.4. Integrating Research and Testing

To bring the application to the students, institutions and all stakeholders. The research team applied the application to the university's public relations activities and curriculum promotion activities. At the same time, the research team collected the survey data on attitudes and academic achievement of students in the upper secondary school that represent the academic relationship between academic achievement and institutions [7].

The survey found that the Grade Point Average (GPA), and school types had an impact and affects the attitudes toward the

selection institutions, which is consistent to student's attitude, their academic achievement and research's factors.

3.5. Operations, Maintenance and Evaluation

This step is the final stage of the research process, which is a key topic in the research process. The goal of this step is to publish and recommend to relevant organizations, which will provide the application and use it in the student's performances development, development of learning achievement and promoting lifelong learning.

The process at this stage consists of promoting the application and evaluation of applications in all dimensions, including finding the level of satisfaction toward the application, and evaluate the performance of the application, as elements as shown in Figure 10.

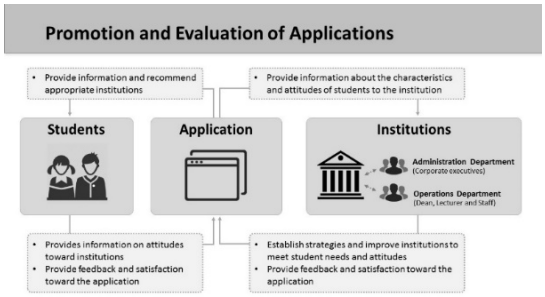


Figure 10. Process of Promotion and Evaluation of Applications

From Figure 10, it shows the process of promotion and evaluation of applications, which consist of two dimensions, including the student dimension and the institution dimension.

To carry out the full procedure as planned, the research team had planned operation consists of four stages: 1) promoting and introducing the application, 2) providing results and information from the application, 3) collecting feedback and satisfaction from the users, and 4) summary of suggestions and analyze the level of satisfaction. The details of each step is explained in the next section.

3.5.1. Promoting and Introducing the Application

The samples were selected to promote the application consists of 4 groups, including 186 high school students at Phadungnaree School, Maha Sarakham province, 191 university students at Rajabhat Mahasarakham University, 12 head of institutions at Phadungnaree School and Rajabhat Mahasarakham University, and 42 teachers and lecturers at Phadungnaree School and Rajabhat Mahasarakham University.

3.5.2. Providing Information from the Application

After presenting and introducing the application, when the user provides information, attitudes, and level of satisfaction to the application. The application performs data synthesis and presents all results to the user.

For example, students enter the attitude toward the key factor. After that, the application calculates the predictive function and introduces the appropriate institution to the student. Another example is that the application provides an overview of the user feedback to the university, each of which factors the institution can use to improve its potential. At the same time, the research team was able to gather information, attitudes and opinions to further improve and develop the research.

3.5.3. Collecting feedback and satisfaction

Gathering feedback and satisfaction levels on the application, the research team has designed and constructed a questionnaire to evaluate and ask satisfaction, which consists of four key areas to ensure compliance with software quality including 1) functional requirement test, 2) functional test, 3) usability test, and 4) security test.

However, user and volunteer participation is strictly voluntary. Apart from the results and the predictions of the application, the participants do not receive any compensation or items for participation. Participants were asked to honestly score statements on a 5-point Likert scale, ranging from 1 to 5, with 1 indicating little to no agreement and 5 indicating high agreement with statements.

3.5.4. Analyze the Level of Satisfaction

At this stage, data analysis from the sample is based on simple statistical analysis, which includes the average and standard deviation. Then analyze the results and summarize key points from the recommendations and suggestions.

4. Results and Discussion

According to the research team has designed the research protocol into five parts, The process has been completed in four stages [5, 7, 24]: 1) defining research problems [24], 2) research design, 3) implementing research and developing tools [5], and 4) integrating research and testing [7]. There is still another important step, the final step is operation, maintenance and evaluation.

Therefore, this section is divided into two parts: reporting on the results of the final stage of the research methodology and discussion topics.

4.1. Results of operations, maintenance and evaluation

The reported results of operations are presented survey results were compiled from four sample groups: 1) high school students, 2) university students, 3) head of institutions, and 4) secondary teachers and university lecturers. It has a variety of views and opinions as summarized in Table V.

Table V. Summary of sample groups

Sample Groups	Number and Source
Gender	228 female (52.90%) and 203 male (47.10%)
High School Students	186 students (43.16%) from Phadungnaree School, Maha Sarakham province, Thailand
University Students	191 students (44.32%) from Rajabhat Mahasarakham University, Maha Sarakham province, Thailand
Head of Institutions	12 school administrators (2.78%) from Phadungnaree School and 10 administrators, department heads and deans from Rajabhat Mahasarakham University.
Secondary Teachers and University Lecturers	42 teachers (9.74%) from Phadungnaree School and lectures from Rajabhat Mahasarakham University.
Total:	**431 Samples**

From Table V, it shows the details of the sample and the data collected. The data obtained from the survey is based on a

questionnaire with questions on four main perspectives as shown in Table VI.

Table VI. Four main perspectives of the questionnaire

Perspectives and Functions	Sub-topics and Details
Functional Requirement Test	• Stage 1: The system's ability to predict and recommend institutions • Stage 2: System capabilities in service and user management • Stage 3: System capabilities in data management • Stage 4: System capabilities in reporting and performance
Functional Test	• Stage 5: The accuracy of the results of the processing • Stage 6: The speed of processing of the application • Stage 7: Reliability of the program • Stage 8: Coverage of programs developed for the enterprise
Usability Test	• Stage 9: Ease of use and user friendliness • Stage 10: Appropriateness of color, size and font of the application • Stage 11: The appropriateness of using text to describe and convey meaning • Stage 12: The appropriateness of using the symbols or images • Stage 13: Standards and design criteria for the user interface • Stage 14: Appropriateness of interaction with users
Security Test	• Stage 15: Controlling access to systems, authorizing, and data access control • Stage 16: System failure prevention and data entry

After designing the questionnaire and conducting the survey on selected samples, survey results are analyzed and summarized in Table VII.

Table VII. Overview of application satisfaction

Perspectives and Functions	Sample Groups				Total:
	High School Students	University Students	Head of Institutions	Teachers and Lecturers	
Functional Requirement Test					
• Stage 1	4.36 (0.51)	4.55 (0.65)	4.33 (0.49)	4.31 (0.52)	4.44 (0.58)
• Stage 2	3.55 (0.57)	4.23 (0.76)	4.00 (0.60)	3.74 (0.59)	3.91 (0.74)
• Stage 3	3.62 (0.55)	4.42 (0.68)	4.00 (0.60)	3.83 (0.58)	4.00 (0.72)
• Stage 4	3.95 (0.28)	4.35 (0.74)	3.83 (0.39)	3.93 (0.51)	4.12 (0.57)
Functional Test					
• Stage 5	4.24 (0.44)	4.42 (0.60)	4.50 (0.52)	4.43 (0.50)	4.35 (0.53)
• Stage 6	4.81 (0.41)	4.50 (0.66)	4.75 (0.45)	4.71 (0.51)	4.66 (0.56)
• Stage 7	3.76 (0.51)	4.33 (0.76)	3.50 (0.52)	3.91 (0.62)	4.02 (0.70)
• Stage 8	3.84 (0.42)	4.33 (0.70)	3.83 (0.39)	3.91 (0.48)	4.06 (0.61)
Usability Test					
• Stage 9	4.90 (0.30)	4.61 (0.64)	4.92 (0.29)	4.98 (0.15)	4.78 (0.50)
• Stage 10	4.59 (0.49)	4.48 (0.59)	4.67 (0.49)	4.50 (0.51)	4.54 (0.54)
• Stage 11	4.12 (0.38)	4.47 (0.60)	4.25 (0.45)	4.10 (0.37)	4.28 (0.52)
• Stage 12	4.00 (0.27)	4.38 (0.67)	4.00 (0.43)	4.14 (0.35)	4.18 (0.53)
• Stage 13	4.76 (0.47)	4.41 (0.70)	4.83 (0.39)	4.69 (0.47)	4.60 (0.61)
• Stage 14	4.54 (0.54)	4.41 (0.67)	4.67 (0.49)	4.50 (0.51)	4.49 (0.60)
Security Test					
• Stage 15	4.18 (0.51)	4.43 (0.66)	3.92 (0.52)	4.17 (0.49)	4.28 (0.59)
• Stage 16	4.21 (0.50)	4.52 (0.64)	4.17 (0.72)	4.19 (0.55)	4.35 (0.60)
Total:	**4.22 (0.61)**	**4.43 (0.67)**	**4.26 (0.63)**	**4.25 (0.60)**	**4.32 (0.65)**

From Table VII, it can be seen that the overall satisfaction of the application is at a high level (4.32). The highest satisfaction level was the student in the University, with the satisfaction level at 4.43. The second highest satisfaction level was the group of head of institutions, with the satisfaction level at 4.26. And the lowest level of satisfaction was high school student's satisfaction, at 4.22.

However, the recommendations from the survey can be summarized as important issues, such as how to publish the information to the stakeholders, how to use of data and the results of research into concrete action, future research will be carried out commercially or not, and how to monitor and evaluate learners in the long term.

4.2. Discussion

Refer to research process and research objectives, this research has completed all the stages from problem study to research evaluation. At every stage, the research team publishes the results in a step-by-step manner. All the published research, it has been expertly investigated and reviewed by a specialist. So the research results are truly qualitative, beneficial and useful. However, in the research discussion, the results were based on recommendations from experts who had received research feedback and comments from the participants from the survey to the discussions.

From the feedback and comments, it was found that the recommendations of the experts encouraged the research team to develop in-depth research that related to the research in education data mining. Because, most educators use knowledge-based learning processes for learners, there is a lack of tools that support and track student achievement. While the viewpoint of the general user, it is consistent with the ideas of the experts who need the tools and technology that can support, monitor, and predict students' achievement.

In future research, the research team clearly recognizes the required for substantial studies and development into concrete actions. Research focused on tracking the educational achievement of students at all levels, including the connections and relationships between the boundaries of each education level, resolving dropout rates during education, and increasing the promotion and development of lifelong learning.

5. Conclusions

After the research team has implemented and completed five main processes, including 1) the process of requirement definition, 2) the process of system and software design, 3) the process of implementation and unit testing, 4) the process of integration and system testing, and 5) the process of operation and maintenance.

The information that has been used in the research is composed of four groups: 186 high school students at Phadungnaree School, Maha Sarakham province, 191 university students at Rajabhat Mahasarakham University, 12 deans at Phadungnaree School and Rajabhat Mahasarakham University, and 42 secondary teachers and university lecturers at Phadungnaree School and Rajabhat Mahasarakham University.

The results indicated that the total satisfaction are highest, which is equal to 4.32, and the group with the highest level of satisfaction is group of university students, which is equal to 4.43. It can be concluded that the application is appropriate and it is advisable for students who wish to study at university.

By referring to the purpose of the research, the team can conclude that this research has been very successful. The developed applications are appropriate and should be widely disseminated. In the future, the research team has the idea to use this application to promote sustainable learning that is consistent with the personality of the individual.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This article has received financial support, research information, experts and resources from Phadungnaree School, Rajabhat Mahasarakham University, the University of Phayao, and Siam Crystal Consulting Co., Ltd. Bangkok, Thailand. The authors would like to thank the researchers, participants, and technicians for their efforts toward the completion of this research.

References

[1] H. B. Olaiya, "Transforming our world: The 2030 agenda for sustainable development & international decade for people of African descent," 2016.

[2] J. Castaño Muñoz, C. Redecker, R. Vuorikari, and Y. Punie, "Open Education 2030: planning the future of adult learning in Europe," *Open Learning: The Journal of Open, Distance and e-Learning*, vol. 28, no. 3, pp. 171–186, 2013.

[3] M. Tight, "Lifelong learning: opportunity or compulsion?," *British Journal of Educational Studies*, vol. 46, no. 3, pp. 251–263, 1998.

[4] M. Sharples, "The design of personal mobile technologies for lifelong learning," *Computers & Education*, vol. 34, no. 3, pp. 177–193, 2000.

[5] K. Pupara, W. Nuankaew, and P. Nuankaew, "An institution recommender system based on student context and educational institution in a mobile environment," in *Computer Science and Engineering Conference (ICSEC), 2016 International*, 2016, pp. 1–6.

[6] D. Dougherty and J. Sharkey, "Reconnecting Youth: Promoting emotional competence and social support to improve academic achievement," *Children and Youth Services Review*, vol. 74, pp. 28–34, 2017.

[7] W. Nuankaew, P. Nuankaew, S. Bussaman, and P. Tanasirathum, "Hidden academic relationship between academic achievement and higher education institutions," in *Digital Arts, Media and Technology (ICDAMT), International Conference on*, 2017, pp. 308–313.

[8] N. Omar, M. M. Mohamad, and A. N. Paimin, "Dimension of Learning Styles and Students' Academic Achievement," *Procedia - Social and Behavioral Sciences*, vol. 204, pp. 172–182, Aug. 2015.

[9] M. M. Hernández *et al.*, "Elementary students' effortful control and academic achievement: The mediating role of teacher–student relationship quality," *Early Childhood Research Quarterly*, vol. 40, pp. 98–109, rd 2017.

[10] E. C. M. van Rooij, E. P. W. A. Jansen, and W. J. C. M. van de Grift, "Secondary school students' engagement profiles and their relationship with academic adjustment and achievement in university," *Learning and Individual Differences*, vol. 54, pp. 9–19, Feb. 2017.

[11] M. Yavuz and F. B. Kiyici, "Teachers' Opinions Regarding the Effects of the Usage of out-of-school Learning Environments on Students' Academic Achievement and Anxiety towards Science," *Procedia - Social and Behavioral Sciences*, vol. 106, pp. 2532–2540, Dec. 2013.

[12] I. Han and W. S. Shin, "The use of a mobile learning management system and academic achievement of online students," *Computers & Education*, vol. 102, pp. 79–89, Nov. 2016.

[13] E. B. Costa, B. Fonseca, M. A. Santana, F. F. de Araújo, and J. Rego, "Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses," *Computers in Human Behavior*, vol. 73, pp. 247–256, Aug. 2017.

[14] S. H. Chan, Q. Song, L. H. Rivera, and P. Trongmateerut, "Using an educational computer program to enhance student performance in financial accounting," *Journal of Accounting Education*, vol. 36, pp. 43–64, Sep. 2016.

[15] S. Art-in, "Current Situation and Need in Learning Management for Developing the Analytical Thinking of Teachers in Basic Education of Thailand," *Procedia - Social and Behavioral Sciences*, vol. 197, pp. 1494–1500, Jul. 2015.

[16] N. Satchakett and S. Art-in, "Development of Heuristics Problem Solving, and Learning Achievement of Grade 9 Students by Using Learning Management Focusing on Heuristics Approach in Thailand," *Procedia - Social and Behavioral Sciences*, vol. 116, pp. 1011–1015, Feb. 2014.

[17] Cabi and S. Yalcinalp, "Lifelong Learning Considerations: Relationship Between Learning Styles and Learning Strategies in Higher Education," *Procedia - Social and Behavioral Sciences*, vol. 46, pp. 4457–4462, Jan. 2012.

[18] P. Nuankaew and P. Temdee, "Online Mentoring Model by Using Compatible Different Attributes," *Wireless Pers Commun*, vol. 85, no. 2, pp. 565–584, Nov. 2015.

[19] M. Laal, A. Laal, and A. Aliramaei, "Continuing Education; Lifelong Learning," *Procedia - Social and Behavioral Sciences*, vol. 116, pp. 4052–4056, Feb. 2014.

[20] S. Natek and M. Zwilling, "Student data mining solution–knowledge management system related to higher education institutions," *Expert Systems with Applications*, vol. 41, no. 14, pp. 6400–6407, Oct. 2014.

[21] L. Briz-Ponce, A. Pereira, L. Carvalho, J. A. Juanes-Méndez, and F. J. García-Peñalvo, "Learning with mobile technologies – Students' behavior," *Computers in Human Behavior*, vol. 72, pp. 612–620, Jul. 2017.

[22] I.-Y. Yoo and Y.-M. Lee, "The effects of mobile applications in cardiopulmonary assessment education," *Nurse Education Today*, vol. 35, no. 2, pp. e19–e23, Feb. 2015.

[23] I. Simonova and P. Poulova, "Learners Preferences in Mobile-Assisted Higher Education," *Procedia Computer Science*, vol. 104, pp. 174–182, Jan. 2017.

[24] P. Nuankaew, W. Nuankaew, and T. Thamma, " The Recommended System for the Relationship between Educational Programs and Students' Interests," in *Digital Arts, Media and Technology (ICDAMT), International Conference on*, 2016, pp. 308–313.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1565-1576 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Planning an Availability Demonstration Test with Consideration of Confidence Level

Frank Müller*, Bernd Bertsche

Institute of Machine Components, University of Stuttgart, 70569 Stuttgart, Germany

ARTICLE INFO

Article history:
Received: 30 May, 2017
Accepted: 29 July, 2017
Online: 21 August, 2017

Keywords:
Average Availability
Availability Demonstration
Availability Demonstration Test
Availability Inference
Bootstrap Markov Process
Bootstrap Monte Carlo Simulation
Bootstrap Renewal Process
Confidence Interval
Confidence Level
Inverse Bootstrap Monte Carlo Simulation
Point-Availability
Repairable System
Steady-State Availability
Time-Dependent Availability

ABSTRACT

The full service life of a technical product or system is usually not completed after an initial failure. With appropriate measures, the system can be returned to a functional state. Availability is an important parameter for evaluating such repairable systems: Failure and repair behaviors are required to determine this availability. These data are usually given as mean value distributions with a certain confidence level. Consequently, the availability value also needs to be expressed with a confidence level.

This paper first highlights the bootstrap Monte Carlo simulation (BMCS) for availability demonstration and inference with confidence intervals based on limited failure and repair data. The BMCS enables point-, steady-state and average availability to be determined with a confidence level based on the pure samples or mean value distributions in combination with the corresponding sample size of failure and repair behavior. Furthermore, the method enables individual sample sizes to be used. A sample calculation of a system with Weibull-distributed failure behavior and a sample of repair times is presented.

Based on the BMCS, an extended, new procedure is introduced: the "inverse bootstrap Monte Carlo simulation" (IBMCS) to be used for availability demonstration tests with consideration of confidence levels. The IBMCS provides a test plan comprising the required number of failures and repair actions that must be observed to demonstrate a certain availability value. The concept can be applied to each type of availability and can also be applied to the pure samples or distribution functions of failure and repair behavior. It does not require special types of distribution. In other words, for example, a Weibull, a lognormal or an exponential distribution can all be considered as distribution functions of failure and repair behavior.

After presenting the IBMCS, a sample calculation will be carried out and the potential of the BMCS and the IBMCS investigated.

1. Introduction

This paper is an extension of work originally presented in the course of the Annual Reliability and Maintainability Symposium (RAMS 2017) [1].

Availability is an important parameter for describing a repairable system. This system metric shows the quality of both of its major influences: reliability and maintainability. High availability is typically demanded from the customer, e.g., if capital goods are concerned. The required availability target is frequently specified in combination with a certain confidence level. The confidence level defines the probability that the actual population availability is at least the specified availability.

To demonstrate that the system or product will operate at a certain performance level under specified operating conditions, a demonstration test is required. Several reliability demonstration tests have been proposed, such as in [2] or [3]. For the demonstration of a repairable system's quality, an availability demonstration test needs to be carried out. This test shows that the system or product will operate with the stated level of availability.

Descriptions of failure and repair behavior, in addition to their demonstration through tests, are important topics for specifying

*Corresponding Author: Frank Müller, M. Sc., Institute of Machine Components, University of Stuttgart, Pfaffenwaldring 9, 70569 Stuttgart, Germany, Phone: +49 (0)71168569949, Email: frank.mueller@ima.uni-stuttgart.de

availability. Reliability and maintainability are usually demonstrated on the basis of limited sample sizes within a reliability test. The evaluation of the test thus yields a mean-value failure or repair distribution with a certain confidence level. The availability based on this limited information also includes a confidence level. Consequently, the confidence level of availability also needs to be considered within the course of an availability demonstration test.

Several methods which allow for the availability inference with confidence level have been published. Relatively recent approaches can be found in [4] and [5]. However, these works either require special types of distribution, or are restricted to steady-state availability. In [6], the author published three methods for availability prediction with confidence level which are not restricted to special distribution types and which allow for the prediction of time-dependent and steady-state availability with a confidence level.

Many different reliability demonstration tests have been proposed to demonstrate that a system will operate with a certain reliability [2]. However, only limited work has been conducted to develop an availability demonstration test. As a first publication, Coppola [7] provided an overview of the topic of availability demonstration. In [8], the authors proposed confidence limits on availability for exponentially distributed failure and repair times and a non-parametric approach. The approach put forward by [9] also requires exponentially distributed failure and repair times. In [10], the researcher present an initial procedure for conducting an availability demonstration tests under the assumption that the failure and repair behavior are both exponentially distributed. However, all of these works require special distribution types, e.g., exponential distribution, or they only allow for the demonstration of steady-state availability. Ref [1] provided initial concepts and methods for the availability demonstration of time-dependent and average availability with confidence level. In [11], [12], the authors provided further approaches for availability assessment based on Monte-Carlo simulation. These works does not require confidence levels and the number of samples are not predefined in advance.

To the authors' knowledge, no complete procedure has been published so far concerning the availability demonstration test with consideration of confidence levels for a specific point in time or time interval that provides a test plan. The known methods only demonstrate availability based on a predefined number of given failure and repair times or do not include confidence levels at all.

In this paper, we present a new procedure named the "inverse bootstrap Monte Carlo simulation" (IBMCS) which facilitates the availability demonstration with consideration of confidence levels, which is based on the basic concept for availability demonstration published in [1]. The procedure yields a test plan consisting of the required number of failures that must be observed in order to demonstrate a certain availability value for a point in time or as an average value for a specific time interval. Furthermore, the ratio of total observed repair times to total observed failure times will be provided. The new procedure is not restricted to special distribution types and has the potential to be applied to general availability scenarios.

Firstly, several methods for calculating availability are briefly introduced: the Markov process as an analytical method, the renewal process as a numerical method and the Monte Carlo simulation which provides approximate solutions. Several methods for determining confidence levels are also briefly outlined. Furthermore, the bootstrap Monte Carlo simulation (BMCS) for availability inference and the availability demonstration are described in greater detail. The bootstrap Markov process (BMP) and the bootstrap renewal process (BRP) are also named briefly. A sample calculation of BMCS is presented for a system with Weibull-distributed failure behavior and a given sample of repair times. All three methods are in the field of simulation-based methods like the Monte Carlo simulation based approaches.

After this, the new procedure for conducting availability demonstration tests with consideration of confidence levels is presented on the basis of the BMCS. The procedure yields a test plan comprising the required number of failures that must be observed in order to demonstrate a certain availability value. The availability value can be given as a specific availability for a point in time or as an average availability – specified in combination with a confidence level. Furthermore, the procedure yields the required ratio of total observed repair times to total observed failure times.

Finally, a sample availability demonstration test is presented that takes confidence levels into consideration. In the presented example, the failure behavior conforms to a Weibull distribution, whereby the repair times are lognormally distributed. Finally, the potential of the BMCS and the IBMCS alike is investigated.

2. Fundamentals of Reliability Engineering

In this section, several fundamentals of reliability engineering are summarized. Firstly, the basic definitions of statistics and probability theory are presented before the most common lifetime distributions for a reliability description are outlined.

2.1. Reliability and Failure Probability

The failure probability $F(t)$ is the complement of the reliability with $F(t) = 1 - R(t)$. The reliability $R(t)$ of a component or system is defined as the probability that the component or system will not fail prior to time t when operating under prescribed functional and environmental conditions [13].

Besides the reliability and the failure probability as the cumulative distribution function (cdf), the probability density function (pdf) $f(t)$ and the failure rate $\lambda(t) = f(t)/R(t)$ are often required. The mean lifetime $MTTF$ (Mean Time To Failure) is defined as the expected value of lifetime τ [13]:

$$MTTF = E(\tau) = \int_0^\infty t\, f(t)\, dt = \int_0^\infty \big(1 - F(t)\big)\, dt \quad (1)$$

Equation (1) yields the $MTTR$ (Mean Time To Repair) if the distribution is based on repair times instead of failure times.

2.2. Lifetime Distributions

Several lifetime distributions are available in reliability engineering for describing the reliability or failure behavior of a system or component. In the following section, the most commonly used lifetime distributions are described: the exponential distribution and the Weibull distribution.

The exponential distribution is often used for electronic components [13]. Its pdf is defined as [14]:

$$f(t) = \lambda \cdot e^{-\lambda \cdot t},\ t \geq 0 \text{ and } \lambda > 0 \quad (2)$$

The only parameter of the exponential distribution is the constant failure rate λ. Starting from an initial value, the density of the exponential distribution decreases constantly according to an inverse exponential function. The mean lifetime described is given as [13]:

$$MTTF = 1/\lambda \tag{3}$$

The Weibull distribution is the most common lifetime distribution used in mechanical engineering. The pdf for the three-parameter Weibull distribution is defined as [14]:

$$f(t) = \frac{b}{(\eta - t_0)}\left(\frac{t-t_0}{\eta-t_0}\right)^{b-1} e^{-\left(\frac{t-t_0}{\eta-t_0}\right)^b}, t \geq t_0 \geq 0 \tag{4}$$

Here, b is the shape parameter and η the scale parameter (characteristic lifetime). The failure-free time t_0 determines the point in time from which failure occurs. It is called referred to as a two-parameter Weibull distribution, if $t_0 = 0$. The failure rate $\lambda(t)$ of a Weibull distribution is a function in time.

3. Maintenance and Availability

The following section outlines the fundamentals of repairable systems. Firstly, the basic definitions of availability and repairable systems are summarized. Afterwards, three methods for calculating repairable systems are presented: the Markov process, the renewal process and the Monte Carlo simulation. Whereas the Markov process provides an analytical solution, the renewal is a numerical method. The Monte Carlo simulation as a simulation-based method provides approximate solutions.

3.1. Availability and Repairable Systems

Availability is an important parameter for evaluating repairable systems. A repairable system can be returned to a function state following a failure. Considering a stochastic point process with two possible states, the operational state is assigned number 1. If the system fails, the state is assigned number 0. A state indicator can be established with these definitions:

$$c(t) = \begin{cases} 1 & \text{if the system is "operational" at time } t \\ 0 & \text{if the system is "failed" at time } t \end{cases} \tag{5}$$

Here, the pdf $f(t)$ describes the transition from 1 to 0 (failure) and the pdf $g(t)$ the transition from 0 to 1 repair. Table 1 presents further descriptions of the failure and repair behavior.

The availability is defined as the probability that the system will operate satisfactorily at time t [13]. The time-dependent or point-availability is defined as the expected value of the state indicator according to (5). The following applies [13]:

$$A(t) = P(c(t) = 1) = E(c(t)) \tag{6}$$

Table 1. Description of the failure and repair behavior.

Failure behavior		Repair behavior	
Failure density	$f(t)$	Repair density	$g(t)$
Failure probability	$F(t)$	Repair probability	$G(t)$
Failure rate	$\lambda(t)$	Repair rate	$\mu(t)$
Reliability	$R(t)$	-----	
Expected lifetime	$MTTF$	Expected repair time	$MTTR$

The asymptotic value A_∞ with

$$A_\infty = \lim_{t\to\infty} A(t) = \frac{MTTF}{MTTF+MTTR} \tag{7}$$

is referred to as steady-state availability. Finally, the average availability $A_{Av}(t)$ can be calculated using [13]:

$$A_{Av}(t) = \frac{1}{t}\int_0^t A(x)dx \tag{8}$$

3.2. Markov Process

The Markov method (also called Markov model) is a method for analyzing repairable systems. It provides as an analytical method exact calculation results. The objective of the model is to determine the availability of the repairable system or component. The method is subject to a number of requirements [15]. These assumptions limit modelled maintenance actions and simplify calculations [13]:

- The unit to be observed switches continually between the states of "operational" and "repair", i.e., the state of the system or component can only be "operational" or "repair".
- After each maintenance action or repair, the repaired unit is as good as new.
- The times required for operation and repair for each unit observed are continuous and stochastically independent.
- The influence of any switch devices is not taken into consideration.

The Markov method is based on the Markov process [15], a stochastic process with a limited number of states. As a direct result of the Markov property, only systems whose elements possess constant failure and repair rates can be investigated. Thus, the failure and repair behavior need to be exponentially distributed.

For an individual system or component with the failure rate λ and repair rate μ, the availability $A(t)$ can be determined as the probability that the item assumes the state "operational". The following applies [13]:

$$A(t) = \frac{\mu}{\mu+\lambda} + \frac{\lambda}{\mu+\lambda} \cdot e^{-(\lambda+\mu)\cdot t} \tag{9}$$

If t tends to infinity ($t \to \infty$), the availability converges with the steady-state availability A_∞:

$$A_\infty = \lim_{t\to\infty} A(t) = \frac{\mu}{\mu+\lambda} = \frac{MTTF}{MTTF+MTTR} \tag{10}$$

3.3. Renewal Process

One additional method for evaluating repairable systems is the renewal process. If the time for renewal, i.e., the repair duration, is not disregarded, it is termed an alternating renewal process. In this case, the life and repair time succeed each other alternately (see Figure 1). The associated graph of the state indicator of an alternating renewal process of a single item switches alternately between 1 (operational) and 0 (repair) [13]. At time $t = 0$, the item begins operating in a new condition. When the lifetime $\tau_{1,i}$ comes to an end, it is called the point of failure $T_{1,i}$. In the ensuing repair status, the component is either repaired or replaced. The point of renewal $T_{0,i}$ ends the repair duration $\tau_{0,i}$. All points of failure constitute the embedded 1-renewal process and all points of renewal the embedded 0-renewal-process [15].

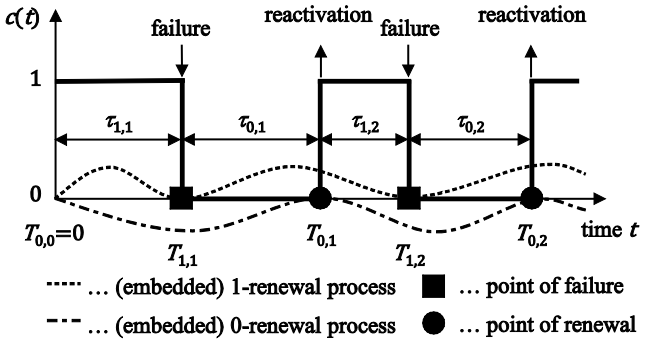


Figure 1. Alternating renewal process [13].

The renewal density function (rdf) $h_0(t)$ and $h_1(t)$ for the renewal points described by $T_{0,i}$ and $T_{1,i}$ are given as [16]:

$$h_0(t) = f * g(t) + \int_0^t f * g(t-t')h_0(t')dt' \qquad (11)$$

$$h_1(t) = f(t) + \int_0^t f * g(t-t')h_1(t')dt' \qquad (12)$$

Here, the operator $*$ denotes convolution. The renewal functions can be determined by integration of the rdfs [17]:

$$H_0(t) = F * g(t) + \int_0^t f * g(t-t')H_0(t')dt' \qquad (13)$$

$$H_1(t) = F(t) + \int_0^t f * g(t-t')H_1(t')dt' \qquad (14)$$

There are several possibilities for describing availability with the aforementioned definitions [17]. The time-dependent availability can be determined as:

$$A(t) = 1 + H_0(t) - H_1(t) \qquad (15)$$

The difference between the renewal functions is equal to the unavailability. The steady-state availability can be determined according to (7) by applying the expected values $MTTF$ and $MTTR$ of the life and repair distributions.

The alternating renewal process has no restrictions with regard to the failure and repair rates. Both constant and time-dependent failure or repair rates, e.g., with Weibull- or lognormally distributed failure and repair times, are allowed. In general, the point-availability can only be calculated numerically based on the renewal process. An analytical solution can only be derived for special types of cdfs like the exponential or Erlang distribution [16].

3.4. Monte Carlo Simulation

Besides the Markov process and the renewal process, the availability of a system or component can be calculated with the help of the Monte Carlo simulation. The Monte Carlo simulation is based on the principle of random sampling. It is a numerical method for the approximate solution of an analytical problem [18]. Thus, no exact values can be determined by Monte Carlo simulation but approximate values. The Monte Carlo simulation is frequently used for reliability analysis and facilitates the calculation of reliability parameters of very complex systems. In contrast to the Markov process, the Monte Carlo simulation has no restrictions with regard to the distribution functions or parameters [19]. The basic principle for calculating the point-availability with the help of the Monte Carlo simulation is illustrated in Figure 2 based on [20].

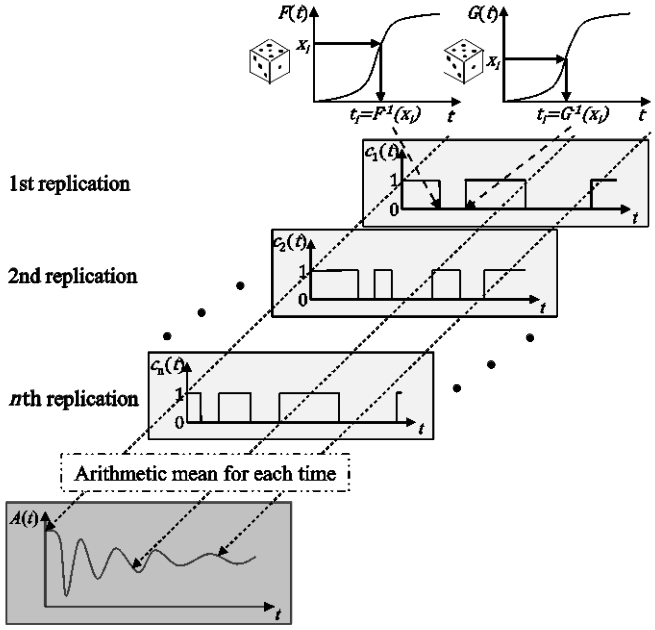


Figure 2. Monte Carlo simulation of (point-) availability [20].

Firstly, random failure and repair times are generated based on the given cdfs of failure and repair behavior. Correspondingly, pseudorandom numbers within the bounds 0 and 1 are generated. These random numbers are interpreted as the failure $F(\xi)$ or repair $G(\xi)$ probabilities. The failure and repair times t^* can subsequently be determined based on these random failure and repair probabilities and using the inversion method [21]:

$$t^* = F^{-1}(\xi) \text{ o}\rho\ t^* = G^{-1}(\xi) \qquad (16)$$

By alternatively selecting one of the randomly generated failure and repair times t^*, one trajectory of the state indicator is established. As for the renewal process, the starting condition $c(t=0) = 1$ is assumed: The component is thus starting in a new condition at time $t = 0$.

After establishing one state indicator throughout the entire observation time, the procedure is repeated n times. n is given as the number of Monte Carlo replications. As a result, n trajectories of the state indicator are calculated. Finally, the time-dependent availability $A(t)$ can be derived as the arithmetic mean of all generated state indicators for each time.

It should be mentioned that the calculated availability represents an approximate solution. The statistical quality of the values determined by means of the Monte Carlo simulation can be estimated. According to the central limit theorem [21], (17) is valid for the random number $\bar{\tau}$ according to (18) [22].

$$\lim_{n\to\infty} P\left(\left|\bar{\tau} - E(\tau) \le \frac{x\sigma}{\sqrt{n}}\right|\right) = \frac{1}{\sqrt{2\pi}}\int_{-x}^{+x} e^{-\left(\frac{u^2}{2}\right)} du \qquad (17)$$

where $-\infty < x < \infty$

$$\bar{\tau} = \frac{1}{n}\sum_{i=1}^{n}\tau_i \qquad (18)$$

This means that the estimated random number normally follows an asymptotic distribution, with the expected value $E(\tau)$ and the variance σ^2/n. Based on this thesis, the statistical quality of the mean value determined with the help of the Monte Carlo simulation can be specified as a level of statistical quality.

4. Confidence Intervals

Selected fundamentals for determining confidence levels are presented in this chapter. After the basic definition, the bootstrap method is presented in greater detail.

4.1. Basic Definition of Confidence Level

The confidence level P_A of availability A_L can be determined in an analogous manner to the reliability [23]. According to (19), it can be calculated using the integral of the density with the considered value A_L as a lower limit:

$$P_A = P(A_L \leq A \leq 1) = \int_{A_L}^{1} f(A')dA' \qquad (19)$$

The calculation is based on the density $f(A)$ of the availability. The confidence level describes the fact that a minimum availability value A_L is reached with a probability of $P_A = 1 - \alpha$. One inherent part of a certain confidence level is the specification of the probability P_A. As a sample, the 95% confidence level implies that the observed value is at least the certain value A_L in 95 out of 100 cases. Usually, a confidence interval lies symmetrically to the median. This means that a 90% confidence interval is bounded by the 5% and 95% confidence level.

There are several methods for determining the confidence intervals of a distribution function. Table 2 shows the most commonly used methods, which can be divided into "analytical methods", "approximate methods" and "simulation-based methods". The table itself is based on [24].

An analytical description of the availability function is required in order to determine a confidence interval with analytical or approximate methods. However, if general scenarios are considered or the Monte Carlo simulation is used, this analytical description is not present or modelled implicitly into an algorithm. Even in the case that the analytical formulation is absent, the bootstrap methods are suitable for determining confidence levels of availability. However, the confidence level determined by bootstrapping is also an approximate solution.

Table 2. Most commonly used methods for determining a confidence interval.

Analytical methods	Approximate methods	Simulation-based methods
• Beta-binomial distribution • Likelihood ratio Interval • Bayes confidence interval	• Fisher matrix • Wald method • Wilson score interval	• Bootstrapping: Non-parametric bootstrapping Parametric bootstrapping

4.2. Bootstrapping

Bootstrapping [25] is a method of resampling and suitable for statistical evaluation, especially if the parameters of a given sample cannot be determined using other analytical methods [26]. The method was first introduced by Bradley Efron [27]. There are two different types of bootstrapping according to Figure 3: non-parametric and parametric bootstrapping.

Non-parametric bootstrapping starts with a given sample $T = \{t_1, t_2, ..., t_n\}$ with sample size n. First, a bootstrap sample $\tilde{T}$ is established by drawing n random numbers from T with replacement and without regard to the order. Afterwards, using well-known standard parameter estimators such as the Maximum Likelihood Method (MLE) [13], the parameters θ of the bootstrap sample are estimated so that the cdf obtained by bootstrapping $\tilde{\Phi}(t)$, named as realization, can be identified.

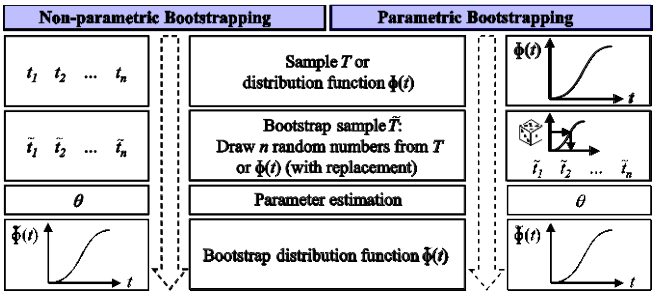


Figure 3. Non-parametric (left) and parametric (right) bootstrapping [28].

In contrast to non-parametric bootstrapping, the parametric bootstrap method is based on a given distribution function $\Phi(t)$. If this cdf is not given in advance, it can be estimated using well-known estimators such as the MLE or by employing the non-parametric bootstrap method. The first step of parametric bootstrapping is to generate n random failure times from the given distribution function. This is performed in the same manner as for the Monte Carlo simulation: with the help of pseudorandom numbers and the inversion method. The distribution function by means of bootstrapping $\tilde{\Phi}(t)$ will subsequently be determined in the same manner as non-parametric bootstrapping, i.e., using standard parameter estimators.

Several bootstrap confidence intervals are available for analyzing the confidence level by bootstrapping [2]. The evaluation can be carried out as [29]:

- An empirical bootstrap confidence interval
- A standard bootstrap confidence interval
- A percentile bootstrap confidence interval
- A bootstrap-*t*-confidence interval
- A BCα confidence interval
- etc.

Irrespective of the type of bootstrap confidence level, the bootstrap method needs to be carried out B (number of bootstrap replications) times. As a consequence, B realizations of the distribution function $\tilde{\Phi}_i(t)$, $i = 1(1)B$ can be identified. This array of trajectories subsequently has to be evaluated. Figure 4 shows the procedure for determining a confidence level using the bootstrap method.

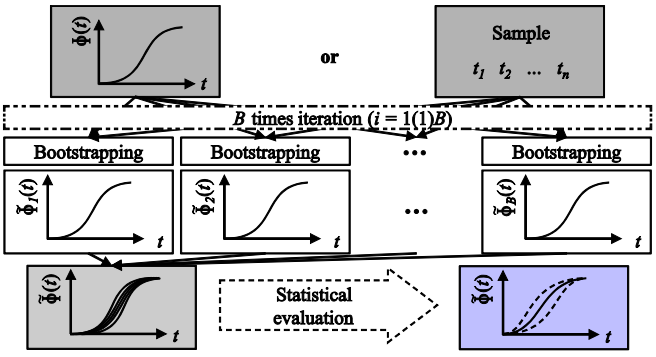


Figure 4. Determining a confidence level using bootstrapping.

In the following, only the "empirical bootstrap confidence interval", the "standard bootstrap confidence interval" and the "percentile bootstrap confidence interval" are presented in greater detail. If an empirical bootstrap confidence interval is calculated, the trajectories are evaluated by performing an empirical bootstrap density function $\varphi\left(\tilde{\Phi}(t_j)\right)$ for each time t_j. Based on the percentiles of these density functions, the confidence level of the original distribution function can be determined.

It is likewise possible to calculate a "standard bootstrap confidence interval". Under the assumption that the realizations $\tilde{\Phi}_i(t)$ are roughly normally distributed, the upper and lower confidence limit of the 100(1-2α)% standard bootstrap confidence interval can be calculated for each time t_j as:

$$\left(\hat{\Phi}(t_j) - z_\alpha \cdot sd\left(\hat{\Phi}(t_j)\right), \hat{\Phi}(t_j) + z_\alpha \cdot sd\left(\hat{\Phi}(t_j)\right)\right) \quad (20)$$

$\hat{\Phi}(t_j)$ is the arithmetic mean value of the B realizations $\tilde{\Phi}_i(t)$, where $sd\left(\hat{\Phi}(t_j)\right)$ is given as the standard deviation of these realizations. z_α is the αth quantile of the standard normal distribution.

On the other hand, the percentile bootstrap confidence interval can be selected based on the given B realizations $\tilde{\Phi}_i(t)$. They have to be sorted in ascending order. Consequently, the $\Phi_i^*(t_j)$ is the ith smallest element. Afterwards, the 100(1-2α)% percentile bootstrap confidence interval can be determined according to (21). Here, $[x]$ equals the largest integer less than or equal to x.

$$\left(\Phi^*_{[\alpha\cdot B]}(t_j), \Phi^*_{[(1-\alpha)\cdot B]}(t_j)\right) \quad (21)$$

In general, the input parameters of the bootstrap parameter can be variously interpreted as failure or repair times. As an example, the given cdf $\Phi(t)$ can either be a failure or a repair distribution function. The confidence interval calculated using the bootstrap method can thus be the confidence interval for a failure distribution function or a repair distribution function. The confidence level of a reliability function or other distribution functions using the bootstrap method is calculated in the same way.

5. Bootstrap Monte Carlo Simulation

This section presents the bootstrap Monte Carlo simulation as a new method for availability inference and demonstration with a confidence level. Using this highlighted method, the confidence level of point-availability as well as the confidence level of average and steady-state availability can be determined.

5.1. Basic Concept

The bootstrap Monte Carlo simulation (BMCS) combines bootstrapping as a method for determining confidence levels with the Monte Carlo simulation. Figure 5 shows the basic principle of BMCS for calculating point-availability with a confidence level. With this basic principle, it is possible to use the information concerning the confidence level of the input data, i.e., of the failure and repair distributions.

The BMCS can be carried out based on pure samples or on the distribution functions (cdfs) of failure and repair behavior. It is also possible to shuffle both types of input data, such as by providing the failure behavior as a cdf and the repair behavior as a sample of repair times or vice versa. The basic procedure also facilitates individual sample sizes for lifetime and repair data. In other words, the sample sizes do not have to be equal.

In the first step of the BMCS, one realization $\tilde{F}_i(t)$ of the failure behavior and one realization $\tilde{G}_i(t)$ of the repair behavior are generated using the bootstrap method and based on the input data. Depending on the type of input data, non-parametric or parametric bootstrapping is employed. After this, these realizations are used to calculate one realization of the point-availability $\tilde{A}_i(t)$ using the Monte Carlo simulation, according to Figure 2. After repeating the entire calculation B times, B realizations of point-availability $\tilde{A}_i(t)$, $i = 1(1)B$ can be calculated. B corresponds to the predefined number of bootstrap replications. Finally, this array of trajectories of point-availability subsequently has to be evaluated.

As shown in the previous chapter, it is possible to determine an empirical bootstrap confidence interval, a standard bootstrap confidence interval, a percentile bootstrap confidence interval or another type of bootstrap confidence interval. If an empirical bootstrap confidence interval is chosen, an empirical bootstrap probability density function $\varphi_j(\tilde{A}(t_j))$ is derived from the individual function values of the realizations of point-availability.

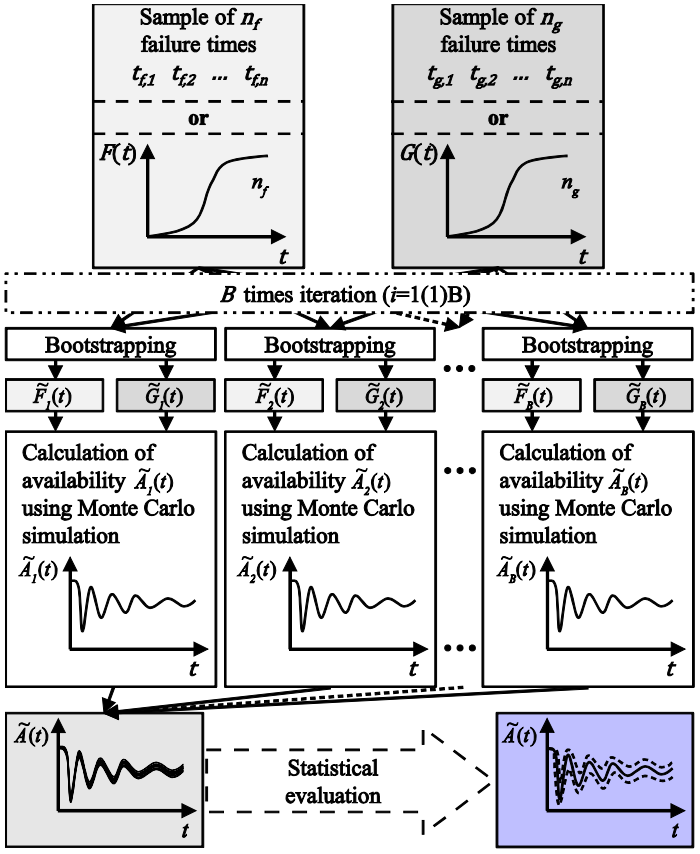


Figure 5. Basic procedure of bootstrap Monte Carlo simulation (BMCS) for calculating point-availability with a confidence level.

The confidence interval can be derived based on the percentiles of these density functions. Figure 6 illustrates this step of the statistical evaluation. On the other hand, the standard bootstrap confidence interval can be determined according to (20) or the percentile bootstrap confidence interval according to (21).

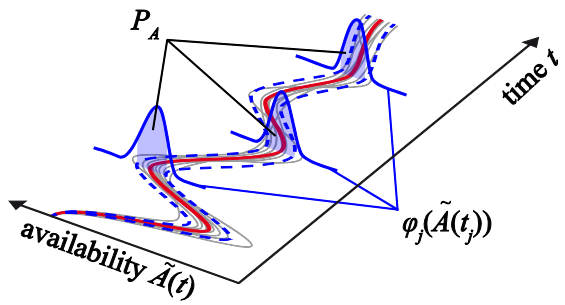


Figure 6. Statistical evaluation of the realizations of point-availability.

The BMCS is an approximation method as the Monte Carlo simulation is used: The calculated confidence interval is thus also an approximation. The BMCS can only provide accurate approximations from a certain threshold number of replications. As shown in [28], the number of bootstrap replications should not be fewer than 200. If at least 200 bootstrap replications are carried out, the approximation of the BMCS is very thorough. On the other hand, the number of Monte Carlo replications should be chosen in the range of 1,000-10,000 Monte Carlo replications. Generally speaking: The higher the numbers of replications, the better the approximation of BMCS. It should be mentioned that the simulation time increases sharply in line with higher numbers of replications.

If the Markov process is used in place of the Monte Carlo simulation to calculate point-availability within the basic procedure according to Figure 5, the bootstrap Markov process (BMP) is established (see [1] or [6], for example). The BMP combines the bootstrap method with the calculation of availability using the Markov method. The BMP therefore has the same restrictions as the Markov process. In other words, the failure and repair behavior needs to be exponentially distributed.

Besides the BMCS and the BMP, the bootstrap renewal process (BRP) [6] combines the bootstrap method with the renewal process [1]. In other words, the availability calculation of the basic procedure according to Figure 5 is carried out using the alternating renewal process. All three methods – the BMCS, the BMP and the BRP – facilitate the availability prediction and demonstration with confidence level, whereby the BMCS has the highest potential to be extended to more general scenarios of availability prediction and demonstration.

5.2. Steady-State and Average Availability with Confidence Level

It is likewise possible to estimate a confidence interval for steady-state and average availability using the bootstrap Monte Carlo simulation provided in Figure 5. The basic concept needs to be amended so that steady-state or average availability is determined for each bootstrap replication i with $i = 1(1)B$, as opposed to point-availability. The calculation of steady-state availability is performed using (7), whereas the average availability is calculated according to (8). In total, B realizations of the value of steady-state availability $\tilde{A}_{\infty,i}$ or B realizations of the curve of average availability $\tilde{A}_{Av,i}(t)$ are generated.

Exactly as with point-availability, both the B realizations of steady-state availability and the B realizations of average availability are evaluated statistically. An empirical, standard or percentile bootstrap confidence interval can thus be constructed.

5.3. Sample Calculation

In the following section, a sample calculation using the BMCS is presented. The sample calculations determine point-availability, steady-state availability and average availability with a 90% confidence level. The availability values are calculated with a 90% confidence interval for the sample failure and repair behavior listed in Table 3.

Table 3. Sample failure and repair behavior.

Failure behavior $F(t)$	**Repair behavior $G(t)$**
• Weibull-distributed • $b = 3.0$ • $\eta = 1{,}119.85$ h • $MTTF = 1{,}000$ h • $n_f = 10$	• Sample ($n_g = 20$) of repair times [h]: 154.0 209.5 246.0 291.5 171.5 217.0 254.0 304.0 183.0 224.0 262.0 320.0 193.0 231.0 271.0 342.5 201.5 239.0 280.5 382.0

The failure behavior $F(t)$ is given as a cdf, whereby the repair behavior $G(t)$ is based on a given sample. As shown in Table 3, the sample sizes of the failure and repair behavior are not equal. The BMCS is carried out with 200 bootstrap and 10,000 Monte Carlo replications. Within the last step of the BMCS, an empirical bootstrap confidence interval is determined.

To begin with, Figure 7 shows the $B = 200$ realizations of the point-availability calculated during the BMCS. Furthermore, the mean availability and the 90% confidence interval of the time-dependent availability are illustrated. Some empirical density functions $\varphi(\tilde{A})$ of the confidence interval are also plotted on the graph. The color of the pointed density functions clarifies the frequency n_B. In other words, this color illustrates the number of availability curves going through the point $(t_i, A(t_i))$ in absolute terms.

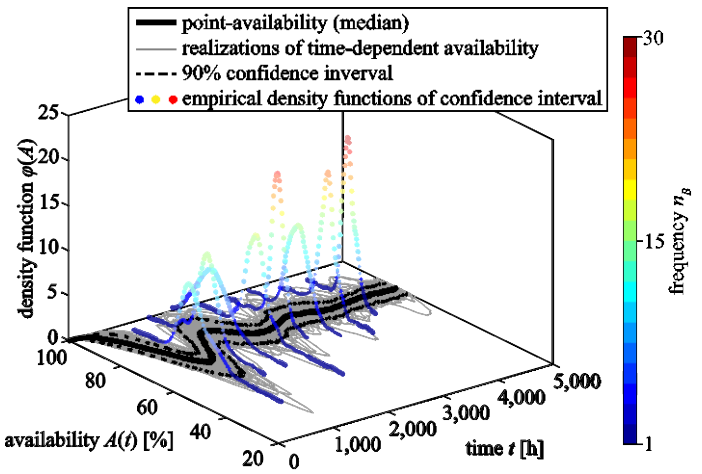


Figure 7. Point-availability with 90% confidence interval and selected density functions of confidence level.

Figure 8 builds on Figure 7 to show the steady-state availability as well as the average availability. Both availability graphs are clarified with a 90% confidence interval. The confidence interval of all graphs expands slightly with shorter durations and converges to a constant range at longer ones. For long operation times, the confidence intervals of all types of availability coincide with each other and lead to a constantly wide range.

Although it is not necessary to carry out this investigation because non-parametric bootstrapping is applied to the sample of the repair behavior, the analysis of the repair times using the Maximum Likelihood method provides a lognormally distributed repair behavior ($\mu = 5.4902$, $\sigma = 0.25$, $MTTR = 250$ h).

As shown in [1] and [6], the confidence interval of all types of availability becomes narrower or smaller as the sample sizes increase. Independently of the sample sizes, the confidence interval of point-availability, steady-state availability and average availability harmonizes very well within a constantly wide range for long operation times. A first verification of the BMCS as well as an investigation about the accuracy of the BMCS is provided by Müller et al. [30]. If the numbers of bootstrap and Monte Carlo replications are high enough, the BMCS provides acceptable results as an approximate method. First parameter studies of the input parameters of the BMCS are carried out in [1], [20] or [30].

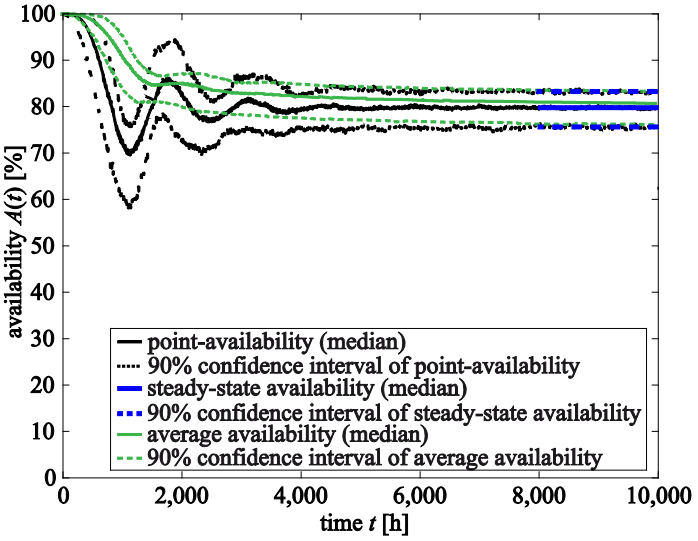


Figure 8. Point-availability, steady-state availability and average availability with 90% confidence interval.

6. Availability Demonstration Test with Consideration of Confidence Level

In this section, the inverse bootstrap Monte Carlo simulation (IBMCS) as a new procedure is presented. The IBMCS enables an availability demonstration test with consideration of confidence levels to be conducted. Firstly, the basic concept of IBMCS is presented in detail before a sample calculation is carried out.

6.1. Inverse Bootstrap Monte Carlo Simulation

An availability demonstration test needs to be carried out in order to demonstrate the quality of a repairable system. This test provides the demonstration that the system will operate with a stated level of availability. As a consequence of the availability description with confidence level using the BMCS, the availability demonstration tests also need to be specified with a confidence level. The inverse bootstrap Monte Carlo simulation (IBMCS) enables such availability demonstration tests with consideration of confidence levels. The new procedures provide a test plan comprising the required number of failures that must be observed to demonstrate a certain availability value $A(t^*)$ at a specific point in time t^*. The IMBCS can be seen as the inversion of the BMCS.

Figure 9 shows the basic concept of the inverse bootstrap Monte Carlo simulation.

The procedure begins by defining the input data. The objective is to determine the required number of observed failure and repair times, i.e., the required sample size n_f and n_g of failure and repair behavior to demonstrate the availability value $A_{objective}(t^*)$ at time t^* with confidence level $P_{A,objective}$. It is assumed that the failure behavior $F(t)$ and repair behavior $G(t)$ of the repairable system or product is known, but not its corresponding sample sizes. Furthermore, the ratio of sample sizes of failure and repair behavior is predefined:

$$n_{ratio} = \frac{n_f}{n_g}$$

Besides the ratio of the sample sizes, the minimum $n_{f,min}$ and maximum number of observed failure time $n_{f,max}$ are also predefined. These numbers correspond with the minimum and maximum number of reliability tests that could be carried out. Frequently, the maximum number of possible tests is fixed by other limitations, e.g., the number of available test benches. The minimum number is usually fixed by the corporate strategy or normative rules. These requirements frequently require a certain minimum number of reliability tests to be carried out.

Finally, the calculation parameters B (number of bootstrap replications of the BMCS) and n_{MC} (number of Monte Carlo replications of the BMCS) are required.

In the first step of the IBMCS, a BMCS is carried out based on the input data. In other words, an availability value $A_{min}(t^*)$ with confidence level $P_{A,objective}$ is calculated using the BMCS and under the assumption of the minimum sample size of failure behavior $n_{f,min}$. The same calculation is carried out under the assumption of the maximum sample size $n_{f,max}$, so that the availability value $A_{max}(t^*)$ with confidence level $P_{A,objective}$ can be determined.

It should be noted that the corresponding sample size of the repair behavior can be calculated for each step of the IBMCS using the predefined value of ratio n_{ratio} according to (22) using the actual sample size of failure behavior.

After determining both values $A_{min}(t^*)$ and $A_{max}(t^*)$, it must be verified whether the objective availability value $A_{objective}(t^*)$ lies within these ranges. In other words, a check is conducted as to whether (22) is valid or not.

$$A_{min}(t^*) \leq A_{objective}(t^*) \leq A_{max}(t^*) \qquad (22)$$

There is no solution if (22) is not valid, since the objective availability value does not lie within the interval of possible availability values $[A_{min}(t^*), A_{max}(t^*)]$. Values which are not included in this interval cannot be achieved with the predefined limitations of minimum and maximum failure time figures.

If the calculation can be advanced further, as (22) is valid, a binary search is performed to determine the required number of observed failure and repair times. Therefore, in the first step, a new sample size number is chosen according to (23), where $n_1 = n_{f,min}$ and $n_2 = n_{f,max}$.

$$n_{f,new} = \frac{n_1 + n_2}{2} \qquad (23)$$

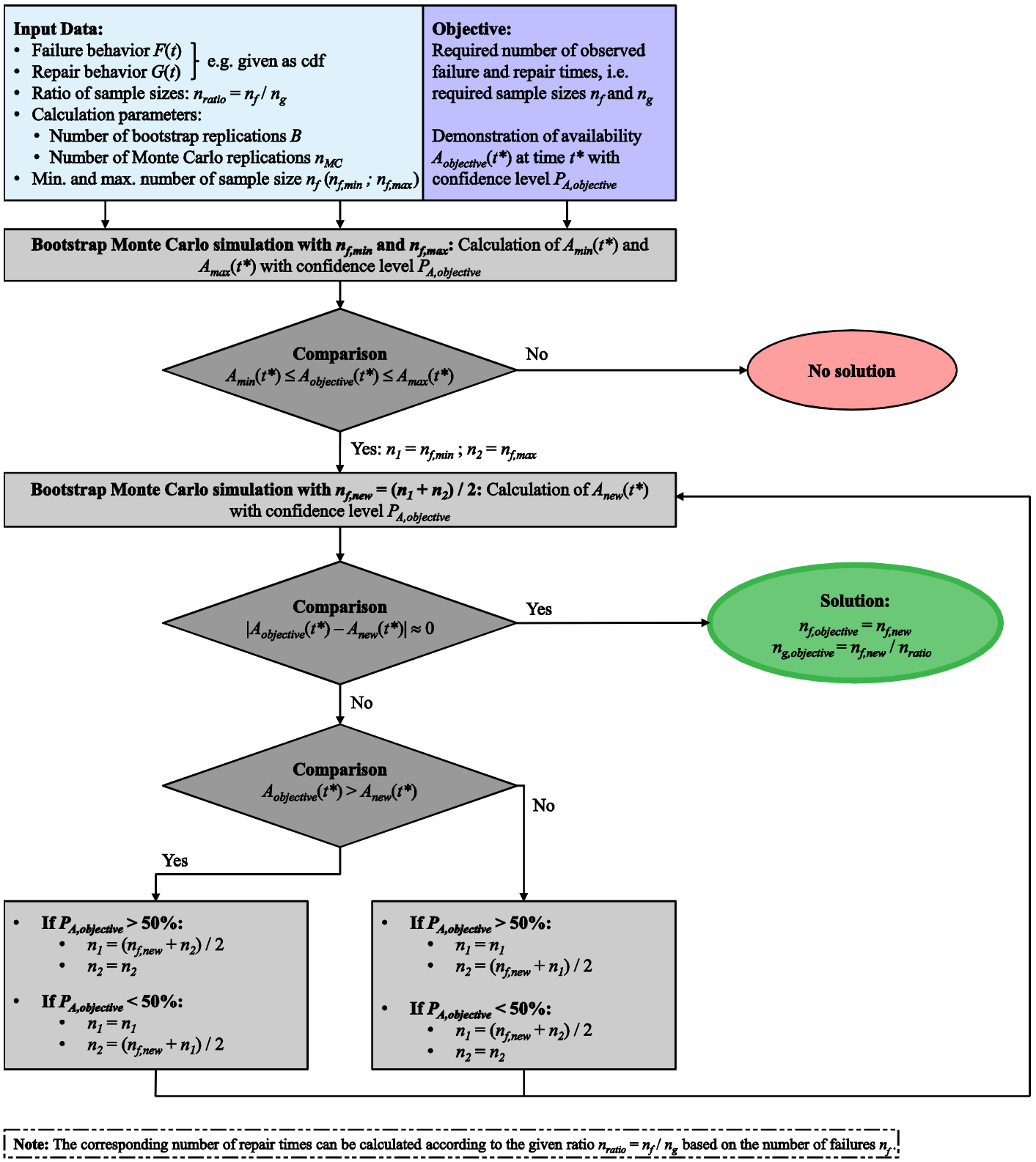


Figure 9. Basic procedure of the inverse bootstrap Monte Carlo simulation (IBMCS).

The new value $n_{f,new}$ is derived as the mean of the previous sample sizes due to the binary search. After this, a new calculation using the BMCS gives the availability value $A_{new}(t^*)$ with confidence level $P_{A,objective}$ based on the new sample size $n_{f,new}$. If this value is not approximately equal to the required demanded availability value $A_{objective}(t^*)$, the sample size requires further adaptations. Therefore, two case distinctions have to be made. One distinction investigates if the objective availability value

$A_{objective}(t^*)$ with confidence level $P_{A,objective}$ is greater than the actual calculated availability value $A_{new}(t^*)$ with confidence level $P_{A,objective}$.

Furthermore, the second distinction has to be made with regard to the demanded confidence level $P_{A,objective}$. As shown in Figure 10, the availability value with a constant confidence level increases in line with increasing sample sizes if the confidence level is lower than 50%. On the other hand, if $P_A > 50\%$, the availability value decreases as sample size n increases.

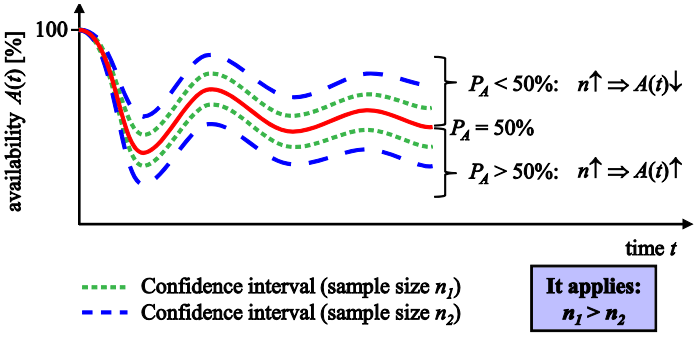


Figure 10. Confidence intervals with varying sample sizes.

After performing both case distinctions, the sample size of the failure behavior is adapted according to the equations shown in Figure 9. The corresponding number of observed repair times is calculated according to (22). A new availability value $A_{new}(t^*)$ with confidence level $P_{A,objective}$ based on the new sample sizes is subsequently determined using the BMCS.

These calculation steps are repeated until the actual determined availability value $A_{new}(t^*)$ with confidence level $P_{A,objective}$ and corresponding sample sizes $n_{f,new}$ and $n_{g,new} = n_{f,new}/n_{ratio}$ is approximatively equal to the objective availability value $A_{objective}(t^*)$ with confidence level $P_{A,objective}$. If this is the case, the required number of failures and repair actions that must be observed in order to demonstrate a certain availability value is determined as:

$$n_{f,objective} = n_{f,new} \tag{24}$$

$$n_{g,objective} = \frac{n_{f,new}}{n_{ratio}} \tag{25}$$

As a derivation of the IBMCS, it is possible to conduct an availability demonstration test with consideration of the confidence level. It is therefore possible to demonstrate a specific availability value for a point in time and a given confidence level. The procedure yields the required number of failures as well as the required number of repair actions that must be observed within a reliability test of the investigated repairable system or product with a known failure and repair behavior type. The availability test plan can subsequently be conducted and several availability requirements can be demonstrated with a confidence level.

The IBMCS can be adapted further so that it is not the value of point availability $A_{objective}(t^*)$ at time $t^*$with confidence level $P_{A,objective}$ is demanded within the objective, but rather the value of average availability $A_{Av,objective}(t^*)$ at time $t^*$with confidence level $P_{A,objective}$. In this case, the point-availability value is not calculated, but rather the average availability value within the inherent steps of BMCS. The IBMCS procedure is carried out in the same manner.

If the BMP or BRP is used within the basic procedure according to Figure 9 as opposed to the BMCS, the inverse bootstrap Markov process (IBMP) or the inverse BRP (IBRP) are established.

6.2. Sample Calculation

In the following section, a sample calculation of the IBMCS is presented. The procedure is applied to the system already investigated in the previous sample calculation. In other words, the failure behavior of the system is given from previous investigations as being Weibull-distributed with $b = 3.0$, $\eta = 1{,}119.85$ h and $MTTF = 1{,}000$ h. The repair behavior is given as being lognormally distributed with $\mu = 5.4902$, $\sigma = 0.25$ and $MTTR = 250$ h.

The IBMCS is carried out with 200 bootstrap replications and 10,000 Monte Carlo replications. It is assumed that an availability of 75% should be present at an operating time of 5,000 h. The availability should be indicated with a confidence level of 95%. It is possible to carry out a maximum of 100 reliability tests. In other words, the maximum number of sample sizes for the failure behavior is 100. The ratio between both sample sizes is predefined as 2, i.e., the sample size of the repair behavior should be half of the sample size of the failure behavior.

On the basis of the IBMCS, it is possible to determine the required sample size of failure behavior to $n_f = 16$ and the sample size of the repair behavior $n_g = 8$. Assuming that the failure behavior is known and constant as given, 16 failures need to be observed during the reliability test in order to demonstrate the demanded availability of 75% with a 95% confidence level at a time of 5,000 h. In other words, the analysis of the given failure behavior must be based on 16 failure times. Conversely, the analysis of the repair behavior must simultaneously be based on 8 observed repair times.

Figure 11 shows point-availability with a 90% confidence interval. The confidence interval is determined with input data from $n_f = 16$ and $n_g = 8$. As illustrated in this figure, the requirements of 75% availability with a 95% confidence level at an operating time of 5,000 h can be demonstrated with the given sample sizes.

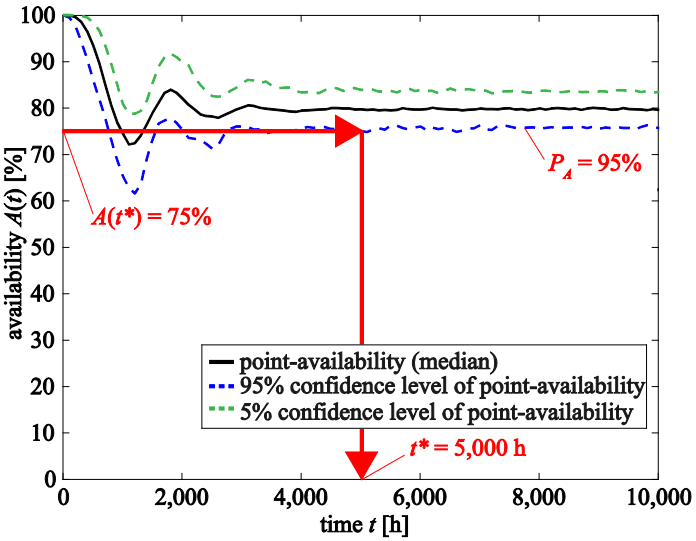


Figure 11. Point-availability with 90% confidence interval based on 16 failure and 8 repair times.

7. Potential of the Bootstrap Monte Carlo Simulation and the Inverse Bootstrap Monte Carlo Simulation

Both the bootstrap Monte Carlo simulation (BMCS) and the inverse bootstrap Monte Carlo simulation (IBMCS) have a high potential to be applied to more general availability scenarios. Beyond the consideration of preventive maintenance, the BMCS and the IBMCS are not restricted to selected maintenance strategies. It is possible to investigate technical systems with general maintenance strategies.

As given in [31], different maintenance strategies are conceivable. In general, the maintenance strategy defines the type and planning of the intended maintenance activities [31]. Possible strategies include corrective, preventive or condition-based maintenance actions. Furthermore, combined strategies are possible. Besides the aspect of maintenance, the BMCS and IMBCS have a high potential to take system aspects, failure dependencies or limited maintenance resources into consideration.

The BMCS and the IBMCS are very innovative in terms of the possibility of applying them to more general scenarios. This also includes transferring the procedures to predict and demonstrate reliability with a confidence level, e.g., for systems with periodical maintenance (see [28] or [30]).

8. Summary & Conclusions

In this paper, the bootstrap Monte Carlo simulation (BMCS) is highlighted as a new method for availability inference and demonstration with a confidence level. Besides the BMCS, the bootstrap Markov process (BMP) and the bootstrap renewal process (BRP) are briefly outlined. The BMCS facilitates analysis based on a sample of failure and repair times or on a mean value distribution with corresponding sample sizes of failure and repair behavior. The BMCS has no restrictions concerning the distribution function.

The inverse bootstrap Monte Carlo simulation (IBMCS) is presented in detail alongside the BMCS. The IBMCS enables the planning of availability demonstration tests with consideration of confidence levels. The procedure is based on the BMCS. Using the IBMCS, it is possible to determine the required number of failures and repair actions that must be observed in order to demonstrate a certain state of availability with a predefined confidence level.

Firstly, the fundamentals of reliability engineering and a basic definition of repairable systems and availability were given. After this, the bootstrap method for determining a confidence level of a distribution function was shown in more detail. After describing the BMCS in detail, a sample availability calculation with a confidence level was carried out. Based on the BMCS, the IBMCS was derived. The basic procedure was outlined on the basis of a sample calculation, too. Finally, the potential of the BMCS and the IBMCS was investigated.

The BMCS and the IBMCS have high potential for application to more general scenarios concerning general systems. Besides corrective maintenance actions, other maintenance strategies can be included within the procedure. Both the BMCS and the IBMCS are very innovative. In addition to availability with a confidence level, further parameters of reliability engineering, e.g., reliability, can be integrated into the BMCS and IBMCS.

Conflict of Interest

The authors declare that there are no conflicts of interest.

References

[1] F. Müller, P. Zeiler, B. Bertsche, "Availability Demonstration with Confidence Level Based on Reliability and Maintainability" in Proc. of Annual Reliability and Maintainability Symposium (RAMS 2017), Orlando, USA, 2017.

[2] C.E. Ebeling, An Introduction to Reliability and Maintainability Engineering, Mc Graw-Hill: New York, 1997.

[3] A. Krolo, B. Bertsche, "An Approach for the Advanced Planning of a Reliability Demonstration Test based on a Bayes Procedure", in Proc. of Annual Reliability and Maintainability Symposium (RAMS 2003), Tampa, USA, 2003.

[4] J.-H. Lim, S.W. Shin, D.K. Kim, D.H. Park, "Bootstrap Confidence Intervals for Steady-State Availability" in Asia-Pacific Journal of Operational Research, Vol. 21, No. 3, 2004, pp. 407-419.

[5] J.-C. Ke, Y.-K. Chu, "Nonparametric Analysis on System Availability: Confidence Bound and Power Function" in Journal of Mathematics and Statistics, Vol. 3, No. 4, 2007, pp. 181-187.

[6] P. Zeiler, F. Müller, B. Bertsche, "New methods for the availability prediction with confidence level" in Proc. of European Safety and Reliability Conference (ESREL 2016), Glasgow, Scotland, 2016.

[7] A. Coppola, "Some observations on demonstrating availability" in RAC Journal (the Journal of the Reliability Analysis Center), Vol. 6, No. 1, 1998, pp. 17-18.

[8] W. Wang, D.B. Kececioglu, "Confidence Limits on the inherent availability of equipment" in Proc. of Annual Reliability and Maintainability Symposium (RAMS 2000), Los Angeles, USA, 2000.

[9] V.P. Singh, S. Swaminathan, "Sample Sizes for System Availability", in Proc. of Annual Reliability and Maintainability Symposium (RAMS 2002), Seattle, USA, 2002.

[10] J.S. Usher, G.D. Taylor, "Availability Demonstration Testing" in Quality and Reliability Engineering International, Vol. 22, 2006, pp. 473-479.

[11] B. Le, J. Andrews, C. Fecarotti, "A Petri net model for railway bridge maintenance" in Proc. of the Institution of Mechanical Engineering Part O: Journal of Risk and Reliability, Vol. 231(3), 2017, pp. 306-323.

[12] G.A. Pérez Castañeda, J.-F. Aubry, N. Brinzei, "Stochastic hybrid automata model for dynamic reliability assessment" in Proc. of the Institution of Mechanical Engineering Part O: Journal of Risk and Reliability, Vol. 225(1), 2011, pp. 28-41.

[13] B. Bertsche, Reliability in Automotive and Mechanical Engineering, Springer: Berlin, Heidelberg, 2009.

[14] F. Beichelt, Reliability and Maintenance. Networks and Systems, CRC Press: Boca Raton, 2012.

[15] F. Beichelt, Applied Probability and Stochastic Processes, CRC Press: Boca Raton, 2016.

[16] A. Fritz, "Berechnung und Monte-Carlo Simulation der Zuverlässigkeit und Verfügbarkeit technischer Systeme (Calculation and Monte Carlo simulation of reliability and availability of technical systems)", PhD Thesis, University of Stuttgart, 2001.

[17] A. Fritz, P. Pozsgai, B. Bertsche, "Notes on the Analytic Description and Numerical Calculation of the Time Dependent Availability" in M. Nukulin, N. Limnios (eds), Proc. International Conference on Mathematical Methods in Reliability (MMR 2000), Bordeaux, France, 2000, pp. 413-416.

[18] R.Y. Rubinstein, D.P. Kroese, Simulation and the Monte Carlo Method, John Wiley & Sons, Inc.: New Jersey, 2008.

[19] E. Zio, The Monte Carlo Simulation Method for System Reliability and Risk Analysis, Springer: London, 2013.

[20] F. Müller, "Integration der Aussagewahrscheinlichkeit in die Berechnung und Simulation der Verfügbarkeit (Integration of the Confidence Level in the

Calculation and Simulation of Availability)", Master Thesis, University of Stuttgart, 2015.

[21] VDI-Guideline 4008-6, Monte-Carlo-Simulation, VDI: Düsseldorf, 1999.

[22] P. Poszgai, "Realitätsnahe Modellierung und Analyse der operative Zuverlässigkeitskennwerte technischer Systeme (Close-to-Reality Modelling and Analysis of the Operational Reliability Characteristics of Technical Systems)", PhD Thesis, University of Stuttgart, 2006.

[23] P. Zeiler, B. Bertsche, "Component reliability allocation and demonstration test planning based on system reliability confidence limit" in Proc. European Safety and Reliability Conference (ESREL 2015), Wroclaw, Poland, 2015.

[24] G. Yang, Life Cycle Reliability Engineering. John Wiley & Sons, Inc.: Hoboken, New Jersey, 2007.

[25] B. Efron, R.J. Tibshirani, An Introduction to the Bootstrap. Champ & Hall / CRC: Baca Raton, London, New York, Washington, 1993.

[26] C.E. Lunneborg, Data Analysis by Resampling: Concepts and Applications, South Western, USA, 2000.

[27] B. Efron, "Bootstrap Methods: Another Look at the Jackknife" in The Annals of Statistics, Vol. 7, No.8, 1979, pp. 1-26.

[28] F. Müller, P. Zeiler, B. Bertsche, "Bootstrap-Monte-Carlo-Simulation von Zuverlässigkeit und Aussagewahrscheinlichkeit bei periodischer Instandhaltung (Bootstrap Monte Carlo Simulation of Reliability with Confidence Level regarding Periodical Maintenance)" in 28. VDI-Fachtagung Technische Zuverlässigkeit 2017, Entwicklung und Betrieb zuverlässiger Produkte, VDI-Berichte 2307, VDI: Düsseldorf, 2017, pp. 69-81.

[29] W.Q. Meeker, L.A. Escobar, Statistical Methods for Reliability Data. John Wiley & Sons, Inc.: New York, Chichester, Weinheim, Brisbane, Singapore, Toronto, 1998.

[30] F. Müller, P. Zeiler, B. Bertsche, "Bootstrap Monte Carlo Simulation of Reliability and Confidence Level with Periodical Maintenance" in Forschung im Ingenieurwesen, 2017, DOI 10.1007/s10010-017-0220-6, pp. 1-11 (online first article).

[31] P. Zeiler, B. Bertsche, "Availability modelling and analysis of an offshore wind turbine using Extended Coloured Stochastic Petri Nets" in Proc. of European Safety and Reliability Conference (ESREL 2014), Wroclaw, Poland, 2014.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1577-1584 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

A Reliable, Non-Invasive Approach to Data Center Monitoring and Management

Moises Levy*[1], Jason O. Hallstrom[2]

[1] *Department of Computer & Electrical Engineering and Computer Science, Florida Atlantic University, 33431, FL, USA*

[2] *Institute for Sensing and Embedded Network Systems Engineering, Florida Atlantic University, 33431, FL, USA*

ARTICLE INFO

Article history:
Received: 30 May, 2017
Accepted: 29 July, 2017
Online: 21 August, 2017

Keywords:
Data centers
DCIM
Infrastructure Monitoring
Management
Performance
Efficiency
Internet Of Things
Control Systems
Wireless Sensors
Sensing

ABSTRACT

Recent standards, legislation, and best practices point to data center infrastructure management systems to control and monitor data center performance. This work presents an innovative approach to address some of the challenges that currently hinder data center management. It explains how monitoring and management systems should be envisioned and implemented. Key parameters associated with data center infrastructure and information technology equipment can be monitored in real-time across an entire facility using low-cost, low-power wireless sensors. Given the data centers' mission critical nature, the system must be reliable and deployable through a non-invasive process. The need for the monitoring system is also presented through a feedback control systems perspective, which allows higher levels of automation. The data center monitoring and management system enables data gathering, analysis, and decision-making to improve performance, and to enhance asset utilization.

1. Introduction

There is significant room for improvement in current practices that address data center performance relative to benchmarks. A novel data center infrastructure monitoring approach was introduced in a paper originally presented at the 7th IEEE Annual Computing and Communication Workshop and Conference (CCWC) in 2017 [1], for which this work is an extension.

A data center is defined as a facility with all the resources required for storage and processing of digital information, and its support areas. Data centers comprise the required infrastructure (e.g., power distribution, environmental control systems, telecommunications, automation, security, fire protection) and information technology equipment (e.g., servers, data storage, and network/communication equipment). Data centers are complex and highly dynamic. New equipment may be added, existing equipment may be upgraded, obsolete equipment may be removed, and both prior and current generation systems may be in use simultaneously. Increasing reliance on digital information makes data centers indispensable for assuring data is permanently available and securely stored.

Data centers can consume 40 times more energy than conventional office buildings. IT equipment by itself can consume 1100 W/m^2 [2], and its high concentration in data centers results in higher power densities. By 2014, there were around 3 million data centers in the U.S., representing one data center for every 100 people [3]. These data centers consumed approximately 70 billion kWh, or 1.8% of total electricity consumption [4]. Data center energy consumption experienced a nearly 24% increase from 2005-2010, 4% from 2010-2014, and is projected to grow by 4% from 2014-2020. In 2020, U.S. data centers are projected to consume about 73 billion kWh [4]. Emerging technologies and other energy management strategies may reduce projected energy demand.

Technical issues and human errors may threaten a data center. Power and environmental issues are especially important, and might affect performance. Human errors can be diminished through monitoring, automation, and control systems. Intangible losses, such as business confidence and reputation are difficult to

* Corresponding Author: Moises Levy. Florida Atlantic University, USA
Email: mlevy2015@fau.edu

quantify. When problems arise, data centers may shut down for a period of time. The cost of downtime depends on industry, potentially reaching thousands of dollars per minute. The average cost associated with data center downtime for an unplanned outage is approximately $9,000 per minute, an increase of 60% from 2010 to 2016 [5].

According to legislation, standards, best practices, and user needs, data centers must satisfy stringent technical requirements in order to guarantee reliability, availability, and security, as they have a direct impact on asset utilization and data center performance. In order to comply with such requirements, equipment needs continuous monitoring, control, and maintenance to assure it is working properly. A data center infrastructure management (DCIM) system can be summarized as collectively comprising a suite of tools that can monitor, manage, and control data center assets [6].

The emerging Internet of Things (IoT) makes it possible to design and implement a low-cost, non-invasive, real-time monitoring and management system. IoT can be understood as a global infrastructure enabling the interconnection of physical and virtual "things" [7]. Gartner forecasts more than 20 billion connected things by 2020 [8]. According to Cisco, 50 billion devices will be connected to the internet by 2020, and 500 billion by 2030 [9]. Each device includes sensors collecting and transmitting data over a network. These new trends help to reevaluate how to monitor and manage data center infrastructure and IT equipment. IoT makes it possible for a data center monitoring and management system to be deployed rapidly through a non-invasive process, helping to establish an integrated strategy towards performance improvement, while lowering risk.

2. Background

Since the deregulation of telecommunications in the U.S. (Telecommunications Act of 1996), a number of standards and best practices related to telecommunications, and more recently to the data center industry have emerged. Appropriate standards are essential in designing and building data centers, including national codes (required), local and country codes (required), and performance standards (optional). Governments have enforced regulations on data centers depending on the nature of the business. Different entities have also contributed to generate standards, best practices, and recommendations for the industry. These organizations include: U.S. Department of Energy; U.S. Environmental Protection Agency; U.S. Green Building Council; European Commission; Japan Data Center Council; Japan's Green IT Promotion Council; International Organization for Standardization; International Electrotechnical Commission; American National Standards Institute; National Institute of Standards and Technology; the Institute of Electrical and Electronics Engineers; National Fire Protection Association; Telecommunication Industry Association; Building Industry Consulting Service International; American Society of Heating, Refrigerating and Air-Conditioning Engineers; the Green Grid; and the Uptime Institute.

Recent standards for data centers make general recommendations that should be taken into consideration for data center infrastructure management. The ANSI/BICSI 002 standard "Data Center Design and Implementation Best Practices" [10] provides a chapter for DCIM with general recommendations for components, communication protocols, media, hardware, and reporting. The ISO/IEC 18598 standard "Information Technology - Automated Infrastructure Management (AIM) Systems - Requirements, Data Exchange and Applications" [11] includes a section on DCIM. Intelligent Infrastructure Management systems comprise hardware and software used to manage structured cabling systems based on the insertion or removal of a patch (or equipment) cord. This standard should increase operational efficiency of workflows, processes, and management of physical networks, helping to provide real-time information related to connectivity from cabling infrastructure (e.g., switches, patch panels).

Legislators have been increasingly concerned about how data centers perform. Federal data centers alone consume about 10% of total U.S. data center consumption. The Energy Efficiency Improvement Act of 2014 (H.R. 2126) [12] forces federal data centers to review energy efficiency strategies, with the ultimate goal of reducing power consumption and increasing efficiency and utilization. Likewise, in 2016, the Data Center Optimization Initiative required federal data centers to reduce their power usage below a specified threshold by September 2018, unless they are scheduled to be shutdown as part of the Federal Data Center Consolidation Initiative. To help data centers comply with legislation faster and more effectively, manual collection and reporting must be eliminated by the end of 2018. Automated infrastructure management tools must be used instead, and must address different metric targets, such as energy metering, power usage effectiveness, virtualization, server utilization, and facility monitoring [13]. The U.S. Data Center Usage Report of June 2016 [4] highlights the need for future research on performance metrics that better capture efficiency. Most recently, the Energy Efficient Government Technology Act (H.R. 306) [14] of 2017 required the implementation of energy efficient technology in federal data centers, including advanced metering infrastructure, energy efficient strategies, asset optimization strategies, and development of new metrics covering infrastructure and IT equipment.

Data center infrastructure management systems have been driven mainly by private industry. Previous experience in building automation systems, building management systems, energy management systems, facilities management systems, and energy management and control systems have enabled private industry to deploy similar solutions for the data center sector. However, these are not solutions explicitly designed for data centers. The leading companies who offer data center infrastructure management tools include Nlyte Software, Schneider Electric, and Emerson Network Power [15]. In addition, there are several reports from private industry with general guidelines and recommendations [16-24].

Academic work in data center infrastructure management includes research for specific areas, such as energy management systems [25]; distribution control systems [26]; modeling and control [27]; power efficiency [28-29]; power consumption models [30-31]; power management [32]; data center networks [33]; wireless network optimization [34-36]; thermal and computational fluid dynamics models [37-38]; wireless sensors for environmental monitoring [39]; power monitoring [40]; corrosion detection [41]; and improvement strategies, such as big data [42], genetic algorithms [43], and machine learning [44]. No previous work presents guidelines to develop a comprehensive data center monitoring and management system, considering recent advances in technology and the mission critical nature of this highly dynamic sector.

3. Data Center Monitoring and Management

3.1. Concept of Operations

This section describes quantitative and qualitative characteristics and objectives of the data center monitoring and management system. The system is envisioned to be reliable, non-invasive, wireless, sensor-based, low-cost, and low-power. It provides real-time measurements that enable the creation of a real-time data map for each parameter. It measures power usage across equipment, environmental and motion parameters across the physical space of the data center, and IT resource usage. This underscores the need for developing innovative approaches to reliable and non-invasive real-time monitoring and management.

Data center performance analysis requires understanding of facilities and IT equipment. Table 1 outlines the main components of a data center, categorized by power delivery and power consumption (e.g., IT equipment, environmental control systems, and other).

Table 1: Components of a data center

Component	Description
Power delivery systems	Different sources of energy Generators Switch gear Automatic transfer switches Electrical panels UPS systems Power distribution units Remote power panels Distribution losses external to IT equipment
IT equipment	Servers Storage systems Network / communication equipment
Environmental control systems	Chillers Cooling towers Pumps Ventilation Computer room air handling units Computer room air conditioning units Direct expansion air handler units
Other	Structured cabling systems Automation systems Lighting Security Fire systems

Real-time data collection can add value in several areas, including data center infrastructure control and management; performance optimization; enhancement of availability, reliability and business continuity [45]; reduction of total cost of ownership [46]; optimization of asset usage; and predictive analytics. A real-time monitoring and management system can highlight areas to reduce energy consumption, increase efficiency, improve productivity, characterize sustainability, and boost operations. The practice may also help to avoid or defer expensive facility expansion or relocation, via optimization of the current facility.

At least two classes of sensing devices are needed for data gathering: power sensing devices and environmental / motion sensing devices. For power sensing, we focus on non-invasive methods. Environmental data such as temperature, humidity, airflow, differential air pressure, lighting, moisture, motion, vibration, and cabinet door closure measurements are also required. Table 2 summarizes the requirements for these sensing devices.

Table 2: Requirements for sensing devices

Requirement	Description
Reliability	Due to the dynamic nature of data centers, the equipment shall be reliable and easy to install, relocate, or replace.
Wireless	Wireless data transmission using a radio transceiver reduces the use of cables in the data center. To avoid interference with existing equipment in the data center, wireless monitoring devices should have low-power transmissions, low data transmission rates, and non-overlapping frequency ranges [47].
Real-time	Sensing data in real-time helps to build situational awareness tools, such as reports and dashboards, which lead to immediate and more informed actions, reducing downtime and operating expenses.
Low-power	Low-power sensing devices enable an extended lifecycle before replacing batteries, reducing maintenance.
Battery operated	Battery operated devices eliminate the risk of connecting devices to an outlet, which could conflict with critical loads.
Low-cost	Low-cost microcontrollers and sensors make monitoring technologies available to all data centers.

Obtaining real-time data can be challenging, especially in existing data centers that lack adequate instrumentation [48]. For sampling purposes, several factors must be considered. If the sampling rate is too high, excessive data could be collected. Conversely, if the sampling rate is too low, relevant data for real-time monitoring and prediction may be lost. Collecting the right data and understanding its nature is more important than simply collecting more data [45]. We must ensure that required information is not lost. In many cases, new data may be required; the system should be flexible, allowing addition of new sensors and retrieval of additional data from equipment. After sensing and gathering the required data, it must be stored, locally or remotely, for real-time processing and future analysis. Data analysis leads to more informed decisions.

3.2. Data Collection and Transmission

For data collection, different classes of sensing devices are required. It is also necessary to gather IT equipment resource utilization data through processor measurements.

Power Measurements:

For power consumption measurements, all data center IT equipment and its supporting infrastructure must be considered. This task could be accomplished in many ways, but we are interested in reliable and non-invasive methods. Most data center components, such as uninterruptible power supplies, power distribution units, remote power panels, electrical panels, and cooling systems offer real-time data through existing communication interfaces. To avoid additional cabling, the use of wireless transmission must be the first option.

A standard protocol should be used to access data on remote devices, such as the Simple Network Management Protocol (SNMP) over TCP/IP. In addition, appropriate security measures must be implemented, since the data retrieval feature could create a security vulnerability. For equipment without a communication interface, one option is to install a specific power meter with the required communication interface. It will be an additive component, applicable across different device types, reducing operational risk. Another alternative is to replace existing equipment with new units that include metering and communication interfaces.

Software developed for data collection must gather and filter data collected from sensing devices and transmit the data to one or more storage devices. All monitored devices must be evaluated to define the required parameters and sampling rates.

Environmental / Motion Measurements:

The described system measures environmental and motion parameters across the physical space of the data center. Equipment cabinets, and various measurement points at different heights must be considered to capture relevant environmental variables. Different areas without equipment, such as below the raised floor, the aisles, and other points, must also be taken into consideration.

Table 3. Parameters to be monitored

Parameter	Description	Sensing
Temperature	Acceptable range required by equipment, and to optimize cooling systems and enhance efficiency.	Across the data center. In each IT cabinet: front of the cabinet (bottom, middle, and top) may be sensed for cold air inlet, and back or top for exhaust.
Humidity	Acceptable range required to reduce electrostatic discharge problems (if humidity is low), and condensation problems (if humidity is high).	Across the data center. In each IT cabinet: in the middle of the air inlet.
Airflow	Equipment and infrastructure require airflow, as specified.	Below the raised floor. Above perforated tiles.
Differential air pressure	Differential air pressure analysis could help to optimize airflow and cooling systems. If differential pressure between two areas is low or high, it indicates the airflow rate is also low or high.	Across the data center. Sensors are installed between two areas with varying air pressure (e.g., below and above the raised floor, inside and outside air plenum areas, between hot and cold aisles).
Water	Water sensing is required to prevent failures from cooling equipment, pipes, or external sources.	Near cooling system, water pipes, equipment, or potential risk areas.
Vibration	In case of earthquakes or excessive vibration, vibration sensors can trigger appropriate alerts. The sensors could also detect unauthorized installation, removal, or movement of equipment.	In each cabinet.
Lighting	Different levels of lighting according to human occupancy or security requirements.	Across the data center.
Security	Motion sensors could detect movement around cabinets. Contact closure sensors to enhance security systems. Airflow, temperature, and humidity in equipment could be affected if a cabinet door is opened, or if people are around the perforated tiles. Contact closures for access doors may be checked and tied to temperature variations nearby.	Door contact closures for cabinets and access doors, motion sensors, and surveillance cameras located nearby.
Fire systems	Fire alarms must be considered.	Data from fire system.

Each sensed parameter covers a specific range, and depending on the equipment, the range varies. Newer technologies allow a broader range of environmental parameters. ASHRAE Technical Committee 9.9 has created a set of guidelines regarding the optimal and allowable range of temperature and humidity set points for data centers [49-51]. Manufacturers have also established the environmental optimal operating ranges for their equipment, which enables a more detailed analysis of operational ranges. The standard ANSI/BICSI-002 [10] recommends three different lighting levels (level 1 for security purposes, level 2 when unauthorized access is detected, and level 3 for normal conditions).

The parameters to be monitored across the data center are described in Table 3. All parameters must be analyzed to define sampling rates. Contingent on specific requirements, additional inputs or sensors may be added.

As shown in Figure 1, after sensor data is acquired, transmitters are responsible for relaying data from all sensing devices to a storage device. The transmitter must be capable of storing data for an established period of time in case transmission problems occur, to avoid losing data.

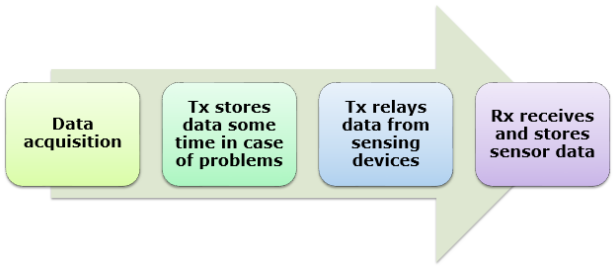


Figure 1: Data collection and transmission workflow

Given the limited (battery) energy available to each device, one data acquisition challenge is to maximize network lifetime, and to balance the use of energy among nodes. Network lifetime can be defined as the time the network is able to perform tasks, or as the minimum time before any node runs out of energy and is unable to transmit data. A multihop routing protocol should be used, ensuring there is an active path between each network device and the network's sink [52].

The following aspects associated with the wireless sensor nodes deserve attention: device location, which depends on data center layout; new device additions or relocations, where communication is not disrupted; failure diagnostics (e.g. energy problems, disconnections); and redundancy requirements across critical tasks.

Processor Measurements:

Enhancements by some IT equipment manufacturers include the ability to directly access some measurements related to power, surrounding environment, and resource utilization for each device. These measurements could include parameters such as power utilization, air inlet temperature, airflow, outlet temperature, CPU usage, memory usage, and I/O usage. Platform level telemetry helps transform data center monitoring and management, allowing direct data access from the equipment processor. Comparison of processor measurements to environmental and power measurements, as well as analysis of the IT resource usage, can improve decision-making [53].

With IT resource utilization data available, IT resource optimization through workload allocation and balancing, as well as through other control strategies, is feasible. These measurements may not be available for all equipment. Querying

this information is unlikely to interfere with IT equipment's primary tasks.

3.3. Redundancy

Redundancy can be understood as the multiplication of systems to enhance reliability. Since data center components require maintenance, upgrade, replacement, and may fail, redundancy is required to reduce downtime. Critical components of the monitoring and management system may also fail, and sensitive data could be lost. To evaluate the need for monitoring and management redundancy, several factors must be considered.

The data center monitoring and management system should be designed according to the reliability tier of the data center and its related redundancy level [54]. If the data center presents redundancy (e.g., N+1, 2N, 2N+1), the monitoring and management system should provide the same redundancy level to preserve reliability. Conversely, if a data center has no redundancy in its IT equipment and supporting infrastructure (e.g., electrical, mechanical), redundancy may not be required. In addition, for the evaluation of the required redundancy of monitoring and management devices, mission critical IT equipment in the data center should be identified.

Data centers can use thousands of sensors to gather required data. For that reason, it is convenient to have general estimations of the number of sensing devices required for a given data center. Although the best way to calculate this is to analyze each data center in detail, there are guidelines to make preliminary estimates.

The first step is to estimate how many cabinets fit in the data center layout. Initially, the data center layout and information about the equipment must be examined. Most IT equipment is installed in a cabinet or open rack, with supporting rails 19in wide (Electronic Industry Alliance industry standard); adding space required for panels, doors, wiring and airflow increases the width to between 24in and 30in. The typical depth ranges from 36in to 48in, depending on the equipment. Standalone equipment may require special cabinets, which can increase the previous estimates. Considering a raised floor grid of 24in x 24in (industry standard), the cabinet's area can be estimated as 24in x 48in, or 8 ft^2. A similar area per cabinet is required for hallways, infrastructure facilities, and other non-IT equipment. The second step is to estimate how many sensing devices are required. Based on the previous step, for each cabinet, or for each 16 ft^2, at least one environmental / motion sensing device will be required. Power sensing devices must also be considered; the total count can be estimated as the total number of elements (e.g., electrical facilities, cooling and ventilation systems) that support IT equipment. In addition, depending on spatial distribution, at least one transmitter should be in place to gather and transmit sensed data for storage and further processing. The final step is to review redundancy requirements and to estimate the total number of sensing and transmitter devices needed.

3.4. Data Storage, Processing, and Management

As shown in Figure 1, after gathering data, it is sent to the storage device, locally or remotely, to support future processing and reporting. The large amount of data collected must be organized to guarantee that it can be integrated, shared, analyzed, and maintained. A metadata management system is suggested. Metadata is understood as structured information that describes an information resource. It is, in effect, data about data. A metadata management system ensures that metadata is added correctly and provides mechanisms to optimize its use. A software interface must be developed to process and manage the information. Figure 2 shows a diagram of the different elements involved in data storage, processing, and management.

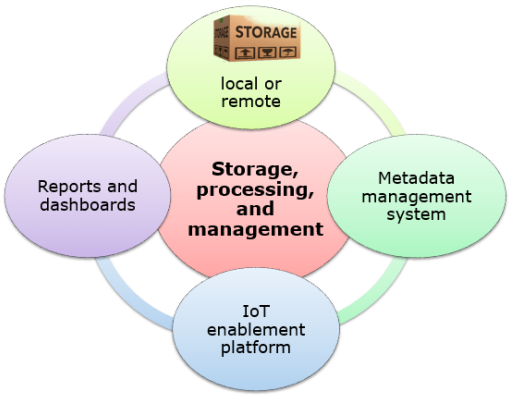


Figure 2: Storage, processing, and management

The use of an IoT enablement platform is a viable option to connect monitoring devices anywhere to the cloud. The key added value of these platforms is to simplify connectivity, data collection, and device management.

Reports and dashboards help to visualize a data center's past and present performance. Dashboards are graphical user interfaces customized to user requirements, used to display key performance indicators to make them easier to read and interpret. Visualizing results in real-time supports responsive improvement. Warnings may be used for undesirable events. Furthermore, forecast reports can be used to predict how the data center will perform in the future.

Numerous reports can be generated from the data gathered by the monitoring and management system, such as real-time data reports for existing conditions; historical reports showing behavior with baselines and trends; statistical reports; and analytical reports showing asset utilization, performance, and other key indicators. In summary, different reports and dashboards translate data into information and strategic knowledge.

Proactive strategies are more desirable than reactive actions. Strategies may be undertaken based on what has been learned. Results will be measured as the system continues to collect data, reflecting the effectiveness of prior actions. This feedback often constitutes the basis for new actions [45].

3.5. Metrics

Data centers can be evaluated in comparison to the goals established or to similar data centers through the use of different metrics, as the performance cannot be understood with the use of a single metric [48]. Inconsistencies or variations in measurements can produce a false result for a metric, and for that reason, it is very important to standardize metrics. After thoroughly reviewing existing data center metrics, four areas are identified through which data center performance can be measured: efficiency, productivity, sustainability, and operations [55]. Table 4 presents general concepts for these metrics.

Table 4. Performance measurements

Metric	Concept
Efficiency	Due to the high-energy consumption of the data center sector, efficiency has been given substantial attention. Efficiency indicators show how energy efficient site infrastructure, IT equipment, environmental control systems, and other systems are.
Productivity	It gives a sense of work accomplished and can be estimated through different indicators, such as the ratio of useful work completed to energy usage, or useful work completed to the cost of the data center. Useful work can be understood as the sum of weighted tasks carried out in a period of time, such as transactions, amount of information processed, or units of production. The weight of each task is allocated depending on its importance.
Sustainability	It is defined as development that addresses current needs without jeopardizing future generations' capabilities to satisfy their own needs [56]. Measurements for sustainability include carbon footprint, the greenhouse gas footprint, and the ratio of green energy sources to total energy.
Operations	Measurements gauge how well managed a data center is. This must include an analysis of operations, including site infrastructure, IT equipment, maintenance, human resources training, and security systems, among other factors.

The process to gather data to calculate metrics can easily be automated, as many parameters are updated automatically through the monitoring and management system. This allows some key indicators to be updated in real time. Other parameters are not easily automated. Instead, they derive from audits, human observation, evaluation, and analysis, and must be updated if and when a process changes. This is the case for some aspects of sustainability and operations. In addition, new variables will be measured, and new metrics will be created, as needs and technologies evolve.

Data center performance is the basis for designing and implementing plans aimed at improving key indicators. Asset usage is an important component of performance measurements, as it reflects how adequately the different data center assets are being used at specific points in time. Improving specific performance indicators leads to asset usage optimization.

3.6. New Tools for Data Analytics

Considering the large amount of data collected, new tools must be developed to help users employ historical and real-time data, together with know-how, to undertake assertive actions. Predictive modeling enables simulation and forecasting of how a data center will perform in the future. As shown in Figure 3, combining real-time data from the monitoring and management system, know-how, and new tools (e.g., simulation, modeling, machine/deep learning) will lead to predictive analytics for the data center and its components. Simulations show how different components could perform or fail, allowing more time to evaluate and mitigate problems.

Simulation is an important tool for model validation, and for analyzing different scenarios. Models are safer, faster, and more affordable than field implementation and testing. Simulations help to explore how systems might perform under specific scenarios, including disturbances, equipment failure, different layout configurations, or specific control strategies. Historical data must be used to validate and calibrate the models developed for each specific data center. To transform data center predictive modeling from probabilistic to physics-based, tools such as computational fluid dynamics models for airflow, temperature, humidity, and other parameters must be used. Simulations can also help to understand how adaptive and resilient a data center is. Resiliency can be understood as the ability of a data center and its components to adapt existing resources to new situations.

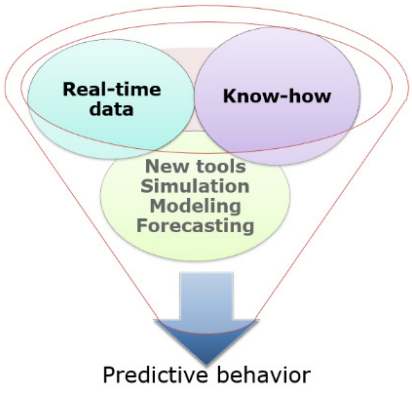


Figure 3: Predictive Analytics

To recognize new insights or predict future behavior, tools such as predictive analytics and machine/deep learning are needed. These tools involve algorithms that can learn from and make predictions about data without being explicitly programmed. Models should be constantly updated with real-time data to make predictions based on new realities. These tools lead to planned scenarios, rather than emergencies or failures, enabling more proactive decision-making.

3.7. End-to-end Resource Management

Data center optimization strategies should include all assets - power and cooling equipment, and IT equipment resources - to support end-to-end resource management. Resource management is iterative, enabling data center managers to plan and make adjustments. Examples include physical space planning for replacing, upgrading, or relocating equipment; detection of unused or idle IT equipment; capacity planning; optimization of cooling systems; power system optimization; IT resource optimization, including servers, storage, and networking equipment; workload allocation to control power and environmental parameters; workload balance among IT equipment; and virtualization. Comprehensive control strategies allow performance optimization.

4. Data Center Modeling and Control Systems

With increasing levels of automation in the data center, the role of control systems is growing. Data center monitoring systems are the main components of any such feedback control system; they deliver the required information so the control system can take appropriate actions. The control system comprises components or processes interconnected to achieve a function with multiple inputs and outputs. Optimization is a desirable objective, considering data center components are dynamic and change continuously with time, or with any other independent parameter.

From a control systems point of view, the data center can be described as a closed loop system with multiple inputs and multiple outputs, controllers and actuators, as shown in Figure 4. The desired values are generally based on standards, best practices, and/or user requirements. The system or process to be controlled is the IT equipment in the data center. Various devices and sensors are required to measure the variables to be controlled, such as power, temperature, humidity, air flow, differential air pressure, closure, motion, vibration, and IT equipment resource utilization.

The actuator represents infrastructure equipment (e.g., cooling system, ventilation, security, fire system) and IT equipment resource management (e.g. virtualization, workload transfer, load balancing). The actuator will influence the process, and by consequence, the associated outputs / controlled parameters. The controller, as its name suggests, computes the control signals. The multiple outputs / controlled parameters are compared to the multiple inputs or desired values. Based on the error signal, or the difference between the desired values and the outputs, the controller takes corrective actions, affecting the actuators, so that the output values track the reference values within a desired margin.

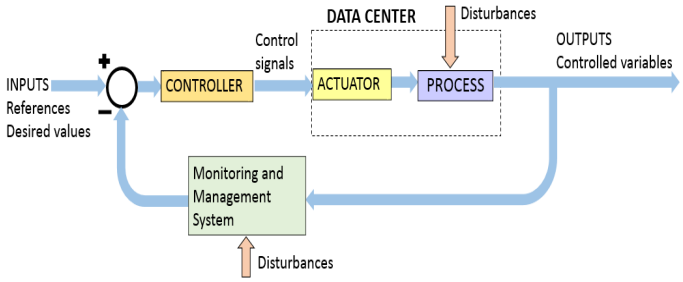


Figure 4: Data center feedback control loop block diagram

Disturbances that can affect the behavior of the processes, undesired or unforeseen, should be taken into consideration. Feedback is required for adaptive and robust control. Feedback control allows corrective action regardless of the disturbance sources and decreases the sensitivity of the controlled parameters to process changes and other disturbances.

Each of the different data center components can be modeled through a mathematical description using a macro level approach, with inputs and outputs. After identifying all the measurements and parameters required, the data center monitoring and management system must gather all data inputs and outputs from various components, from which the key indicators will be calculated. The monitoring and management system allows reliable feedback control, measuring the controlled variables to adjust the manipulated variables; and also enables validation and calibration of the proposed models.

5. Conclusions

The described data center monitoring and management approach allows rapid deployment of a reliable, real-time monitoring system to improve operations, to enhance performance, and to reduce risk of failures. Using battery operated, low-power, wireless sensing devices, and retrieving data directly from equipment, information is collected through a non-invasive, continuous process.

Collected data can be used in many ways to make more informed decisions, to reduce downtime and cost of operations, to improve performance and key indicators, to optimize asset usage, to implement more informed actions and strategies, and to develop predictive behavior models. Data center professionals must think critically to determine what data is needed for each particular case, also considering future scenarios. This practice prevents missing important measurements and helps to avoid collecting unnecessary data.

Going forward, data center monitoring and management systems must include end-to-end resource management, covering both the IT equipment and supporting infrastructure. The ability to directly access measurements related to power, environment, and IT resource utilization opens new windows for optimization. Future data centers may avoid full time on-site staff, by relying on monitoring and management systems.

References

[1] M. Levy and J. O. Hallstrom, "A New Approach to Data Center Infrastructure Monitoring and Management (DCIMM)," *IEEE CCWC 2017. The 7th IEEE Annual Computing and Communication Workshop and Conference*. Las Vegas, NV, 2017.

[2] S. Greenberg, E. Mills, and B. Tschudi, "Best Practices for Data Centers : Lessons Learned from Benchmarking 22 Data Centers," *2006 ACEEE Summer Study on Energy Efficiency in Buildings*. Pacific Grove, CA, pp. 76–87, 2006.

[3] U.S. Department of Energy. Office of Energy Efficiency & Renewable Energy, "10 Facts to Know About Data Centers," 2014. [Online]. Available: http://energy.gov/eere/articles/10-facts-know-about-data-centers.

[4] Ernest Orlando Lawrence Berkeley National Laboratory, "United States Data Center Energy Usage Report. LBNL-1005775," 2016.

[5] Ponemon Institute LLC, "Cost of Data Center Outages. Data Center Performance Benchmark Series," 2016.

[6] ECO Datacenter Expert Group - Association of the German Internet Industry, "Data Center Infrastructure Management Market Overview and Orientation Guide," 2014.

[7] "Series Y: Global information infrastructure, internet protocol aspects and next-generation networks. Next generation networks – Frameworks and functional architecture models. Overview of the Internet of Things," Recommendation ITU-T Y.2060, 2012.

[8] C. Geschickter and K. R. Moyer, "Measuring the Strategic Value of the Internet of Things for Industries. ID: G00298896," 2016.

[9] Cisco Systems Inc., "Cisco and SAS edge-to-enterprise IoT analytics platform. White paper," 2017.

[10] "Data Center Design and Implementation Best Practices," ANSI/BICSI 002, 2014.

[11] "Information technology - Automated infrastructure management (AIM) systems - Requirements, data exchange and applications," ISO/IEC 18598, 2016.

[12] The Senate of the United States. 113th Congress 2nd session, *H.R. 2126, Energy Efficiency Improvement Act of 2014*. 2014.

[13] T. Scott, "Data Center Optimization Initiative (DCOI). Memorandum," 2016.

[14] The Senate of the United States. 115th Congress (2017-2018), *H.R. 306, Energy Efficient Government Technology Act*. 2017.

[15] Gartner Inc., "Magic quadrant for Data Center Infrastructure Management Tools," 2016.

[16] Enterprise Management Associates (EMA), "Quantifying Data Center Efficiency : Achieving Value with Data Center Infrastructure Management. White Paper," 2013.

[17] Forrester Research Inc, "The total economic impact of the Emerson DCIM solution," 2013.

[18] IDC Research Inc, "White Paper, Datacenter Infrastructure Management (DCIM): Bringing Together the World of Facilities and Cloud Computing," 2011.

[19] Sunbird Software, "DCIM Software and IT Service Management - Perfect Together DCIM : The Physical Heart of ITSM. White Paper," 2015.

[20] Sunbird Software, "EO 13514 and Smart Federal Government Data Center Management. White Paper," no. July, pp. 1–8, 2015.

[21] Datacenter Research Group. A division of Datacenter Dynamics, "Monitoring efficiently? White Paper," 2011.

[22] Raritan Inc, "Virtualization and Data Center Management :," 2010.

[23] N. Rasmussen, "Electrical efficiency modeling for data centers. White paper # 113.," *Schneider Electric's Data Center Science Center*. 2011.

[24] T. K. Nielsen and D. Bouley, "How Data Center Infrastructure Management Software Improves Planning and Cuts Operational Costs. White Paper #107, Revision 3," *Schneider Electric's Data Center Science Center*. 2012.

[25] Y. Nakamura, K. Matsuda, and M. Matsuoka, "Augmented Data Center Infrastructure Management System for Minimizing Energy Consumption," *2016 5th IEEE International Conference on Cloud Networking (Cloudnet)*. Pisa, pp. 101–106, 2016.

[26] M. Wiboonrat, "Distribution control systems for data center," *2015 IEEE/SICE International Symposium on System Integration (SII)*. Nagoya, pp. 789–794, 2015.

[27] L. Parolini, B. Sinopoli, B. H. Krogh, and Z. K. Wang, "A cyber-physical systems approach to data center modeling and control for energy efficiency," *Proc. IEEE*, vol. 100, no. 1, pp. 254–268, 2012.

[28] S. McCumber and S. J. Lincke, "Tooling to improve data center efficiency," *2014 IEEE International Technology Management Conference, ITMC 2014*. Chicago, IL, pp. 1–5, 2014.

[29] S. Kaxiras and M. Martonosi, "Computer Architecture Techniques for Power-Efficiency," in *Computer Architecture Techniques for Power-Efficiency*, vol. 3, no. 1, Morgan & Claypool, 2008, pp. 1–207.

[30] M. Dayarathna, Y. Wen, and R. Fan, "Data Center Energy Consumption Modeling: A Survey," *IEEE Commun. Surv. Tutorials*, vol. 18, no. 1, pp. 732–794, 2016.

[31] S. Rivoire, P. Ranganathan, and C. Kozyrakis, "A comparison of high-level full-system power models," *Conference on Power aware computing and systems (HotPower 2008)*. San Diego, CA, pp. 1–5, 2008.

[32] L. Ganesh, H. Weatherspoon, T. Marian, and K. Birman, "Integrated Approach to Data Center Power Management," *IEEE Trans. Comput.*, vol. 62, no. 6, pp. 1086–1096, 2013.

[33] W. Xia, P. Zhao, Y. Wen, and H. Xie, "A Survey on Data Center Networking (DCN): Infrastructure and Operations," *IEEE Commun. Surv. Tutorials*, vol. 19, no. 1, pp. 640–656, 2017.

[34] K. Hong, S. Yang, Z. Ma, and L. Gu, "A Synergy of the Wireless Sensor Network and the Data Center System," *2013 IEEE 10th International Conference on Mobile Ad-Hoc and Sensor Systems*. Hangzhou, pp. 263–271, 2013.

[35] H. Shukla, V. Bhave, S. Sonawane, and J. Abraham, "Design of efficient communication algorithms for hotspot detection in data centers using Wireless Sensor Networks," *2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*. Chennai, pp. 585–590, 2016.

[36] M. Wiboonrat, "Data Center Infrastructure Management WLAN Networks for Monitoring and Controlling Systems," *The International Conference on Information Networking 2014 (ICOIN2014)*. Phuket, pp. 226–231, 2014.

[37] C. D. Patel, R. Sharma, C. E. Bash, and A. Beitelmal, "Thermal considerations in cooling large scale high compute density data centers," *ITherm 2002. Eighth Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (Cat. No.02CH37258)*, vol. 2002. San Diego, CA, pp. 767–776, 2002.

[38] C. D. Patel, C. E. Bash, and C. Belady, "Computational fluid dynamics modeling of high compute density data centers to assure system inlet air specifications," *IPACK'01. The Pacific Rim/ASME International Electronic Packaging Technical Conference and Exhibition*. Hawaii, 2001.

[39] M. G. Rodriguez, L. E. Ortiz Uriarte, Y. Jia, K. Yoshii, R. Ross, and P. H. Beckman, "Wireless sensor network for data-center environmental monitoring," *2011 Fifth International Conference on Sensing Technology*. Palmerston North, pp. 533–537, 2011.

[40] R. Khanna, D. Choudhury, P. Y. Chiang, H. Liu, and L. Xia, "Innovative approach to server performance and power monitoring in data centers using wireless sensors (invited paper)," *2012 IEEE Radio and Wireless Symposium*. Santa Clara, CA, pp. 99–102, 2012.

[41] X. Liu, Y. Hu, F. Wang, Y. Pang, and H. Liu, "A Real-Time Evaluation Method on Internet Data Centre's Equipment Corrosion by Wireless Sensor Networks," *2016 International Symposium on Computer, Consumer and Control (IS3C)*. Xi'an, pp. 622–627, 2016.

[42] M. Hubbell *et al.*, "Big Data strategies for Data Center Infrastructure management using a 3D gaming platform," *2015 IEEE High Performance Extreme Computing Conference (HPEC)*. Waltham, MA, 2015.

[43] R. Khanna, H. Liu, and T. Rangarajan, "Wireless Data Center Management: Sensor Network Applications and Challenges," *IEEE Microwave Magazine*, vol. 15, no. 7, pp. S45–S60, 2014.

[44] R. Khanna and H. Liu, "Machine learning approach to data center monitoring using wireless sensor networks," *2012 IEEE Global Communications Conference (GLOBECOM)*. Anaheim, CA, pp. 689–694, 2012.

[45] J. Koomey and J. Stanley, "The Science of Measurement: Improving Data Center Performance with Continuous Monitoring and Measurement of Site Infrastructure," 2009.

[46] J. Koomey, K. Brill, P. Turner, J. Stanley, and B. Taylor, "Uptime Institute white paper: A Simple Model for Determining True Total Cost of Ownership for Data Centers," 2007.

[47] U.S. Department of Energy Federal Energy Management Program. By Lawrence Berkeley National Laboratory, "Data Center Energy Efficiency Measurement Assessment Kit Guide and Specification," 2012.

[48] The Green Grid, "PUE™: A Comprehensive Examination of the Metric. White paper # 49," 2012.

[49] ASHRAE Technical Committee 9.9, "2011 Thermal Guidelines for Data Processing Environments – Expanded Data Center Classes and Usage Guidance," 2011.

[50] ASHRAE Technical Committee 9.9, "Data Center Networking Equipment – Issues and Best Practices," 2012.

[51] ASHRAE Technical Committee 9.9, "Data Center Storage Equipment – Thermal Guidelines , Issues , and Best Practices," pp. 1–85, 2015.

[52] S. Hussain and O. Islam, "An Energy Efficient Spanning Tree Based Multi-hop Routing in Wireless Sensor Networks.pdf," *IEEE Wireless Communications and Networking Conference*. Kowloon, pp. 4386–4391, 2007.

[53] J. L. Vincent and J. Kuzma, "Using Platform Level Telemetry to Reduce Power Consumption in a Datacenter," *2015 31st Thermal Measurement, Modeling & Management Symposium (SEMI-THERM)*. Folsom, CA, 2015.

[54] Uptime Institute, "Data Center Site Infrastructure Tier Standard : Topology," 2012.

[55] M. Levy and D. Raviv, "A Novel Framework for Data Center Metrics using a Multidimensional Approach," *15th LACCEI International Multi-Conference for Engineering, Education, and Technology: Global Partnerships for Development and Engineering Education*. Boca Raton, FL, 2017.

[56] United Nations, "Report of the World Commission on Environment and Development: Our Common Future," 1987.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1585-1591 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

An Ontology-Based Temporal Scheduling and Complexity Reduction Model for Timetabling Knowledge Management

Patience Usoro Usip*, Enobong Moses Umoren, Udoinyang Godwin Inyang, Mfonobong Maurice Ntekop

University of Uyo, Computer Science Department, Faculty of Science, 520003, Uyo, Nigeria

ARTICLE INFO

Article history:
Received: 04 June, 2017
Accepted: 21 July, 2017
Online: 24 August, 2017

Keywords:
Temporal Interval Relations
Formal Ontologies
Knowledge Management Systems
Optimal Time Scheduling
Timetabling Models

ABSTRACT

Technological tools such as Semantic Web and ontologies have contributed immensely to the reduction of most managerial complexities on daily basis. However, most resolved complexities are solely dependent on the emerging issues which are not general enough to accommodate some specific domain challenges. Temporal scheduling complexities occur in several domains including timetabling, but inter-departmental courses allocation on a general institutional timetable has not been considered. This poses serious managerial concerns and threats to the academic performance of any institution. This paper is aimed at formalizing an ontology model for analyzing temporal scheduling complexities in an existing schedule and giving optimal possible time scheduling resolutions showing their reduction rates for efficient and intelligent knowledge management. Protégé was used in modeling the domain ontology described in the resulted ontograf. The semantics of the formal ontology model is described based on Allen's interval temporal relations. The use of ontologies make way for easy and intelligent reasoning with knowledge and exposes the need for such knowledge management tool to manage the influx of the numerous data.

1. Introduction

This paper is an extension of work originally presented in the Future Technologies Conference 2016 [1], where the use of ontologies in semantic web was established as an efficient and intelligent tool for managing timetabling knowledge [2]. Domain specific concepts and some complexities were presented in figures 2, 3(a) and 3(b) of the conference paper. The Allen's interval-based temporal relations [3] were considered suitable and introduced for inferencing [4] in the sample rules used to describe possible resolutions to the highlighted complexities in the conference paper. Allen's interval-based temporal relations were introduced to handle time durations and is adapted or reused for formal description of time periods for complex knowledge management as seen in the timetable domain. However, this paper extends works done in the conference paper as it aims at formalizing an ontology model for analyzing temporal scheduling complexities in an existing schedule and providing optimal possible time scheduling resolutions showing their reduction rate for efficient and intelligent knowledge management.

*Corresponding Author: Patience Usoro Usip, Computer Science Department, University of Uyo, Uyo, +2348060546140, patiencebassey@uniuyo.edu.ng, patceeng@gmail.com

In recent years, the explicit formal specifications of terms in any given domain and relations among them have gradually moved from the realm of Artificial Intelligence (AI) laboratories to the desktops of domain experts. The representation of domain concepts in structural ways - definition of concepts (classes and properties) and mapping the relationships among the defined concepts, form the basis of Ontology [5]. From the computer science perspective, it is a formal naming and definition of the types, properties, and interrelationship of the entities that really or fundamentally exist for a particular domain of discourse [6] [7]. For example, in a lecture timetable domain, courses, timeslots, students, lecturers, venues and lectures are some of the concepts from which the timetabling application domain can be described. These concepts and their meanings together constitute ontology for timetable and can be used as common knowledge for communication among educational stakeholders and provides information for the development of a timetable information system.

Ontologies exist in diverse forms - lexicons, dictionaries, thesauri, and logical models described in languages such as first order logic (FOL). Lexicons provide standard lists of words

(vocabulary) in a language with corresponding knowledge of how each word is used. Hence a lexicon can be seen as an index for mapping written form of a word with the information about that word [8]. Dictionaries can be organized to form hierarchies-taxonomies, meronomies, mereologies and merons, according to specific relations. Related terms, linguistics objects are added to any given collection of terms through thesauri. Ontologies provide standardization of the terms used to represent knowledge about a domain in any of these forms. They can as well support inference with FOL or its subset, by deriving new facts from a collection of facts and enforcement of consistency. It also allows for sharing common understanding of the structure of information among people or software agent; reuse of domain knowledge; making domain assumption explicit; separating domain knowledge from operational knowledge; and analyzing domain knowledge. These considerations are clearly useful for knowledge management, especially when large amounts of knowledge are being processed. Ontology is increasingly used in various fields such as Knowledge Engineering, Artificial Intelligence(AI) and computer science, in applications related to knowledge management (KM), natural language processing.

KM involves the acquisition, creation, use, representation, organization and advancement of knowledge in its many forms. As a requirement for effective KM, an understanding of how individuals, groups and organizations use knowledge is needed. One major aspect of an organization that require effective management of knowledge is timetable scheduling. This is because an organization's general timetable exhibit different levels of temporal scheduling complexities depending on the available resources. The threats posed by these complexities ranges from poor organization performance as seen in the domain of discourse (poor academic performance). Knowledge management is thereby required in attempt to reduce these complexities. Automation of timetabling process as seen in existing systems [9] [10] [11] does not resolve these complexities. This paper demonstrates the application of ontologies in timetable KM. It is aimed at analyzing any existing temporal schedule to discover all complexities and providing possible temporal schedules with reduced complexities (in time and space requirement) that tends towards optimality. This is achieved by utilizing Allen's temporal interval relations described in the following sections and evaluating the resulting optimal temporal schedule for proof and the need for adoption.

2. Formal Ontologies and Temporal Interval Relations

The time ontology has been considered by several authors with unique contributions to issues relating to temporal (time) representation and reasoning among others. Hobbs and Pan [12] considered temporal relations (TR) involving two subclasses of time— time instant and time interval. Allen's interval relations [13, 3] centers on various time intervals and possible reasoning involving co-operating concepts. As proposed by Allen [3], a framework for temporal reasoning, and all the relations proposed dealt with the directionality of time. In his proposal, intervals are the only temporal primitives in the temporal logic. Allen aimed at illustrating natural language sentences and to represent plans. To achieve the thirteen (13) basic relations between time intervals, with six of the intervals being the inverse of the other six: before, after, finishes, finished-by, overlaps, overlapped-by, starts, started- by, during, contains, meets, met-by and equal [3]. The basic interval temporal relations (Figure 1). These temporal relations depict and relate the actions and plans described in this paper.

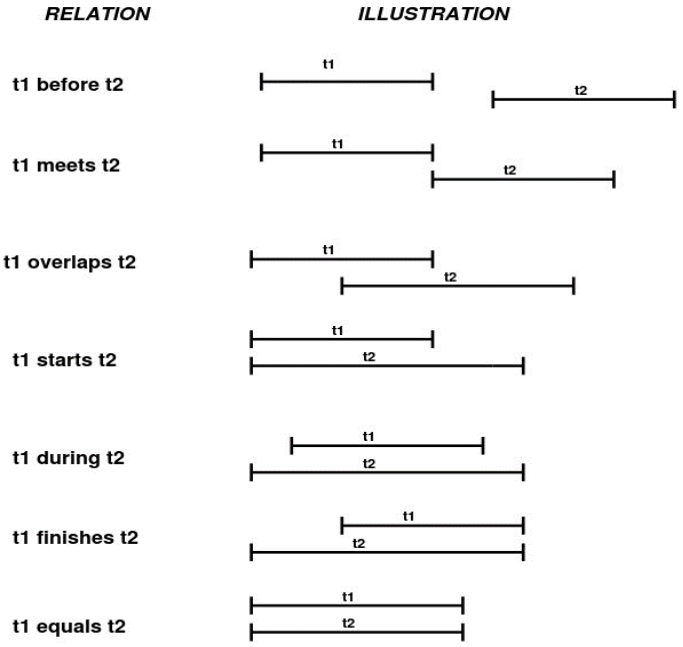


Figure 1: Allen's interval based temporal relations

This paper is targeted at formalizing the basic temporal ontology model for reasoning and inference towards an optimal time scheduling, after an analysis of an existing timetable to unravel all the complexities and evaluating feasible solutions towards achieving optimality. An illustration of how moment and point works on interval-based temporal logic is given in "moment and point in an interval based temporal logic". Moment is a non-decomposable period where the time is corresponding to instantaneous events while a point (a zero duration time) is where reasoning about the beginning and ending of event do arises. The axiomatization of the stated theory of time was done in terms of the simple relationship "meets" and subsumes the interval-based theory proposed in [3] thereby extending the theory to point-like time periods.

3. Timetabling Complexities

Timetable analysis reveals the major concepts - departments, timeslots, courses, students, venue and lecturers as objects responsible for complexities in timetable implementations. Three categories of courses exist in any typical higher institution; departmental, faculty-based and university-based courses. This paper considers the complexities arising from their inter-dependencies in scheduling the intra-departmental and inter-departmental (faculty-based and university-based) courses. Amongst other timetabling competency questions, are the following in the optimization of the general temporal scheduling ontology. They include:

i. Is the course offered by other departments, other than the host department?
ii. Are other courses for the same level in all cooperating departments allocated to the same time?
iii. Are students in higher levels in the cooperating departments also offering the course?
iv. Is the course lecturer teaching another course allocated at the same time?
v. Is the venue for course far apart from venue of the preceding or succeeding courses (courses before or after)? That is, are the respective venue for teaching the courses on different campuses? or are they within the same campus but widely separated?

Each of these concepts has peculiar attributes. These concepts relate in diverse ways and some of the relations exist in reverse form. The relations include *owns, has, can-be, offers, teaches, holds-in, lectures-in*, and *are-assigned-to*. In consideration of the fact that most resources of the university (such as lecturer, venue, and courses) are shared and in some other cases limited, departments compete for these resources. A university has a given number of programmes running on specified number of campuses. Campuses or lecture venues widely separated far apart from each other will exhibit a high level of complexity with numerous inter-dependencies. A sample model in fig. 1 with the departmental general timetable ontology for each department was described in [1] and shown in the resulting ontograf (Figure 2) from protégé. Almost all the components of the ontology are shared, making it needful for interoperability amongst relating departments. An attempt to link the ontologies for all the relating departments will make the resulting ontology cumbersome and more complicated largely because components of the ontology are not owned by a department, the number of inter-dependencies will generate some form of conflict of interest.

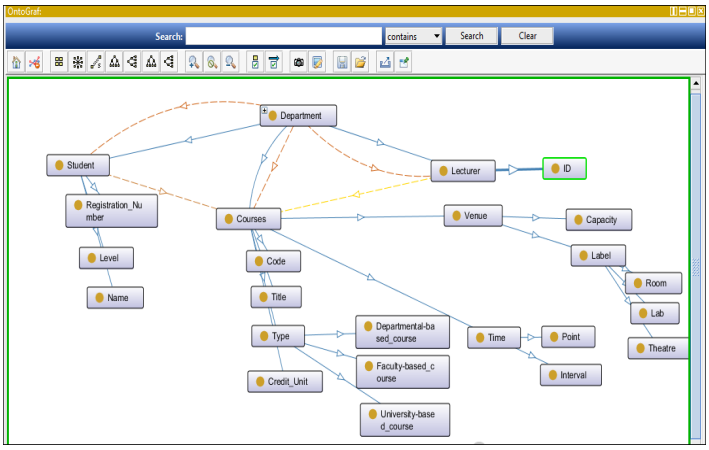


Figure 2: An Ontograf of a Departmental Timetable Ontology

Suppose course c_i is to be taught by lecturer L_l, scheduled to hold at venue v_1 at time t_i and another course c_j to be taught be the same lecturer L_l at same time t_i in another venue v_2, exposes some form of complexities. Again, suppose c_i and c_j are both scheduled to hold at v_1 at the same time t_i bring to bear another level of complexity that is predominant in the domain. That is a groups of students offering a particular course holding at time t_1 at campus U_1, and also offering another course c_j holding at time t_2 at campus U_2, where time t_1 meets time t_2. The meeting time does not give allowance for the participating students to travel from campus U_1 to campus U_2 (or from venue vi to venue v2 with campus U_i). A general model of these complexities is shown in Figure 3.

A view of the various complexities based on the Allen's interval relations results in the following temporal complexity types: overlapping time (which also includes starts, ends and during relations) for courses taught by same lecturers and offered by same level of students. Meeting time complexity also results for courses offered by inter campus students. Other complexity types, not handled in this paper, include venue-clash and carrying capacity check.

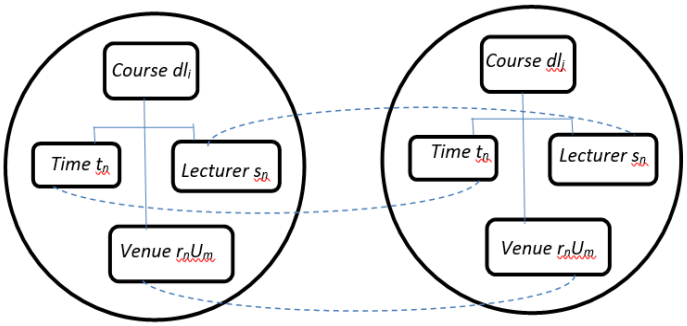


Figure 3: A course allocation complexity model

where *d* is the department's code, *l* is the level of students, t represents the timeslot, s is the lecturer, r the room number while U is the campus. *i,j,n,m>0* are instances of the various objects

From the general course allocation complexity model in Figure 3, several levels of complexity are defined to include:

- Departmental Complexity (Regular)
- Departmental Complexity (Carryover)
- Faculty-Based (Inter-Departmental) Complexity (Regular)
- Faculty-Based (Inter-Departmental) Complexity (carryover)
- Inter-Faculty/Same Campus Complexities (Regular)
- Inter-Faculty/Same Campus Complexities (Carryover)
- Inter-Faculty/Inter-Campus Complexities

A generalized timetable ontology will expand the departmental timetable ontology is obtained from a mapped departmental timetable ontology as modeled in [1]. This require the mapping of domain resources ranging from, past results, course registration list, departmental brochure for details of all level courses and lecturers, campus location (or distance apart), etc.

Table 1 gives a justification of complexities in all the cases in consideration and the possible resolution based on interval temporal relations.

4. System Framework

Figure 4 gives the architecture of the proposed Ontology-Based Temporal Scheduling Framework showing the relevant components for the achievement of its set goal of producing an optimal time schedule.

Table 1: Complexity description by type and possible time scheduling resolution

Case number	Case description	Complexity type	Resolution
1.	Carryover /repeating students in the host department.	Overlapping- Time	If carryover students are offering course c_i at time t_1, then time t_2 for course c_j (time for their level course should not overlap).
2.	Carryover / repeating students in the serviced department.	Overlapping- Time	If carryover students are offering course c_i at time t_1, then time t_2 for course c_j (time for their level course should not overlap).
3.	Regular students (students at that level) in the services department.	Meeting-Time	If course c_i is scheduled at time t_1 in campus U_1, (for e.g 10-12) and course c_j is also scheduled at time t_2 in campus U_2, where the end-time of course1 c_i meets with the start-time of course c_j, the course should be rescheduled.
4.	Assigning the same lecturer to teach at different campuses, without giving sufficient time interval.	Meeting-Time	A lecturer assigned for course c_1 at time t_1 in campus U^1, and also assigned for course c_j at time t_2 in campus U_2, where the end-time of course c_i meets with the start-time of course c_j, should be rescheduled.
5.	students at same level and carryover/repeaters student in the services department.	Starts/ Finishes	If course c_i and course c_j are of different department, then course c_j should not start at the same time with course1 and finishes after course c_i
6.	Students in their regular year of studies and carryover student of different department, but the same faculty.	During	If course c_i and course c_j are of different department , d_i and d_j respectively, but of the same faculty, then course c_i should not hold during course c_j
7.	Students (both regular students and carryover student) at another campus, offering a course holding at a different campus.	Location Campus location (distance apart)	If course c_i is scheduled for time t_1 in campus U_1, and course c_j is also scheduled for time t_2 in campus U_2, then time t_1 should not meet t_2
8.	The class size and the carrying capacity of the lecture venue.	Space (carrying capacity)	If the number of students offering course c_i is greater than the venue assigned, then course c_i should be re-scheduled for a bigger venue.

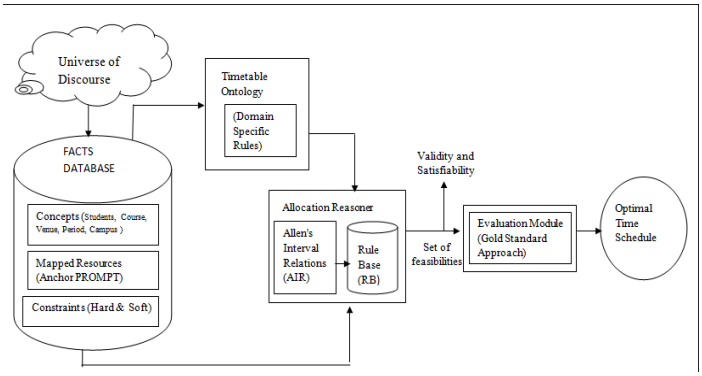


Figure 4: Architecture of the Proposed Ontology-Based Time Scheduling System

The major components of the system as depicted in Figure 4, are described as follows:

(i) Universe of Discourse: the universe of discourse where the domain knowledge under consideration is obtained. It consists of all the domain concepts — courses, venues, universities, timeslots, lecturers, students, students' status, departments, faculties, levels, campuses, locations, timetable and other aspects of the timetabling domain

(ii) Fact Database: this includes the values and instances of the domain concepts, the resources, relations and a mapping of facts to concepts and resources. It also has the hard and soft constraints.

(iii) Time Ontology; this component stores the domain specific rule (owns, has, offers etc.) which will be considered while mapping the resources.

(iv) Allocation Reasoner: this is where reasoning and inferencing take place based on the Allen's interval-based temporal relations. The sets of rules arising from the competency questions and some already established constraints, are also part of this component. the Allocation Reasoner works in collaboration with the domain specific rules.

(v) Ontology Evaluation Approach: Criteria from the gold or golden ontology evaluation standard helps in assessing the set of feasibilities obtained from the Allocation Reasoner towards optimality of the timetable.

(vi) Optimal Time Schedule: This gives the resulting possible time schedule showing the resolutions made towards optimality.

To manage the complexities called for the optimized general time schedule with an allocation reasoner with interval-based temporal relations operating in-between the cooperating departmental timetable ontologies. Shared components of the optimized general timetable ontology. The interval-based temporal relations/rules serve as the instrument for managing the identified complexities in the timetabling system. With the interval-based temporal relation allocation of courses to venues, allocation of courses to timeslots (time duration), and assignment of lecturers to courses/venues are efficiently done.

4.1 Proposed Complexity Reduction Rules

From competency questions (i) to (v) identified in section 3.0, rules R1 to R12 addresses the complexities as well as adopting the Allen's interval relations in the proposed resolutions. It is assumed that all departments within the same faculties are domicile in the same geo-location:

R1: IF student of department d_i *offers* course c_i AND department d_i NOT *owns* course c_i THEN time t_1 for course c_i **NOT *overlaps*** time t_2 for course c_j in department d_2.

R2: IF course c_i *holds* at time t_1 and course c_j holds at time t_2 AND t_1 ***equals*** t_2 AND students offer c_i and c_j THEN t_1 and t_2 ***overlaps***.

R3: IF course c_i *holds* at time t_1 and course c_j holds at time t_2 AND time t_1 is ***during*** time t_2 AND students offer course c_i and course c_j THEN t_1 and t_2 ***overlaps***.

R4: IF course c_i *holds* at time t_1 and course c_j holds at time t_2 AND t_1 ***starts*** with t_2 AND students offer c_i and c_j THEN t_1 and t_2 ***overlaps***.

R5: IF course c_i *holds* at t_1 and course c_j holds at time c_2 AND t_1 ***finishes*** at t_2 AND students offer c_i and c_j THEN t_1 and t_2 ***overlaps***.

R6: IF c_i and c_j *belongs* to the same department, d_1, THEN scheduled c_j to t_i AND c_j to t_j such that $t_i \neq t_j$ (**NOT *overlaps***).

R7: IF c_i and c_j are of different department, d_i and d_j respectively, but of the same faculty, THEN time for c_i should **NOT *overlaps*** with c_j or c_i **NOT *during*** c_j.

R8: IF t_i for c_i and t_2 for c_j ***meets***, THEN t_j for c_j will be changed and scheduled for another time (**NOT *meets***)

R9: IF c_i and c_j are courses of different faculty, THEN their time **NOT *meets*** and **NOT *overlaps***

R10: IF time for c_i and c_j are ***equals*** AND venue assigned for c_i is the same venue assigned for c_j, THEN (c_i OR c_j) should be rescheduled for different time.

R11: IF time for c_i and c_j are ***equals*** or ***meets***, AND the lecturer assigned for c*i* in campus U_1 is also assigned for c_j in U_2 within the time interval of one hour, THEN c_j should be re-scheduled for another time.

R12: IF c_i is scheduled for t_i in U_1 AND c_j scheduled for t_j in U_2 THEN t_i **NOT *meets*** tj.

R13: IF carryover/repeating students are offering c_i scheduled for time t_i THEN t_j, time for their level course, c_j **NOT *overlaps*** t_i.

R14: IF c_i in campus U_1 and course c_j in U_2, THEN time t_i for c_i and t_j for c_j **NOT *meets*** or **NOT *overlaps.***

R15: IF number of students offering c_i is greater than the venue carrying capacity, THEN re-schedule c_i.

Formally expressing the above rules in FOL results in axioms A1 to A15. Recalling from the conceptualization of the domain concepts: **student, lecturer, time**, **venue** and **campus** with domain-specific relations: *owns, has, offers, holds* and *teaches* does not resolve time scheduling complexities when considering inter-departmental course allocation. This called for the adoption of the Allen's interval relations which includes temporal relations such as ***overlaps. meets, equals, during*** and their negations for a more explicit representation. Axioms A1 to A15 constitute the core of the ontology reasoner that results in the optimal time scheduling process.

A1: $\forall$*Student, Course1, Time1, Dept1* $\exists$ *Course2, Time2, Dept2.*
offers(Student, Course1) $\wedge$ *owns(Dept1, Course1)* $\wedge$ *offers(Student, Course2)* $\wedge$ *owns(Dept2,Course2)* $\Rightarrow$ *cooperates(Dept1, Host,Dept2, Serviced).*

A2: $\forall$*Student, Course1, Time1,Course2, Time2.*
holds(Course1, Time1) $\wedge$ *holds(Course2, Time2)* $\wedge$ *equals(Time1, Time2)* $\wedge$ *offers(Student, Course1)* $\wedge$ *offers(Student, Course2)* $\Rightarrow$ *overlaps(Time1, Time2).*

A3: $\forall$*Student, Course1, Time1, Course2, Time2.*
holds(Course1, Time1) $\wedge$ *holds(Course2, Time2)* $\wedge$ *during(Time1, Time2)* $\wedge$*offers(Student, Course1)* $\wedge$ *offers(Student, Course2)* $\Rightarrow$ *overlaps(Time1, Time2).*

A4: $\forall$*Student, Course1, Time1, Course2, Time2.*
holds(Course1, Time1) $\wedge$ *holds(Course2, Time2)* $\wedge$ *starts(Time1, Time2)* $\wedge$*offers(Student, Course1)* $\wedge$ *offers(Student, Course2)* $\Rightarrow$ *overlaps(Time1, Time2).*

A5: $\forall$*Student, Course1, Time1, Course2, Time2.*
holds(Course1, Time1) $\wedge$ *holds(Course2, Time2)* $\wedge$ *finishes(Time1, Time2)* $\wedge$*offers(Student, Course1)* $\wedge$ *offers(Student, Course2)* $\Rightarrow$ *overlaps(Time1, Time2).*

A6: $\forall$ *Course1, Time1, Dept1,Course2, Time2.*
owns(Dept1, Course1) $\wedge$*owns(Dept1, Course2)* $\wedge$*holds(Course1, Time1)* $\wedge$ *holds(Course2, Time2)* $\wedge$*equals(Time1, Time2)* $\wedge$*equals(Level1, Level2)*
$\Rightarrow \neg$*overlaps(Time1, Time2).*

A7: $\forall$ *Course1, Time1, Dept1, Fac1, Course2, Time2, Dept2. Fac2.*
owns(Dept1, Fac1, Course1) $\wedge$*owns(Dept2, Fac2, Course2)*
$\Rightarrow \neg$*overlaps(Time1,Time2)* $\vee \neg$*during(holds(Course1, Course2)) .*

A8: $\forall$ *Course1, Time1, Course2, Time2,* $\exists$ *t.*
holds(Course1, Time1) $\wedge$*holds(Course2, Time2)* $\wedge$ *finishes(Time1, t)*$\wedge$ *starts(Time2, t)* $\Rightarrow$ *meets(Time1, Time2).*

A9: $\forall$ *Course1, Time1, Dept1, Fac1, Course2, Time2, Dept2. Fac2.*
owns(Dept1, Fac1, Course1) $\wedge$*owns(Dept2, Fac2, Course2)* $\wedge\neg$ *equals (Fac1, Fac2)*
$\Rightarrow \neg$*meets(Time1, Time2)*$\wedge\neg$*overlaps(Time1, Time2).*

A10: $\forall$ *Course1, Time1, Venue1, Course2, Time2, Venue2.*
holds(Course1, Time1) $\wedge$*holds(Course2, Time2)* $\wedge$*equals (Venue1, Venue2)* $\Rightarrow \neg$*equals(Time1, Time2).*

A11: $\forall$ *Course1, Time1, Let, Camp1, Course2, Time2, Camp2.*
holds(Course1, Time1, Camp1) $\wedge$*holds(Course2, Time2, Camp2)* $\wedge$*equals (Time1, Time2)*$\wedge$*equals (Camp1, Camp2)* $\wedge$*teaches (Let, Camp1)*$\wedge$*teaches (Let, Camp2)*
$\Rightarrow \neg$*overlaps (Time1, Time2)*$\vee\neg$*meets (Time1, Time2)*

A12: $\forall$ *Course1, Time1, Camp1, Course2, Time2, Camp2.*
holds(Course1, Time1, Camp1) $\wedge$*holds(Course2, Time2, Camp2)* $\wedge$*equals (Time1, Time2)*$\wedge$ *equals (Camp1, Camp2)*
$\Rightarrow \neg$ *meets (Time1, Time2)*

A13: $\forall$*Student, Course1, Time1, Dept1, Course2, Time2.*
offers(Student, Course1, Time1) $\wedge$*offers(Student, Course2, Time2)* $\wedge$*owns(Dept1, Course1)* $\wedge$ *equals(Time1, Time2)* $\wedge$ *equals(Level1, Level2)* $\Rightarrow \neg$*overlaps(Time1, Time2).*

A14: $\forall$*Student, Course1, Time1, Dept1, Course2, Time2.*
offers(Student, Course2, Course2) $\wedge$ *holds(Course1, Camp1)* $\wedge$ *holds(Course2, Camp2)* $\Rightarrow \neg$*meets(Time1, Time2)*$\neg$ *overlaps (Time1, Time2).*

A15: ∀Student, Course1,Venue1, Time1, nos, ∃Venue2, ccap.

offers(Student, Course1, Venue1) ∧ holds(Course1, Venue1, Time1) ∧ (nos > ccap) ⇒ holds(Course1, Venue2)

4.2 Weighting of Complexities

The severity levels of complexities are denoted by 0%, 50% 70% and 100% respectively based on the sources of complexity (time, space and location).

1. *Overlapping Time*:

Overlapping time complexities includes five relations in Allen's Interval-based Temporal Relations: s*tarts, finishes, overlaps, during* and e*quals*. In this paper, these five temporal relations are classified into two levels of complexities namely complete and partial. Complete overlap occurs if two or more courses offered by a group of Students begins and ends at the same time (Equals Relation – See Table 2) or the Start and End times of these course is within the start and end time of c_j (During Relation – See Table 3), then a complete overlap of weight 100% has occurred. However, the During relation will not apply in this domain since the university system does not have 4-hour lecture period. For example, instances of *equals* and *during* relations are given in Tables 2 and 3 respectively.

Table 2: equals Relation

8:00 – 9:00 AM	9:00 – 10:00 AM	10:00 – 11:00 AM	11:00–12:00 NOON
c_i			
c_j			

Table 3: during Relation

8:00 – 9:00 AM	9:00 – 10:00 AM	10:00 – 11:00 AM	11:00– 12:00 NOON
c_i			
	c_j		

A partial overlap is described using Allen's Interval-based Temporal Relations as follows; *overlaps*, *starts* and *finishes* shown in Tables 4, 5 and 6 respectively. These relations under partial overlap are assigned a complexity of 50%. Examples of partial overlaps are given in Tables 4, 5 and 6

Table 4: overlaps Relation

8:00 – 9:00 AM	9:00 – 10:00 AM	10:00 – 11:00 AM	11:00– 12:00 NOON
c_i			
	c_j		

Table 5: starts Relation

8:00 – 9:00 AM	9:00 – 10:00 AM	10:00 – 11:00 AM	11:00–12:00 NOON
c_i			
c_j			

Table 6: finishes Relation

8:00 – 9:00 AM	9:00 – 10:00 AM	10:00 – 11:00 AM	11:00– 12:00 NOON
c_i			
	c_j		

In the resolved overlap, the desired Allen's interval-based temporal relation for the resolution of overlapping time complexity is *before* in Table 7 and Table 8. When course c_i is *before* course c_j a 0% complexity is recorded.

Table 7: before Relation (A)

8:00 – 9:00 AM	9:00 – 10:00 AM	10:00 – 11:00 AM	11:00– 12:00 NOON
c_i			
		c_j	

Table 8: before Relation (B)

8:00 – 9:00 AM	9:00 – 10:00 AM	10:00 – 11:00 AM	11:00– 12:00 NOON
c_i			
			c_j

2. Meeting Time Complexity

If courses c_i and c_j hold in geographically dispersed Venues (R_nU_m, $R_{n+1}U_{m+1}$) and it takes a student time t_n to move from venue R_nU_m to $R_{n+1}U_{m+1}$, there exist a time complexity if course c_j starts immediately course c_i ends as shown in Table 9.

Table 9: Meeting Time Complexity (100%)

8:00 – 9:00 AM	9:00 – 10:00 AM	10:00 – 11:00 AM	11:00 – 12:00 NOON
c_i			
		c_j	

The 70% is for time to reach the venue and 100% is if there is additional time for the student to settle in the class before the actual lecture start time, an additional 30% weight is added. A 70% and a 100% resolution of the Meeting time complexity is as shown in Tables 10 and 11

Table 10: 70% Resolved Meeting Time Complexity

8:00 – 9:00 AM	9:00 – 10:00 AM	10:00 – 11:00 AM	11:00 – 12:00 NOON
c_i			
			c_j

Table 11: 100% Resolved Meeting Time Complexity

8:00 – 9:00 AM	9:00 – 10:00 AM	10:00 – 11:00 AM	11:00 – 12:00 NOON
c_i			
			c_j

5. Implementation Results and Discussions

The timetable showing schedules for two departments located at different campuses are used for the implementation of the proposed system.

These two departments offer common courses — CSC 211, STA 211, CSC 111. The natures of complexity are (meeting and overlapping times) clearly. Analysis of existing system shows the course under study, CSC 211 with Geo-informatics students in 300 level, who may be carrying over the course. It also handles the distance apart between the two campuses, being the venue for these two courses. Another case considers a complete overlap where CSC 211 and STA 211 from two different departments at two different venues, though in the same campus. Here, students of both departments are offering CSC 211. The analysis of these complexity levels with given assigned weights on the identified complexities is as shown in Table 12.

Table 12: Complexity weights of the existing time schedules

Case No	Complexity Type	Complexity Weight (%)
1	Meeting Time	100
2.	Overlapping Time (*equals*)	100

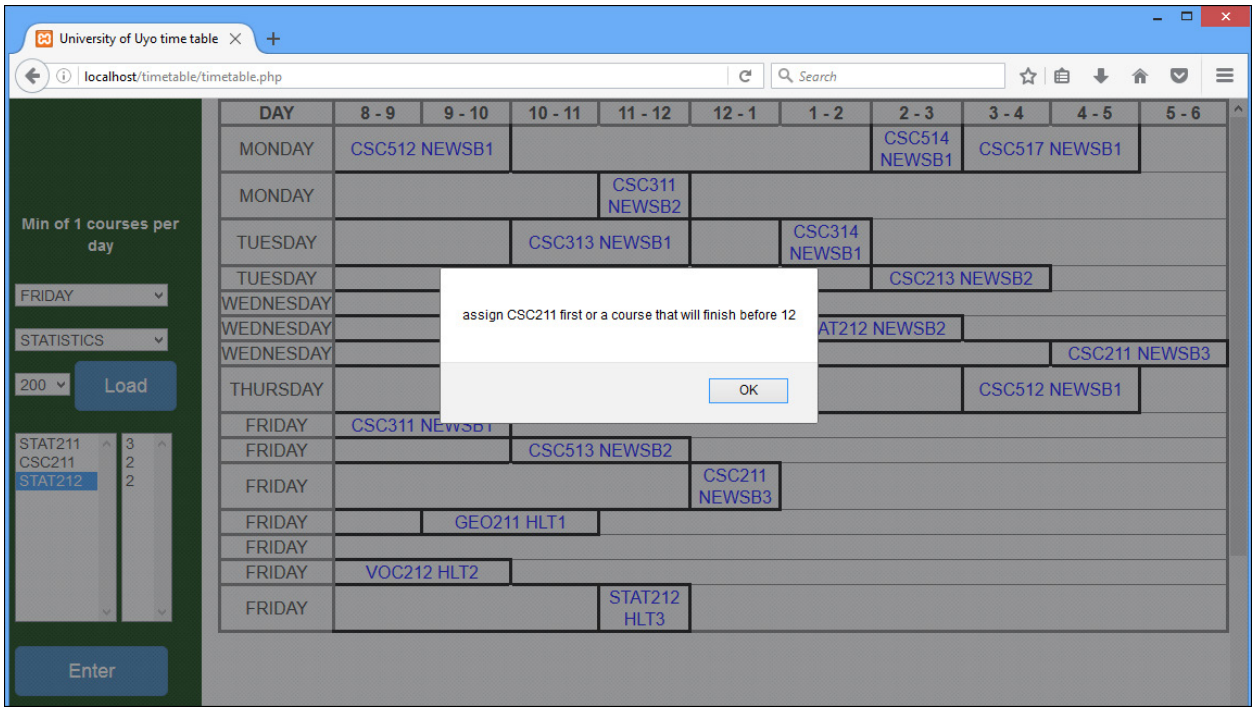


Figure. 5: Active Resolution of Meeting Time Complexity.

5.1 Resolutions with proposed Time Scheduling system

To resolve the time schedules, the complexity weights of the two cases should be 30% and 0% for meeting time and overlapping time respectively. The screen shots in Figure 5 gives the resolution of the meeting time complexity for the scheduled GEO 211 and CSC 211. The screen shot also shows a prompt to reschedule the STA 212 course as shown on the white pop-up window. That is the overlapping time complexity resolution, where the prompt calls for the relocation of the statistics (STA) course resulted. This is because CSC 211 is offered by many other departments, thereby having increased level of inter-dependency than the STA 212 course.

6. Conclusion

The proposed optimized time scheduling system analyzes any given existing time schedule and identifies the complexity types in existence. It provides the users the window to reschedule any time with identified complexity to obtain an optimized time schedule as shown in this paper. The rules in the reasoner are based on Allen's interval-based temporal relations as well the domain specific relations. Knowledge management as seen in the university timetabling complexities is possible with the mapping of the departmental timetable ontologies shown in the ontograf with all other required resources in the databases of fact. As a further work, the formal model that describes how the embedded semantics in the given rules are used for inferencing will be implemented on the ontology.

References

[1] P. U. Usip and M. M. Ntekop. "The use of ontologies as efficient and intelligent knowledge management tool", *In Proceedings of Future Technologies Conference. 6-7 December 2016, San Francisco*. pp 626-631. **DOI:** 10.1109/FTC.2016.7821671

[2] M. E. Saleh. Semantic-based query in relational database using ontology. Journal on Data and Knowledge Engineering. 2011.

[3] J. Allen. Maintaining knowledge about temporal intervals. *Communications of the ACM* 26. 1983. (11): 832-843.

[4] P. U. Usip and E. Umoren. Towards a formal interval-based temporal ontology for an optimized timetable management. In Proceedings of Transition from Observation to Knowledge to Intelligence Conference (TOKI 2016), Lagos. pp. 293-304

[5] K. H. Shafa'amri and J. O. Atoum. A framework for improving the performance of ontology matching techniques in semantic web. International Journal of Advanced Computer Science and Applications, 2012. 3(1) 8-14.

[6] B. Coppin. Artificial Intelligence Illuminated. Jones and Bartlett Publishers, Inc., 2004. Sudbury

[7] D. Waltz. Artificial Intelligence: realizing the ultimate promises of computing, AI magazine. 1997) 18, 49–52

[8] I. M. Jurisica and J. Eric. Ontologies for knowledge management: an information systems perspective. Knowledge and Information Systems. 2004. 6, 380–401. DOI 10.1007/s10115-003-0135-4

[9] D. Abramson. Constructing school timetable using simulated annealing: sequential and parallel algorithm. 1991.

[10] E. K. Burke and S. Petrovic. Recent research directions in automated timetabling. *European Journal of Operational Research*. 2002. 140. 266-280.

[11] A. O. Modupe, O. E. Olusayo and O. S. Olatunde. Development of a university lecture timetable using modified genetic algorithm approach. International Journal of Advance Research in Computer Science and Software Engineering. 2014. 4. 163-168.

[12] J. R Hobb and F. Pan. *An ontology of time for semantic* web. 2004. 3, 66-85.

[13] M. Hamalatha, V. Uma and G. Aghila. Time ontology with reference event based temporal relation (RETR). *International Journal of web and Semantic Technology* (IJWSSY). 2012. 3.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1592-1598 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

COTS – Harsh Condition Effects Considerations from Technology to User Level

Kirsten Weide-Zaage[*,1], Guillermo Payá-Vayá[2]

[1]*RESRI Group, Institute of Microelectronic Systems, Leibniz Universität Hannover, 30176, Hannover, Germany*

[2]*AS Group, Institute of Microelectronic Systems, Leibniz Universität Hannover, 30176, Hannover, Germany*

ARTICLE INFO

Article history:
Received: 15 June, 2017
Accepted: 23 July, 2017
Online: 24 August, 2017

Keywords:
Radiation
Simulation
SRAM
Metallization
Reliability
Migration
Packaging
Circuit

ABSTRACT

Radiation hardened devices are mostly extremely expensive. The continuously downscaling of microelectronic structures and the unavoidable presence of particle radiation on ground and in space leads to unwanted failures in electronic devices. Furthermore it is expected that in the next few years around 8000 new satellites will be launched around the world. Due to the enormous increasing need for Rad-Hard devices, there will be more focus on Commercial Of The Shelf (COTS) devices, which costs are lower. Also nowadays microelectronics for automotive systems are tested to withstand radiation especially SEU-single event upsets. It is clear that SEU cannot be ignored anymore especially in the application of unmanned autonomous vehicles and systems. Reliability testing is expensive and extremely time consuming. The use of COTS-Commercials of the shelf is the ultimate goal to reach. In this paper, an overview of radiation effects on different CMOS technologies used in COTS devices is given. These effects can be considered while selecting different functional equivalent COTS devices implemented with different technologies. Moreover, an overview of software techniques used in programmable commercial devices to reduce the radiation effects is also described.

1. Introduction

This paper is an extension of work originally presented at the SMTA Pan Pacific Symposium [1].

It is well known, that CMOS components are radiation sensitive and mostly have a variability concerning radiation response. The most commercial devices can withstand doses in the order of 5-10 krad and a minority can survive up to 100 krad [2]. To identify the more Rad-Hard Components radiation tests are needed. To get information about the radiation response especially single event upsets (SEU), the influences of the technology node, metallization scheme, new materials as well as package and shielding materials have to be investigated and evaluated. It has to be said, that concerning packaging materials, design and technology many steps were taken to avoid radiation sensitivity.

The procurement of COTS parts for use in a radiation environment is a problem. Technically the same components are produced in different foundries of the companies. A product can be changes without the knowledge of the consumer if the datasheet is not affected.

Ideally, the purchaser would like to have a homogeneous quantity for the purchase of several components, which originate from the same process line produced simultaneously. Then the characterization of the radiation behavior of a small sample of parts is representative for the whole lot. In cases where the contract volume is interesting for the trader or manufacturer, it may be possible to negotiate a better traceability with the manufacturer / trader. Another possibility is the use of qualified automotive components for space applications. Here the problem is that air and space applications have almost smaller quantities compared to automotive and the manufacturers only deliver big quantities.

The increase of high integration of chip and package influences the costs and the time-to-market in parts negatively. The system complexity results in integration of analog and digital components on a single chip and the use of new packaging concepts.

The environment as well as new materials like Pb-free solder or Platinum as metallization material, the use of organic materials

[*]Corresponding Author: Kirsten Weide-Zaage, RESRI Group, Institute of Microelectronic Systems, Leibniz Universität Hannover, 30176, Hannover, Germany | Email: weide-zaage@ims.uni-hannover.de

and new dielectrics influence the reliability and quality of the devices and systems and can yield, in the worst case, to a complete re-design (Figure 1).

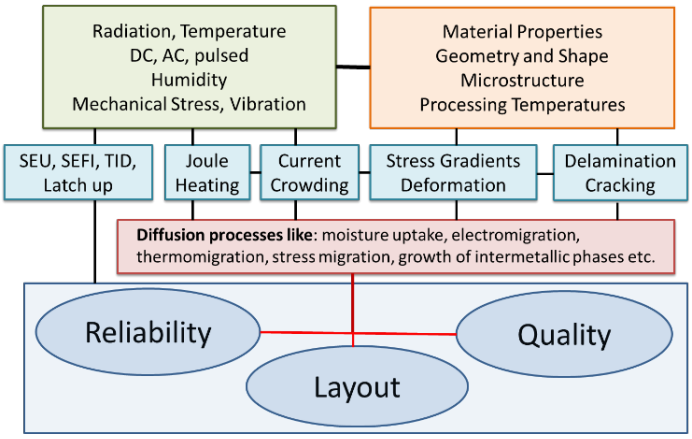


Figure 1. Reliability, quality and layout dependence of some physical failures in the devices and components.

In this paper several considerations about radiation effects are given in Section 2. After the description of the different CMOS related failures and a discussion of the different technology and materials, some simulation ideas will also be considered in the frame of high temperature application. Furthermore migration effects resulting from harsh conditions are discussed. Looking to ultra-scaled memories, it is found that the sensitivity against radiation is directly coupled with the reduction of the size and the distance between two circuits. This leads directly to a reduction of the critical charge which can induce a bit flip [3]. Then section 3 introduces different methods used in programmable COTS devices to reduce or mitigate the harsh condition effects. Finally this paper is concluded in Section 4.

2. Technology Node and Material

The quantity of products from the different technology nodes is shown in Figure 2. It is found, that a big portion of the products is still manufactured in 180 nm and 130nm technology node [4].

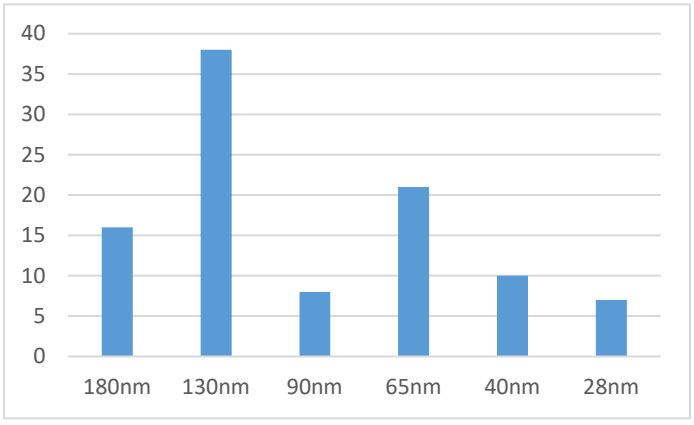


Figure 2. Quantity of products in (%) of the different technology nodes.

Generally the number of the materials used in the Front-End manufacturing of the devices is in the two-digit range. Every material and processing step can have more or less influence on the radiation sensitivity of the device. A look on the semiconductor processing as well as the materials and the packaging for the technology nodes beyond the 65nm node is necessary caused by the fact that materials etc. change. In the 22nm technology node with Hafnium as gate material new possible radiation risks may occur. ^{174}Hf is an alpha particle emitter. Materials used in the semiconductor industry with their natural occurrence are shown in Table 1. An investigation of SEU related to these materials shows that due to the high half-life there should not be a problem. However, due to the ultra-high integration the portion of these material increases drastically and therefor an SEU risk cannot be excluded anymore.

Table 1. Alpha-Emitter with half-live and natural occurrence.

Alpha Emitter	Half-live (a)	Natural occurrence [%]
^{174}Hf	2.0 x 10^{15}	0.162
^{152}Gd	1.08 x 10^{14}	0.2
^{144}Nd	2.29 x 10^{15}	23.8
^{148}Sm	7.0 x 10^{15}	11.3
^{147}Sm	1.06 x 10^{11}	15.0
^{90}Pt	6.5 x 10^{11}	0.01

2.1. Alpha Particles and Packaging

The emissivity of alpha particles in the different package materials are given in Table 2 [4-8]. Especially in Flip-Chip applications the emissivity's as well as the range in the materials are important. The highest emissivity is found in Pb based solder followed by Sn based solder material.

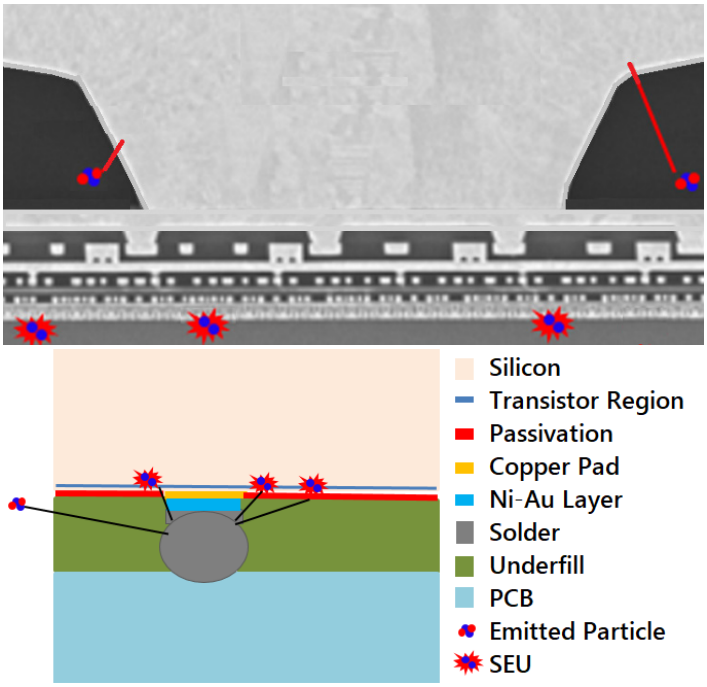


Figure 3. Schematic view of the interaction between particles emitted from the solder with the transistor level.

In Figure 3 the interaction between particles emitted from the solder with the transistor level is shown schematically. The risk for alpha emission from the Sn is not eliminated due to the fact that trace contamination of Pb in Sn occurs which cannot eliminated easily [9].

Table 2. Emissivity (α/cm²-hr) of different materials.

Material	Emissivity (α/cm²-hr)
Mold Compound	0.024→ < 0.002
Underfill Flip Chip Pb	0.004→ 0.0007
Eutectic Pb based Solder	7300→ < 0.0009
DIP Leadframe	0.00677
Alloy 42 (Hitachi)	8
Au-plated Alloy 42	4
Sn	>1000 → <1
AlSiC (Lanxide)	215
ZIP Leadframe	0.00258
256K DIP	0.00124
64K DIP	0.00109
Metal package lid	0.030
Ceramic package lid	3.10
Plastic (Epoxy)	0.00080
Ceramic DIP	0.03230
Ceramic LCC	0.02530

With concern to alpha particle induced SEU several conditions must be fulfilled so that an alpha particle causes a soft error. These are:

- An alpha particle-emitting radioisotope must be present in the packaging material and at a distance within 50µm.
- The alpha particle must be emitted in the direct direction to the active layer.
- The alpha particle must have enough energy or exceed the critical charge in order to influence the active layer in such a way that a bit flip occurs.

Alternatively, the influence of the alpha particles can be eliminated by:

- a better design, i.e. more distance to the active layer
- shielding material such as die coatings or thin polyimide films
- the choice of Ultra-Low-Alpha housing materials

There are different package categories defined related to radiation. The categories are given in Table 3 [10].

Table 3. Categories of the different packages related to radiation.

Category	Emissivity (α/cm²-hr)
Standard	10 – 0.01
Low-alpha (LA)	< 0.01
Ultra-low-alpha (ULA)	< 0.001
Hyper-low-alpha (HLA)	< 0.0005

2.2. CMOS Related Failure Mechanism

The following error mechanisms can be distinguished in CMOS devices:

- **SEU**: Single Event Upset as a singular event temporarily interfering with the system, e.g. wrong calculation information or a bit flip in a memory. If the bit is corrected, the system continues to operate normally.
- **SEFI**: Single Event Functional Interrupts occurring in control loops and bringing the system to an undefined state, a test mode, or a standstill that should be reset by a reset.
- **SEL**: In a single-event latch-up of a parasitic PNPN structure, a high-energy particle can cause a short between two transistors. This remains as long as the system is connected to the power supply. If the current is switched off and the system is restarted, the SEL is over. The energy for an SEL at the processor is 80 MeV/cm²/mg and in the CMOS-RAM 50 MeV/cm²/mg.
- **SET**: A single-event transient is a high-energy particle that induces a surge current that travels through the circuit and can disturb the module in a manner similar to an electrostatic discharge (ESD).
- **Single event snapback**, which can occur similarly to the SEL but does not require a PNPN structure.
- **TID** total ionizing can as a long-term impact on CMOS chips by trapping holes near the boundary layer between the dielectric (SiO2) and silicon.

2.3. Simulation of Iteration

A precise prediction of the radiation response in the new technologies and circuits is crucial to ensure a trouble-free operation. Simulation may support this and help to reduce the costs in terms of stress tests. Three dimensional physical simulations of ionizing radiation effects in the semiconductor structure, which are coupled to an external circuit simulator are one way to do so [11].

The reaction on a radiation event is in technology nodes beyond 100nm closely related to the structure of the device caused by the presence of interconnects, bias, and different metallization layers affecting the performance [11-13].

A physical-based modeling approach that includes detailed high-energy physics-based simulations of nuclear reactions coupled with advanced component and circuit simulators is a need here. With tools such as e.g. GEANT4, a refinement of TCAD simulations with e.g. SILVACO can be done. Approaches to this are described in [14, 15].

GEANT4 is an on C++ based open source software package composed of tools that can be used to simulate the interaction of particles with matter. This software package was developed at CERN. The calculation of decay, track, energy of the particles, elastic and inelastic bumps etc. is possible. Furthermore beside NIST (National Institute of Standards) materials user materials can be defined.

A Monte Carlo simulation using GEANT4 to determine the path of the particles through the transistor as well as the deposited energy was carried out. This data was then used to determine the generation rate for electron hole pairs (3.6 eV) and the radial distribution. The radial distribution is energy-dependent and complex analytical models can be used for their determination [16]. In the following simulation with the TCAD Tool SILVACO, the resulting characteristics can then be determined. In Figure 4 the simulation flow in terms of radiation using GEANT4 and SILVACO TCAD is shown.

It has to be remarked that GEANT4 is highly complex and requires a long training period and knowledge about the physics

and coupling. All user interfaces have to be programmed for the special needed cases. This makes it's use for a fast application uncomfortable.

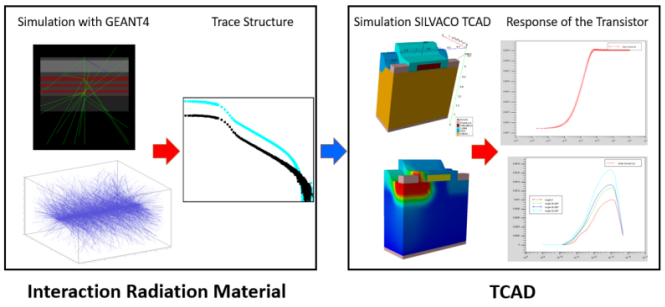


Figure 4. Simulation flow in terms of radiation.

2.4. Multilevel Metallization for Harsh Application

In CMOS devices used under harsh temperature environment tungsten metallization is used. To determine the reaction of the tungsten metallization layer on neutron radiation a metallization scheme was modeled (Figure 5). A B-doped active layer with a thickness of 100nm followed by a 500nm SiO2 including a Tungsten interconnect was positioned on the silicon substrate. The thickness of the first two Tungsten metallization and the dielectric was set to 500nm as well. The third level was set to 750nm followed by 2µm Aluminum power line covered with SiO2 passivation.

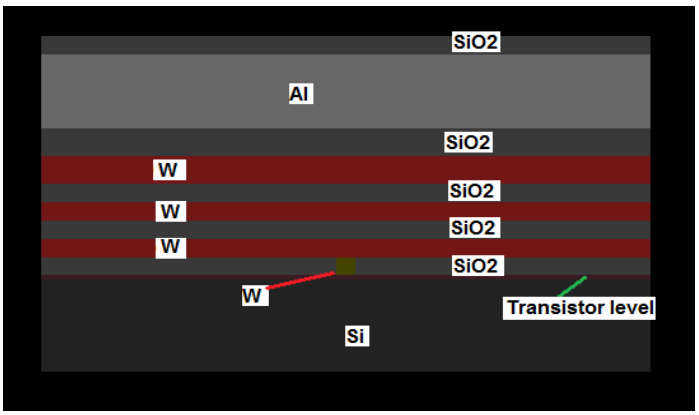


Figure 5. Model in GEANT4 with tungsten metallization and aluminum power line.

In Figure 6 the reaction after neutron radiation is shown. It was found that a lot of reactions occur in the Tungsten interconnects. This corresponds to observations in [17].

Under high temperature exposure coupled with high voltage or current, beside radiation effects migration effects also play an important rule. Depending on the stress conditions electromigration (EM) or stress migration (SM) are the main driving forces for these failures.

In the case of migration in metallization electro- and stress migration are the main failure mechanism. In the case of solder bumps thermomigration can occur as well [18, 19]. Stress migration occurs preferred at locations with high stress gradients, with a high concentration of vacancies or microstructure defects and at locations with active diffusion pathways with a low activation energy [20].

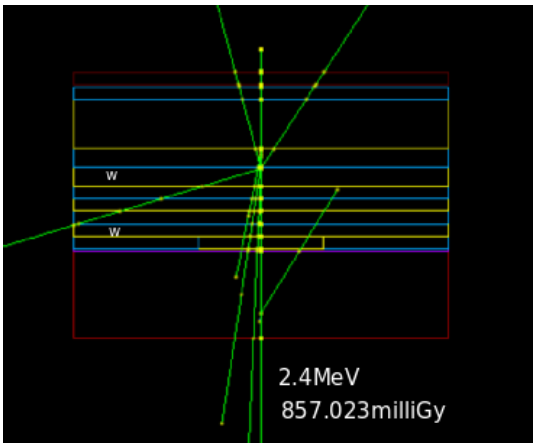


Figure 6. Reaction after neutron radiation with 2.4MeV.

It appears from the literature that the choice of relevant materials is the most discussed issue to enhance device performances as the reduction of interconnect is the "standard" evolution of all systems and cannot be prevented. In particular, the choice of the conducting material has a high influence on the specific resistance, on the homogeneity of thermal gradients, and on hot spots [18].

Migration effects are strong depending on local current densities, heating and stress gradients and the processing temperatures have to be considered [21, 22]. The mass flux describes the strength of the electromigration effect in the metallization structure. The mass flux represents the mass that flows through unit area per unit time. The mass flux of the different mechanisms are given in equation 1-3.

$$\overrightarrow{J_{el}} = \frac{D \cdot N}{k_B \cdot T} \cdot Z^* \cdot e \cdot \rho \cdot \left(\vec{j} - \overrightarrow{j_{th}}\right) \quad (1)$$

$$\overrightarrow{J_{TM}} = -\frac{N Q^*}{k_B T^2} D_0 \exp\left(-\frac{E_A}{k_B T}\right) grad\, T \quad (2)$$

$$\vec{J}_S = -\frac{N\Omega}{k_B T} D_0 \exp(-\frac{E_A}{k_B T}) \quad (3)$$

In equation 1-6, j is the local current density, j_{th} the threshold current density N is the number of the activated particles, k_B the Boltzmann constant, T the temperature, eZ* the effective charge, ρ the resistivity, D_0 the self-diffusion constant, Q* the heat of transport, grad T the temperature gradient, Ω the atomic volume, grad σ_H the stress gradient and E_A the activation energy for the specific mechanism. The resulting divergences are given in equation 4-6.

$$div\left(\overrightarrow{J_{el}}\right) = \left(\frac{E_A}{k_B T^2} + \frac{\alpha_T \rho_0}{\rho} - \frac{1}{T}\right) \cdot \overrightarrow{J_{el}} \cdot grad(T) \quad (4)$$

$$div\left(\overrightarrow{J_T}\right) = \left(\frac{E_A}{k_B T^2} - \frac{2}{T}\right) \cdot \overrightarrow{J_T} \cdot grad(T) - \frac{QND}{k_B T^2} \cdot divgrad\, T \quad (5)$$

$$div\, \overrightarrow{J_{SM}} = \left(\frac{E_A}{k_B T^2} - \frac{1}{T}\right) \cdot \overrightarrow{J_{SM}} \cdot grad\, T - \frac{D\, N\, \Omega}{k_B T^2} divgrad\, \sigma_H \quad (6)$$

For divergences bigger than zero a void can occur and for divergences smaller than zero a hillock can grow.

The different material properties for metallization materials as well as SnAgCu (SAC) as solder material are given in table y. It can be seen that the activation energy for tungsten as well as for platinum is higher compared to Al or Cu. The migration process is in the range of x magnitudes smaller compared to aluminum. The use of tungsten helps in this case to decrease migration effects due to the high activation energy for that process.

Table 4. Material properties for the different metallization materials [23-25].

Material	ρ[μΩcm]	κ[W/mK]	E_A[eV]	Z*	Q*[eV]
Al	3.16	237	0.7	-10	-0.104
Cu	1.74	395	0.9	-5.5	-0.312
W	5.5	174	1.7	12.4	0.009
Pt	10.5	73	1.0	0.3	0.677
SnAgCu	13.2	58.7	0.863	20	-0.009

2.5. SILVACO TCAD

With SILVACO TCAD the behavior of a transistor after SEU exposure with different energies as well as different LET Doses and a radius of 5nm with a cylindrical way through the transistor were calculated. In Figure 7 the dimensions of the transistor is shown.

The incidence of the SEU began at the drain vertical trough the transistor. The drain current depending on the transient time is shown in Figure 8. The higher the LET dose the sharper the profile of the drain current occurs. Once the onset of the SEU, a prompt charge collection will occur followed by the diffusion charge collection.

Electron-hole pairs with high carrier concentrations are created. If the ionization track traverses or comes close to the depletion region, carriers are rapidly collected by the electric field creating a large current/voltage transient. The 'prompt' collections phase follows, indicated in Figure 8. A funnel is created which enhances the efficiency of the drift collection by extending the depletion region deeper into the substrate. The size of the funnel is a function of the substrate doping. This phase is completed within a nanosecond and followed by a phase where diffusion begins to dominate the collection process. Additional charge is collected as electrons diffuse into the depletion region in a time scale of hundreds of nanoseconds until all excess carriers have been collected, recombined, or diffused away from the junction area [26].

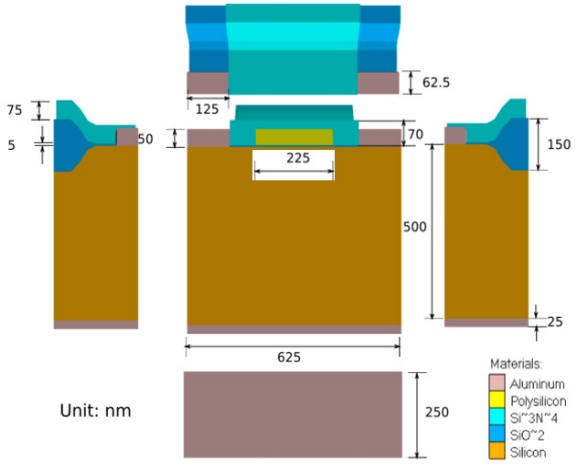


Figure 7. Dimensions of the transistor in SILVACO.

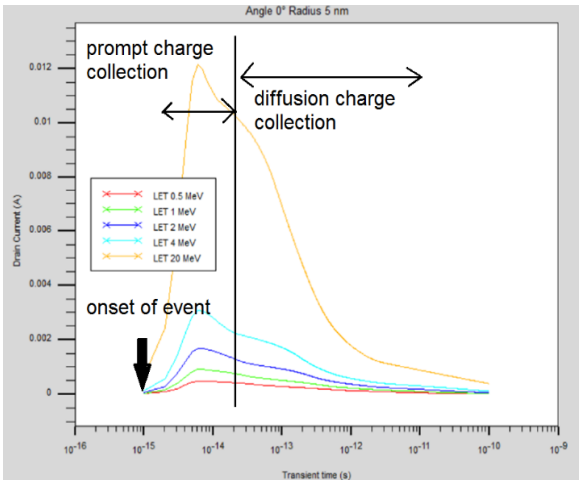


Figure 8. Drain current depending on the transient time for different LET.

2.6. Stacked Dies in 3D-Integration

In 3D integration several memories are stacked one by the other. In Figure 9 a possible scenario of a SEU interaction in 3D integration is shown. This might be an additional problem caused by the fact, that not only one silicon die (die 4) is exposed by alpha emitters from the solder. The die below (die 3) will also be affected due to the strong silicon thinning for stacking the IC's. Due to the high penetration rate of alpha particles in silicon SEU's can occur in die 3.

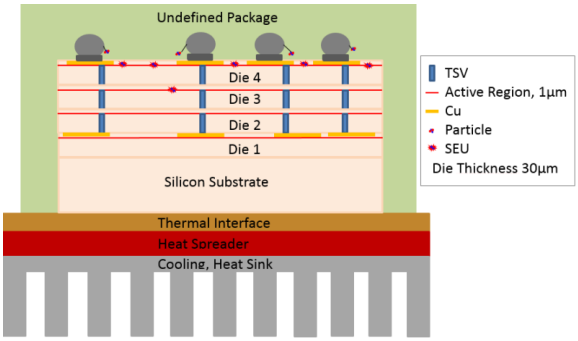


Figure 9. Scheme of stacked dies in 3D integration with possible scenario of SEU interaction.

2.7. SRAM in Different Technology Nodes

After dispensing with boron phosphorus silicate glass (BPSG) in the production processes below the 0.25µm technology node, the proportion of the soft error rate could be reduced by a factor of 10 [27]. From literature it is found that BPSG is also used for technology nodes below 250nm [28]. As a conclusion, it can be seen that thermal neutrons have an influence on the SEU even in small technology nodes. It is well known that commercial of the shelf (COTS) SRAMs were investigated regarding their radiation hardness.

The SRAMs differ concerning the transistor processing in the different technology node of 90nm and 130nm in the gate length as well as the number of the wiring planes. The neutron-induced failure on eight different SRAM components with different neutron energy was investigated and the single event upset (SEU) rate was measured. The test set up and the measurement results are described in [29]. It was found that various sources deal with the

analysis of conventional SRAMs of these technology nodes. In [30] SRAMs of the technology nodes 90nm to 350nm are reported. The components were not designated, so that a conclusion regarding manufacturers is not possible. A comparison between the different technologies with respect to their critical charge, shows no linearity between the technology nodes and the critical charge. The consideration of the cell size also shows no linearity with respect to the critical charge. This finding is relevant for the possible establishment of a standard model for checking on the basis of the critical charge.

3. Harsh Condition Effects Mitigation on Programmable COTS Devices

The use of programmable COTS devices in systems for satellites and space missions [31] implies the use of fault tolerant mechanisms at different levels (i.e., system, application, and instruction level) capable of detecting and correcting the previously mentioned (see Section 2.2.) CMOS related failure mechanisms.

In the system level, the duplication or triplication of programmable COTS devices is the most usual technique [32]. Bi-Modular Redundancy (Bi-MR) architectures are mainly used in fail-stop systems, since a duplex mechanism cannot intrinsically correct failures, requiring extra mechanisms that use time redundancy at the application level for diagnostic and recovery. The telemetry module of the Ariane 5 uses a Bi-MR based on COTS digital signal processors [31]. When results on both DSP differ, the results of one DSP are used and the second one is reinitialized. Tri-Modular Redundancy (Tri-MR) architectures also enable the correction of SEUs by reinitializing, for example, this processor which results differ. The Proton100K computer used by Space Micro [33] extended the concept of Tri-MR by applying time redundancy at the application level. The SCS750 space-qualified board [35] implements three IBM PowerPC 750FX working in triplex mode and includes a Radiation Tolerant FPGA, which is in charge of comparing the results among the PowerPC processors. Increasing the number of duplicated modules can increase even more the failure correction rate. For example, the flight control system implemented by Airbus A320/A330/A340 made use of eight computers [34]

In the application level, time redundancy is mostly used consisting in executing two consecutive times the target application on the same processor. Then, the failures are detected when the results of both executions differ. This technique is commonly used in conjunction with Bi-MR or Tri-MR architectures as it is mentioned above.

Replication techniques can also be applied at instruction level. In this case, each instruction is executed twice sequentially on the same processor, followed by a conditional branch. In case the results mismatch, the condition of the following branch instruction is true and the corresponding recovering subroutine is executed. This technique requires a huge memory and can drastically reduce the processing performance. A first software implementation of this technique was presented in [36] and tested on an Intel 8051 microcontroller. In [37], the authors present a new methodology that permits easy combination of hardware and software soft errors mitigation techniques and allows the automatic generation of protected source code, called hardened code.

4. Conclusion

Influences concerning radiation hardness on the components were discussed. The ultra-high integration with new materials and processes influences the SEU risk. For many applications the transition from Pb-free to Sn-based solder materials is done. The risk for alpha emission from the Sn is not eliminated due to the fact that trace contamination of Pb in Sn. Different metallization layers and Tungsten as local interconnect influences the radiation sensitivity. Simulation investigations with tools such as e.g. GEANT4 with a refinement of TCAD simulations e.g. SILVACO can help to understand the radiation impact on the devices. The availability as well as the utilization of such a tool is difficult. BPSG should be history below the 250nm nodes. It is reported, that in 250nm and beyond BPSG is still or again present.

In the case of high temperature and high voltage or current, application migration effects can occur in the metallization or solder material. Migration effects are strong depending on local current densities, heating and stress gradients and the processing temperatures have to be considered by simulation.

Finally examples how to mitigate harsh condition effects on programmable COTS devices are presented on three different design levels.

References

[1] Weide-Zaage, K.; Eichin, P.; Chen, C.; et al: COTS – Radiation Effects Appraoches and Considerations, SMTA Pan Pacific Symposium, 2017

[2] F. Faccio: “COTS for the LHC radiation environment the rules of the game”, [alicedcs.web.cern.ch]

[3] S. Mitra, P. Sanda, N. Seifert: “Soft Errors: Technology Trends, System Effects and Protection Techniques”, IEEE International On-Line Testing Symposium (IOLTS), 2007.

[4] Kumar, S; Agarwal, S.; Jung, J.P.: “Soft Error Issue and Importance of low Alpha Solders for Microelectronics Packaging”, Rev. adv. Mater. Sci. 34 (2013) 185-202.

[5] Baumann, R.C., Smith, E.B., “Neutron-Induced Boron Fission as a Major Source of Soft Errors in Deep Submicron SRAM Devices”, Proc. IEEE International Reliability Physics Symposium, 2000, pp: 152-157.

[6] Ditali, A.; Hasnain, Z.; “Monitoring alpha particle sources during wafer processing”, Semiconductor International, Vol. 16, No.7, 993, pp: 136-140.

[7] Levine, S.W.: “Alpha-emission measurements of lids and solder preforms on semiconductor packages”, IEEE Trans. On Comp., Hybrids and Manu. Tech., Vol. CHMT-2, No.4, 1979 pp: 391-395.

[8] Meieran, E.S.; Engel, P.R; May, T.C.; “Measurement of alpha particle radioactivity in IC device packages”, Proc. IEEE International Reliability Physics Symposium; 1979, pp: 13-22.

[9] B. M. Clark, D. Grove, T. Unvar: “Alpha radiation Dynamics in Electronics Packaging Structures”, Chip Scale Review, July/August 2013 pp.34-37.

[10] Reynard, J (2002). "Integration of fluorine-doped silicon oxide in copper pilot line for 0.12-µm technology". Microelectronic Engineering 60: 113. doi:10.1016/S0167-9317(01)00586-X.

[11] M. Turowski, A. Raman, G. Jablonski: “Mixed Mode Simulation and Analysis of Digital Single Event Transients in Fast CMOS ICs”, Proc. MIXDES 2007, Ciechocinek, Poland, pp. 433-438.

[12] C. L. Howe, R. A. Weller, et al, “Role of Heavy-Ion Nuclear Reactions in Determining On-Orbit Single Event Error Rates,” IEEE Trans. Nucl. Sci., vol. 52, Issue 6, part 1, Dec. 2005, pp. 2182-2188.

[13] S. Kobayashi, D. Ball, et al: “The effect of metallization layers on single event susceptibility,” IEEE Trans. Nucl. Sci., vol. 52, no. 6, pp. 2189–2193, Dec. 2005.

[14] M. Turowski, A. Fedoseyev, A. Raman: ”Single event upset modeling with nuclear reactions in nanoscale electronics”, Conf. MIXDES, 2008, pp. 443-448.

[15] M. Raine: “Etude de l’ effet de l’ energie des ions lourds sur la sensibilite des composants electroniques”, Universite Paris Sud - Paris XI, 2011.

[16] A. Akkerman, J. Barak: “Ion-Track Structure and its Effects in Small Size Volumes of Silicon”, IEEE Trans. on Nucl. Sci., Vol. 49 No. 6, 2002, pp: 3022-3032.

[17] H. Zhang, S. Wang, et al: “The effect of heavy metal in CMOS on neutron induced single event upset simulated with Geant4“.

[18] L. Meinshausen, K. Weide-Zaage, H. Fremont, Electro- and thermo-migration induced failure mechanisms in Package on Package, Microelectronics Reliability 52, pp. 2889-2906, 2012.

[19] Tu, K. N., Recent Advances on Electromigration in Very-Large-Scale-Integration of Interconnects, Journal of Applied Physics, Vol. 94, no. 9, pp. 5452—5473,, 2005

[20] E. T. Ogawa, J. W. McPherson, et.al: “Stress-induced voiding under vias connected to wide Cu metal leads”, 2002 Proc. Conf. 40th IEEE (Dallas, TX: IEEE) pp. 312–31.

[21] K. Weide-Zaage, J. Zhao, J. Ciptokusumo, et.al: “Determination of Migration Effects in Cu-Via Structures with Respect to Process Induced Stress”, Microelectronics Reliability 48, 2008, pp.1393–1397.

[22] B. L. Park, et al.: “Mechanisms of Stress-Induced Voids in Multi-Level Cu Interconnects,” IEEE 2002 International Interconnect Technology Conference, 2002, pp. 130-132.

[23] R. C. Weast: “CRC Handbook of Chemistry and Physics”, 62. Edition CRC Press, Boca Raton, Florida 1981.

[24] H. Weaver: “ Elektro- und Thermotransport in Metallen”, J.A. Barth Leipzig 1973.

[25] B. Chao, S.-H. Chae, et. al.: “Investigation of diffusion and electromigration parameters for Cu–Sn intermetallic compounds in Pb-free solders using simulated annealing“, Acta Met., 55, 2007, pp.2805-2814.

[26] R. Baumann: „Radiation-Induced Soft Errors in Advanced Semiconductor Technologies“, IEEE Trans. Dev. Mat. Rel., Bd. 1 Vol. 5, No. 3, 2005.

[27] Kumar, S; Agarwal, S.; Jung, J.P.: “Soft Error Issue and Importance of low Alpha Solders for Microelectronics Packaging”, Rev. adv. Mater. Sci. 34 (2013) 185-202.

[28] H. Zhang, H. Jiang. Et al: ”Thermal Neutron-Induced Soft-Error Rates for Flip-flop Designs in 16nm Bulk FinFET Technology", Proc. IEEE International Reliability Physics Symposium; 2017, 3D-03.

[29] Zabel, J.; Gudehus, B.; Kaiser, A.; Krüger, K.: Microanalysis of Radiation-sensitive Semiconductors. Annual Military Scientific Research Report 2011, Federal Ministry of Defense, March 2012.

[30] M. Olmos, R. Gaillard, et al: “Investigation of thermal neutron induced soft error rates in Commercial SRAMS with 0.35µm to 90nm Technology”, IEEE 44th Annual International Reliability Physics Symposium, IRPS, San Jose, 2006.

[31] M. Pignol. COTS-based applications in space avionics. In *Proceedings of the Conference on Design, Automation and Test in Europe* (DATE '10). European Design and Automation Association, 3001 Leuven, Belgium, Belgium, pp.1213-1219, 2010

[32] M. Pignol, "How to cope with SEU/SET at system level?," *11th IEEE International On-Line Testing Symposium*, 2005, pp. 315-318. doi: 10.1109/IOLTS.2005.34

[33] D. Czajkowski, and M. McCartha, Ultra low-power space computer leveraging embedded SEU mitigation. In Proc. IEEE Aerospace Conf (Vol. 5, pp. 2315-2328), 2003

[34] D. Briere and P. Traverse, "AIRBUS A320/A330/A340 electrical flight controls - A family of fault-tolerant systems," *FTCS-23 The Twenty-Third International Symposium on Fault-Tolerant Computing*, Toulouse, France, 1993, pp. 616-623.doi: 10.1109/FTCS.1993.627364

[35] R. Hillman, G. Swift, P. Layton, M. Conrad, C. Thibodeau and F. Irom, "Space processor radiation mitigation and validation techniques for an 1,800 MIPS processor board," *Proceedings of the 7th European Conference on Radiation and Its Effects on Components and Systems, 2003. RADECS 2003.*, Noordwijk, The Netherlands, 2003, pp. 347-352.

[36] M. Rebaudengo, M. S. Reorda and M. Violante, "A new software-based technique for low-cost fault-tolerant application," *Annual Reliability and Maintainability Symposium, 2003.*, 2003, pp. 25-28. doi: 10.1109/RAMS.2003.1181897

[37] A. Martinez-Alvarez, S. Cuenca-Asensi, F. Restrepo-Calle, F. R. Palomo Pinto, H. Guzman-Miranda and M. A. Aguirre, "Compiler-Directed Soft Error Mitigation for Embedded Systems," in *IEEE Transactions on Dependable and Secure Computing*, vol. 9, no. 2, pp. 159-172, March-April 2012. doi: 10.1109/TDSC.2011.54

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1599-1608 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

A Secure Automated Elevator Management System and Pressure Sensor based Floor Estimation for Indoor Mobile Robot Transportation

Ali Abduljalil Abdulla[*,1], Hui Liu[1], Norbert Stoll[2], Kerstin Thurow[1]

[1]*Center for Life Science Automation (celisca), University of Rostock, Rostock 18119, Germany*

[2]*Institute of Automation, University of Rostock, Rostock 18119, Germany*

ARTICLE INFO

Article history:
Received: 16 June, 2017
Accepted: 29 July, 2017
Online: 24 August, 2017

Keywords:
Mobile robot
Multi-floor
Automated Elevator
Floor Estimation
Pressure Sensor

ABSTRACT

In this paper, a secure elevator handling system is presented to enable a flexible movement of wheeled mobile robots among laboratories distributed in different floors. The automated handling system consists mainly of an ADAM module which has the ability to call the elevator to the robot's current floor and to request the destination floor. The LPS25HP pressure sensor attached to an STM32F411 microcontroller is utilized as a height measurement system to estimate the robot's current floor inside the elevator. The ultrasonic sensor is used to recognize the elevator's door status. Many challenges have to be solved to realize a stable height measurement system based on pressure sensor readings. The difference of the pressure sensor readings before and after soldering is realized by comparing the reading after soldering with an accurate barometric reading. In addition, the sensor output signal shows oscillation and wide variation of the same floor pressure sensor readings at different times. The oscillation in the output signal has been handled using a first order FIR smoothing filter. The first order filter was selected to balance between the stability and the elapsed time to receive the updated values. An auto-calibration stage is established to maintain the wide variation in the atmospheric pressure readings by calibrating the sensor readings with the robot's current floor before entering the elevator. An error handling management system is utilized to guarantee a stable automated elevator management system performance. Many experiments to assess and verify the performance of the automated elevator management system and robot's current floor estimation are reported. The experimental results show that the proposed methods and sub-systems developed for the mobile robot are effective and efficient in providing a transportation service in multiple-floor life sciences laboratories.

1. Introduction

This paper is an extension of work originally presented at the International Conference on Mechatronics - Mechatronika (ME 2016) [1].

Over the last years, mobile service robots have played critical roles in many applications, including (but not limited to) medical [2, 3] and cleaning services [4, 5], agriculture [6, 7], operations in hazardous environments [8, 9], construction and demolition [10], space [11] and military applications [12], materials transportation [13], support services for the elderly and handicapped [14], and entertainment [15]. In contrast to certain service mobile robots used for surveillance and cleaning services, which can utilize random movement in performing their tasks, the majority of mobile robot applications requires a sophisticated navigation ability in order to move in the desired manner. Thus, an autonomous navigation system is a fundamental functional requirement. Mobile robots in automated laboratories are usually used for the transportation of samples and other material between sub-systems located in different places. The mobile robot has to solve four basic problems in order to achieve highly effective autonomous navigation, which are mapping, localization, path planning, and action. The mobile robot must either climb stairs or use an elevator to deal with complex building structures with

[*]Corresponding Author: Ali Abduljalil Abdulla; Center for Life Science Automation (celisca), University of Rostock, Rostock 18119, Germany, and College of Engineering, University of Mosul, Mosul, Iraq,
E-mail: Ali.abdulla@celisca.de

multiple floor working environments. Legged robots and special designs of wheeled mobile robots [16] can climb stairs. However, climbing stairs during transportation tasks adds further challenges to the robot movement related to load capacity and balance preservation. Thus the elevator offers a suitable solution to this problem. An elevator handling system (EHS) includes many operations which must be accomplished in order to travel successfully on an elevator. The elevator entrance door and its status (open or closed) need to be recognized, the entry button must be detected, the destination floor button distinguished, and the current floor identified. Chen et al. presented a mobile service robot which has the abilities to both plan its own path as well as to call the elevator at each floor[17]. For the elevator calling operation, wireless control was designed. An Xbee module is employed to control the servo motor to push the elevator entry button. An RFID is utilized to approximate the current floor when the robot is inside the elevator. Experimental results demonstrated that the proposed mobile robot can ride the elevator and complete delivery tasks. This technique has many constraints, which would prevent its employment in life sciences laboratories. Firstly, each floor requires an electronic circuit consisting of an Xbee module, servomotors, microcontroller and power supply positioned beside the elevator for the call button pressing operation. Also, an RFID antenna would be required on each floor for current floor estimation, and finally this technique does not handle destination floor calling. Maneerat et al. presented a simple method for current floor estimation [18]. It is based on received signal strength (RSS) which is acquired from the number of wireless network sources installed on every floor. The main limitation of this method is the costs since it requires at least four wireless network sources on each floor to realize an acceptable success rate. Chen et al. presented a height measurement system using the MS5534B pressure sensor [19]. This sensor was attached to an LPC1114 microcontroller with the necessary peripherals attached to build the hardware platform. An experiment was conducted in a hotel covering 37 floors to examine the stability of the proposed system. However, this system does not take into consideration the wide variation in pressure sensor readings for one location at different times. Xia et al. presented a pedestrian floor estimation for smartphones [20]. A multi-reference barometer floor positioning method used multiple barometer sensors in each floor to determine a floor reference, and then the smartphone reading was compared with the floor reference to identify the current floor accurately. The method can overcome the problem of daily changes in pressure but the system will be expensive since it needs extra sensors for each floor, especially when applied with the highest number of floors and Floor-positioning servers. Frank et al. developed a motion estimation system to identify motion types, including walking, elevator, escalator, and stairs [21]. They use a Bayesian network to classify motion based on IMU and pressure sensors. The system estimates the kind of movement with value from 73.8 for falling detection to 95.5 for jumping. At CELISCA, researchers are working to develop the automated life science laboratory of the future. For a facility such as a laboratory, a system consisting of several H20 mobile robots has been established which allows samples, labwares and materials to be transported between individual automated sub-systems as support for laboratory staff or to replace them outside of working hours. In multi-floor transportation systems for mobile robots, the elevator aspect is a major topic to be handled. Two different working strategies are developed to guarantee secure glass elevator handling using automated and vision detection algorithms. Orders are sent to the elevator over Wi-Fi signals using the former, while the latter deals with the detection of entry and internal buttons and elevator door status, and controls the kinematic arm for the button-pressing operation. A pressure sensor is utilized as a stable floor estimation approach for both the automated and vision-based elevator handling.

In this paper, a new method is presented to provide a transportation mobile robot with a secured elevator facility to move in a multi-floor building environment. The automated elevator with its management system uses a Wi-Fi signal for data transmission (order/status exchange). The automated elevator management system translates the number of the called floor into a specific pin and port. Then the pin and port number with the “on” order are sent to the ADAM module. After 100ms, an “off” order is sent. The selected ADAM module has four digital inputs and four relay outputs. The module is connected to the elevator entry button of four floors in a laboratory building. The main limitation of the automated elevator is that it does not provide any feedback about the elevator's current floor or its door status, and thus pressure sensor based current floor estimation and an ultrasonic sensor based elevator door status reader are used. To utilize pressure as a current floor reader, many challenges must be considered. Firstly, differences in pressure sensor readings may occur before and after soldering. This problem is solved by applying a one-point calibration technique. Secondly, readings of the same floor pressure sensor vary at different times. A smoothing filter with an FIR structure is utilized to overcome small variations in sensor readings, while an adaptive calibration method handles wide variations during the week. Finally, the integration of the proposed elevator operation handling system in a real mobile robot transportation system in life sciences laboratories is initialized. Figure 1 shows the installed pressure sensor with its microcontroller board on the H20 Mobile robot.

Figure 1 Installed Pressure Sensor with its Microcontroller board

This paper is organized as follows: the system structure is described in section 2, while section 3 explains the automated elevator management system, floor estimation based on the pressure sensor and the related smoothing filter, calibration stages,

and integration into a real transportation system. The hybrid method to select a suitable elevator handling method is detailed in section 4. Section 5 describes four experiments conducted for system validation. The final section summarizes the results.

2. System Structure

At celisca, Rostock, Germany researchers have been developed a hierarchal management system to handle the processes of establishing a fully automated life science laboratories. This system has four control layers. The first layer has been divided into the Transportation and Assistance Control System (TACS), and Process Control Adapter System (PCAS). The TACS has a middle control layer which is the Robot Remote Center (RRC). This layer was established to manage the transportation tasks of mobile robots. It manages the robot by receiving a transportation order from the TACS then forwarding it to the Robot Board Computer (RBC) of an available robot which has the highest battery voltage. The Multi-Floor System (MFS) is the core component of the RBC level, and was developed to perform transportation tasks in multiple floor environments. As shown in Figure *2* the MFS is established to realize the functions of a multiple floor navigation system with a Robot Arm Kinematic Module (RAKM), Elevator Handling System (EHS), and Collision Avoidance System (CAS).

A multiple floor navigation system including mapping, indoor localization, path planning, an Internal Management Automated Door Controlling System, as well as the communication system, and Internal Battery Charging Management System has been reported earlier [22]. The MFS is integrated with the Elevator Handling System (EHS) to control elevator operation. The EHS consists of two sub-systems to handle the elevator operation as follows: elevator handling based computer vision [23] with kinematic arm solution for pressing operation [24]; and an automated elevator management system (AEMS). The Automated Elevator (AE) utilizes a WISE 4060 ADAM socket to call the elevator. An LPS25HB pressure sensor and STM32F411 microcontroller are used in the current floor estimation system hardware platform and are configured and programmed to overcome the inability of the AE to give any feedback about elevator's current floor. The LPS25HB sensor is employed to measure the atmospheric pressure and the STM32F411 microcontroller to receive the raw data from the pressure sensor board, to filter them, and to send them to the MFS over USB cable.

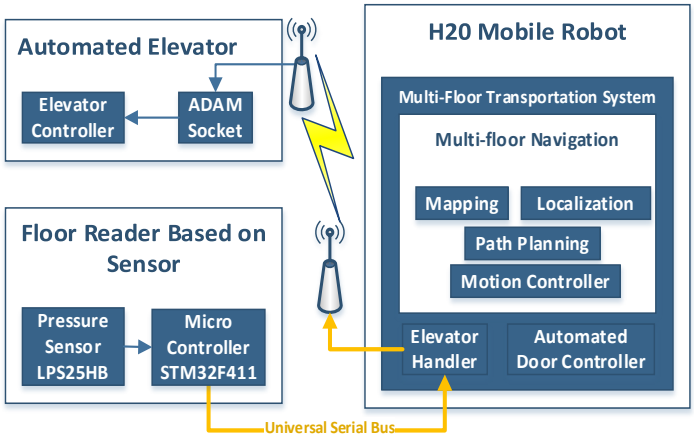


Figure 2 System Structure

3. Method Description

3.1. Automated Elevator Management System

The elevator is the key for the wheeled mobile robot to move among laboratories distributed on different floors in a complex building structure. Thus, it is necessary to establish a secure system to deal with the elevator. The system developed earlier was based on computer vision to deal with elevator operations (entry button detection, internal buttons recognition, and current floor estimation) [22], and the H20 arm solution [24] was developed for the button pressing operation. The weakness of the H20 arm joints can decrease the usability of this system. Thus, the AEMS is developed as a secure solution. This system controls the AE over Wi-Fi socket. The AE consists mainly of an ADAM module for calling the elevator and requesting the destination floor. WISE 4060 ADAM socket is a digital acquisition and control system over a computer network. It has the ability to collect data, and to remotely control the devices required to accomplishing integration between automated and enterprise systems through network technology. The ADAM mmodule consists mainly of a microcontroller, power circuit, Wi-Fi adaptor, four insulated digital inputs and four insulated high-speed relay outputs, and two watchdog timers (one for communication and the other for the system). It can be accessed using the IEEE 802.11b/g/n WLAN network either directly to the device (outdoor coverage area: 110m) or through a router/repeater to extend the coverage area. This module is embedded inside the elevator cabin and is connected directly to the elevator controller of the required floors to establish the AE as demonstrated in Figure 3. The AEMS is developed to connect the MFS with the AE. It translates the required elevator destination floor into a specific hardware port and pin numbers.

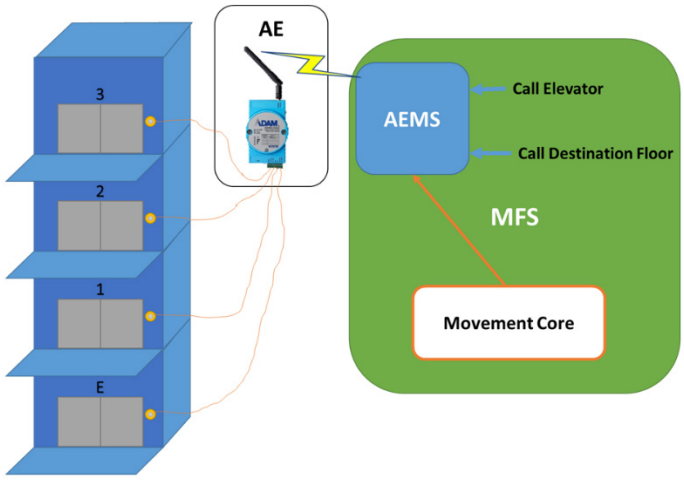


Figure 3 Automated Elevator with Adam Socket and Automated Elevator Management System

The AEMS calls the elevator to the robot's current floor when the mobile robot needs the elevator to move to another floor during a transportation task. To identify the robot's current floor outside the elevator, the indoor localization system developed earlier is utilized [22]. Inside the elevator, the MFS needs a logical solution to specify the destination floor. At this level, the movement core depends on the transportation task status (Grasp Position Done, Place Position Done, and Charge Position Done) to determine the current destination. When the robot enters the elevator, the movement core checks the current destination floor based on the current intermediate goal, as clarified in Table 1. For example, if

the grasping operation has been completed, the placing operation floor has to be requested. Then the elevator is directed to go to the required destination.

Table 1 Destination Floor Selection Strategy.

Grasp Position Done	Place Position Done	Charge Position Done	Entry Button	Internal Button
Not yet	Not yet	Not yet	Current Floor	Grasp Floor
Done	Not yet	Not yet	Current Floor	Place Floor
Done	Done	Not yet	Current Floor	Charge Floor

The reconnection function is developed to handle the weakness of the Wi-Fi signal which connects the AEMS with the AE. As mentioned above, the AE is embedded in the elevator passengers cabin while the AEMS is a part of MFS which is the core of the RBC. Thus, the weakness in the Wi-Fi signal is related to the mobile robot's large navigation area. This function starts to receive the updated elevator status. In case of missing data, the function closes the socket to AE and then initializes the connection again till a stable connection is realized.

3.2. Current Floor Estimation

The elevator is shared with human users in the building. Many people can use it simultaneously, and the elevator controller will respond to all requests based on the priority for moving direction (up/down).

For mobile robots, the current floor estimation technique is a topic aspect to consider. This technique is employed to inform the robot about its current floor inside the elevator. Earlier a vision system is utilized to recognize the current floor number. This was installed in the glassy elevator shaft for each floor to estimate the current floor [23]. This method has a high success rate in recognizing the correct floor in cloudy weather conditions. However, in some situations (such as a human forming an obstacle between the robot and the floor number indicator, or the disability of the vision sensor to recognize the floor number indicator at some times of the sunny day because of sunlight reflecting over the number mark) the robot completely fails to estimate its current floor. These limitations make this method an inappropriate solution for multiple floor transportation task since incorrect current floor estimation can make the robot lose its way to the destination. Thus, a new innovative method based on a height measurement system is developed to estimate the elevator current floor for mobile robot applications to enable a secure navigation system.

As a hardware platform for the height measurement system, the LPS25HB pressure sensor and STM32F411 microcontroller were configured and programmed to sense the environment and detect the current floor position. Many challenges have to be solved to use the pressure sensor as a floor estimation system. A soldering drift, which is defined as the difference between the accuracy of the sensor before and after soldering, appeared when the pressure sensor was attached to the STM32F411 microcontroller. A one-point calibration technique was used to solve the soldering drift problem by comparing the pressure sensor readings after attachment with a precision barometer. The difference was calculated and added as an offset to each pressure sensor reading. In addition, absolute digital barometer (pressure sensor) readings at the same floor of the building keep changing during the day due to various weather conditions. The oscillation in the output signal and the wide variations in barometric readings would reduce the applicability of this technique for floor detection. Two methods were applied to deal with the variations in pressure. Firstly, a smoothing filter with a finite impulse response (FIR) structure was used to solve the problem of small variations in pressure sensor readings. The FIR design diagram is shown in Figure 4, and equation (1).

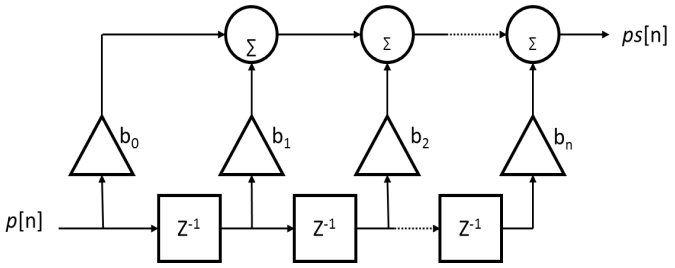


Figure 4 Finite impulse response filter

$$Ps[n] = b_0 p[n] + b_1 p[n-1] + ... + b_n p[n-N] \quad (1)$$

Where:

p[n]: is the input pressure reading.

Ps[n]: is the output smooth pressure value after the filter.

b_i: Filter coefficient.

N: Filter order.

The FIR filter waschosen due to its stability, linearity, and simplicity. A practical analysis was performed to choose the filter in order to achieve a balance between the time required and the stability of pressure readings. While the stability of the sensor readings is enhanced by increasing the filter order, the time required to complete the smoothing filter is also increases. In real transportation task applications, the elapsed time to estimate the current robot's floor is an essential aspect to manage since the robot must estimate its current floor, distinguish the elevator door status, and leave the elevator before the elevator door closes. Thus, a first order smoothing filter has been selected. The smoothing filter is embedded inside the microcontroller and the MBED development environment is used to program the microcontroller using the C language. Figure 5 shows the implemented flowchart for a microcontroller which starts by initializing it with the necessary setting, then collecting the pressure sensor data, applying the smoothing filter to these data with a special procedure on the first occasion to stabilize the output, and finally sending it over USB to the robot computer every 1s. Secondly, an adaptive calibration method is used to calibrate the sensor readings to the robot's current floor before entering the elevator in order to overcome the wide variations in daily pressure readings. According to an experiment conducted over one week (the second experiment reported in the results section), the variation can reach up to approximately one hundred meters. The indoor localization method developed earlier, which is based on passive landmarks for a multi-floor environment, was utilized to identify the robot's current floor number outside the elevator.

The calibration value is calculated as follows. Firstly, the multiple floor localization method developed earlier is used to find the current floor information outside the elevator based on the landmark ID located near the elevator. Then the pressure sensor reading is converted into the height using equation 2. The extracted

height is calibrated with the current floor height to realize a constant number which is the calibration value. Then the calibration value is added to current floor height. Finally, the extracted height is compared with the height range of each floor to identify the current robot's floor (see Figure 6). The reconnection function is developed to maintain a stable connection to the sensor board by continuously checking the incoming message. In the case of missing data, the serial port will close and initialize again till receiving the updated data.

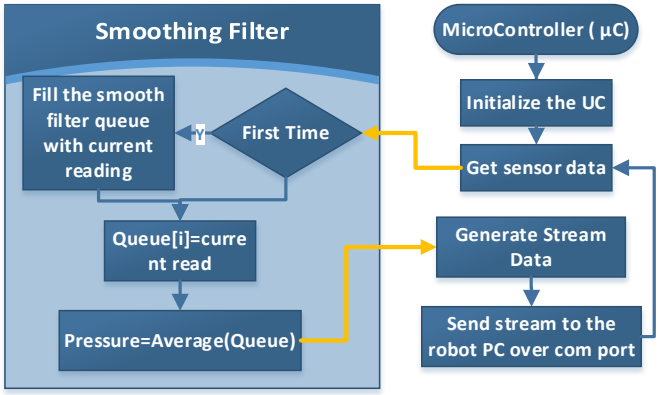


Figure 5 Microcontroller flowchart

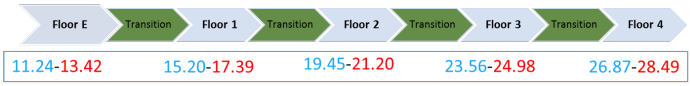
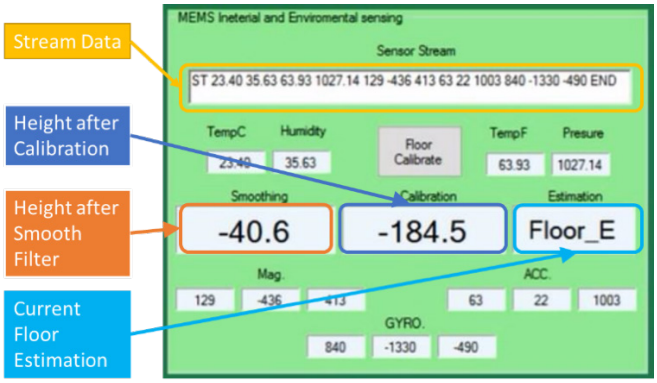


Figure 6 The Range of Height Measurements for each Floor

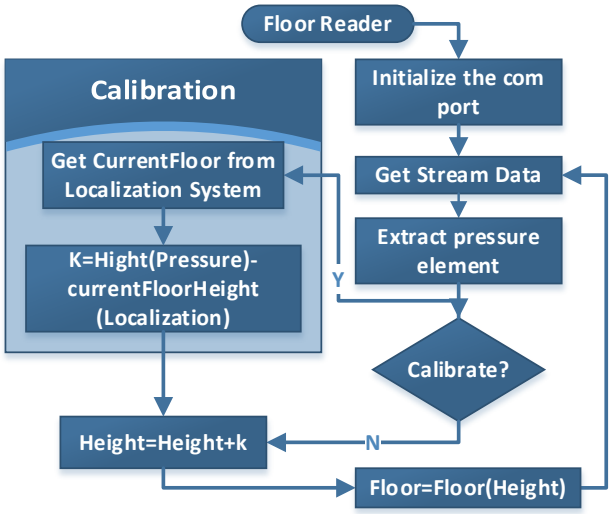
The MFS is embedded with the required GUI and coded with C# to handle the connection to the microcontroller; the incoming data and calibration stage are shown in Figure 7.

During a multi-floor transportation tasks, the ultrasonic sensor is used to recognize the elevator door's status. When the destination floor matches the estimated robot's current floor and the elevator's door status is recognized as "open", the robot leaves the elevator to complete the transportation task process.

$$Height(p) = \frac{1-\left(\frac{p}{1013.25}\right)^{0.190284} * 145366.45}{3.2808} \quad (2)$$



(a) Developed graphical user interface



(b) Floor estimation flowchart inside multi-floor system

Figure 7 Embedded floor estimation inside MFS

4. Elevator Handling Error Management System

The proposed detailed robot-elevator interface control strategy is illustrated in Figure 8. The EHS classifies the error according to its appearance space as either outside or inside elevator error handling. The elevator handling system has two methods for handling outside elevator errors. If the mobile robot fails to reach the required elevator button pushing area accurately which is usually due to wheel slip, the position and orientation correction function checks the robot's position after movement and tries to correct it three times. Next, the elevator handling system based on computer vision and an arm solution starts. After the robot has found the button's position as X, Y, Z coordinates, the values are then passed to the robot arm kinematic module for button pressing operation. The ultrasonic sensor is then utilized to distinguish the elevator door status. If the door is recognized as closed, the EHS controls the robot kinematic arm again to press the detected button. If previous attempts have failed, the EHS selects the AEMS over the Wi-Fi socket to open the door. If the functions listed above have failed in handling the elevator correctly, the MFS reports to the higher level controller RRC, terminates the current transportation operation, and directs the robot to the charging station.

The error handling system inside the elevator starts by monitoring whether or not the destination floor has been reached by continuously checking the current floor reader. If it does not reach the destination floor after a specified number of attempts, the MFS sends a warning alarm to the RRC to initiate the appropriate procedures. After finding the destination floor, the next stage of error handling is enabled to check the elevator door status, and another warning alarm is sent to the RRC if the door is still closed after a defined number of attempts. The final error handling maneuver is to check if the floor reached matches the required destination floor based on the landmarks installed outside the elevator. If the landmark is missing or a wrong floor number landmark is read, the error handling system returns the robot back to the elevator and chooses the destination floor again. Finally, if the robot fails to reach the destination floor after reaching the maximum allowed number of attempts, an error message is sent to the RRC. These steps are arranged as a flowchart in Figure 9.

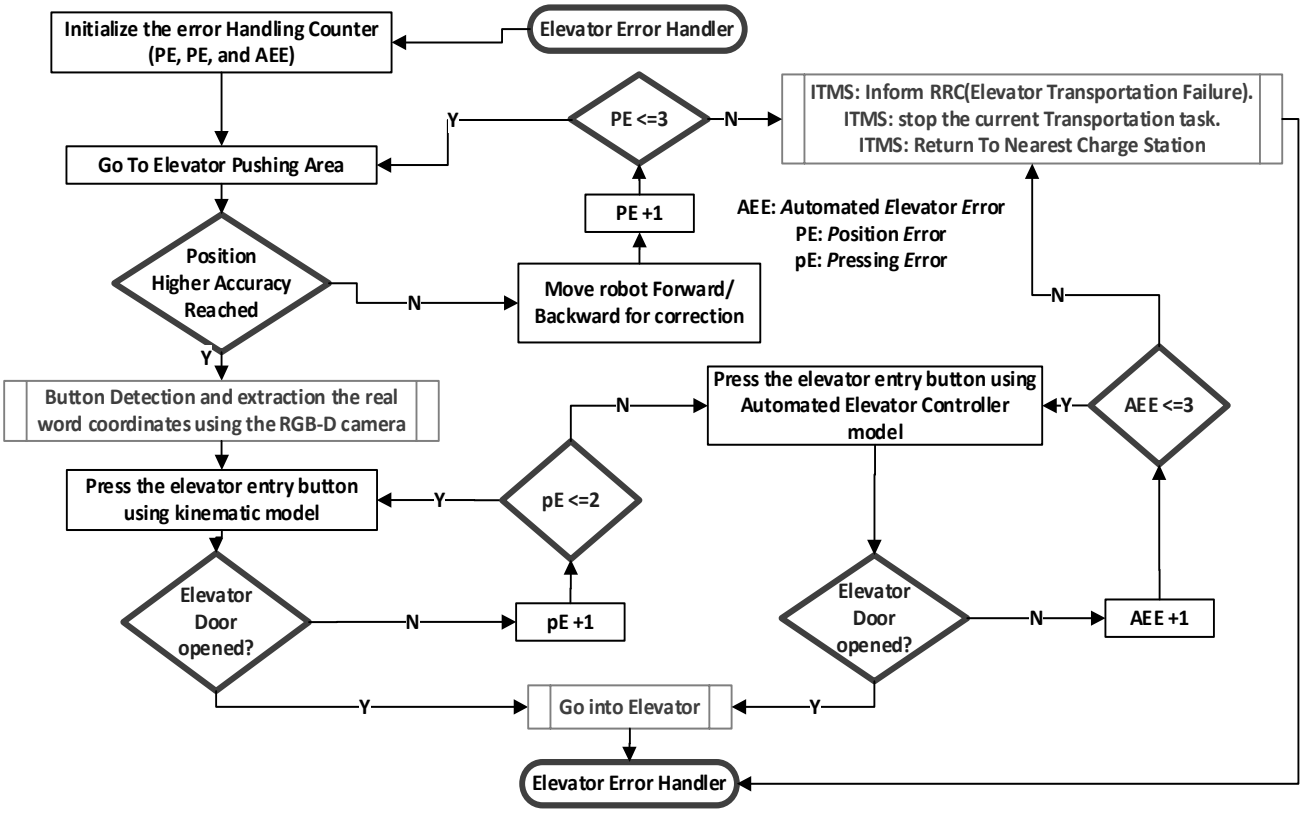


Figure 8 Outside Elevator Error Handling Flow Chart.

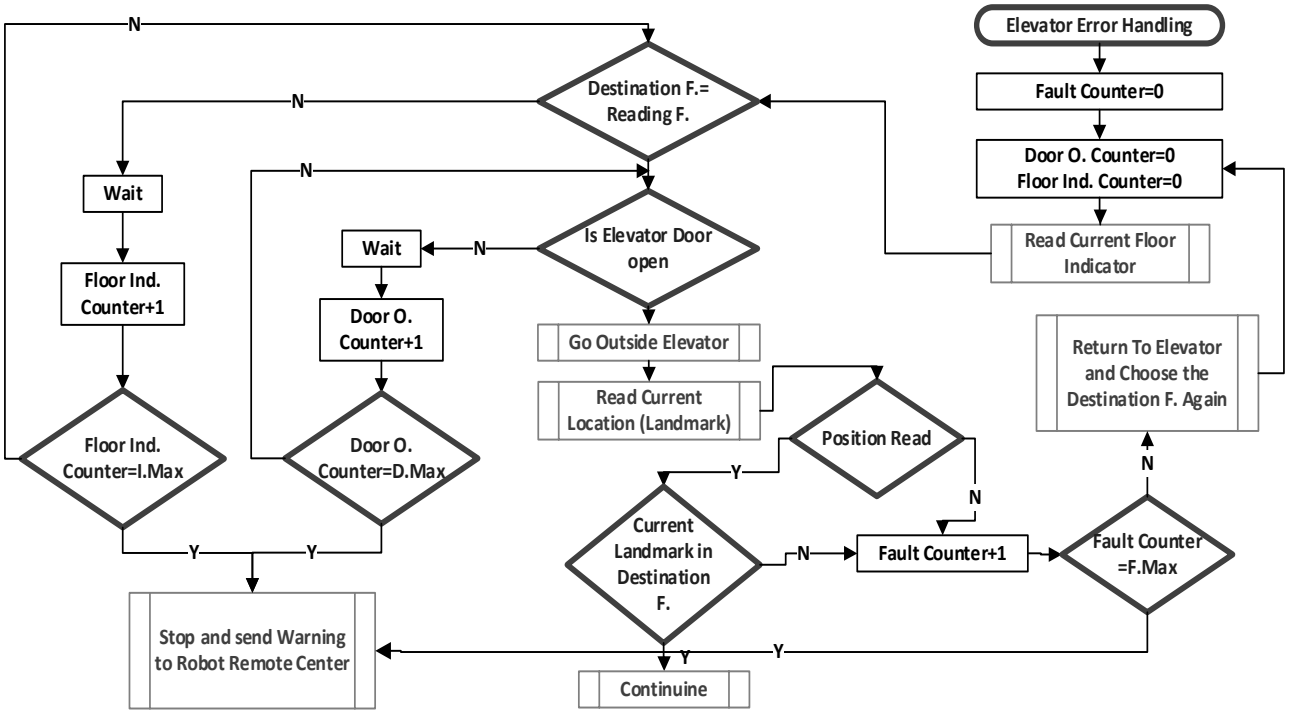
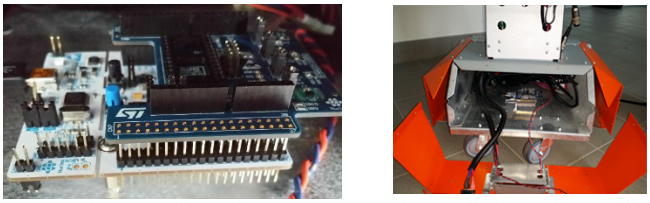


Figure 9 Inside Elevator Error Handling Flowchart

5. Results

Four sets of experiment have been performed to prove the efficiency of the developed system. The first experiment was conducted to check the stability of the pressure sensor before and after smoothing filter application. The smoothing filter was applied with different parameters to find an acceptance performance level to achieve a balance between filter performance and elapsed time. The second experiment was executed to validate the stability of atmospheric pressure readings over time. Then the integration into a real mobile robot-based transportation system was examined, and finally the system's extension for multiple robots was verified as follows.

The IKS01A1 MEMS inertial and environmental sensor board has the pressure sensor LPS25HB. This board was attached to the STM32F411 microcontroller board as shown in Figure 10, and the developed system for data acquisition was written in the C# language. Pressure sensor readings were collected from five floors with one hundred pressure readings for each floor. The pressure sensor data before applying smoothing are shown in Figure 11. Next, the smoothing filter with first order was applied to the acquired pressure sensor readings (see Figure 12), and finally the second order smoothing filter was applied to the collected data to validate performance (see Figure 13). Table 2 shows the average and possible tolerance for pressure sensor readings before and after smoothing filter application. From Table 2, it can be seen that the oscillation of the pressure sensor readings for the same floor at different times can exceed the distance between two floors, which makes decisions about floor identification impossible. Thus, the acquired data has to be filtered, and the second order filter gives higher smoothing performance than the first order smoothing filter, but the required time can cause the robot to miss the destination. Thus the first order smoothing filter was chosen to determine the floor within an acceptable time.



(a) Microcontroller and Sensor Board

(b) Installation Inside Mobile Robot

Figure 10 Floor Estimation Hardware

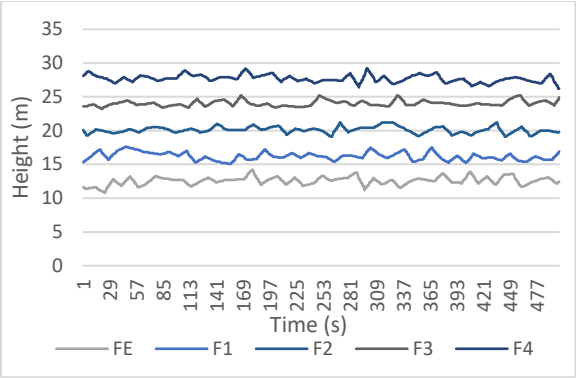


Figure 11 Pressure Raw Data

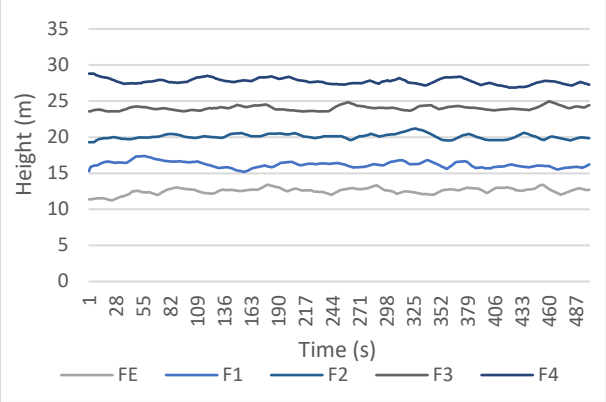


Figure 12 First Order Smoothing Filter

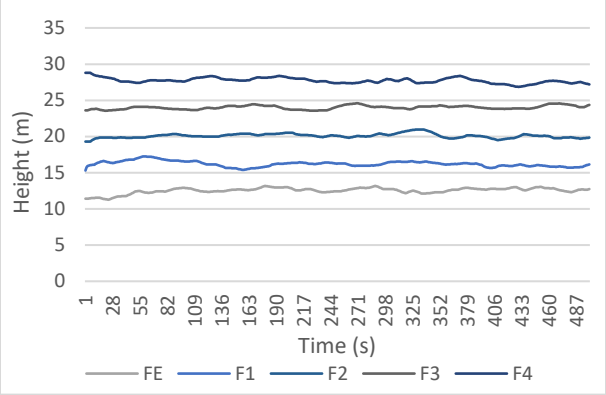


Figure 13 Second Order Smoothing Filter

Table 2 Height measurement in meter with different filter order

Floor	Raw		First Order		Second Order	
	Average	+-	Average	+-	Average	+-
4	27.696	1.29	27.694	0.81	27.694	0.768
3	24.032	0.965	24.03	0.71	24.038	0.52
2	20.108	1.015	20.11	0.83	20.116	0.73
1	16.211	1.245	16.208	1.097	16.209	0.928
e	12.586	1.59	12.598	1.09	12.609	0.803

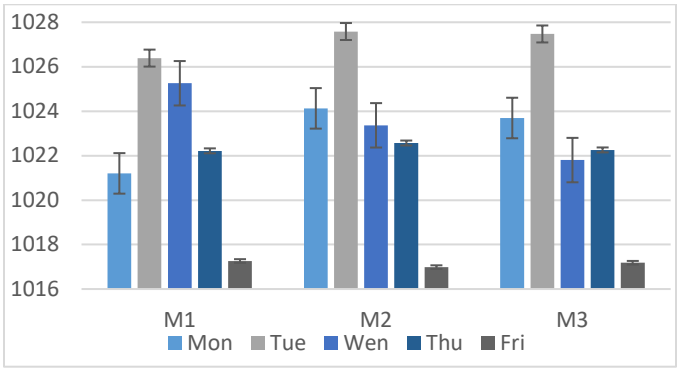
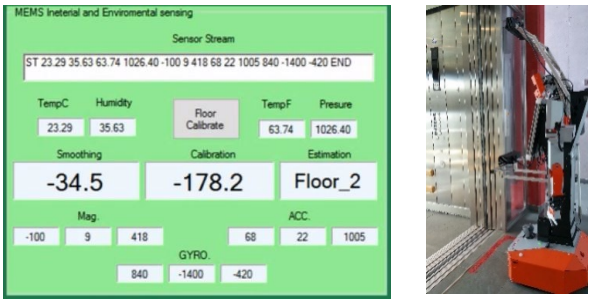


Figure 14 Daily Pressure

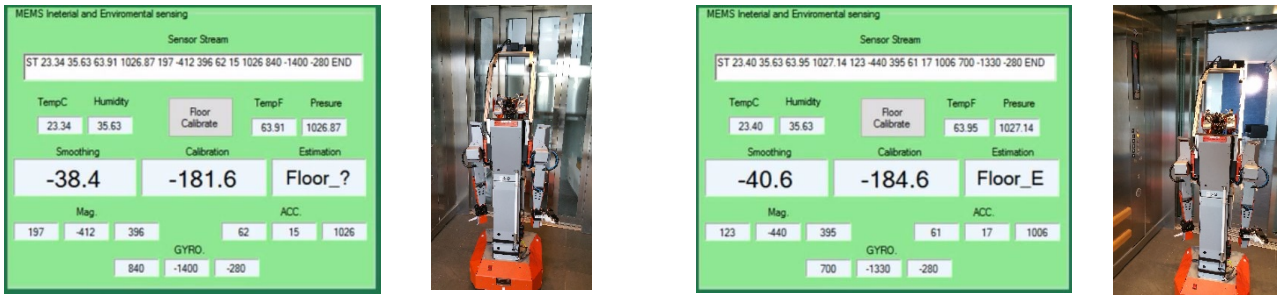
The second experiment was conducted to check the variance in daily pressure and the possibility of using the developed floor estimation method directly. During five working days, pressure sensor readings were taken and recorded. Three hundred readings were taken each day distributed equally among three times (morning, midday, and afternoon). From Figure 14, it can be noticed that the pressure sensor gives wide variation even on the same day, and thus the pressure data require an adaptive calibration stage.

The third set of experiments was conducted in order to evaluate the performance of the AEMS and current floor estimation technique based on the height measurement system. A transportation task between laboratories located on the second and third floors was executed one hundred times in the CELISCA building. The transportation task started from the second floor, moving to the third floor for placing and finally returning to the second floor for charging. The AEMS calls the elevator to the robot's current floor for it to move to the third floor.

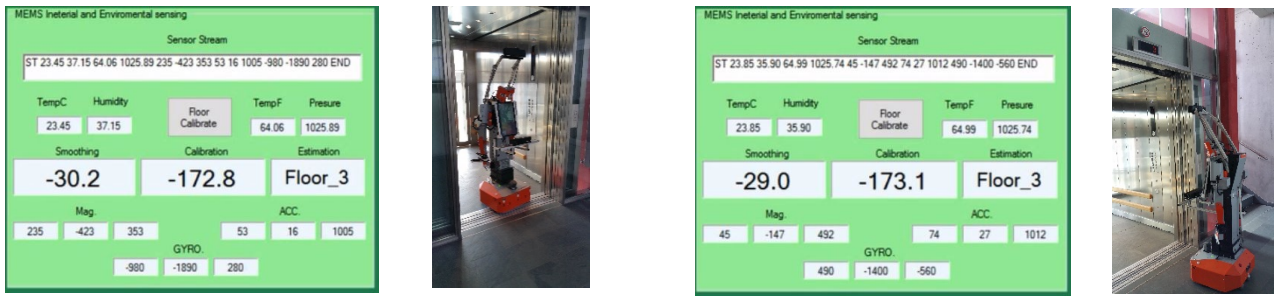
(a) Robot Reaches Elevator Before Calibration



(b) After Calibration



(c) Elevator in movement (d) Wrong Floor



(e) Leave the elevator at 3rd flor (f) Call the elevator from 3rd floor

(g) Wrong Floor (h) Leave the elevator at 2nd floor

Figure 15 Transportation Experiment

Inside the elevator, the ground floor was chosen by another passenger before the robot called the 3rd floor. This stage was included to examine the current floor estimation method. At first, the elevator controller directed the elevator to the ground floor and the elevator's door opens. At this stage, the robot stays in the elevator since the detected current floor estimation does not match the destination floor. The robot waits until the detected current floor matches the destination floor and the elevator's door is recognized as open. Also, the same procedures were repeated when the robot had to return to the 2nd floor for charging, where the 4th floor was selected by another passenger before the robot called its destination. Each time, the robot succeeding in leaving the elevator at the correct floor. The developed system's stability and reliability is proven with a 100% success rate. Figure 15 demonstrates the experiment steps in detail.

The final experiment was executed to validate the possibility of extending the method of current floor estimation based on the height measurement system to other mobile robots. The same kind of sensor board with the same processing stages (a one-point calibration stage, first order FIR smoothing filter, and an adaptive calibration stage) was embedded for three other mobile robots. The procedures used in the third experiment were repeated ten times for each robot to validate their ability to estimate the current floor. Each robot succeeded in estimating its current floor inside the elevator without any change needed in the sensor board or the processing stages.

6. Conclusion

In this paper, a secured method for the elevator handling system based automated elevator is presented. To handle the elevator, four main operations should be considered which are elevator requesting, destination floor calling, current floor estimation, and elevator's door status recognition. The automated elevator handling system mainly consists of WISE 4060 ADAM modules. A Wi-Fi socket is established to connect the automated elevator management system with the ADAM module. This module can call the elevator to the current floor and request the current floor when the robot enters the elevator. However, this module does not provide any feedback about the robot's current floor inside the elevator which may make the robot lose its way to the destination. Thus, an innovative and stable current floor estimation based on a height measurement system is presented. The height measurement system consists of the LPS25HB pressure sensor attached to the STM32F411 microcontroller. To utilize the pressure sensor as a height measurement system many aspects must be considered which as follows: firstly, a soldering drift,

which is defined as the difference between the accuracy of the sensor before and after soldering. A one-step calibration stage is employed to solve this problem. This technique compares the pressure sensor reading after soldering with a precision barometer. The difference is added to each pressure reading. Secondly, the oscillation of the output signal and the wide variations in barometric readings during the week. The first order FIR smoothing filter is employed to overcome the oscillation in the pressure sensor output signal while an adaptive calibration stage is used to handle the wide variation. The ultrasonic sensor is used to recognize the elevator's door stats. An error management system is developed to guarantee a secured elevator handling system. A series of experiments has been conducted to evaluate the variation in the pressure sensor readings and to validate the performance of the presented systems. The results of the experiments prove the efficiency of the automated elevator management sensor and robot's current floor inside the elevator with a success rate reaches to 100% for each.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The authors would like to thank the Ministry of Higher Education and Scientific Research in Iraq for the scholarship provided by Mosul University, as well as the Federal Ministry of Education and Research Germany for the financial support (FKZ: 03Z1KN11, 03Z1KI1).

References

[1] A. A. Abdulla, H. Liu, N. Stoll, and K. Thurow, "An automated elevator management and multi-floor estimation for indoor mobile robot transportation based on a pressure sensor," in 2016 17th International Conference on Mechatronics - Mechatronika (ME), 2016, pp. 1–7.

[2] A. V. Savkin and C. Wang, "A framework for safe assisted navigation of semi-autonomous vehicles among moving and steady obstacles," Robotica, vol. 35, no. 5, pp. 981–1005, 2017.

[3] V. Subbian, J. J. Ratcliff, J. M. Meunier, J. J. Korfhagen, F. R. Beyette Jr., and G. J. Shaw, "Integration of new technology for research in the emergency department: Feasibility of deploying a robotic assessment tool for mild traumatic brain injury evaluation," IEEE J. Transl. Eng. Health Med., vol. 3, 2015.

[4] D. Leidner, A. Dietrich, M. Beetz, and A. Albu-Schäffer, "Knowledge-enabled parameterization of whole-body control strategies for compliant service robots," Auton. Robots, vol. 40, no. 3, pp. 519–536, Mar. 2016.

[5] A. Ravankar, A. A. Ravankar, Y. Kobayashi, and T. Emaru, "On a bio-inspired hybrid pheromone signalling for efficient map exploration of multiple mobile service robots," Artif. Life Robot., vol. 21, no. 2, pp. 221–231, Jun. 2016.

[6] J. Zhang, D. Chen, S. Wang, X. Hu, and D. Wang, "Design and experiment of four-wheel independent steering driving and control system for agricultural wheeled robot," Nongye Gongcheng XuebaoTransactions Chin. Soc. Agric. Eng., vol. 31, no. 18, pp. 63–70, 2015.

[7] J. Bell, B. A. MacDonald, H. S. Ahn, and A. J. Scarfe, "An Analysis of Automated Guided Vehicle Standards to Inform the Development of Mobile Orchard Robots," IFAC-Pap., vol. 49, no. 16, pp. 475–480, Jan. 2016.

[8] L. Cragg, H. Hu, and N. Voelker, "Modularity and mobility of distributed control software for networked mobile robots," Springer Tracts Adv. Robot., vol. 30, pp. 459–484, 2007.

[9] B. Siemiatkowska, B. Hrasymowicz-Boggio, and M. Wisniowski, "The application of mobile robots for building safety control," J. Autom. Mob. Robot. Intell. Syst., vol. 10, no. 2, pp. 9–14, 2016.

[10] D. Schmidt, C. Hillenbrand, and K. Berns, "Omnidirectional locomotion and traction control of the wheel-driven, wall-climbing robot, " Robotica, vol. 29, no. 7, pp. 991–1003, Dec. 2011.

[11] M. Yim, K. Roufas, D. Duff, Y. Zhang, C. Eldershaw, and S. Homans, "Modular Reconfigurable Robots in Space Applications," Auton. Robots, vol. 14, no. 2–3, pp. 225–237, Mar. 2003.

[12] D. Culler and J. Long, "A Prototype Smart Materials Warehouse Application Implemented Using Custom Mobile Robots and Open Source Vision Technology Developed Using EmguCV," Procedia Manuf., vol. 5, pp. 1092–1106, Jan. 2016.

[13] K. Kashiwazaki et al., "A car transportation system using multiple mobile robots: iCART II," in 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2011, pp. 4593–4600.

[14] S.-N. Yu, J.-H. Jang, D.-H. Kim, J.-Y. Lee, and C.-S. Han, "A gait-assistive mobile robot based on a body weight support and autonomous path tracking system," Proc. Inst. Mech. Eng. Part C J. Mech. Eng. Sci., vol. 226, no. 3, pp. 828–841, Mar. 2012.

[15] I.-H. Kuo, C. Jayawardena, E. Broadbent, and B. A. MacDonald, "Multidisciplinary Design Approach for Implementation of Interactive Services," Int. J. Soc. Robot., vol. 3, no. 4, pp. 443–456, Nov. 2011.

[16] T. He, M. Bando, M. Guarnieri, and S. Hirose, "The development of an autonomous robot system for patrolling in multi-floor structured environment," Int. J. Autom. Technol., vol. 6, no. 1, pp. 13–21, 2012.

[17] Chen, L.-K. and Hsiao, M.-Y., "Control of service robot by integration of multiple intermittent sensors," in AMPT 2013, Taipei; Taiwan, 2014, vol. 939, pp. 609–614.

[18] K. Maneerat, C. Prommak, and K. Kaemarungsi, "Floor estimation algorithm for wireless indoor multi-story positioning systems," in 2014 11th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), 2014, pp. 1–5.

[19] L. Chen, B. Sun, and X. Chang, "MS5534B Pressure Sensor and Its Height Measurement Applications," in 2011 International Conference on Information Technology, Computer Engineering and Management Sciences (ICM), 2011, vol. 1, pp. 56–59.

[20] H. Xia, X. Wang, Y. Qiao, J. Jian, and Y. Chang, "Using Multiple Barometers to Detect the Floor Location of Smart Phones with Built-in Barometric Sensors for Indoor Positioning," Sensors, vol. 15, no. 4, pp. 7857–7877, Mar. 2015.

[21] K. Frank, E. M. Diaz, P. Robertson, and F. J. F. Sánchez, "Bayesian recognition of safety relevant motion activities with inertial sensors and barometer," in 2014 IEEE/ION Position, Location and Navigation Symposium - PLANS 2014, 2014, pp. 174–184.

[22] A. A. Abdulla, H. Liu, N. Stoll, and K. Thurow, "A New Robust Method for Mobile Robot Multifloor Navigation in Distributed Life Science Laboratories," J. Control Sci. Eng., vol. 2016, Jul. 2016.

[23] A. A. Abdulla, H. Liu, N. Stoll, and K. Thurow, "A Robust Method for Elevator Operation in Semioutdoor Environment for Mobile Robot Transportation System in Life Science Laboratories," in International Conference on Intelligent Engineering Systems (INES), 2016 IEEE International, Budapest, Hungary, 2016, pp. 45–50.

[24] M. M. Ali, H. Liu, R. Stoll, and K. Thurow, "Intelligent Arm Manipulation System in Life Science Labs Using H20 Mobile Robot and Kinect Sensor," in IEEE International Conference on Intelligent Systems (IS'16), Sofia, Bulgaria, 2016, pp. 382–387.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1609-1615 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Call Arrival Rate Prediction and Blocking Probability Estimation for Infrastructure based Mobile Cognitive Radio Personal Area Network

Neeta Nathani*[1], Gopal Chandra Manna[2]

[1]*Research Scholar, Department of Electronics Engineering, G.H. Raisoni College of Enginnering, Nagpur, Maharashtra, 440016, India*

[2]*Retd. Chief General Manager, Indian Telecommunication Service, BSNL, Jabalpur, Madhya Pradesh, 482002, India*

ARTICLE INFO

Article history:
Received: 16 June, 2017
Accepted: 29 July, 2017
Online: 24 August, 2017

Keywords:
Cognitive-Radio Base Station
Blocking Probability
Grade of Service
Poisson's model

ABSTRACT

The Cognitive Radio usage has been estimated as non-emergency service with low volume traffic. Present work proposes an infrastructure based Cognitive Radio network and probability of success of CR traffic in licensed band. The Cognitive Radio nodes will form cluster. The cluster nodes will communicate on Industrial, Scientific and Medical band using IPv6 over Low-Power Wireless Personal Area Network based protocol from sensor to Gateway Cluster Head. For Cognitive Radio-Media Access Control protocol for Gateway to Cognitive Radio-Base Station communication, it will use vacant channels of licensed band. Standalone secondary users of Cognitive Radio Network shall be considered as a Gateway with one user. The Gateway will handle multi-channel multi radio for communication with Base Station. Cognitive Radio Network operators shall define various traffic data accumulation counters at Base Station for storing signal strength, Carrier-to-Interference and Noise Ratio, etc. parameters and record channel occupied/vacant status. The researches has been done so far using hour as interval is too long for parameters like holding time expressed in minutes and hence channel vacant/occupied status time is only probabilistically calculated. In the present work, an infrastructure based architecture has been proposed which polls channel status each minute in contrary to hourly polling of data. The Gateways of the Cognitive Radio Network shall monitor status of each Primary User periodically inside its working range and shall inform to Cognitive Radio- Base Station for preparation of minutewise database. For simulation, the occupancy data for all primary user channels were pulled in one minute interval from a live mobile network. Hourly traffic data and minutewise holding times has been analyzed to optimize the parameters of Seasonal Auto Regressive Integrated Moving Average prediction model. The blocking probability of an incoming Cognitive Radio call has been assessed as a function of instantaneous occupancy and the blocksize of trunk servers.

1. Introduction

The mobile communication network is faced with serious challenges to ensure optimum Quality of Service (QoS). In order to survive in this present economic situation, the network operators must put into consideration that the available resources must be effectively used with optimum QoS. This leads to design of effective network and effective network planning. It can be achieved through parameters like call arrival rate, optimum holding time, blocking probability, etc. These parameters are based on available resources and traffic load in the mobile communication network. The proposed work is aimed at developing a traffic model that will predict blocking probability for voice calls in Global System for Mobile Communication (GSM) network.

To have interference free communication, it is imperative that the SUs are able to distinguish between the channels those are currently occupied by PUs and transmit result to Cognitive Radio-Base Station (CR-BS). The CR-BS predicts the availability of vacant channels depending on the past sensing results which directs a channel requesting Secondary User (SU) to associate

*Corresponding Author: Neeta Nathani, G.H.R.C.E. Nagpur, Maharashtra, India, +919300986070, E-mail: neeta_nathani@yahoo.com

with vacant channels accordingly. The proposed model shall assess daily traffic pattern based on monthly observation, hourly traffic pattern based on daily observation and minutewise traffic pattern based on hourly traffic observation particularly in and around Primary User (PU) busy minutes when scarcity for PU channels are at peak.

Assume the SUs in the area are covered through ISM band and available with sufficiently high Carrier-to-Interference and Noise Ratio (CINR) so that this part of the network is not a bottleneck. The procedure for initial selection of channel is described as follows:

(a) The Gateway (GW) node scans the area where SU allocation is to be made. It informs to CR-BS about the PU's activities in the area along with their activities like idle/dedicated and since when it was dedicated.

(b) Listing of all the GWs under the CR-BS is done.

(c) The coverage of the GWs is estimated from different CR-BS's and is updated in the CR-BS table. The coverage, signal strength, signal quality, etc. are also to be noticed in the table.

(d) Hourwise prediction of weekly occupancy is done at day = dd, hour = hh using static database from dd = 1 to (dd-1) that holds occupancy records of PU on monthly basis. The estimated value of occupancy obtained at hour = hh helps to estimate minutewise occupancy of PU at time t, i.e., at minute = mm.

(e) A dynamic database for minutewise occupancy of PU is obtained and updated. Whenever SU initiates a call request at an instant, the computation of blocking probability is done for each channel.

(f) The list of channels are available for use by the SU at minute = mm+1 which have blocking probability < 0.5.

Blocking probability (BP) at a moment of call demand can be assessed by channel occupancy rate amid last clock hour, e.g., 9am-10am for call request at a moment between 10am-11am, and so on, as in traditional teletraffic hypothesis. This estimation has been additionally enhanced through expectation models taking 9am-10am hour information for a few earlier days. The CR-BS might keep up and occasionally refresh database of channel occupancy during (a) every hour of a day, and (b) minutewise channel occupancy during each hour. Spectrum detecting is a key capacity of intellectual radio to keep the unsafe impedance with PUs and recognize the accessible range for improving utilization of spectrum use. However, the performance in detection techniques is regularly traded off with multipath blurring, shadowing and recipient instability issues. To alleviate the effect of these issues, cooperative spectrum sensing (CSS) has been appeared to be a viable technique to enhance the detection performance by exploiting spatial diversity [1,2].

The CSS is grouped into three classifications as; centralized, distributed, and relay-assisted [3]. In the centralized CSS, the master node collects the sensing information from all the other nodes within the network and the information is shared with all the CR nodes. In distributed CSS, each CR node acts as an independent node for sharing information between various users. Relay assisted nodes are helpful for extending the coverage area of the network.

In all the three detecting techniques listed above, the agreeable CR nodes with fusion centre takes decision about vacant channels. It doesn't unequivocally specify which SU might get benefit from which CR node in its physical range. Likewise, CR node takes choice of channel accessibility relying upon the then circumstance which may not be sane. Further, CR nodes are not anticipated that would keep up a database for choice support on the selected channels. To comprehend these restrictions in the adhoc CSS, a CR-BS design is proposed. In the CR-BS engineering, a Cognitive Radio- Base Station Controller (CR-BSC) is provisioned which is relied upon to pronounce the channel accessibility in view of the present channel occupancy and the previous history of occupancy of the channels. In this architecture,

(i) SU works as probe for occupied primary and secondary channels accessible in the environment with location data through GPS and transmit to CR–BS for intermittent updation of channel status database.

(ii) To decide if a channel is involved or not, a filtering recurrence is set and for all useful purposes, this examining recurrence is set as one minute. Consequently, when a SU makes a channel demand to CR-BS, CR-BS surveys the accessibility of each channel as a combination of current occupancy in addition to expected vacant time in future from its occupancy history. In this way, occupancy history is atomized in minutes.

(iii) A database is set up in CR-BSC which stacks information in minutes, collected in hours and in days.

A thorough forecast model is utilized to assess the normal vacant time from the time of demand by SU. This count is done at CR-BSC for all accessible vacant channels and the best one is chosen for allocation to the SU.

Several researchers have recommended methods for spectrum sensing, filtering and dynamic spectrum access for use in CR. Probability of channel availability, duration etc. has also been modeled. These models do not identify individual channels available for cognitive radio use effectively. These aspects have been discussed in section 2 viz. Literature Survey. The CR network architecture modeled in the present work is based on observation of 50 cells data for 36 days of PU network. The background study for present work include channel occupancy during a typical day and holding time distribution and has been discussed in Section 3. Section 4 discusses about architecture for infrastructure based CR. Collection of PU data and its call arrival rate prediction has been discussed in Section 5 followed by the blocking probability in Section 6. The relevance of the proposed work is concluded with results in Section 7 followed by the future work in Section 8.

2. Literature Survey

Cognitive Radio SUs accesses unallocated channels to PUs in licensed bands. Secondary users need to dynamically detect and access unused channels of PUs. Spectrum sensing consumes considerable energy, and hence, predictive methods for concluding gross availability of spectrum holes and idle channels can be used to schedule scan frequency and optimize energy consumption of secondary users.

The authors of the paper [4] propose a CR network architecture made out of an arrangement of SUs with or without a secondary Base Station in which CSS is utilized. At the point when environment is highly dynamic, the detected data may not be valuable because of user's mobility, channel blurring, and so on. The CSS is not an ideal strategy, as individual sensor nodes may encounter heterogeneous false alarm and missed detection because of their diverse locations.

Xiukui Li and Seyed A. (Reza) Zekavat [5] have accepted that different channels which are licensed to PUs, from which SUs can choose any idle channel for correspondence. A method is recommended that empowers SUs to assess direct accessibility in cognitive radio systems when SUs exist together with PUs. The call arrival s of PUs is monitored by the SUs. The simulations are done with the PU data to study the effect of the proposed prediction technique on the communication and have been compared with Seasonal Auto Regressive Integrated Moving Average (SARIMA) model. This model does not specify about the channel updation technique.

The author in [6] discusses about an infrastructure-based CR network architecture integrated in a PU network operating in licensed spectrum band where, the CR network has a centralized network entity such as a base station and associated CR nodes. The proposed architecture overcomes the drawbacks of the Poisson's modeling.

A two phase joint power and spectrum allocation scheme for a Cooperative Cognitive Radio Network (CCRN) has been discussed in [7]. The PU leases unused licensed spectrum to various SUs utilizing Frequency Division Multiple Access (FDMA) technique. This paper considers the design with respect to PU-SU. This paper lacks spectrum sensing and is, thus, not a pure CR situation.

The authors [8,9] have presented different traffic forecast procedures and have discussed about the process of assessing channel accessibility through predicting traffic patterns of PUs for cognitive radios in an infrastructure of CR network as clarified in [5] above. In this model, SUs can forecast the call arrival rate and call holding time of primary users that utilized this channel. At that point, as per the prediction as well as estimation results, secondary users can assess the probability that the channel would be accessible for a given time period or not. By comparing the assessed probability with some threshold, secondary users can choose whether to utilize this channel. This model doesn't unequivocally express the need of infrastructure. The sensing, computation and prediction needs high processing ability and more power and are thus hard to implement in practice.

An infrastructure based CR network that empowers cooperative spectrum sharing over different service providers by means of static CR nodes have been proposed in [10,11]. At the point when all the channels of a service provider are accessed or it is over-burden, it can utilize the available channels that are authorized to the service providers that are under loaded. This is not a pure CR subject.

In this manner, it is seen from above literature surveys that no document is sufficiently complete to quantify the availability of vacant channels with adequately free holding time to accommodate a secondary user call. In the present work, the prediction of availability of vacant channels for the proposed holding time of the secondary user is done. These aspects have been incorporated into the present review and results are required to be very helpful to CR building architects.

3. Background Studies

3.1 Daily Traffic Pattern of PU

Conventionally, a telecom operator analyzes the total traffic on hourly basis and identifies the busy hour where total traffic is maximum. Hourly data arranged on weekly basis does not give a clear picture of the peak hour as it contains many peaks. The traffic variations within a clock hour are not predictable in weekly analysis. Thus, data arranged on daily basis is taken to estimate the behavior of hourly traffic. The minutewise collected occupancy data was taken on hourly basis for 50 cells of different channel numbers. Figure 4.3 depicts daily occupancy pattern for the locations with 60 channels (8 RFs) and differently loaded at various traffic places. The results are similar for other Radio Frequency (RF) counts also. The figure indicates double heaps in the channel occupancy. The heap pattern shows near parabolic nature from 05:00 to 15:00 hours and 19:00 to 23:00 hours.

In this analysis, minutewise gathered occupancy information was figured on hourly basis for 50 cells of varying channel numbers, e.g., each cell with 7channels/14 channels/28 channels/60 channels for 1/2/4/8 Radio Frequencies (RFs) of GSM framework. Channels with same number of RFs are some where lightly loaded and highly loaded in some different areas. Areas with 8 RFs (60 channels) and distinctively stacked at various activity areas are shown in figure 1. The outcomes shown are comparable for other RF counts also. The figure shows double heaps in the channel occupancy. The heap design appears near to parabolic in nature from 07:00 to 15:00 hours and 17:00 to 22:00 hours. An observation of instant occupancy together with last two observations might show peak time location along with the occupancy at peak time. This will help to regulate scanning frequency to optimize battery life of the SU. Based on the channel occupancy pattern, SU shall work in three scan states for power consumption optimization given below:

(i) State-I: This is part of the day during which primary traffic is too low, hence any SU request for channels can be serviced by CR-BS system without blockage. The State-I is depicted in figure 1 from 00:00 to 05:00 hours (from points A to B), and from 15:00 to 19:00 hours (from points D to E). During these periods the occupancy is very less and hence, the SU can save

its battery power as no scan (scan frequency =0) is required. The SUs may go to sleep mode keeping only its receiver active for receiving paging message from CR-BS system only. In this state, SU RF is active in paging Listen mode from CR-BS but scanner remains inactive. GPS remains inactive while SU is static but restarts when SU is in motion. Manual interrupt is active for switching to dedicated mode for SU.

(ii) State-II: This is part of the day during which primary traffic goes on increasing, hence any SU request for channels can be serviced only when vacant PU channels as are sufficient. As depicted in Figure 1, from 05:00 to 11:00 (from points B to C) and from 19:00 to 20:00 hours (from points E to F), the channel occupancy shows a positive slope at the beginning which requires SUs to periodically scan the environment and update to CR-BS at highest rate to support channel hopping to active SUs.

(iii) State-III: In this state, the scanning frequency becomes active after every few seconds as specified by CR-BSC. The State-III is depicted in figure 1 from 11:00 to 15:00 hours (from points C to D) and 20:00 to 23:00 hours (from points F to G).

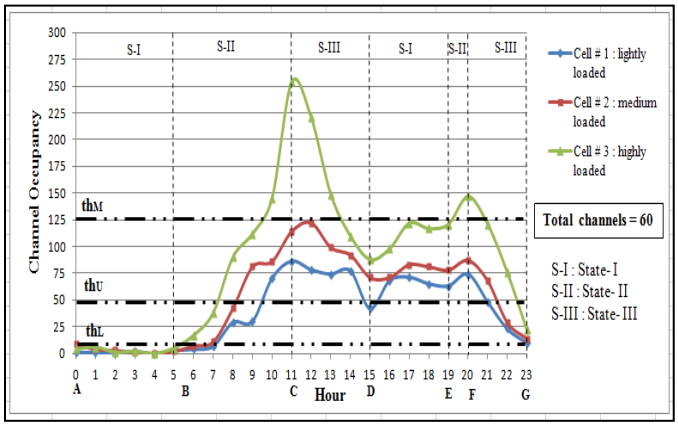


Figure 1. Daily Channel Occupancy Pattern for Three Cells With 60 Channels

(iii) State-III: In this state, the scanning frequency becomes active after every few seconds as specified by CR-BSC. The State-III is depicted in figure 1 from 11:00 to 15:00 hours (from points C to D) and 20:00 to 23:00 hours (from points F to G).

3.2.Holding Time Distribution

The data was collected at busy hour for 60 channels and minutewise occupancy was computed. An analysis of 60 minutes calls is practically taken for calculation purpose. The holding time distribution is shown in figure 3. It is observed that 58.5% calls are lying within 1 minute, 75% calls within 2 minutes, 82.9% calls within 3 minutes and 86.4% calls within 4 minutes 90.3% calls within 5 minutes. It is observed that holding time is nearly zero for a large number of calls which are due to transaction of Short Message Service (SMS) and Unstructured Supplementary Service Data (USSD) calls.

Study of figure 2 reveals that holding time of 86.4% calls are 4 minutes and the peak occupancy is 173. So, traffic carried is 173*(8/60) = 23 Erlangs, whereas total traffic maximum traffic carrying capacity is 60 Erlangs, i.e., actual loading is (23/60)*100 = 38.4%. Hence, there is enough vacant space to accommodate SU calls even in busy hour.

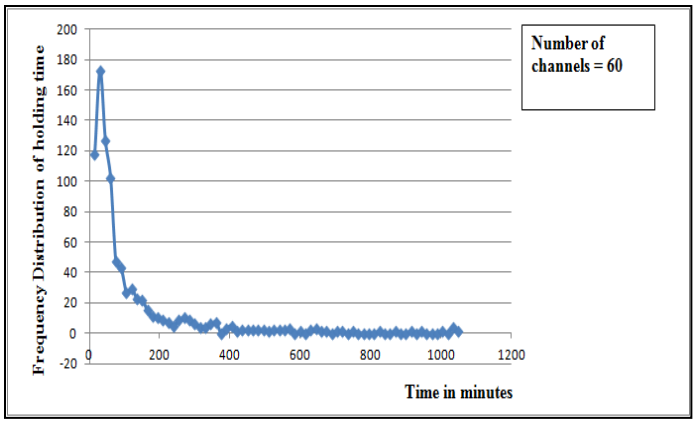


Figure 2. Holding Time Distribution Pattern in Busy Hour

4. CR Architecture and Model

The present document shall use the SU and Cognitive Radio-User Equipment (CR-UE) interchangeably. In previous study [12,13], it has been established that more than 20% of the completely licensed bandwidth is practically vacant in a saturated market environment. This is well above the requirement of 1/8th part of the band to get access to the whole of the bandwidth at a time and adequate to take additional Media Access Control(MAC) level overhead required for C. There is no immediate necessity of these channels by the licensed operators which can be safely deployed as common control channel for cognitive radio purpose. This paper also assumes that the communication shall take place in licensed band channels only so that the UEs can be used as SUs with little modification in access protocol.

The evolved network architecture as shown in figure 3 contains CR nodes which may be combined in exclusive clusters or Personal Area Networks (PANs). The clusters will have a Cluster Head (CH) or GW. The CR-GW shall internally coordinate with sensor nodes using standard 6LowPAN based protocols and CR-MAC protocol for GW to CR-BS communication where it will use the vacant channels of licensed band. The GW will be able to handle multi channel multi radio for communication with BS. It will also have a dedicated RF scanner for monitoring licensed channels and inform channel status change to BS. Unique code will be assigned to all CR-BS. According to capacity of CR-BS in radio side, CR-BS can control number of SUs. In the backbone side, it provides interface to Mobile Switching Center (MSC) through Media Gateway (MGW).

At the time of access for traffic data, GW-SU transmitter ensures that the access of PU channel do not interfere with PU functioning and CINR at BS. Network operators define various traffic accumulation data counters at BS. These parameters are polled at regular intervals for dynamic record of various RF parameters usually on hourly basis (T) and stored in Operations and Maintenance Center for Radio (OMC-R) [fusion centre in present case]. One hour time is too long for critical ON/OFF activity observation. Hence, channel status acquisition information need to be polled and updated at fusion centre at rate >> λ (call arrival rate) of the channels, or, in other words, polling time (τ) <<1/ λ, i.e. channel occupancy repetition time. In addition to periodic polling, both (i) the status of the channel change from free state to occupied state and, (ii) SU channel

allocation request should be interrupt driven and get instantly registered in Fusion Centre. Channel allocation shall be made by fusion centre from available list of channels for each channel based on: a) instantaneous channel status i.e. free state; b) amount of time passed since the channel was free; c) channel free time behavior prediction during the hour from data warehouse and d) User call holding time prediction during the hour from data warehouse. Eligible list of channels is thus prepared in order of merit and allocation is made by the switching centre depending upon the user service type request and QoS eligibility of the user.

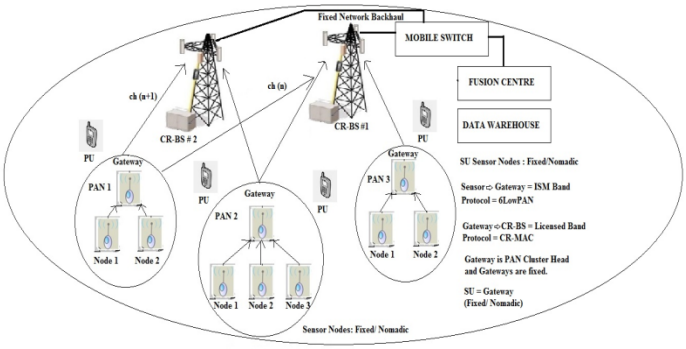


Figure 3. Cognitive Radio Cellular and Sensor Network

The proposed architecture also assumes that the communication shall take place in licensed band channels only so that the UE can be used as SUs with little modification in access protocol. PUs will communicate through BSC whereby SUs are located in between different PUs. The SU shall communicate through CR-BSC. The CR-BSC contains transceiver assembly based on Software Defined Radio (SDR) architecture, an exclusive wideband RF receiver and a Global Positioning System (GPS). Also it is expected that every SUs need to have in built GPS facility. The job of SU transceiver is to monitor channel occupancy information and the signal levels for all PUs. The CR-BS receiver collects the information such as channel occupancy from SUs. Thus the CR-UE shall take the role of static CR node for cooperative sensing with exception that it is mobile. On the other hand, the idea of swapping the dynamic database among CR nodes is missing. Instead, the database acquired by an SU during a specific duration is sent to a central node i.e., CR-BSC through CR-BS.

Secondary infrastructure might stay associated with CR-BSC for cognitive call handling function. CR-BSC shall hold the tables and decision logic for allocation of channels to SUs. In like manner, the spare channels and the assigned frequencies are distributed by CR framework in order to have minimum interference.

5. Call Arrival Rate Prediction

CR infrastructure and CR users are same as licensed user framework with the exception of that CR frameworks needs to follow the mandatory guideline that: (a) exclusive vacant channels of PU are to be utilized and returned when PU is active on this channel, (b) it won't cause any noise to PU framework. In this way, SU design might contain some extra logic than PU framework; else they are basically the same. Accordingly, it is basic to examine altogether the PU behavior for drawing conclusions about CR traffic handling effectively.

A review was completed for 42 days for 7 channels of various geographic and business significance of the city e.g., local location, market, play area and so forth, out of which 35 days (7 weeks) information was utilized for development of the model and one week (7 days) information was forecasted and verified with actual days information from 36th day to 42nd day to guarantee goodness of fit.

In this analysis, 35 days information of every cell was gathered and arranged on hourly basis and one specific hour was chosen. It was expected that hourly information might follow a weekly pattern and then a period of 7 days was taken after. SARIMA model was applied to forecast 7 days information for that specific hour in view of 35 days observed information of that hour. The prediction of traffic pattern for a week follows the following relationship:

Forecast Occupancy = $F_i = S_i * T_j$

where, i = 1 to 7 as one week is considered as a seasonal unit,

j = 1 to n,

n= number of days of observation = 35 in present analysis.

$T_i = \{A_0 - A_1 * D_j\}$,

S_i = Seasonal Coefficients;

A_0& A_1= Intercept Coefficients obtained from SARIMA modeling.

A lag of 7 was observed to be ideal fit for predicted data with actually obtained result. In this way SARIMA model of (7,7,7) was found to be best fit after application on 7 channels. Such perception for an estimated busy hour has been plotted in Figure 4.

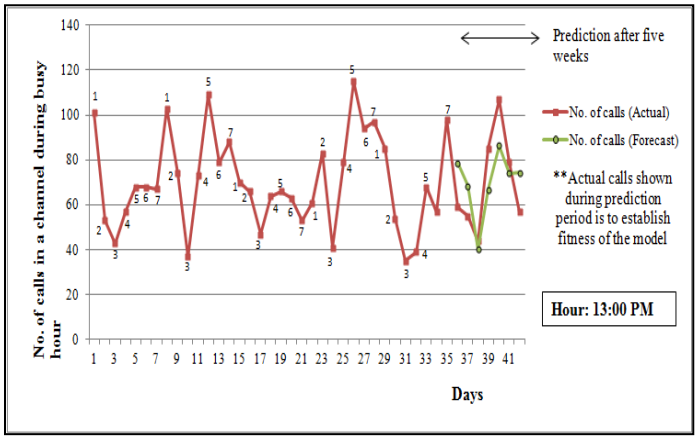


Figure 4.Forecast of Channel Availability during Busy Hour: 13:00 PM of Each Day of Five Weeks Observation Using SARIMA Approach

Investigation of the plots for week after week observation demonstrates reliable conduct from week to week as shown by digits 1 to 7 in the figure 4. This information is stored and updated in data warehouse of the CR-BSC for long term forecast.

The prediction of availability of vacant primary channels or in other words, occupancy of primary channels during the holding time of SU is essential, in particular, in and around peak occupancy. Despite the fact that, the parabolic model gives an idea of peak occupancy time and peak occupancy count from static perception however it can't be completely depended upon [14]. Thus,

prediction mechanism is essential about busy hour and at least for the duration of the holding time.

The probability that no call arrives ($P_i(0)$) during the holding time of SU (τ) was calculated using the relation:

$P_i(0) = \exp -[(\lambda/60)* \tau]$

where, λ = call arrival rate of PU,

τ = holding time of SU

It is found in figure 5, even if we assume that call arrival rate of PU = 8 with τ = 3 minutes, the SU call will mature with probability of 0.67.

6. Blocking Probability

The determination of QoS provided by a particular network configuration is required for an efficient design of communication networks [15]. Whenever a customer makes a telephone call, the routing equipment will decide to reroute the call to alternative equipment, to receive the call or to reject it entirely. Call rejection is occurred because of heavy traffic loads (congestion) on the system which results in either call being delayed or lost. The blocking probability is defined as the probability that the call attempts cannot be handled successfully due to unavailability of the radio resources (i.e. unavailability of free circuits or other switching equipments).

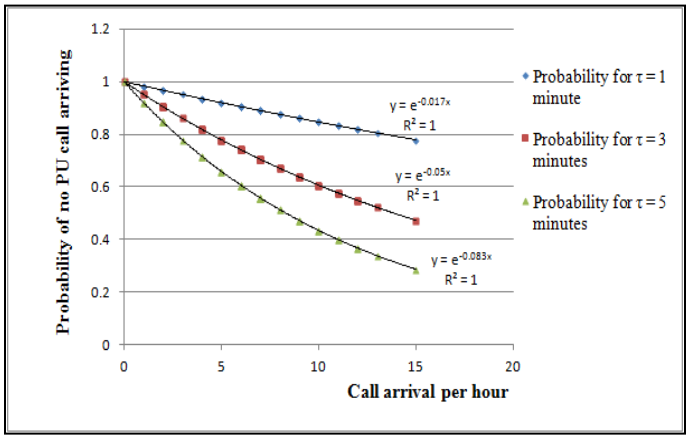


Figure 5. Graph of Probability of No Call Arriving Of PU, $P_i(0)$ vs Number of Vacant Channels with Different Holding Time Of SU (τ)

Efficient design of communication networks requires the ability to determine the quality of service provided by a particular network configuration. A common quality of service measure is the blocking probability, which is the probability that a new call will not be admitted to the network due to insufficient network resources [16].

Using the Poisson theorem, the probability that a call will be blocked can be calculated as:

$$P(c,a) = 1 - \sum_{k=0}^{c-1} \frac{a^k}{k!} e^{-a}$$

where, P(c,a) = the proportion of offered load that is blocked using the Poisson model given "c" servers and "a" offered load.

For a Lost Call system, the Grade of Service (GoS) for CR shall be measured by using modified Poisson's Model, as proposed in this work is given by the equation:

$$P(c,N) = 1 - \sum_{k=0}^{c-1} \frac{N^k}{k!} e^{-N}$$

where, k = 0 to (c-1) with c = total number of trunked channels, $N=N_p+N_s$, N_p = Count of PUs in the system, $N_s = \sigma N_p$ + offset = Count of SUs in the system, where, $0 < \sigma \leq 1$. A portion of the PU, σ (known as SU factor) can be considered for the calculation of the blocking probability of a secondary call combined with PUs traffic in the system. Also, 0<offset<1 such that N_s is an integer of higher value. These values of GoS help to determine whether the channel allocation to SU shall be successful or fail.

At the point when a cellular system is planned, the blocking probability P should be less than some pre-determined value. An observation for varying channels was made with c= 60 servers and is depicted in figure 6.

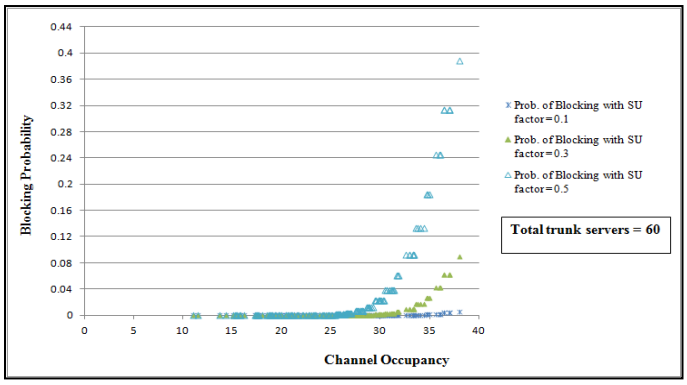


Figure 6.Call Blocking Probability with Trunk Servers (c) = 60 using Poisson's Model

7. Discussions on Results and Conclusion

The study establishes that there are scanty occupancy of PU channels about 16 hours of a day during which large quantum of SU can be allocated without problem. The performance of SU can be further improved in case, the traffic holding time is less than 1 minute. The SARIMA prediction model with a periodicity of 7 ideally fits the predicted data. It is also established that when the PU channel occupancy is 50%, the CR-BS shall provide voice channel to SU with blocking probability ≤ 0.02, where industry standard for blocking is 0.02.

8. Future Work

The SARIMA method is a better method for prediction of channel occupancy rate which takes into account the auto regression and moving average method along with the seasonal parameters. But still there is a scope of further improvement in the results of prediction by including three exponential smoothening

factors viz. (a) the level (or mean) that can be smoothed to give a local average value for the series of data, (b) the trend that can be smoothed, and (c) each seasonal sub-series (i.e., all the values of Monday, all the values of Tuesday, etc. for weekly data) that can be smoothed separately to give a seasonal estimate for each of the seasons. Some other methods like Holt-Winters (HW) additive technique may provide more predictive results.

Conflict of Interest

The authors whose names are listed immediately below certify that they have NO affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

References

[1] Niranjan Muchandi, Rajashri Khannai, "Cognitive Radio Spectrum Sensing: A Survey", IEEE International Conference on Electrical, Electronics and Optimization Techniques (ICEEOT), 3- 5 March 2016, ISBN: 978-1-4673-9939-5/16, Pg 3233-3237.

[2] Tabakovic, Zeljko, "A Survey of Cognitive RadioSystems", Post and Electronic Communications Agency, Jurisiceva 13, 2013.

[3] Navpreet Kaur, Inderdeep Kaur Aulakh, "A Survey of Cooperative Spectrum Sensing in Cognitive Radio Networks", International Journal on Recent and Innovation Trends in Computing and Communication, Vol. 3, issue 11, ISSN: 2321-8169, Pg. 6313-6316.

[4] Beibei Wang and K. J. Ray Liu, "Advances in Cognitive Radio Networks: A Survey", IEEE Journal of selected topics in Signal Processing, Vol. 5, No. 1, February 2011, Pg. 5-23.

[5] Xiukui Li and Seyed A. (Reza) Zekavat "Cognitive Radio Based Spectrum Sharing: Evaluating Channel Availability via Traffic Pattern Prediction", Journal of Communications and Networks, Vol. 11, No. 2, April 2009, Pg. 104-114.

[6] Berk Canberk, Ian F. Akyildiz, and Sema Oktug:, " Primary User Activity Modeling Using First-Difference Filter Clustering and Correlation in Cognitive Radio Networks", IEEE/ACM Transactions on Networking, Vol. 19, No. 1, February 2011, Pg. 170-183.

[7] Rahul Bajpai, Aditya Trivedi, " Joint Power and Spectrum Allocation for Cooperative Cognitive Radio Network with End User Mobility", International conference on Wireless and Optical Communications Networks (WOCN), 2014 Eleventh International Conference, 11-13 Sept. 2014, IEEE, Pg. 1-5.

[8] Xiukui Li and Seyed A. (Reza) Zekavat, "Traffic Pattern Prediction Based Spectrum Sharing for Cognitive Radios", Lecture notes in Electrical and Electronic Engineering, Published: November 1, 2009 under CC BY-NC-SA 3.0 license ISBN 978-953-307-021-6.

[9] X. Li S.A. (Reza) Zekavat, "Spectrum sharing across multiple service providers via cognitive radio nodes", IET Communication, Vol. 4, Iss. 5, 2010, Pg. 551–561,

[10] R. Kaniezhil and Dr. C. Chandrasekar, " Multiple Service providers sharing Spectrum using Cognitive Radio", International Journal of Scientific & Engineering Research, Volume 3, Issue 3, March-2012, Pg. 1-7.

[11] R. Kaniezhil, Dr. C. Chandrasekar, S. Nithya Rekha, "Channel Selection for Selective Sensing using Cognitive Radio Nodes", International Journal of Computer Applications, Vol.39, No.3, February 2012, Pg. 20-25.

[12] Neeta Nathani, G.C.Manna, S.B. Mule, "An Empirical Assessment of Quasi-Permanently Vacant Channels in Mobile Communication Bands for Cognitive Radio", ICACT Transactions on Advanced Communications Technology (TACT) Vol. 3, Issue 1, ISSN: 2288-0003, Vol. 3, Issue 1, January 2014, Pg. 389-394.

[13] Neeta Nathani, G C Manna, S B Mule, "Assessment of Quasi-Permanently Vacant Channels in Mobile Communication Bands for Cognitive Radio", ICACT 2013, January 2013, ISBN: 978-89-968650-0-1, Pg. 484-488.

[14] Neeta Nathani, G.C. Manna, S B Mule, "Dynamically Available Channel Model for Cognitive Radio in GSM Band", Proceedings of National Conference on Trends in Signal Processing & Communication (TSPC'14), 12th –14th April 2014, ISBN: 978-93-83842-40-7, Pg. 8-12.

[15] Neeta Nathani. G.C.Manna, "Quality of Service challenges for mobile Cognitive Radio", 2015 International Conference on Computer Science and Environmental Engineering, May 17-18, Beijing, China, ISBN: 978-1-60595-240-6, Pg. 967-998.

[16] Neeta Nathani, G.C.Manna, S.S.Dorle, "Network Architecture Model of Infrastructure based Mobile Cognitive Radio System in Licensed Band with Blocking Probability Assessment", IEEE Future Technologies Conference (FTC 2016), 6-7 December 2016, ISBN: 978-1-5090-4171-8/16, Pg. 231-236.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1616-1625 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Soft Handoff Evaluation and Efficient Access Network Selection in Next Generation Cellular Systems

Moses Ekpenyong[1*], Daniel Asuquo[1], Samuel Robinson[1], Imeh Umoren[2], Etebong Isong[2]

[1]*Department of Computer Science, University of Uyo, P.M.B. 1017, 520003, Uyo, Nigeria*

[2]*Department of Computer Science, Akwa Ibom State University, Ikot Akpaden, Nigeria*

ARTICLE INFO

Article history:
Received: 03 June, 2017
Accepted: 04 August, 2017
Online: 24 August, 2017

Keywords:
Cellular network
Soft handoff
Machine learning

ABSTRACT

The increased motivation (by service providers) to offer user-centric and seamless communication services – that satisfies users' quality of experience (QoE), has manifested a myriad of challenges in the field of wireless communication; and given the increased traffic capacity and sudden explosion of cellular devices, communication systems are constantly threatened by performance related issues – including soft handoff. Although intelligent techniques have evolved to provide solutions to these issues, they are yet to flourish in the area of soft handoff. This contribution therefore proposes a framework that integrates two components: (i) machine learning methodologies: self-organizing map (SOM) and pattern classification – for robust performance evaluation of available soft handoff data; (ii) multiple attribute decision making mechanisms (MADM): the Analytical Hierarchy Process (AHP) – which result feeds the Technique for Order of Preference by Similarity to the Ideal Solution (TOPSIS) – for efficient access network selection.

Implementation of component one of the design revealed that SOM enabled a precise visualization of handoff features that influenced the system performance; and the error levels of training, validation and test dataset, with number and percentage of correct and incorrect classifications, were obtained from our pattern classifier.

Implementation of component two of the design for four heterogeneous (access) networks indicated that although network two (N2) was selected as best access network by TOPSIS and network three (N3) by Synthetic Extent Analysis (SEA) – a method adopted in a related paper, for a particular application; both TOPSIS and SEA selected N1 as second best alternative access network and network four (N4) as third best alternative network, despite the issue of ranking abnormality in TOPSIS. Further, AHP and TOPSIS can effectively be applied as MADM algorithms in handoff decision framework for selecting the best available network for handoff.

1. Introduction

Handoff is an essential element of cellular communications, as efficient handoff algorithms are a cost-effective way of enhancing the capacity and quality of service (QoS) of cellular systems. Traditional handoff mechanisms are mostly based on received signal strength (RSS) from current access point (AP) or base station (BS). The method to decide handoff based on RSS is considered as the simplest method to handoff decision [1], but on the other hand, it may not have sufficient reliability because of RSS fluctuation [2]. Further, each network in a heterogeneous network has different RSS threshold; hence, RSS-based method results in inefficient handoff, unbalanced load, and frequent service interruption. The heterogeneous network concept was introduced to satisfy the demands of network traffic capacity and data rate [3], and consists of multiplatform networks with various radio access technologies. Conventionally, a mobile user may roam and accomplish handoff using single criteria such as RSS, but the co-existence of different networks with different RSS thresholds degrades the network performance. As such researchers exploit other parameters such as signal to interference and noise ratio (SINR), current location of mobile user, and network conditions for initiating handoffs.

*Corresponding Author: Moses Ekpenyong, Department of Computer Science, University of Uyo, P.M.B. 1017, 520003, Uyo, Nigeria
Email: mosesekpenyong@uniuyo.edu.ng

The integration of diverse but complementary cellular and wireless technologies in the next generation wireless networks requires the design of intelligent handoff decision algorithms to enable mobile users equipped with contemporary multi-interfaced mobile terminals to seamlessly switch network access and experience uninterrupted service continuity anywhere and anytime. Recently, various schemes have been implemented to enhance the quality of multimedia traffic [4], but such schemes are yet to offer good estimates of the traffic load and throughput.

The continuous developments and evolving generations of the cellular network technologies are influenced heavily by increased traffic capacity; high data consumption; and demand for better service quality – sufficient to satisfy end-users' quality of experience (QoE), but, state-of-the-art solutions are yet to parallel these challenges. Next generation cellular systems are expected to leverage the challenges of previous generations and provide ubiquitous connectivity, zero latency and high-speed connection. It is also expected (in the future) that devices should integrate heterogeneous radio access technologies (RATs) in order to improve the network performance as well as satisfy users' quality of experience.

Vertical handoff which occurs between base stations (BSs) with diverse wireless network interfaces, enables mobile subscribers to seamlessly make connections with the new BS before terminating with the previous BS. Network selection for subscribers with multiple connection interfaces however, creates a huge challenge for heterogeneous wireless environment. This raises the concern for innovative approaches to access network selection to ensure: seamless and always best connected (ABC) service to mobile subscribers, maximized resource utilization, and increased revenue to network operators. One promising solution is the use of multiple attribute decision making (MADM) methodology to handle the multiple attribute requirements of each access network for efficient handoff.

In this paper, we engage the prioritization of various handoff decision variables – to improve on the overall network performance. The current work is an extension of our FTC'16 paper [5]. First, we subject the available (handoff) input data to two machine learning procedures: (i) unsupervised analysis of inherent patterns; and (ii) supervised classification of the system performance. We then proceed further to demonstrate the use of two MADM techniques – analytic hierarchy process (AHP) and Technique for order of Preference by Similarity to the Ideal Solution (TOPSIS) – for the network selection problem. The AHP is used to determine the weights of decision criteria, which are later fed into TOPSIS and used to assist the network to rank the access networks. Our focus is on the handoff decision phase and aimed at satisfying mobile subscriber through the ABC concept. Further, the performance of TOPSIS in network ranking is compared with the synthetic extent analysis using decision criteria values obtained in [5].

This paper is structured as follows: A critical review of literature on related works is done in section 2; Section 3 presents the materials and methods required for accomplishing the study; Section 4 discusses the results obtained; and Section 5 offers conclusion to the study.

2. Related Works

Macro-diversity can change the transmission power required to service a mobile station with regards to its location. Although this contributes to improving the system performance, it can also cause unexpected call-dropping after handoff. For multimedia traffics which require high data rates, macro-diversity is necessary. In [6] an analytical soft handoff model considering macro-diversity effect was developed. Using analytic and simulation method, a performance evaluation of blocking probability, handoff failure probability and the region-transition failure probability was carried out. Their results were also compared with previous analytic models to illustrate the effect of macro-diversity on the system performance. The system was found to be useful for managing the determination of optimal resource allocation, cell configuration, and admission strategy. In a similar research, Chung and Cho [7], had proposed an analytical model of soft handoff when multimedia services are served in CDMA based intelligent transport system (ITS). They evaluated the performance of soft handoff scheme using analytical and simulation method, and considered key performance metrics such as blocking probability, handoff failure probability and carried traffic – obtained for the system management parameters considered in [6]. Ma, Cao, Liu and Trivedi [8] investigated the features of cellular geometry in CDMA systems with soft handoff, which distinguishes the controlled area of a cell from the cell coverage area. Important characteristics of the cellular configuration were used to propose a new design for efficient call admission control (CAC) in CDMA systems and the construction of a continuous-time Markov chain (CTMC) model for CAC in CDMA with soft handoff queue. Closed-form solutions were also obtained with algorithms that provide stability in loss probabilities computation and the determination of the optimal number of guard channels. To apply the loss formulas, their proposed modelling techniques were used to evaluate and compare the performance of conventional and proposed soft handoff schemes. Numerical results showed the effectiveness of the proposed Markov chain models as well as benefits of the new soft handoff scheme. A neural network (NN) approach was proposed in [9] to optimize the operation of cellular networks, where an analytical equation was derived to establish the effect of essential handoff contributory factors, with data obtained from base stations of an operational network carrier, and trained using a back-propagation NN algorithm. Their results showed that the network performance improved with large data size. A self organizing map was then used to visualize the existing system for the purpose of improving further, its performance.

Heterogeneous multi-radio networks are emerging with architectures that compromise hierarchical deployments of increasingly smaller cells [10]. These deployments may employ multi-radio access technologies to communicate with the network infrastructure, and calls for a dynamic access network selection. Various schemes have been implemented for access network selection during handoff in heterogeneous wireless networks, but most of them are usually optimization procedures and only useful for the reduction of handoff latency instead of improving QoS. One of the important aspects of seamless communication for ubiquitous computing is the dynamic selection of best access network. Ajuja, Singh and Khanna [11] considered bandwidth availability as the dynamic parameter for network selection in heterogeneous environments. A bootstrap approximation based algorithm is deployed in temporal and spatial domains to check its robustness. Numerical results showed that their algorithm reduced the estimation error, overhead, and improved reliability, compared to the existing algorithm.

To assure the required QoS for various applications run by the mobile user (MU) and avoid frequent handoff errors in heterogeneous networks, an AHP method for network selection was introduced in [12]. AHP performs decision selection on the basis of different handover parameters such as QoS, communication cost, availability and reputation of a network, etc, and combines the average of each criterion to decide an appropriate network for handoff. The AHP method has been employed by different researchers for efficient network selection in WLANs, but the method has not achieved the desired success in cellular networks [12]. A similar scheme was proposed to enhance the QoS of a network through the optimization of different handoff parameters such as data rate and handoff latency [1], and the decision for handoff was performed using a hybrid of fuzzy logic and analytic hierarchy approaches. A hierarchal scheme for discovering network selection rule based on media independent handoff (MIH) optimization has been proposed in [13]. In their experiment, a number of APs/BSs in a particular zone were connected to a zone media independent information system (MIIS) server attached to both the local and global MIIS servers. The energy consumption rate of a MU directly depends on the available networks scanned during a handoff process. Different schemes based on the energy efficient network selection for multimedia based applications have been proposed in [14, 15]. The schemes adopted the concept of adapt-or-handoff for balancing the multimedia traffic during a handoff process, thereby saving energy consumed due to the insignificant degradation in quality. Hence, energy consumption during the selection of networks is important and should be considered carefully. In order to optimize the energy consumption of multiple interfaces, Doppler, Ribeiro and Kneckt [16] have proposed a two-step scanning mechanism for obtaining band/channel information within a specific area. In their scheme, the MU scans highly reachable channels and provides each MU with the network density information as it enters a new area.

One major issue in handoff is how to optimize the decision time in selecting the access network with the best signal quality among the available one at the exact time needed by the mobile users. It is therefore necessary that mobile network operators provide good control and decision-making algorithms that optimize the soft handoff process – to avoid false initiation alarm, and select the best available network. Various MADM techniques for handoff decision optimization abound in the literature (c.f. [17-20]). However, no one technique is superior as there exist drawbacks in them such as ranking abnormality, subjective experts' judgment, uncertainties in human preference and penalizing behaviour towards poor attributes. In this paper, AHP and TOPSIS are combined for performance evaluation of network selection in heterogeneous wireless environment during handoff. Furthermore, the selection of these models is based on the following evaluation criteria [21]: internal consistency and logical soundness, transparency, ease of use, data requirements are consistent with the importance of the issue being considered, realistic time and manpower resource requirements for the analytical process.

The idea with AHP is to form a comparison matrix for each pair of criteria or alternatives such that the paired comparisons produce weighting scores that measure how much importance the criteria or alternatives have with each other. It begins by decomposing a complex decision problem into a hierarchical structure comprising the goal to be achieved at the top of the hierarchy, followed by criteria needed to achieve the goal and possibly the alternatives for each criterion at the bottom of the hierarchy. The major steps in AHP hierarchal structure are presented in [22-23]. The method tries to optimize decision making when decision makers are faced with conflicting qualitative and quantitative data. It optimizes decision making by enabling decision makers, in its final ranking, to choose the best solution among several alternatives across multiple criteria to attaining a specific goal. The ability of the method to capture both subjective and objective data in a quantitative manner for evaluation as well as the provision of a mechanism for checking the consistency ratio of the evaluation measurements makes it a veritable tool for decision making. The use of TOPSIS as a classical MADM algorithm is aimed at identifying an alternative that will have the shortest distance from the positive ideal solution and the farthest distance from the negative ideal solution [24]. The algorithm calculates positive and negative ideal solutions based on the attribute values available for each alternative.

In [25] a modified ''ELECTRE'' algorithm that provides complete ranking of the networks in application scenarios where the utility of some attributes is non-monotonic was presented. In [26], they compared the performance of "PROMETHEE" with "AHP" in terms of consistency, ranking abnormality, robustness, and accuracy. In [27], a comparison among different decision making techniques was made to check their pros and cons in terms of selection of network for sensor based applications. Results indicate that AHP appears more stable, less prone to risk and penalizing judgment behaviour. Also, more than 80-85% of the time, the results made by AHP, TOPSIS, Simple Additive Weighting (SAW) and Multiplicative Exponent Weighting (MEW) were the same. In [28], an access network selection algorithm that uses AHP and TOPSIS was proposed, but crucial QoS parameters such as RSS and available bandwidth resources were not considered in the selection process. However, it has been shown that the available bandwidth, received signal quality, battery life of the MT, security level, cost of service per byte, and distance between MT and BS are important parameters when considering vertical handoff decisions [29-31].

A QoS mechanism was proposed in [5] for improved handoff decisions in mobile communication networks. The proposed design adopted a hybrid channel perception tool to train observed key performance channel characteristics for system channel behaviour discovery. A HMM framework was adopted to ensure that the available channels were not completely consumed during handoffs. Estimates obtained from the training were then passed to a Fuzzy-based model to select the appropriate access network or base station. Their results showed that the proposed system could sustain handoffs at an average of 80%, compared to the empirical system, which average gave 50%. Further, two multiple attribute decision making (MADM) methods: fuzzy AHP and synthetic extent analysis (SEA) were used to assign weights to certain criteria influencing handoff decision, and the alternative access networks ranked based on weight priority scores in order to select the most optimal access network for a particular application.

3. Materials and Methods

The methodology adopted for the soft handoff evaluation integrates two machine learning tools. The self organizing map (SOM) – an unsupervised classifier is first used to cluster the input data in order to observe the inherent patterns exhibited by the existing system. The stipulated threshold for handoff success rate set by the Nigerian Communications Commission (NCC) – the communication regulatory body in Nigeria, was then used to

generate the target class. Both data classes (input and target) are later fed into a (supervised) feed-forward pattern classifier, to explain the performance of the classification algorithm. The system framework is shown in Figure 1.

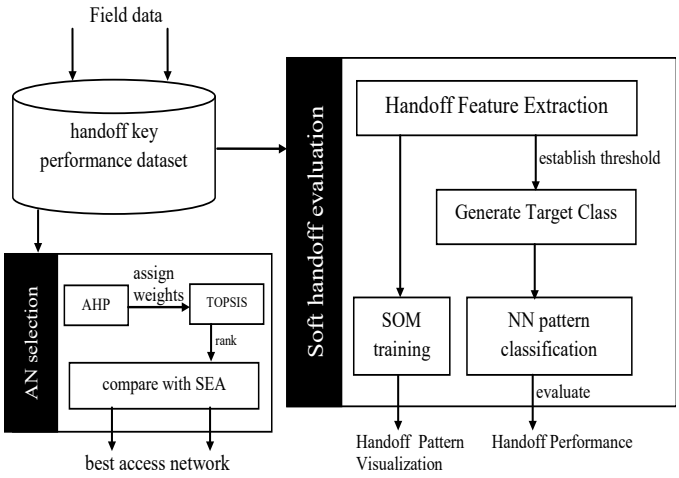


Figure 1. Proposed system framework

To ensure effective comparison between SEA and TOPSIS, the most promising values generated from the membership functions of the triangular fuzzy numbers in the fuzzy-AHP method in [5] were used to construct the AHP pair-wise comparison matrix for the considered criteria necessary for handoff decision. The AHP process in the end assigns weights to criteria which are used by the TOPSIS method to perform final ranking of the access networks. The alternative access network is the one with the highest value of relative closeness to ideal solution. We also compare the access network selection results of SEA technique – obtained in [5] with TOPSIS, to evaluate the efficiency of both selection techniques, in selecting the best access network for handoff.

3.1. Key Performance Dataset

Dataset for the soft handoff evaluation experiment consists of the input class and the target class. Sample data points for the input and generated target classes are presented in Table 1. The input data are real-time measurements acquired from 49 base station controllers (BSCs) of an operational cellular network carrier, at busy hours. These data were gathered over a period of one month, and are required for evaluating the existing system.

The input data extracted for this experiment include: drop call rate (DR); cell availability (CA); base station controller availability (BSCA); carried traffic (CT); block call rate (BR); up-link throughput (UL_TH); down-link throughput (DL_TH); and handoff success rate (HOSR). The target class is made up of three features namely, not satisfactory (C1: class 1); satisfactory (C2: class 2); and very satisfactory (C3: class 3). Data for the target class were obtained through the following classification rule: (if HOSR<90, C1=1, else, C1=0); (if 91≤HOSR≤95, C2=1, else, C2=0); (if 95≤HOSR≤100, C3=1, else, C3=0).

After generating the target class, 7 samples (about 14% of the data) were not satisfactory (fell into class 1), 42 samples (about 86% of the data) were satisfactory (fell into class 2), and none (0% of the data) was very satisfactory (fell into class 3). The analysis indicates that the handoff performance of the existing system was not optimal, as no data fell into the third class.

Table 1. Sample input data points and generated target classes

	Input class								Target class		
SN	DR	CA	BSCA	CT	BR	UL_TH	DL_TH	HOSR	C1	C2	C3
1	1.1200	99.8700	93.3800	402.0800	0.1200	11.7000	33.8100	88.6600	1	0	0
2	0.9000	99.8900	97.5900	569.1900	0.2100	11.3200	30.3200	91.2100	0	1	0
3	1.1800	99.7700	93.7800	507.7200	0.3300	10.7400	28.8800	92.3300	0	1	0
4	0.7400	99.7300	97.9400	645.3700	2.1900	7.1700	23.6100	93.6600	0	1	0
5	0.9000	99.6600	90.7800	615.9900	0.7900	9.7200	25.6600	91.7300	0	1	0
6	0.9700	99.5900	96.1000	889.1500	0.3500	8.4000	22.2700	93.5100	0	1	0
7	1.0300	95.3500	96.0200	1065.5200	0.7700	3.5600	11.8400	92.8100	0	1	0
8	0.6800	95.5000	94.0200	1030.3100	0.3800	8.0200	24.0500	93.6300	0	1	0
9	1.0700	98.0400	91.3200	647.5300	0.9500	8.5000	24.8200	93.3300	0	1	0
10	1.2300	98.0000	97.4600	423.2100	0.0600	10.8300	29.1200	88.9600	1	0	0
11	1.4300	97.9500	95.6100	455.8500	0.3700	14.0200	36.1900	92.0900	0	1	0
12	1.0800	99.8400	95.1700	395.0600	0.1200	8.5200	26.1800	92.7000	0	1	0
13	1.0300	99.9700	92.7100	250.8900	0.0300	10.8900	33.1200	90.3900	0	1	0
14	1.3200	99.7800	96.1300	151.9700	0.3800	12.2700	35.9600	88.6500	1	0	0
15	1.4600	94.6700	96.1900	990.6200	2.9700	9.0700	25.4800	88.5500	1	0	0
16	1.0500	98.5000	98.3300	685.7000	0.1700	10.7900	26.9600	91.4000	0	1	0
17	0.8500	98.4800	95.0300	346.4200	0.1500	11.5300	28.7300	93.4300	0	1	0
18	1.4900	97.3500	97.4300	239.4500	0.0200	11.6200	30.0000	92.3300	0	1	0
19	1.1200	97.3100	65.4500	106.7000	0.0300	6.5600	21.8100	91.5700	0	1	0
20	0.6600	97.3500	94.7300	556.2900	0.0500	9.6000	30.4300	95.2700	0	1	0

3.2. SOM Methodology

One major feature of artificial neural networks (ANNs) is its ability to adapt to an environment by learning to improve its performance at emulating behaviourial characteristics of an object in an unsupervised manner. The basic SOM is as a nonlinear, ordered, smooth mapping of high-dimensional data manifolds onto the elements of a regular, low-dimensional array, and the mapping is implemented as follows:

(i) define the set of input variables x_j as a real vector, $x_j = \{x_1, x_2, \dots, x_n\}^T \in \mathbb{R}^n$;
(ii) associate with each element in the SOM array a parametric real vector: $m_i = \{m_1, m_2, \dots, m_n\}^T \in \mathbb{R}^n$ – also called a model;
(iii) define a distance measure between x and m_i, denoted as $d(x, m_i)$.

In this paper, we adopt the incremental SOM algorithm with a rectangular map topology, as shown in Figure 2.

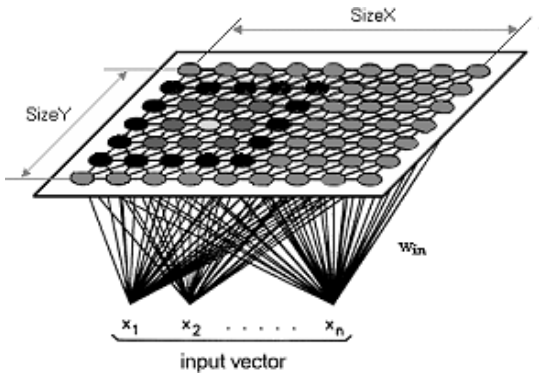


Figure 2. SOM topology; Source [32]

The image of an input vector x on the SOM array is then defined as the array element m_c that best matches with x, where the index $\arg min_i\,(d(x, m_i))$. The main task here is to define mi in such a way that the mapping is ordered and descriptive of the distribution of x. The process in which such mappings are formed is defined by the SOM algorithm. This process is then likely to produce asymptotically converged values for the models mi, the collection of which will approximate the distribution of the input samples $x(t)$, even in an ordered fashion. Vector x may be compared with all the m_i in any metric. In many practical applications, the smallest of the Euclidian distance $|x - m_i|$ can be made to define the best matching node (BMN), defined by the subscript:

$$c = \arg min_i ||x - m_i|| \equiv \big||x - m_c|\big| = |min_i|.$$

3.3. Pattern Recognition Methodology

Pattern recognition is the process of classifying input data into classes based on key features. Two classification methods to pattern classification are identified in the literature and include: supervised and unsupervised classification. This paper adopts the supervised method to pattern classification. The need for pattern recognition in handoff algorithms is essential for evaluating the robustness of communication systems. Pattern recognition application to the handoff process assumes that mobile users travel along same paths within the radio coverage area, and their movements follow the topological structure of the road infrastructure or the architectural structure of the environment. These patterns are then exploited to make handoff decisions and to define the locations where handoff executions are desirable through training of the network – to identify such locations. Pattern recognition can be implemented in a variety of ways and algorithms, and the setting up of a complete pattern recognition based handoff decision system should consider the following tasks:

(i) *Handoff location determination:* defining the location depends on the choice of criteria for what an optimal or good enough position a handoff should occur;
(ii) *Pattern formation:* generating pattern trace of the street, path, etc., with the collected metric samples. The components of the pattern trace are feature matrices associated with the surrounding BSs/APs, which are often referred to as class representatives.
(iii) *Data sample collection:* using data samples of a chosen metric around the chosen handoff location to train and validate the recognized patterns.

In this paper, a two-layer feed-forward network, with sigmoid transfer function in the hidden layer, and a soft-max transfer function in the output layer is proposed. The number of hidden and output neurons are set to 10 and 2, respectively, and the number of elements in the target vector (the number of categories or classes) is 3 (see Table 1). The input data were randomly divided such that 70% of the samples were for training, 15% for validation, and 15% for testing. The neural network architecture showing the pattern recognition components is presented in Figure 3.

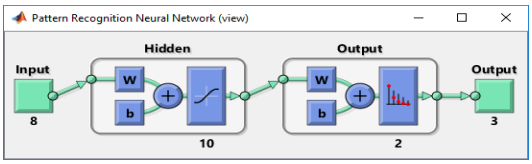


Figure 3. Pattern recognition network architecture

3.4. AHP Methodology

The application of the AHP method to a complex problem usually involves the following major steps [22-23]:

(i) Define the unstructured problem by describing the hierarchy with the goal at the top, followed by criteria at the lower levels and then alternatives at the bottom.
(ii) Construct a set of pair-wise comparison matrices among decision elements (criteria and alternatives) by establishing priorities among them based on a preference scale of 1 to 9 as shown in table 2.

Table 2. AHP preference scale

Scale	Definition
9	Extreme importance
7	Very strong importance
5	Strong importance
3	Moderate importance
1	Equal importance
2, 4, 6, 8	Intermediate values

(iii) Estimate the relative weights (w_k) of the decision elements using eigenvector method; where $w_k = \frac{GM_k}{\sum_{k=1}^{n} GM_k}$, where GM_k is the geometric mean of the kth row.
(iv) Aggregate these relative weights and synthesize them for the final measurements of the decision alternatives.

(v) Compute the consistency ratio (CR) to determine the acceptability or otherwise of the chosen criteria or alternative; where $CR = \frac{CI}{RI}$ (CI is the consistency index, and RI is the random index).

3.5. TOPSIS Methodology

The final weight measurements from AHP are used by TOPSIS for ranking of alternatives. In the TOPSIS method two artificial alternatives are hypothesized, i.e. the positive and negative alternatives. The basic rule behind these hypotheses is that the chosen alternative should have shortest distance from positive ideal solution and longest distance from negative ideal solution. TOPSIS makes the decision of alternative that is very near to the best result and very far from pessimistic alternative. The first step transforms attribute dimensions into non-dimensional attributes, which allows comparisons across criteria. The following steps describe the TOPSIS algorithm for network selection.

(i) Compute root of sum of squares to standardized the overall decision matrix of weights of criteria d_{ij} obtained from AHP method;

(ii) Normalize the decision matrix of criteria d_{ij} acquired from the four candidate networks as listed in Table 1 using (1).

$$z_{ij} = d_{ij} / \sum_{i=1}^{n} d_{ij} \quad (1)$$

(iii) Generate the weighted normalized matrix by multiplying the normalized decision criterion z_{ij} with its assigned weight w_k as in (2)

$$r_{ij} = z_{ij} w_k \quad (2)$$

(iv) Determine the positive ideal solution V^+ and the negative ideal solution V^- as in s (3) and (4)

$$V^+ = (R_1^+, R_2^+, \dots, R_n^+) \quad (3)$$

$$V^- = (R_1^-, R_2^-, \dots, R_n^-) \quad (4)$$

However, for beneficial criteria, $R_i^+ = \max(r_{ij})$ and $R_i^- = \min(r_{ij})$ while for non-beneficial criteria, $R_i^+ = \min(r_{ij})$ and $R_i^- = \max(r_{ij})$.

(v) Determine separation from ideal solution by calculating the similarity distance as expressed in (5) and (6),

$$S_j^+ = \sqrt{\sum_{j=1}^{n} \left(R_i^+ - r_{ij}\right)^2}, j = 1,2, \dots, n \quad (5)$$

$$S_j^- = \sqrt{\sum_{j=1}^{n} \left(r_{ij} - R_i^-\right)^2}, j = 1,2, \dots, n \quad (6)$$

(vi) Determine the relative closeness to the ideal solution by calculating C_j^* as in (7)

$$C_j^* = S_j^- / \left(S_j^+ + S_j^-\right), \; j = 1,2, \dots, n \quad (7)$$

(vii) Rank the access networks according to the descending order of C_j^*.

4. Results

4.1. SOM Visualization

We visualize the feature component planes of the SOM in Figure 4. The component planes show which features has a significant influence on the clustering results. Dark colors represent relatively small values while light colors represent relatively large values. Hence as expected, CA, BSCA, CT and HOSR have high range values, whereas DR, BR, UL_TH, and DL_TH have low range values, as predicted. Interestingly, (DR and BR); (UL_TH and DL_TH); and (BSA and HOSR); exhibit similar patterns, and thus signifying strong correlation between the component planes. In practice, we can say that these variables occur in sequence. For instance, block calls usually occur before drop calls, uplink transmission must occur before downlink transmission, and base station controller availability determines the handoff success rate. Cell availability and carried traffic show dissimilar patterns, indicating in practice that one component cannot determine the other – i.e., the carried traffic does not determine cell availability, or vice versa.

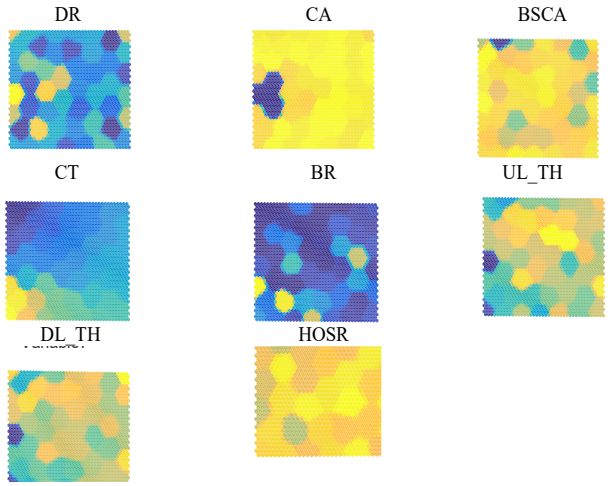


Figure 4. SOM component planes for selected features

4.2. Pattern Classification

Figure 5 shows the performance plots of the neural network for training, validation and test data. We observed that although the error reduced rapidly after the third epoch of training, the validation and test errors were still high even up till the ninth epoch. But the degradation did not cause an over-fit in the training data, and the best validation performance was 0.0559.

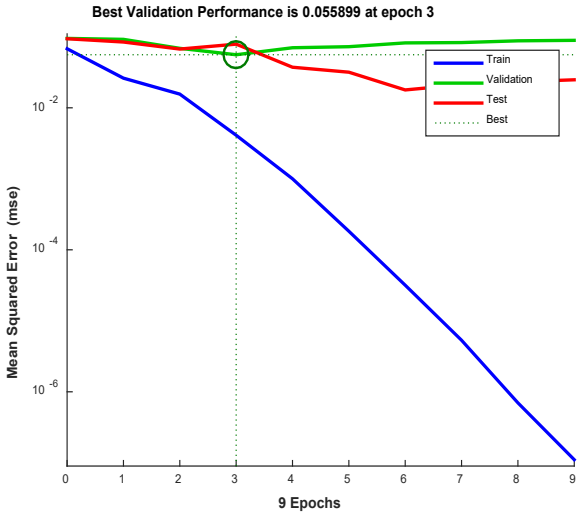


Figure 5. Neural network performance plot for handoff feature data

In Figure 6, the confusion matrix plots for training, validation and test performances of our classifier are shown. We observed that the data were well trained, but some validation and test errors were noticed, as evidenced in Figure 5. The first three diagonal cells in Figure 6 show the number and percentage of correctly classified observations. We concentrate on the overall confusion matrix that takes into account the overall performance of the training, validation and test cases. Hence, 5 BSCs were correctly classified as not satisfactory, corresponding to 10.2% of the overall BSCs within the study area. Similarly, 42 BSCs were correctly classified as satisfactory, corresponding to 85.7% of all BSCs under study. None of the satisfactory cases were incorrectly classified as not satisfactory, 2 of the not satisfactory cases were incorrectly classified as satisfactory – representing 4.1% of all BSCs. Now, out of the 7 not satisfactory predictions, 100% were correct. Also, out of 42 satisfactory predictions, 95.5% were correct, while 4.5% were wrong. Out of 7 not satisfactory cases, 71.4% were correctly predicted as not satisfactory, and 28.6% were predicted as satisfactory. Out of 42 satisfactory cases, 100% were correctly classified as satisfactory and none as not satisfactory. The overall prediction was 95.9%, with 4.1% wrong classifications.

The Receiver Operating Characteristic curve (or ROC curve) shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity), and is presented in Figure 7. It plots the true positive rate against the false positive rate for the different possible cut-points of the classifier. As can be seen, the distortion in the test data plot is not unconnected with the high total misclassification of not satisfactory cases as satisfactory (see Figure 7). Even with this distortion which came as a result of validation and test errors (see Figure 6), the overall ROC seems a more accurate test, as evidence in the closeness of the curves to the left hand border are noticed, except for the third class (C3) which had no recorded data.

Figure 6. Confusion matrix plots for evaluating our pattern classifier

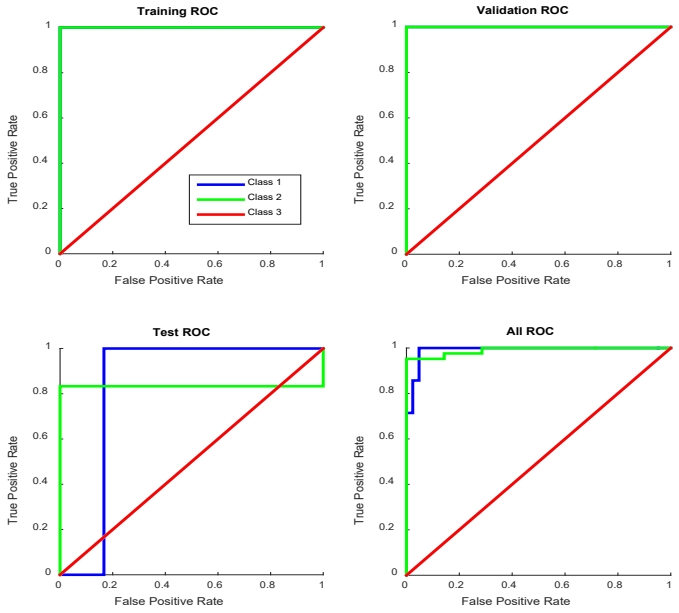


Figure 7. ROC curves showing sensitivity and specificity tradeoffs

4.3. AHP Analysis

To evaluate the performance of synthetic extent analysis and TOPSIS algorithms in network selection, an application scenario of four heterogeneous networks N1, N2, N3, N4 is considered which typically denotes 4G-LTE, 3G-UMTS1, 3G-UMTS2, and WiFi. In such a scenario a MS is assumed to be connected to N1 network and is traversing through an area overlapped by three more networks N2, N3, and N4. So, four networks are available simultaneously to the MS. The MS has to select the best network from the available networks for handoff. For selection process, AHP and TOPSIS algorithms are applied using the most promising values in the membership function of the triangular fuzzy numbers of the Fuzzy AHP method in [5], as AHP values for assignment of weights to criteria. The essence is to comparatively evaluate the performance of synthetic extent analysis with TOPSIS in terms of best network selection. Decision is based on assessing various criteria from each network: Quality of Service (QoS), Available Bandwidth (AB), Security, Cost, Battery Life, Distance between Mobile Node and Base station (DMNBS. The overall decision matrix for the criteria values for network selection are shown in Table 3.

Table 3. Decision matrix showing criteria values

Criteria	N1	N2	N3	N4
QoS	0.29	0.12	0.41	0.18
AB	0.45	0.18	0.27	0.09
Security	0.40	0.13	0.20	0.27
Cost	0.33	0.17	0.39	0.11
B. Life	0.12	0.41	0.18	0.29
DMNBS	0.33	0.13	0.20	0.33

Assignment of Criteria Weights by AHP

Weights are measures of relative importance of criteria. Most of the existing network selection algorithms have employed AHP methodology for assigning weights to the criterion. In AHP, weights are assigned according to the knowledge and perception

of the decision maker. It performs measurement through pair-wise comparisons and relies on the judgment of experts to derive priority scales. The AHP procedure for weight determination consists of the following steps:

(i) Construct pair-wise comparison matrix for the criterion at each level. For each pair, within each criterion award a score, on a scale between 1 and 9 to the better option, and a reciprocal of this value to the other option in the pair. The pair wise comparison matrix of the criteria for handoff calls is shown in Table 4.

Table 4. Citeria-by- criteria pair-wise comparison matrix

Scale	5	7	4	2	3	4
Criteria	QoS	AB	Security	Cost	B.Life	DMNBS
QoS	1.00	0.71	1.25	2.50	1.67	1.25
AB	1.40	1.00	1.75	3.50	2.33	1.75
Security	0.80	0.57	1.00	2.00	1.33	1.00
Cost	0.40	0.29	0.50	1.00	0.67	0.50
B.Life	0.60	0.43	0.75	1.50	1.00	0.75
DMNBS	0.80	0.57	1.00	2.00	1.33	1.00
Total	5.00	3.57	6.25	12.5	8.33	6.25

(ii) Determine the geometric mean of each row for each matrix and normalize the results to obtain the weights for each criterion. This results in table 5.

Table 5. Normalized criteria-by-criteria comparison matrix and weights

	QoS	AB	Security	Cost	B.Life	DMNBS	Weights (w_k)
QoS	0.20	0.20	0.20	0.20	0.20	0.20	0.20
AB	0.28	0.28	0.28	0.28	0.28	0.28	0.28
Security	0.16	0.16	0.16	0.16	0.16	0.16	0.16
Cost	0.08	0.08	0.08	0.08	0.08	0.08	0.08
B.Life	0.12	0.12	0.12	0.12	0.12	0.12	0.12
DMNBS	0.16	0.16	0.16	0.16	0.16	0.16	0.16

(iii) Check the consistency of the pair-wise comparison in the normalized matrix. The original comparison matrix is considered not consistent if the columns in its normalized matrix are not identical. Ideally, if CR is less than or equal to 0.1, the pair-wise comparison is considered acceptable; otherwise the subjective judgment is revised. TOPSIS algorithm requires weights to be assigned to each criterion for further computation of positive and negative ideal alternatives, where the positive ideal alternative is considered to be one with the best value for all attributes.

4.4. TOPSIS Analysis

From the steps in TOPSIS method (see section 3.4), the following results were obtained:

Step 1: Root of sum of squares computation (see Table 6)

Table 6: Root of Sum of Squares for each Criteria

Criteria	N1	N2	N3	N4	RSS
QoS	0.29	0.12	0.41	0.18	0.46
AB	0.45	0.18	0.27	0.09	0.56
Security	0.40	0.13	0.20	0.27	0.54
Cost	0.33	0.17	0.39	0.11	0.55
B. Life	0.12	0.41	0.18	0.29	0.55
DMNBS	0.33	0.13	0.20	0.33	0.52

Step 2: Decision matrix normalization (see Table 7)

Table 7. Normalized decision matrix of criteria

Criteria	N1	N2	N3	N4
QoS	0.63	0.26	0.89	0.39
AB	0.80	0.32	0.48	0.16
Security	0.74	0.24	0.37	0.50
Cost	0.60	0.31	0.71	0.20
B. Life	0.22	0.75	0.33	0.53
DMNBS	0.63	0.25	0.38	0.63

Step 3: Weighted normalized decision evaluation (see Table 8)

Table 8. Weighted normalized decision matrix

Criteria	N1	N2	N3	N4
QoS	0.13	0.52	0.18	0.08
AB	0.22	0.09	0.13	0.05
Security	0.12	0.04	0.06	0.08
Cost	0.05	0.03	0.06	0.02
B. Life	0.03	0.09	0.04	0.06
DMNBS	0.10	0.04	0.06	0.10

Steps 4 and 5: Determination of/and separation from positive and negative ideal solutions (see Tables 9 and 10)

Table 9. Separation from positive ideal solution

Criteria	N1	N2	N3	N4
QoS	0.15	0.0	1.16	0.19
AB	0.0	0.02	0.0081	0.03
Security	0.00	0.0064	0.0036	0.0016
Cost	0.0001	0.0009	0.0	0.0016
B. Life	0.0036	0.0	0.0025	0.0009
DMNBS	0.0	0.0036	0.0016	0.0
	$\sqrt{(0.1537)}$	$\sqrt{(0.0309}$	$\sqrt{(1.1758)}$	$\sqrt{(0.2241)}$
S_j^+	0.392	0.176	1.084	0.473

Table 10. Separation from negative ideal solution

Criteria	N1	N2	N3	N4
QoS	0.0025	0.1936	0.01	0.0
AB	0.0289	0.0016	0.0064	0.0
Security	0.0064	0.0	0.0004	0.0016
Cost	0.0009	0.0001	0.0016	0.0
B. Life	0.0	0.0036	0.0001	0.0009
DMNBS	0.0036	0.0	0.0004	0.0036
	$\sqrt{(0.0423)}$	$\sqrt{(0.1989}$	$\sqrt{(0.0189)}$	$\sqrt{(0.0061)}$
S_j^-	0.2057	0.4459	0.1375	0.078

Step 6: Computation of relative closeness (see Table 11)

Table 11. Relative closeness to ideal solution

Criteria	N1	N2	N3	N4
S_j^+	0.392	0.176	1.084	0.473
S_j^-	0.2057	0.4459	0.1375	0.078
$S_j^+ + S_j^-$	0.5977	0.6219	1.2215	0.551
$S_j^-/(S_j^+ + S_j^-)$	0.344	0.717	0.113	0.142

Step 7: Access network ranking (see Table 12)

Table 12: TOPSIS Ranking of the access networks

S/N	Network	TOPSIS Rank
1	N1	2
2	N2	1
3	N3	4
4	N4	3

4.5. Performance Comparison

From [5], the SEA ranking placed the four networks in the following order as shown in table 13.

Table 13. Access Network ranking by synthetic extent analysis

S/N	Final weight	Access Network	Ranking
1	0.33	N1	2
2	0.11	N2	4
3	0.34	N3	1
4	0.21	N4	3

The results indicate that for the same set of network decision criteria values, the selected network significantly varies for some reasons. The ranking order of Synthetic Extent Analysis is different from that of TOPSIS for handoff decision in heterogeneous networks. Both methods have similar ranking results for selecting N1 access network and N4 access network as second and third best alternative for roaming by mobile subscribers. Whereas TOPSIS considers N3 as negative ideal solution since if selected presents the user with minimum benefit attributes and maximum cost attributes, synthetic extent analysis considers N2 as least network to switch connection to.

5. Conclusion

Diverse technologies deployed by individual operators in heterogeneous wireless environments require innovative network selection methodologies for provision of always best connected (ABC) services to mobile subscribers. Handling multiple attributes or parameters with different relative importance from each network during handoff decision phase is a huge challenge in such network environment. This paper has exploited two machine learning techniques to visualize the performance of handoff in an existing system; and report on the error levels experienced by the network. It also proposed the use of two multiple attribute decision making (MADM) methods – Analytic Hierarchy Process (AHP) and Technique for Order of Preference by Similarity to the Ideal Solution (TOPSIS) for access network selection and compares the performance of TOPSIS and Synthetic Extent Analysis (SEA) [5] – in ranking the alternatives after assignment of weights to network criteria by AHP.

References

[1] M. Zekri, B. Jouaber, and D. Zeghlache. Context aware vertical handover decision making in heterogeneous wireless networks. In Proceedings of the 35th IEEE Conference on Local Computer Networks (LCN '10): 764–768, Denver, Colo, USA.

[2] X. Yan, Y. A. Sekercioglu, and S. Narayanan. A survey of vertical handover decision algorithms in Fourth Generation heterogeneous wireless networks. Computer Networks, 54(11): 1848–1863.

[3] G. Mahardhika, M. Ismail and R. Nordin (2014). Vertical Handover Decision Algorithm Using Multicriteria Metrics in Heterogeneous Wireless Network. Journal of Computer Networks and Communications 2015: 1-8, 2015.

[4] J. Shi, J. Wu, A. Paul, L. Jiao, and M. Gong, Change detection in synthetic aperture radar images based on fuzzy active contour models and genetic algorithms, Mathematical Problems in Engineering, 2014.

[5] M. Ekpenyong, D. Asuquo, E. Isong, I. Umoren, and R. Akpan. Optimizing Soft Handoff Latency and Access Network Selection for Next Generation Mobile Communication Networks, In Proceedings of IEEE Future Technologies Conference (FTC 2016), San Francisco, 213-223, 2016.

[6] Y. Chung and S. Park. Performance evaluation of soft handoff considering macrodiversity effect in multimedia CDMA systems. Telecommunication Systems, 53(2): 179-194, 2013.

[7] Y. Chung and D-H. Cho. Performance evaluation of soft handoff for multimedia services in intelligent transportation systems based on CDMA. IEEE Transactions on Intelligent Transport System, 4(4): 189-197, 2003.

[8] X. Ma, Y. Cao, Y. Liu and K. S. Trivedi. Modeling and Performance Analysis for Soft Handoff Schemes in CDMA Cellular Systems. IEEE Transactions on Vehicular Technology, 55(2): 670-680, 2006.

[9] M. Ekpenyong, J. Isabona and E. Isong. Handoffs Decision Optimization of Mobile Celular Networks. In Proceedings of International Conference on Computational Science and Computational Intelligence, Las Vergas: 698-703, 2015.

[10] S. Andreev, M. Garasimenko, O. Galinina, Y. Koucheryavy, N. Himayat, S-P. Yeh and S. Talwar (2014). Intelligent Access Network Selection in Converged Multi-radio Heterogeneous Networks. IEEE Wireless Communications, 21(6): 86-96.

[11] K. Ahuja, B. Singh and R. Khanna (2016). Network Selection Based on Available Link Bandwidth in Multi-Access Networks. Digital Communications and Networks, 2(1): 15-23.

[12] H. Pervaiz and J. Bigham, "Game theoretical formulation of network selection in competing wireless networks: an analytic hierarchy process model," in Proceedings of the 3rd International Conference on Next Generation Mobile Applications, Services and Technologies (NGMAST '09), 292–297, Cardiff, UK, 2009.

[13] F. Buiati, L. J. G. Villalba, D. Corujo, J. Soares, S. Sargento, and R. L. Aguiar, "Hierarchical neighbor discovery scheme for handover optimization," IEEE Communications Letters, 14(11), 1020–1022, 2010.

[14] R. Trestian, O. Ormond, and G.-M. Muntean. Energy-quality-cost tradeoff in a multimedia-based heterogeneous wireless network environment, IEEE Transactions on Broadcasting, 59(2), 340–357, 2013.

[15] A. Paul, Dynamic power management for ubiquitous network devices. Advanced Science Letters, 19(7), 2046–2049, 2013.

[16] K. Doppler, C. B. Ribeiro and J. Kneckt, On efficient discovery of next generation local area networks, In Proceedings of IEEE Wireless Communications and Networking Conference, 269–274, Cancun, Mexico, 2011.

[17] P. Tran and N. Boukhatem. Comparison of MADM decision algorithms for interface selection in heterogeneous wireless networks. In IEEE 16th International Conference on Software, Telecommunications and Computer Networks, 119–124, 2008.

[18] A. Sgora, D. D. Vergados, P. Chatzimisios. An Access Network Selection Algorithm for Heterogeneous Wireless Environments. In IEEE Symposium on Computers and Communications (ISCC), Riccione (Italy), 890–892, 2010.

[19] M. Lahby, L. Cherkaoui, and A. Adib. Network selection algorithm based on Diff-AHP and TOPSIS in heterogenous wireless networks. In IEEE proceedings of International conference on multimedia computing and systems, 485–490, 2012.

[20] K.S.S., Anupama, S. S. Gowri, and B. P. Rao, Network selection in heterogeneous wireless environment using decision making algorithms-TOPSIS AND PROMETHEE, Journal of Theoretical and Applied Information Technology, 75(2), 2015.

[21] J. Dodgson, M. Spackman, A. Pearman and L. Phillips. *Multi-criteria analysis: a manual*, Department for Communities and Local Government: London, 2009.

[22] T. L. Saaty. Decision making with the analytic hierarchy process, Int. J. Services Sciences, 1(1), 83-98, 2008.

[23] D. E. Asuquo and U. A. Umoh. Analytic hierarchy process for QoS evaluation of mobile data networks, International Journal of Computer Networks & Communications, 7(6), 125-137, 2015

[24] M. Dagdeviren, S. Yavuz, and N. Kılınç. Weapon selection using the AHP and TOPSIS methods under fuzzy environment. Expert Systems with Applications, 36, 8143–8151, 2009.

[25] F. Bari and V. Leung, Application of ELECTRE to Network Selection in a Heterogeneous Wireless Network Environment, *Wireless Communications and Networking Conference,* Hong Kong, 3810-3815, 2007.

[26] K. S. S. Anupama, S. S. Gowri, B. P. Rao, and T. S. Murali. A PROMETHEE Approach for Network Selection in Heterogeneous Wireless Environment, *International Conference on Advances in Computing, Communications and Informatics*, Greater Noida (India), 2560-2564, 2014

[27] H. A.Vine. Comparison between MADM algorithms for Vertical Handoff Decision, Technical Journal, University of Technology and Engineering, Taxila, 1-9, 2010.

[28] A. Sgora1, P. Chatzimisios and D. D. Vergados, Access Network Selection in a Heterogeneous Environment Using the AHP and Fuzzy TOPSIS Methods, Mobile Lightweight Wireless Systems, In Proceedings of the 2nd International ICST Conference (MOBILIGHT'10), May 10–12, 2010,

Barcelona, Spain, LNICST 45, pp. 88–98. Springer, Berlin, Germany 88-98, 2010.

[29] I. Alkhayat and A. Kumar. End-to-End Mobility Solution for Vertical Handoff between Heterogeneous Wireless Networks, IEEE, 2009.

[30] Chandralekha and P. K. Behera. Optimization of Vertical Handoff Performance Parameters in Heterogeneous Wireless Networks", International Journal of Modern Engineering Research (IJMER IJMER, 1(2), 597-601, 2011.

[31] R. A. Saeed, WiFi/WiMAX Seamless Convergence with Adaptive Vertical Handover for Continuity of Internet Access. Advances in Internet of Things, 1: 32-37, 2011.

[32] T. Kohonen. The Self-Organizing Map. In Proceedings of the IEEE, 78: 1464–1480, 1990.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1626-1633 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Proposal of an Embedded Methodology that uses Organizational Diagnosis and Reengineering: Case of bamboo panel company

Eva Selene Hernández Gress, Antonio Oswaldo Ortega Reyes*, Jaime Garnica González, José Ramón Corona Armenta

Autonomous University of Hidalgo, Engineering Academic Area, Institute of Basic Sciences and Engineering, 42181, Mexico

ARTICLE INFO

Article history:
Received: 01 June, 2017
Accepted: 28 July, 2017
Online: 24 August, 2017

Keywords:
Systemic Diagnosis
Reengineering
Embedded Methodologies

ABSTRACT

This work is an extension of the Proceedings of the International Conference on Industrial Engineering, Management Science and Applications, which presented some of the phases of Reengineering applied to Bamboo Panel Company; the results were Strategic planning, Systemic Diagnosis and Performance Indicators through the Balanced Scorecard. Now, the main purpose of this article is to present a methodology that embedding Organizational Diagnosis and Reengineering, which emphasizes the incorporation of culture, context, management style, and knowledge as well as inner and outer actors. The results of the proposed methodology applied to the case study are included, up to the moment of the writing of this article. Future work consists on the development of strategies for Innovation as a strategy planned in the Balanced Scorecard and derived from the embedded methodology.

1. Introduction

This paper shows a subsequent development of the work "Systemic Diagnosis and Strategy-Based Performance Indicators. Bamboo Panel Company Case" exposed at the 6th International Conference on Industrial Technology and Management [1]. The paper came from the project called "Product Development of Modular Components with Prefabricated Bamboo, Phase II". It was part of the Program for Research Stimulation, Technological Development and Innovation of the NATIONAL COUNCIL OF SCIENCE AND TECHNOLOGY (CONACYT, in Mexico). This project was developed by the Autonomous University of Hidalgo in Mexico and a private Enterprise.

The project was carried out from March 2016 to April 2017. It included physical prototypes, procedure manuals, mapping of facilities distribution, simulation of the production systems as well as performance indicators systems, among others.

The paper focuses on the section of methodology, especially, the one that concerns Strategic Planning: Systemic Diagnosis and Performance Indicators following the basis of the Balanced Scorecard, since them conflux to the main perspectives of this stage of development: embedding organizational diagnosis and reengineering processes. Other deliverables include prototypes and development of technology, but they shall not be exposed since that is confidential information. For the same reasons, the name of the company is not disclosed. The company, in which the case study was carried out, designs and constructs structures using bamboo as a raw material. This company creates and executes innovative projects, products and services, from the perspective of sustainability, combining artisan techniques with technologies that allow low environmental impact seeking also, a high social impact. A highlight is that the tool of reengineering was followed for the development of the bamboo prefabricated panel processes.

An organizational diagnosis was performed under the systemic approach, in order to identify the problems of the organization and for the creation of performance indicators. For this, it was necessary to create the so called embedded methodology combining reengineering and organizational diagnosis tools considering the enterprise as a system.

There are many diagnostic tools, however, the methodological design for the diagnosis was the one proposed by Ortega [2], which embeds a systemic perspective and incorporates ad hoc tools according to the situation of the organization, in this case, technology, organization, production, environmental conditions and regulatory frameworks. Additionally, the diagnosis comes to be a component of the Strategic Planning that serves to make decisions.

The first step consisted on stablishing the scope of the diagnosis. For this purpose, the different areas of the organization were analyzed following the methodology proposed by Ortega [2] shown in Figure 1.

*Corresponding Author: Antonio Oswaldo Ortega Reyes, Autonomous University of Hidalgo, Engineering Academic Area, Institute of Basic Sciences and Engineering, 42181, Mexico | Email: oswwaldoo@yahoo.com.mx

Figure 1. Dimensions and useful techniques. Source: Author's elaboration

The second step was to define the dimensions and useful techniques of the intervention which were also defined following this author (Figure 2).

Figure 2. Organizational diagnosis.Source: Author's elaboration

The resultant was a diagnostic map divided into External and Internal, aspects that were also included in the methodological proposal (Figure 3).

Then the tool of the Balanced Scorecard (BSC) was used to integrate the areas of the research process in a Strategic Map [1]. The BSC is a model that becomes a very useful tool for strategic management. It is based on the definition of strategic objectives, indicators and strategic initiatives, establishing relationships by a strategic map in four basic perspectives: financial, clients, internal processes and learning and growth, then the strategy should be translated directly into objectives; where it will be measured through indicators and aligned with initiatives. Success at implementing the BSC depends on the participation of people from different levels and areas of the organization. Kaplan and Norton [3] designed the Balanced Scorecard as an instrument to measure results, based on the establishment of financial and non-financial indicators derived from the company's vision, mission and strategy, making it a tool to manage the strategy.

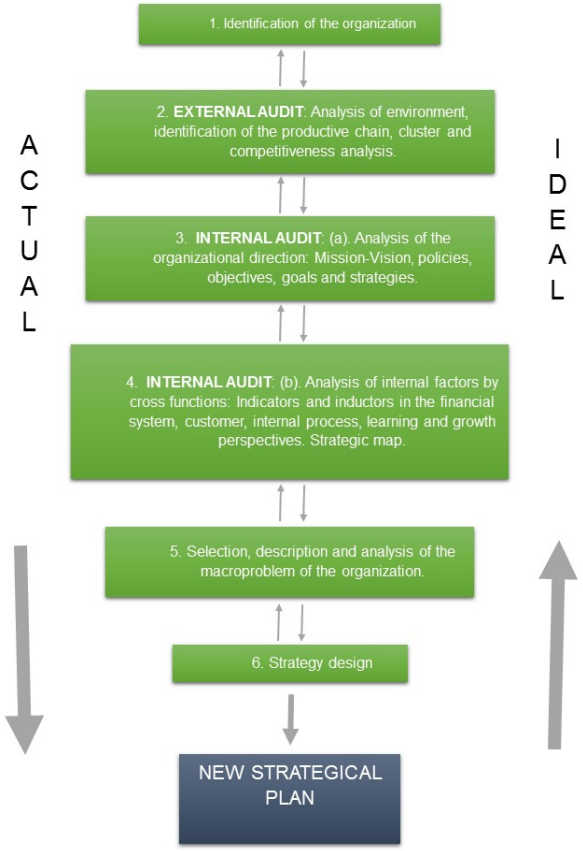


Figure 3. Diagnostic map. Source: Author's elaboration

Some of the benefits of implementing the BSC in organizations are: aligning employees to the company's vision, improving communication to all the staff and compliance, redefining the strategy based on results presented, a translation of the vision and the strategy into actions, which are oriented towards the creation of value; it is integrated with the information of all the areas of the organization and also the capacity of analysis and the decision making is improved. A strategic map as a resultant outcome [4], presents in a simple and coherent way the description of an organization's strategy, in order to establish the objectives and indicators in the financial, customer, internal processes and learning and growth perspectives.

The financial perspective describes the tangible results of the strategy in traditional financial terms, indicators such as the profitability of the investment, shareholder value, revenue growth, unit costs, among others; thus, measuring the creation of value for the organization. The customer perspective, reflects the positioning of the organization in the market, identifying the customer segments and, defines the value proposition for the target customers. Rodiles and Fuentes [5] mention the indicators which are generally considered: customer satisfaction and retention, as well as the acquisition of new customers, customer profitability and market participation where the organization participates. The

internal process perspective identifies internal processes that will have a greater impact on customer satisfaction. Some indicators of this perspective are: productivity, quality and innovation of products and services. The perspective of learning and growth is based on the fact that the formation and growth of an organization come mainly from people, systems and processes. According to Altair Consulting Group [6], the objectives of this perspective are to identify the human capital, systems and organizational climate required to support value creation processes.

Balanced Scorecard has evolved from its early use as a simple performance measurement framework to a full strategic planning and management system. Balanced Scorecard is used by many successful businesses all around the world [7], [8].

2. The Bamboo Industry in Mexico

According to Estrada [9], the Mexican manufacturing industry is investing little in Research and Development activities and the poor resultants of these investments are concentrated in few sectors as: cars, glass, cement as well as office and electronic equipment. Therefore, there is a huge need to encourage new technologies in alternate areas of the economy, so the industrial platform can take off.

In the case of the industry of Bamboo in Mexico the general situation reflects the opportunity for this sector to grow. Nowadays the use of bamboo in Mexico is not very exploited and companies that use it as raw material are limited to decorative purposes only such as walls, floors and partition panels in simple constructions. Nevertheless, there are some states in Mexico with the optimal conditions to have bamboo plantations such as Veracruz, Chiapas and Puebla. Bamboo is known as “one of the most surprising plants in nature”, known as "the plant of the thousand uses" because from it, food, clothing, construction material, pulp for paper and medicines, among others can be obtained. In Mexico, bamboo is used in the construction of rural dwellings, in the manufacture of handicrafts, furniture and home accessories, also for medicinal and ornamental purposes.

For this study, the enterprise was a local producer of Bamboo constructions such as cabins, docks, pergolas, storage rooms and specialized designs which are manufactured in this medium company. The business was created in 2011 and has participated in several government programs with the purpose of developing the technology of its processes and improving its quality and profitability. One of the supports the company received was a special funding from the Mexican National Council of Science and Technology with the participation of a public university, in order to aid the enterprise to create an innovative process for the massive production of bamboo panels. This is the framework that budgeted and allowed the development of this project, which started with an organizational diagnosis process and included a reengineering purpose, embedding both perspectives in the creation of innovative production processes.

3. Problem situation

Nevertheless the organizational diagnosis, was intended to identify the difficulties and opportunity areas of the company, there were some symptoms which pointed to specific problems: lack of a well-defined organizational structure, lack of clarity about the product benefits and its uses, an inexistent trading strategy, non-structured procedures which led the production to artisan and non-standardized products, organizational ends not clearly established, an inappropriate layout, and confusing management styles.

So, the problem statement for this study was that, at the moment of the project, the company did not have formal procedures, documents and a business strategy which allowed it to develop its operations under a quality and innovation style that responded to the requirements of its sector and market.

Therefore, it was necessary to set a way to identify, analyze and combine tools of problem solving, work improvement and processes engineering such as organizational diagnosis and reengineering. Reengineering is a methodology whose principles are used to implement other philosophies such as Total Quality Control [10].

4. Reengineering

Reengineering is “the notion of discontinuous thinking –of recognizing and breaking away from the outdated rules and fundamental assumptions that underlie operations” [11]. Reengineering processes require analysis of managerial efficiency. Welsch, Hilton, Gordon and Rivera mention that the efficiency of an entity is generally recognized as the single most important factor in long-term success. The success of the organization is measured in terms of achieving its goals. For this, it is necessary to implement the activities through the efficient use of human, material and financial resources. That is why it is important to develop and implement the process of planning, organizing, directing and controlling. “Business process improvement is about analyzing the current behavior of process execution in order to identify the process inefficiencies such as agent assignment, resource utilization, and control flow” [12].

To begin with the administrative process, the first phase is planning, which consists of defining what will be done, how to do it and who will. It implies defining the vision, mission, strategies, objectives, goals, actions and initiatives, as well as the allocation of human, material and financial resources for the implementation, monitoring and continuous improvement of the organization. In order to achieve this, it is necessary to develop the strategic plan. This document demands the consideration of the current and desired situation for its construction and defines the strategy.

The holistic view of this managerial approach was understood following the spots of organizational diagnosis.

5. Organizational diagnosis

There have been many scholars who have defined the concept; however, the following definitions were suitable for the theoretical framework of this study. According to Vidal [13], it is “the process of comparison between two situations: the present, which we have come to know through inquiry, and another already defined and supposedly known that serves as a guideline or model”. Porter [14], points out that the competency analysis consists of four diagnostic components: future goals, current strategy, assumptions and capabilities. Savall and Zardet [15] highlight that a socio-economic diagnosis aims “to make the inventory of the dysfunctions and calculate the hidden costs that they originate”, and add that the diagnosis is based on three sources of information: interviews, direct observations, and document analysis. Audirac, et. al. [16] refer that “the diagnosis consists of obtaining valid

information about the organization. It involves collecting and analyzing information about the culture, processes, structure and other essential elements of the organization". Finally, the definition of Ortega [2], established the organizational diagnosis as "the action and effect of specifying the state that an organization keeps from the identification and understanding of its signs and symptoms".

From these definitions and theoretical framework, it was possible to propose a methodological design to diagnose and reengineer the company of this study.

6. Embedded Methodology for the organizational diagnosis in the bamboo industry

The methodological design started from considering the situation of the company as a unique case study. Each case study presents its own features due to its culture, context, management style, knowledge as well as inner and outer actors, which altogether lead to the application of discerned, adjusted analyzes that incorporate these characteristics. Therefore, the methodological design was sustained by different approaches and techniques which have been successfully used in different areas. So, to describe the current conditions of the company, this research used observation, documentary revision, survey and interview, as data collection techniques, with the instruments of field diary, data sheets, questionnaire, focus groups and semi-structured guides [17]. Also, the investigation was considered non-experimental since there was no variables manipulation, and only the facts as they were developed, were observed and described [18].

Therefore, we started form a holistic methodological approach based in the general system theory and particularly in the methodology of soft systems of Checkland [19], that sets the status of a given problem situation, to be then contrasted with the ideal state though the systemic thinking. We designed the diagnostic intervention and we also used benchmarking for the contextual analysis and systemic diagnosis, in order to perform an internal and external auditing of the company. To be able to understand the problems of the organization, a Balanced Scorecard setting was useful in order to align the vision and strategy of the company with the objectives and initiatives. For the application of techniques, a team of five observers was formed, who used audio and video recording equipment as well as observation rooms. The methodological strategy considered a qualitative design applied on March 2016 at the enterprise, which is located in Mexico City, where we worked with local suppliers of this company. The actions performed were:

1. Documentary analysis to formulate a sectorial context at international, national levels through information sources such as sectorial indexes, web pages of companies and videos on internet.
2. Field work through "in situ" inspections that included the applications of the following techniques:
    - Observation guide about planning, organization, management aspects and control of the enterprise.
    - Focus group with the main suppliers of the enterprise to determine the competitiveness of an organization and its productive chain.
    - Technique of Nominal Groups for the revision of the mission and vision, creation of the policies, strategies, objectives and initiatives [20].
    - A focus group with the employees of the company, who also answered the tool of Organizational Health Checkup.
3. Cabinet study for the integration, analysis and synthesis of information collected with the following phases:
    - Design of a cluster for the enterprise.
    - Writing the mission and vision, creating policies, strategies, objectives and initiatives.
    - Drafting of the Balanced Scorecard and creation of the Strategic Map for the Company.

Following these settings, an organizational proposal of the methodology was established dividing the involved elements into endogenous and **exogenous**, considering the actions of organizational diagnosis as inner activities and the reengineering processes as strategic tasks. These, allowed us to include the perspective of intrinsic and extrinsic influence factors in the activity of the enterprise, and embed the theoretical aspects of reengineering and organizational diagnosis as shown in Figure 4.

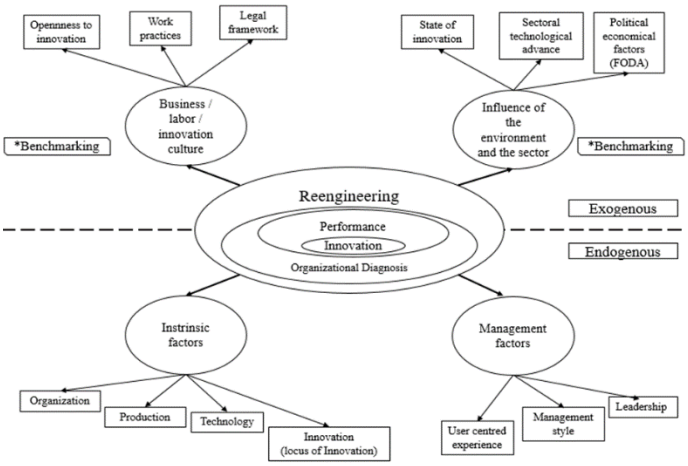


Figure 4.Embedded Methodology for the organizational diagnosis in the bamboo industry. Source: Author's elaboration

Following this design, data gathering and data analysis led us to the following results and discussion.

7. Results and discussion

7.1. Diagnostic Results

To begin with the diagnosis, an evaluation of the environment is required. Environment refers to the set of people, organizations, institutions, objects or elements of any nature, which are outside the company, have some significance for it. For this purpose, the production chain of the enterprise was identified by conducting a focal group interview where suppliers, subcontractors, distribution channels, transport, communication technologies, direct competitors, sector barriers and customers were required.

The mission of the Company is to: Minimize environmental impact and maximize social impact, taking up vernacular construction techniques and enhancing them with innovative technologies.

In general, it is observed that the mission is structured in an appropriate way, although it is to point out that the enterprise should review its geographical scope and develop strategies at local, regional or international levels. As a conclusion, it is clear they are still working in a stage of growth.

The vision of the enterprise is: *To be in 2020 a leading company in the sustainable construction market, through the fusion of vernacular techniques with innovative technological developments.*

The vision and mission were reviewed. In addition, the establishment of strategies and policies was done using a nominal group technique. This session was attended by executives and employees of the Company. The Technique of Nominal Groups allows a working group to reach a consensus on some subject, according to their level of importance and according to priorities established by the group [20].

The vision remained the same; but the mission, was adjusted as follows: *To minimize environmental impacts and maximize social impacts, taking up vernacular construction techniques and enhancing them with innovative technologies, where our employees carry out a specialized monitoring in each of the stages of the cycle life of products.*

After applying the technique of nominal groups the generic strategy for the enterprise was précised as the differentiation since it is desired to:

- Create a product and / or service that is perceived as unique.
- Work in design, brand image, technology, customer service, dealer chain, high quality, etc.
- New product design with high quality materials.
- Intense customer support by providing specialized follow-up at each stage of the product life cycle.

7.2. Balanced Scorecard results

Some specific aspects that were taken into account for the creation of the Integral Chart of the Company contemplating important aspects defined during the Systemic Diagnosis included:

- The alignment of the personal goals of all employees should have with the company's strategy [21]. So employees were asked to write their goals according to the company strategy.
- The value of intangible assets such as knowledge and technology that have a direct impact on income and profits. In this case the investment which is making innovation and technology is of vital importance. Innovation focused companies must offer products and services whose performance exceeds the competitors in which the customer values.

Since the Company is starting with mass production, in the manufacturing aspect, simple indicators such as on-time delivery, customer waiting time, number of reworks, defect rates, cycle times, etc., were presented.

The Balanced Scorecard remained for the company to align the objectives of the employees and directors with the strategy, and to carry out the plan of operation and revision. With this strategy in mind, and the important aspects to consider, the strategic map, the objectives and the initiatives of the Balanced Scorecard, the results and the initiatives developed are in [1] considering the proposal of [4].

7.3. Management style

The Pareto principle establishes a distinction between the aspects that require greater attention and that have greater impact in the organization, discriminating those with less impact. It allows to discriminate the few vital issues of the trivial many. After data gathering with surveys and interviews, a Pareto graphic was formulated showing the most important problems in the management style in Table 1 and Figure 5. This was one of the reasons why the use of Reengineering in the case study is justified because one of the strongest problems they face is organizational and this factor has been shown to have a very strong relationship with reengineering aspects [22].

Table 1. Data for the Pareto Chart, Source: Author's elaboration.

Problems mentioned	Sort Data	Number of Accumulated Defects	% Total	% Accumulated
Organization	10	10	23.80952381	23.80952381
Communication	10	20	23.80952381	47.61904762
Decisions	7	27	16.66666667	64.28571429
Time	6	33	14.28571429	78.57142857
Management of staff	3	36	7.142857143	85.71428571
Planning	2	38	4.761904762	90.47619048
Commitment	2	40	4.761904762	95.23809524
Work environment	2	42	4.761904762	100
	42			

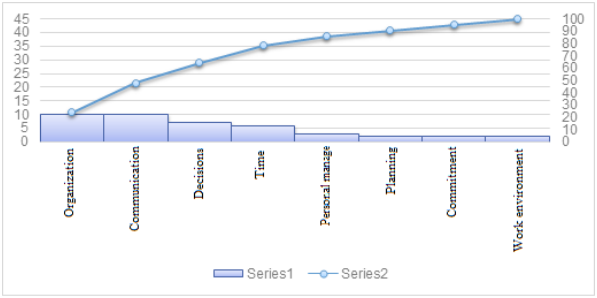


Figure 5. Pareto Chart. Source: Author's elaboration.

Thus, according to the graph of the total of problems detected and confirmed by different techniques, those of greater weight for the correct operation of the company were the organization, the communication and the decision making.

7.4. SWOT Matrix

In addition to the Benchmarking carried out in [1], the SWOT (*strengths*, *weaknesses*, *opportunities*, and *threats)*is a structured

planning method that evaluates those four elements in the organization. It was built with all the elements of the Diagnosis, in this sense 10 employees of the company were asked to vote for the most significant aspects. The results are shown below:

Table 2. SWOT Matrix. Source: Author's elaboration.

Forces	Weaknesses
1. The company is concerned about the environment, which means that it has sustainable products 2. The constant innovation in products. 3. Their products are unique in our country. 4. The architectural projects that have been carried out so far have a harmonious design. 5. The structural and seismic tests performed on constructions with bamboo panels are favorable. 6. Its management for the search of economic funds is good. 7. The presence of the company in associations dedicated to sustainable construction	1. The shared leadership and the different ideas of the owners of the organization. 2. The lack of market certainty for this type of constructions. 3. The lack of customer follow-up. 4. The company do not have defined processes, nor standardize. 5. Its processes of production are handicraft which makes massive production impossible. 6. There is high turnover in staff, which causes them to waste time in training them.
Threats	**Opportunities**
1. Supplier relationships are not narrow, making it difficult to implement some manufacturing systems. 2. The company has no market share. There is not even a market for what is intended to be commercialized. 3. There are companies that are already manufacturing other products with bamboo, especially decorative walls and floors that could venture into the design and construction of houses. 4. Creation of microenterprises with the same turnaround due to unemployment, since the unemployment rate in the country during the second quarter of 2016 was 3.9% of the Economically Active Population and although it is lower than in 2015 with 4.3% , This one is in constant dynamism. 5. The annual inflation rate, which currently stands at 2.88%, and because it is rising, makes it difficult each day to buy its inputs as well as the stability of the prices of its products.	1. The support that many institutions grant to companies that minimize environmental impact (CONACYT and public universities). 2. Currently, until the moment of the investigation, no other company located in Mexico was found on the Internet to do the same. 3. The company has strived to develop sustainable technology and products for the construction of its products. 4. Mexico in many states is suitable for planting bamboo, which could benefit the company by having high quality raw material. 5. The closeness with the United States of America that accepts well the construction with prefabricated wood panels. The possible devaluation of the peso may facilitate exports. 6. Manufacture other bamboo products that are already accepted as decorative walls, partition panels, floors. 7. Carry out a cluster with civil associations and municipal governments that carry out decent housing in marginalized areas. 8, The company is behaving adequately for the business phase in which it is: Growth, dedicating itself to the development and identification of new products and services, expanding facilities and investing in infrastructure 8. Increased demand for housing at the national level, in 2015 was 1 159,480 households and during the current year is estimated to be 1'166,872 households.

This served to evaluate if the strategy of the organization and the objectives set out in the Balanced Scorecard, are correctly defined, in addition to identifying aspects such as management style and influence of the environment.

8. Innovation

The innovation approach followed for the bamboo industry in this study was, the one proposed by Belay-Endalamaw [23], who suggests identifying at first the innovation scope, according to the following typology:

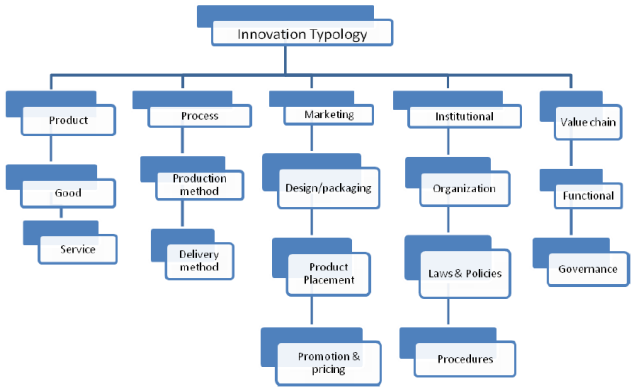


Figure 6. Innovation Typologies [24]

9. Reengineering results

Once the mission, vision, strategy and policies were defined, the organization chart and the job manual were built. Figure 8 shows the organizational chart and in Figure 2, as an example, a position of the manual: project coordinator.

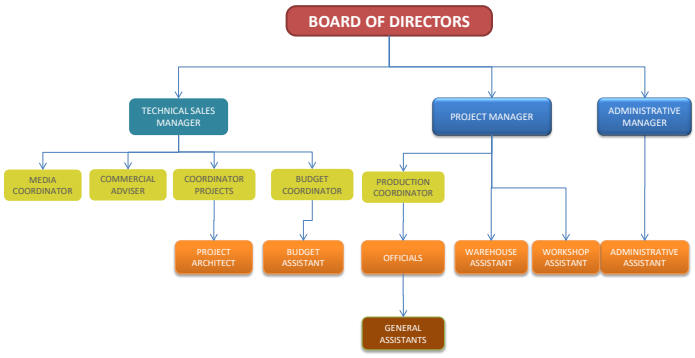


Figure 7. Organization chart of the company. Source: Author's elaboration.

Then, in order to establish the proper features of work positions and to reorganize communication flows and tasks, a manual of positions was developed.

The foregoing supports the development of manual procedures, which serves as a means of communication and coordination between different levels of a company. The procedures manual describes routine work tasks, describing the procedures used within the organization and the logical sequence of each activity, to unify and Control work routines and to avoid their arbitrary alteration. It also helps the supervision of standardization activities, avoiding duplication of functions and unnecessary steps within processes, facilitates the work of the

administrative audit, the evaluation of internal control and its monitoring.

POSITION	PROJECT COORDINATOR	DIRECT BOSS	TECHNICAL SALES MANAGER
AREA	PROJECTS	PERSONNEL IN CHARGE	PROJECT ARCHITECT

SCOPE
Coordinate the design of the projects between architecture and engineering and the operation of the building trades.

Functions and Responsibilities	Profile
1. Coordinate the design of projects 2. Coordinate with the project engineering 3. Deal with customer in design matter 4. Coordinate the building trades	EDUCATION Architect EXPERIENCE In project coordination TRAINING • Office (Word, Excel, Power Point) • Internet and tools • Design programs:Autocad, Sketchup. • Planning Programs: Project ABILITIES: • Agility in media • Capacity for planning and organization • Ability to making decisions • Compromise • Communication • Discipline • Ethics • Order • Proactivity • Responsability

Figure 8. Position of Project coordinator, in the manual of positions. Source: Author's elaboration.

The proper structuring of the manual should reflect the specific activities carried out, as well as the means used to obtain the objectives, while facilitating the execution, monitoring and evaluation of organizational performance. This should be an instrument that supports the process of updating and improvement, by simplifying the procedures that allow the proper and efficient performance of the assigned function. The manual has a general objective, the specific ones by the procedure and the norms of operation. Only the purchasing process is displayed in Figure 9 and Figure 10.

1	Warehouse assistant	Receives a copy of the purchase order	Purchase order CBW01
2		Checks copy of purchase order, identify product and verify matching purchase order, invoice and product	CBW01
4	General Assistant	Unloads and counts the products, entering the quantity on the raw material registration sheet	Raw material registration sheet PBW04
5	Warehouse assistant	Inspects that the raw material meets the required specifications, according to the raw material quality control guide	Raw material quality control guide PBW05
6		Compares the information on the raw material registration sheet with the purchase order information.	Purchase order and raw material registration sheet CBW01, PBW04
7		Do the products and quantities match?	
8		Affirmative. Go to Activity 10	
9		Negative. Return the raw material to the supplier.	
10		Signatures of incidence on the raw material registration sheet and if necessary to return it to the supplier.	Raw material registration sheet PBW04
11		Signatures of conformity on the raw material registration sheet.	Raw material registration sheet PBW04
12		Determines to which warehouse carry the product	
13		Coordinates product income	
14	General Assistant	Transports the goods from the vehicle to the inside of the warehouse	
	Warehouse assistant	Accommodates and classifies the raw material entered into the warehouse	

Figure 9. Example of the procedure of purchasing process, in the Manual of procedures. Source: Author's elaboration.

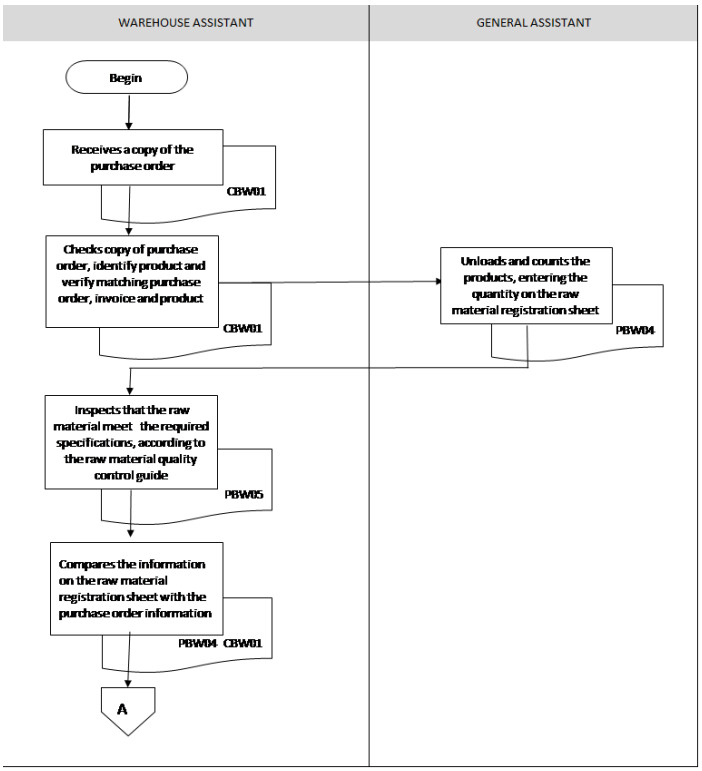


Figure 10. Purchasing process, graphic representation. Source: author's elaboration.

So, after showing evidence of the developments and results of embedding the approaches of organizational diagnosis and reengineering, a summary a table with different approaches considering in both methodologies are shown in Table 3.

10. Conclusions

This project was carried out since March 2016 in a medium company dedicated to the development of bamboo structural panels. In the "Diagnosis", something that was observed is the company did not have a strategic plan, so they proceeded to review their vision and mission. Their policies were generated and the strategy was also determined. Another of the deliverables was the "Performance Indicators", but they were not intended to establish solely the relation to the productive process; they were indicators that really aligned the organization with its strategy to lead it to achieve its vision. The Balanced Scorecard gathers indicators from four perspectives: Customer, Employees, Internal Business Process and Finance and developed clear and measurable objectives in each of these items so that the company can achieve its vision through the differentiation of its product and service [1].

In this article an extension was discussed embedding two methodologies: System Diagnosis and Reengineering, considering exogenous and endogenous factors, innovation culture, influence of the environment, management and another intrinsic values like organization and culture. Some results in a bamboo company are shown in order to validate this methodology. The most important is the use of System Diagnosis and Reengineering to achieve innovation as performance indicator to consolidate the company strategic planning.

Some initiatives to achieve the objectives have already been worked out such as employee training in 5's, revision of the

organization chart and job profile, creation of workshops that bring the Company closer to its suppliers, creation of the manual of processes, among others. However some pending includes the development and implementation of the proper facility layout; implementation of the automated assembly machine, among others

Future work should be aimed at validating the proposal of an inner culture of innovation and its evaluation.

Table 3. Summary of Techniques. Source: Author's elaboration.

	Tools/techniques	Description for application
System Diagnosis	Organizational Health Check up	Employees of the company
	Focus group	Main suppliers of the enterprise to determine the competitiveness of an organization and its productive chain.
	Technique of Nominal Groups	Revision of the mission and vision
	Observation guide	Planning, organization, management aspects and control of the enterprise.
	Interview	Data collection techniques
	Survey	Data collection techniques
	Pareto Graphic	Shows the most important problems in the management style.
	SWOT Matrix	Serves to evaluate if the strategy of the organization and the objectives set out in the Balanced Scorecard, are correctly defined,
Reengineering	Balanced Scorecard	It is based on the definition of strategic objectives, indicators and strategic initiatives, establishing relationships by a strategic map in four basic perspectives: financial, clients, internal processes and learning and growth, then the strategy should be translated directly into objectives.
	Manual procedures	Serves for communication and coordination between different levels of a company
	Manual of positions	Define the proper features of work positions and to reorganize communication flows and tasks
	Innovation	The forms of innovation included in this study were product, process and organizational innovation.

References

[1] Hernández E. et. al."Systemic Diagnosis and Strategy-Based Performance Indicators :Bamboo Panel Company Case". Proceedings of 2017 the 6th International Conference on Industrial Technology and Management (ICITM 2017).Cambridge, UK.ISBN: 978-1-5090-5328-5, pp. 123-133

[2] O. Ortega. "Elementary My Dear Consultant: Sherlock Holmes-style organizational diagnosis". Mexico: Patria Publishing Group, 2015. (Original in Spanish: Elemental mi querido consultor: Diagnóstico organizacional al estilo Sherlock Holmes)

[3] R. Kaplan and D. Norton. "The Balanced Scorecard: Measures that Drive Performance". Harvard Business Review, January-February, 1992, 71-79.

[4] R. Kaplan and D. Norton. Conceptual Foundations of the Balanced Scorecard. Harvard Business School Press, 2010.

[5] Electronic Publication: www.fip.unam.mx/simposio_investigacion2dic04/el_cuadro_extenso.html. 2004. Recovered in October 2016. Authors: Rodiles and Fuentes

[6] ALTAIR CONSULTING GROUP. "Balanced Scorecard".Economía Journal, 3:150, 2005. (Original in Spanish: Cuadro de Mando Integral)

[7] Lesakova L. and Dubcováb K. Knowledge. "Use of the Balanced Scorecard Method in the Businesses in the Slovak Republic". Procedia-Social and Behavorial Sciences. Proceedings of the 3rd International Conference on New Challenges in Management and Organization: Organization and Leadership, 2 May 2016, Dubai, UAE, Elsevier.

[8] Kalendera Z and Vayvay O. "The Fifth Pillar of the Balanced Scorecard: Sustainability". Procedia-Social and Behavorial Sciences .Proceedings of 12th International Strategic Management Conference, ISMC 2016, 28-30 October 2016, Antalya, Turkey, Elsevier.

[9] S. Estrada, "Regional Differences in the Technological Conduct of Mexican Manufacturing Companies: The Case of Guanajuato". Economy, Society and Territory, 2006, 821-869.

[10] D. Kiran "Businesses Processes Reengineering". Chapter in Total Quality Management: Key Concepts and Case Studies, Elsevier, 2017.

[11] Hammer, M. (1990). Reengineering work: don't automate, obliterate. Harvard business review, 68(4), 104-112

[12] Herawan, T., Deris, M.M., & Abawaj,J., (2013). Proceedings of the First International Conference on Advanced Data and Information Engieering (DaEng-2013). Springer Science+Business Media. ISSN 18761119

[13] E. Vidal, Organizational diagnosis: Systemic evaluation of business performance in the digital age. Colombia: ECOE Editions, 2004. (Original in Spanish: Diagnóstico organizacional: Evaluación sistémica del desempeño empresarial en la era digital)

[14] M. Porter. "Capital Disadvantage: America's Falling Capital Investment System".Harvard Business Review, 77:6, November-December 1992.

[15] Savall H. and Zardet V. "Measures and Piloting of Social Responsibility and Company Society, Results of Longitudinal Investigations". Journal of International Institute of Costs, 4, january/june,2009 .(Original in French: MESURE ET PILOTAGE DE LA RESPONSABILITÉ SOCIALE ET SOCIÉTALE DE L'ENTREPRISE -RÉSULTATS DE RECHERCHES LONGITUDINALES)

[16] Audirac, C., C. A,. León, E., V., Domínguez, G., A,. López, G., M. E. y Puerta, N., L.I "ABC of Organizational Development" Mexico :Trillas, 2012. (Original in Spanish: ABC del desarrollo Organizacional)

[17] Rojas, R "Guide to do social research". Mexico: Plaza and Valdes Editors, 2013. (Original in Spanish: Guía para realizar investigaciones sociales).

[18] Hernández, R., Fernández, C. y Baptista, M. del P. "Research Methodology". Mexico: McGraw-Hill, 2010. (Original in Spanish: Metodología de la investigación)

[19] P. Checkland, P. Systems Thinking, Systems Practice. New York: John Wiley and Sons, 1999.

[20] A. Delbecq, A. Van de Ven and D. Gustafson. Group Techniques for Program Planning. Chicago: Scott, Foresman, and Co, 1975.

[21] P. Drucker, P. The Practice of Management. New York: Harper Collins, 1954.

[22] Özge A and Erkut A. "The Effects Of Reengineering, Organizational Climate And Psychological Capital On The Firm Performance. Procedia-Social and Behavorial Sciences. Proceedings of 12th International Strategic Management Conference, ISMC 2016, 28-30 October 2016, "Antalya, Turkey, Elsevier.

[23] Belay, E. T. Towards Bamboo Commercialization in Ethiopia: Analysis of Technology Sources, Innovation and Entrepreneurship Doctor. Technische Universität Dresden,2015.

[24] Weerawardena, J "The Role of Marketing Capability in Innovation-Based Competitive Strategy", Journal of Strategic Marketing11, pp. 15-35, 2003.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1634-1646 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Three-Dimensional EEG Signal Tracking for Reproducible Monitoring of Self-Contemplating Imagination

Qinglei Meng[1*], Deepa Gupta[1], Abenezer Wudenhe[1], Xiaoming Du[2], L. Elliot Hong[2], Fow-Sen Choa[1]

[1]*Department of Computer Science and Electrical Engineering, University of Maryland Baltimore County, 21250, Baltimore, USA*

[2]*Department of Psychiatry, University of Maryland Baltimore, 21201, Baltimore, USA*

ARTICLE INFO

Article history:
Received: 31 May, 2017
Accepted: 12 August, 2017
Online: 24 August, 2017

Keywords:
Electroencephalography (EEG)
Brain waves
Prefrontal lobe
Visual cortex
Dynamic tracing

ABSTRACT

Electroencephalography (EEG) can globally monitor neural activity in millisecond scale, which is critical for identifying causality of human brain functions and mechanisms. However, to obtain accurate EEG stimulation-response relationship one usually needs to repeat multiple-ten times of stimulation-response recording to average out background signals of other irreverent brain activities, making real-time monitoring difficult to be accomplished. In this study, we explored new approaches which don't require repeats. EEG signals were recorded from subjects doing mind tasks including image formation of motor functions or emotional subjects and mathematical calculations in mind. Time stamps in EEG recording were used to mark task completion time. Signals within 300ms or 1,000ms before task completions were analyzed. Using sLoreta 3-D tracking we found that delta-wave activities were mostly located at frontal lobe or visual cortex, isolated with each other. Theta-wave activity tended to rotate around cortex with low spatial correlation. Beta-wave behaved like inquiry types of oscillations between any two regions across cortex and was consistently correlated with each other over different areas. Alpha-wave activity looked like mixture of theta and beta activities. Together with sliding window dynamic connectivity method we confirmed beta waves play key roles in linking different brain areas together for information inquiry. Theta and low Alpha are more likely playing the role of information control, integration, and image formation. With the proposed new method we demonstrated reproducible linkages of subject behaviors with 3-D tracing characteristics along the 3 categories: emotion, math calculation, and motor functions without using event repeats.

1. Introduction

This paper is a full version of 2016 *IEEE Signal Processing in Medicine and Biology Symposium* (Wudenhe, Meng and Choa 2016) conference abstract. Electroencephalography (EEG) is one of the most important neural imaging modalities used for monitoring brain activities [1]. Its ability to achieve millisecond scale time resolution is critical for obtaining causal relationship. However, due to the dynamic nature of brain electrical activity, the same stimuli may generate different responses under different psychological setting. Interpreting EEG signals, extracting EEG biomarkers and correlating them to behaviors have always been challenging since the technique was discovered. Event related potential (ERP) technique has been one of the successful methods that is used to obtain reproducible brain responses to stimuli. However, ERP signals are intrinsically weak and are typically buried under signals produced by other brain activities. To extract ERP signal, one needs to average over many repeats to distinguish it from noises and background signals produced by other brain activities [2]. Furthermore, brain states are high dimensional and complex. Multiple origins or sources can recruit neural activities from different paths and influence the same behavior outcome. That puts serious restrictions on ERP experiment designs and applications for studying event based brain dynamics and causal relationship. So, even though ERP is a useful tool, its ability to study brain dynamics and applications coverage are limited. To eliminate the need of experimental repeats, in this work, we

*Corresponding Author Name: Qinglei Meng, Department of Computer Science and Electrical Engineering, University of Maryland, Baltimore County, Baltimore, MD, 21250, USA. Email: meng2@umbc.edu. Tel: 410-455-6553.

developed EEG 3-D signal tracking methods to monitor brain dynamics without event repeat. We also verify obtained results with EEG based functional connectivity and statistical studies as described in detail in the following.

Earlier studies have shown that EEG signals may play important roles in controlling and initiating brain functions and activities [3, 4]. For example, attention usually requires a high beta wave to synchronize brain oscillators at different locations and alpha wave is used to decouple or disengage attention [4]. Furthermore, brain reticular formation area in the brainstem has been considered where consciousness and attention control initiated [5]. Considering if these descriptions are all correct, we shall likely observe some EEG signals initiated from the brain stem region when monitoring 3-D EEG signal traces.

To investigate brain activities in deep brain region, the sources of EEG signals need to be localized by solving inverse problems. Various techniques have been proposed, such as Low-Resolution Electromagnetic Tomography (LORETA) [6], Variable Resolution Electromagnetic Tomography (VARETA) [7], Brain Electrical Source Analysis (BESA) [8], to establish procedures for source localization. However, most of these programs cannot guarantee unique solution and are not able to accurately localize the EEG sources. One exception is the standardized LORETA (sLORETA) method, which improves the source localization accuracy by reducing spatial resolution. Pascaual-Marqui reviewed different methods of EEG localization [9], in which compared multiple mathematical models and concluded that extending the good localization properties of 2D minimum norm solution to 3D solution spaces with LORETA was the greatest challenge in the development of EEG source localization but was achieved with LORETA. With the help of LORETA, Herrmann and his colleagues investigated error processing of 39 subjects engaging the Eriksen flanker task and found significantly higher brain electrical activity in medical prefrontal areas for incorrect responses and both positive and negative error related components represented different aspects of error processing [10]. An earlier study from Mientus aimed to address whether LORETA was able to detect hypofrontality in schizophrenic patients by recording resting status EEG signals from patients and normal subjects [11]. For comparison, subjects with schizotypal personality and depressive patients were also investigated. Significant increase of delta wave activity was detected in the patients over the entire cortex, while both schizotypal subjects and depressive patients showed much less delta, theta and beta activities in the anterior cingulum. Their conclusions obtained from LORETA analysis are largely consistent with findings from related literatures. Saletu's group reported application of LORETA in diagnosis and pharmacotherapy of depression [12], in which LORETA identifies cerebral generators responsible for the pathogenesis of depression and for the mode of action of antidepressants. LORETA was also used to verify the hypothesis that the highest temporal correlations between 3D EEG current source densities corresponds to anatomical Modules of high synaptic connectivity [13]. These researchers demonstrated that EEG temporal correlations between different brain areas were related to synaptic density as measured by diffusion spectral imaging.

In this study, we apply sLORETA techniques to obtain 3D EEG map, trace them in time and associate these traces with their corresponding behaviors. Since many of these vector traces are long jumps across over large brain regions in a very short period of time, the method provides a big picture about brain dynamics with the understanding that sLORETA has its limitation on spatial-resolution. The proposed method is a much simpler approach to monitor brain dynamics and function mechanism in the sense that no complex experimental arrangement for event repeats are required. It opens the door for real-time monitoring of brain natural responses under any desired psychological setting without worrying about sensory gating effects and how to experimentally set up event-repeats [14].

To verify our conclusions from the monitoring results we have also implemented EEG based sliding-window functional connectivity analysis. We use the technique to obtain correlation statistics and its implications are consistent with 3-D tracing monitoring results.

Correlation between two brain regions is studied by segmenting EEG data followed by calculating correlation coefficients between two electrode channels. Such schemes basically involve breaking down the EEG data into segments followed by extracting features (such as correlation in this case) that are henceforth used to find patterns in the EEG activity [15, 16, 17]. The segmentation approach (as described further in the methods section) has recently adopted by function magnetic resonance imaging (fMRI) researchers to analyze brain dynamics in terms of functional connectivity when subjects are performing tasks [18, 19]. To the best of our knowledge, the combined segmentation and sliding window analysis has not been done with EEG analysis. We describe the implementation methods and experiment details in section 2, report their results in section 3, and finally conclude our studies in section 4.

2. Methods and Experiments

In this section we first describe our experiment design and arrangement and then the methods used to analyze the recorded data with sLORETA and sliding window techniques.

2.1. Experimental setup and design

A 16-channel EEG system was used in this study for EEG signal recording. Figure 1 illustrates the channel locations on cortex map and A1 and A2 are the reference channels on the left and right side brain, respectively. Six subjects participated this study (4 males and 2 females, all aged between 18 and 30). We designed experiments to obtain EEG recordings when subjects were under the following tasks. Subjects were asked (a). to visualize images of some parts of their body, for example, the left and right hands and feet, nose, lips, etc; (b). to form images in their mind of faces of people they love most or they hate most; (c). to perform simple mathematical calculations in their mind (like 15 multiplies 18 or 23 plus 38). When images or calculation results were formed in their mind, they time-marked the moment by finger touching a static electrode at hand which provided an electrical pulse to the EEG recording as a time marker as shown in the bottom tract in Figure 2.

Through the mirror neuron system in human brain, the tasks we selected can activate 3 categories of brain activities: emotion,

math calculation, and motor function without generating EEG artifacts like motor evoked potential (MEP) signals, ...etc. [20]-[23].

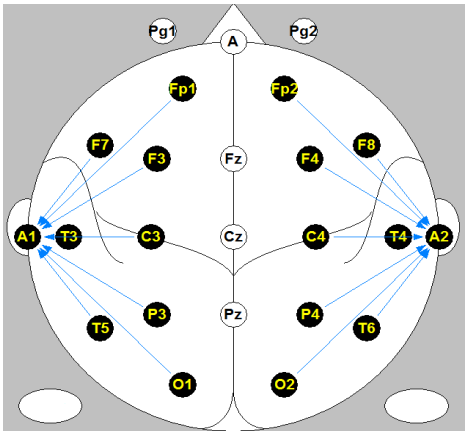


Figure: 1 16-channel EEG cortex and channel locations map

Raw EEG data was recorded by KT88, and then exported as BioSemi format for Matlab with EEGlab toolbox data processing followed by converting data into text files for sLORETA brain activity maps plotting [6]. Since the EEG system had a sampling rate of 100Hz, brain activity maps were recorded every 10ms.

2.2. Brain activity tracking by sLORETA

We processed EEG signals 300ms before the time marker since it takes that much time for human brain to process information and realize task accomplished or decision made. [24]. For example, to track dynamic trace of Delta activities during the 300ms, the filter in sLORETA was firstly set to 0-4Hz low-pass filter to extract delta waves. Then in the real-time EEG signal display in sLORETA, as shown in Figure 3, the cursor was located at the time point 300ms before the time marker (the starting point of the pulse generated in original EEG recording, for which the time was readable from KT88 EEG signals display). From this time point on, one Delta wave activity map was generated for every 10ms until the cursor reached the time marker itself. So during the analyzed 300ms, there was a series of 31 Delta wave activity maps. The same method was applied on Theta, Alpha and Beta waves by setting the filter in sLORETA to 4-8Hz, 8-13Hz and 13-30Hz, respectively.

For each sLORETA brain wave activity map, a corresponding focus point with the strongest brain activity was found (sLORETA provides this function). These focus points were connected and plotted in Matlab at a 10ms increments over the 300ms measured timespan recorded, and viewed considering both the X axis, showing left and right sides of the brain, Y axis, representing the front and back of the brain, and the Z axis representing the top and bottom of the brain. Axis distance range being set to -100 to 100 mm.

2.3. Sliding window analysis

We also analyzed functional connectivity between brain nodes (i.e. 16 EEG channels) during the above described mind imagination tasks by using the sliding window approach [18,19] for delta, theta, alpha and beta wave activity followed by standard deviation analysis as described below.

a. First, EEG data of 1000ms before the marker onset indicating task completion was extracted. This data was filtered for delta, theta, alpha and beta wave followed by processing with ICA to identify components. (We take 1000ms data because the sliding window approach requires larger sample size for computing meaningful results).

b. Then, data points within a time window of first 100ms for one of the wave activity were taken to calculate correlation among all pairs of electrodes for that one window. Next, the window was then shifted in time by 10ms of data points that overlaps between the successive windows. This resulted in 120 correlation values for each of the 91 windows as shown in Figure 4. Here, correlation between two electrodes is given by:

$$\frac{\sum_i (x_i(y_i - \overline{y}) - \overline{x}(y_i - \overline{y}))}{\sqrt{(\sum_i (x_i - \overline{x})^2)(\sum_i (y_i - \overline{y})^2)}}$$

where,

x_i = a time series data point from 1st electrode
y_i = a time series data point from 2nd electrode
$\overline{x}$ = mean of 1st electrode's time series data
$\overline{y}$ = mean of 2nd electrode's time series data
i = 10, 20, 30, … ,100ms

c. Finally, we then computed standard deviation (σ) of these correlation values over the 91 windows for each pair of channels to analyze the fluctuation in their connectivity. Here, if Ais a matrix that contains correlation values over the 91 windows then the standard deviation is defined as:

$$S = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N} |A_i - \mu|^2}$$

where,

μ is the mean of A and i = 1,2,3,…,(N = 91)

The above is done for all 120 pairs of correlated electrodes. Then, the maximum out of the 120 standard deviation values was extracted for narrowing down the analysis to most erratically correlated nodes. This was done for each wave and averaged over all experiment-tasks for that wave as plotted in Figure 5.

The above computation and analysis helps in quantification of the time-varying behavior in terms of functional connectivity and brain dynamics. From the best of our knowledge, the sliding window approach has recently adopted for fRMI data analysis to describe brain dynamics in slower scale compared with EEG signal analysis [18, 19]. Here we applied sliding-window techniques to EEG data analysis and extended the technique with maximum standard deviation analysis to help interpret EEG data. The standard deviation values indicate the degree of fluctuation that exists in the functional connectivity i.e. the correlation values. We compare the selected maximum standard deviation of all waves (Figure 5) as this enables us to test and see the correlation-consistency for each wave among their most fluctuated nodes that are erratically in and out of phase with each other. This is further discussed and analyzed in the results and discussion section.

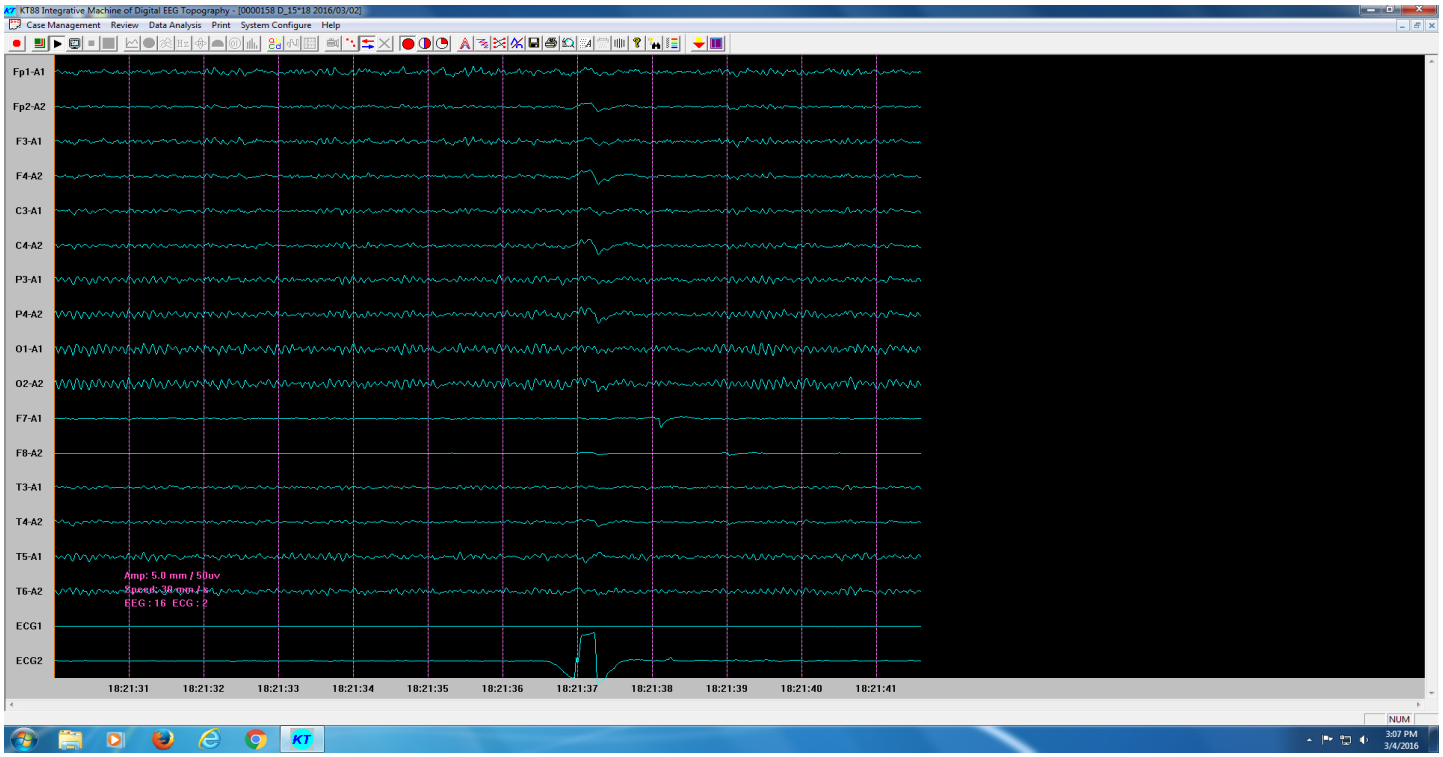


Figure 2: Example of real-time EEG signal on KT88 (The last channel was added to the EEG system as a time marker, once subjects touched the metal clip, a electrical pulse would be generated to mark the time point when images were formed in their mind.)

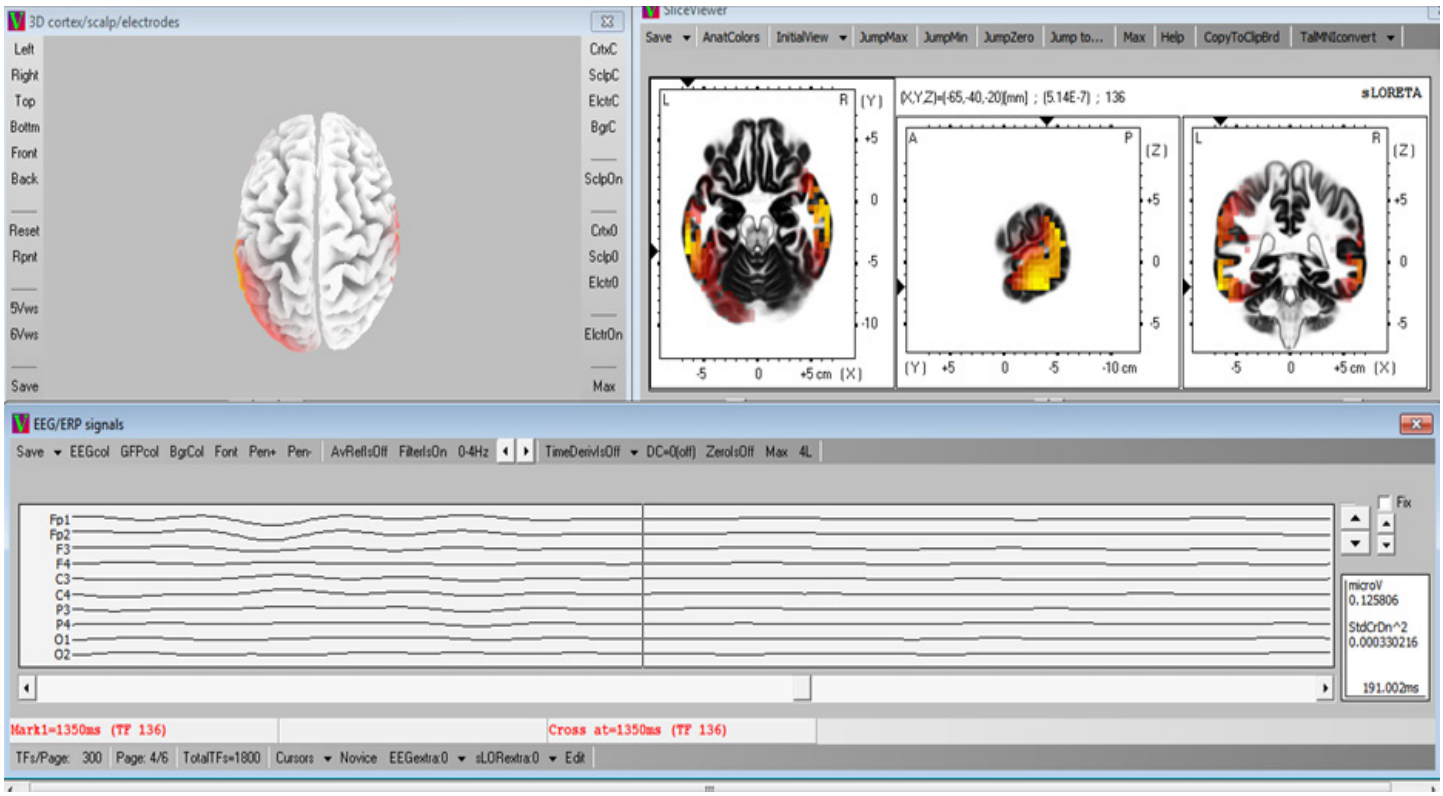


Figure 3: An example of sLORETA brain activity map (The waves under the 2D and 3D maps are EEG signal on time domain. A 0-4Hz low-pass filter was applied to analyze Delta rhythm. The maps are presenting Delta rhythm activities on the time point, where the cursor is located on the time domain signal. The 3D cortex map on the left-top corner can be rotated for different perspectives, and the 3 2D maps are the cross section views of xy, yz and zx planes in a 3D Cartesian coordinate system. They vary according to the positions of cuts, which are shown as the arrows along the x, y and z axis. By relocating these arrows, the brain activity inside the brain can be observed. Red regions represent relatively strong activities and yellow regions represent very strong activities.)

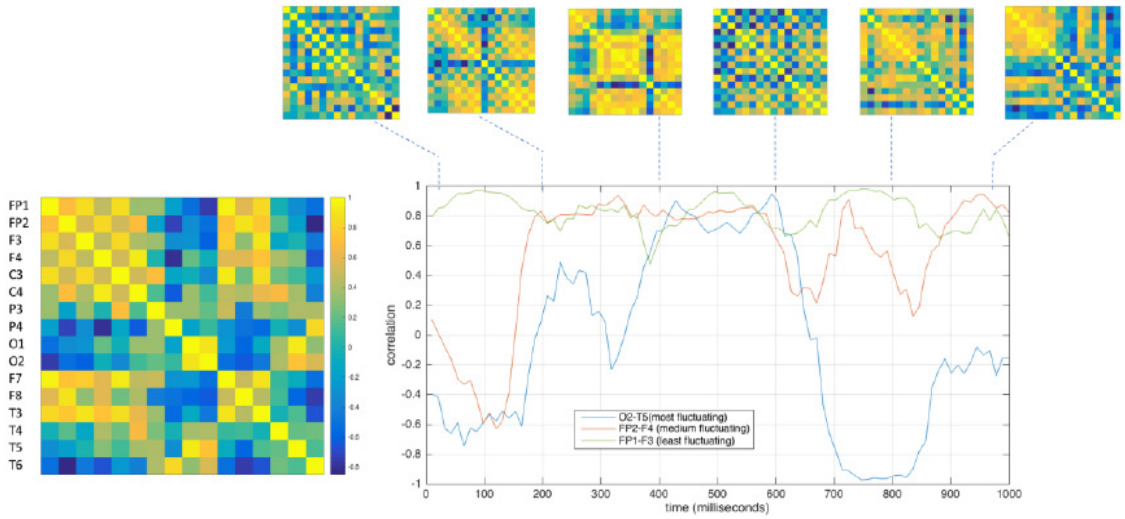


Figure 4: (left) Functional connectivity as correlation matrix between EEG channels from window of 100ms of EEG data. (right) Functional connectivity time series for connections between selected pairs of EEG channels based on maximum (O2-T5 pair in blue), medium (FP2-F4 pair in red) and least (FP1-F3 in green) deviation in their correlation values.

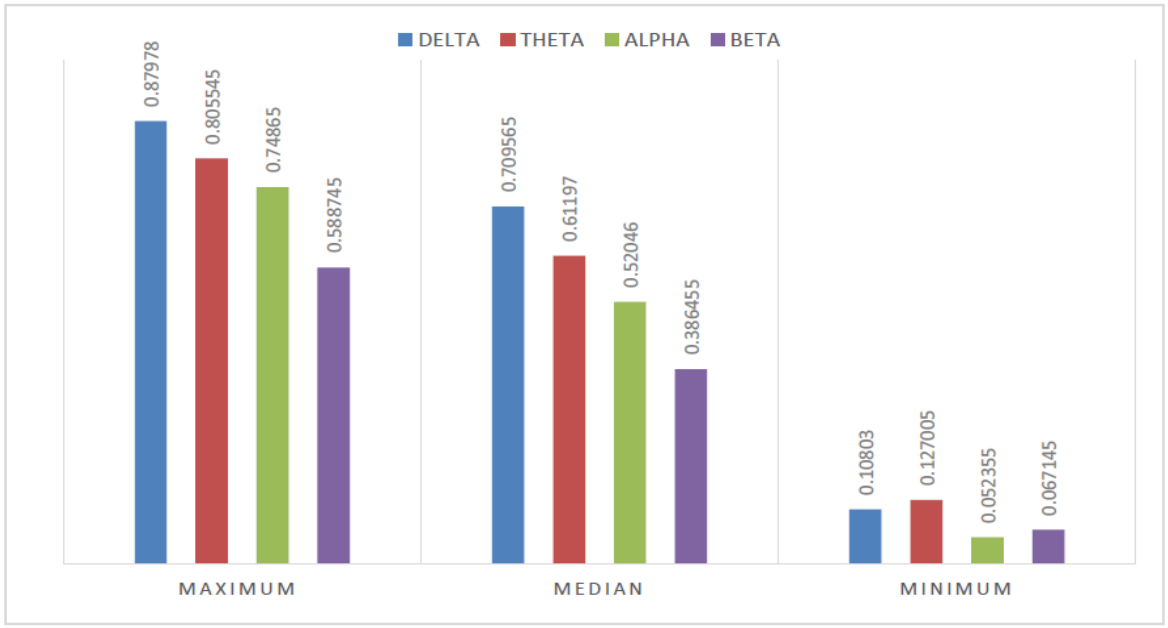


Figure 5: Mean standard deviation (σ) of correlation values that is obtained by the sliding window approach while the subject is performing imagination tasks. This is clustered as most (maximum σ), medium (median σ) and least (minimum σ) fluctuation in phase of the pair of EEG time series channels for each wave rhythm. Overall, correlation of EEG channels is the least fluctuating for Alpha and Beta wave in majority of the cases.

3. Results and Discussion

From cortex maps plotted by sLORETA we found that the Delta oscillations are not only the slowest in frequency, but also slowest in the shift or movement of activated regions within the 300ms before subjects realized image formations in their mind. For most of the time in the 300ms, Delta activities tend to arise at either frontal lobe or visual cortex. A sudden shift of activation between these regions would occur for most cases, and a duration of about 150ms to 180ms was observed between two shifts in the analyzed period. A group of plots in Figure 6 presents such Delta wave activities. We also found that the maximum and medium standard deviation of correlation (deduced from sliding window approach as described in the methods section) was highest for the delta wave. This indicates that the brain regions are not consistently correlated and less interactive with each other for the delta wave activity (Figure 5).

Cortex maps plotted by sLORETA also indicate that the Beta oscillation activities show up at frontal and visual cortex as well for most of the sampled time points, but the spatial shifts can be detected every 10ms as shown in Figure 7, which is much faster than that of the Delta oscillation. This is to say for one certain activity map in the 300ms sampled time series, the excited region is different from both 10ms before or after itself for Beta wave. Since the sampling rate of the EEG system is 100Hz, the Beta activity spatial shifts present a frequency of at least 50Hz,

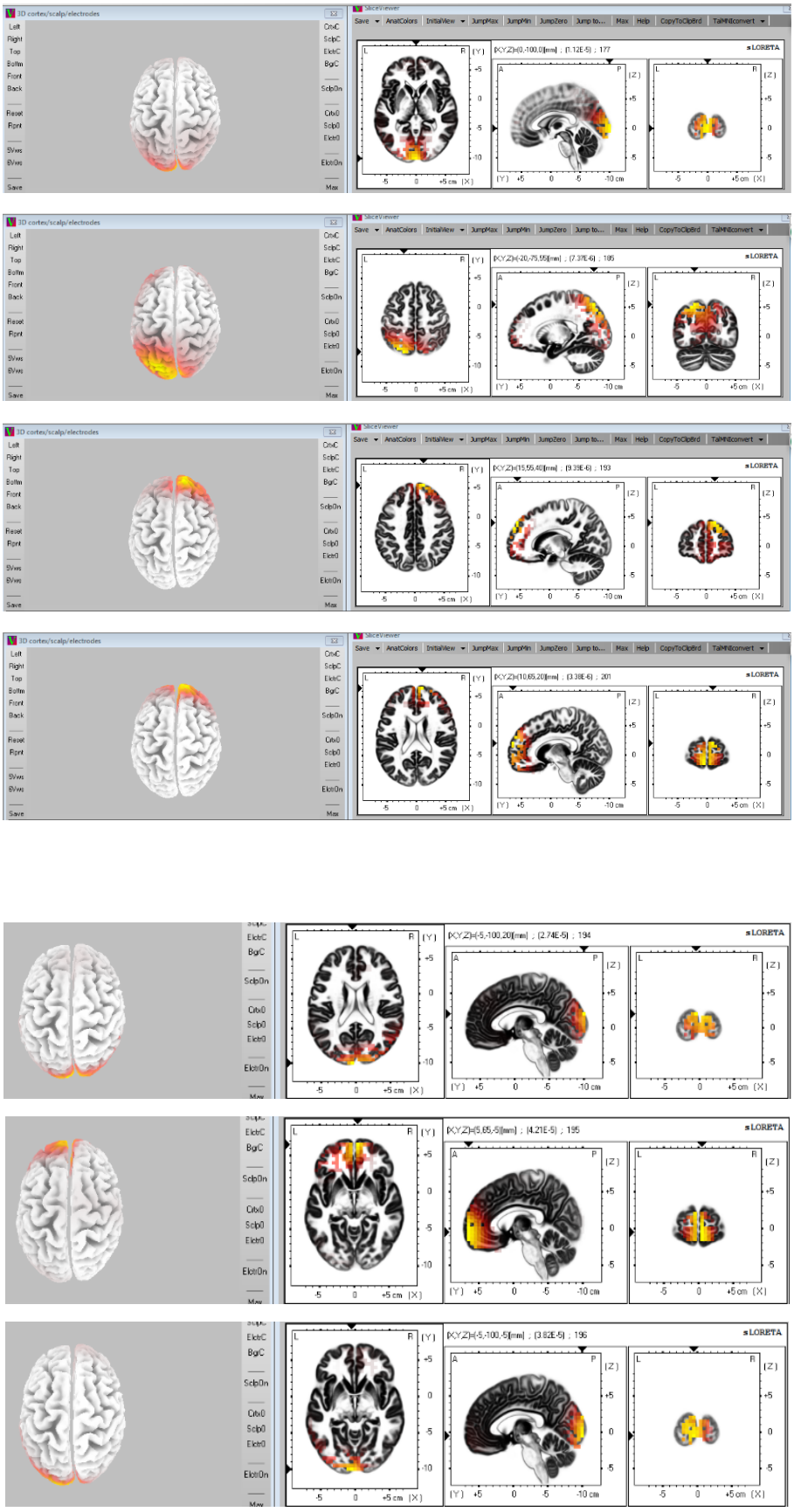
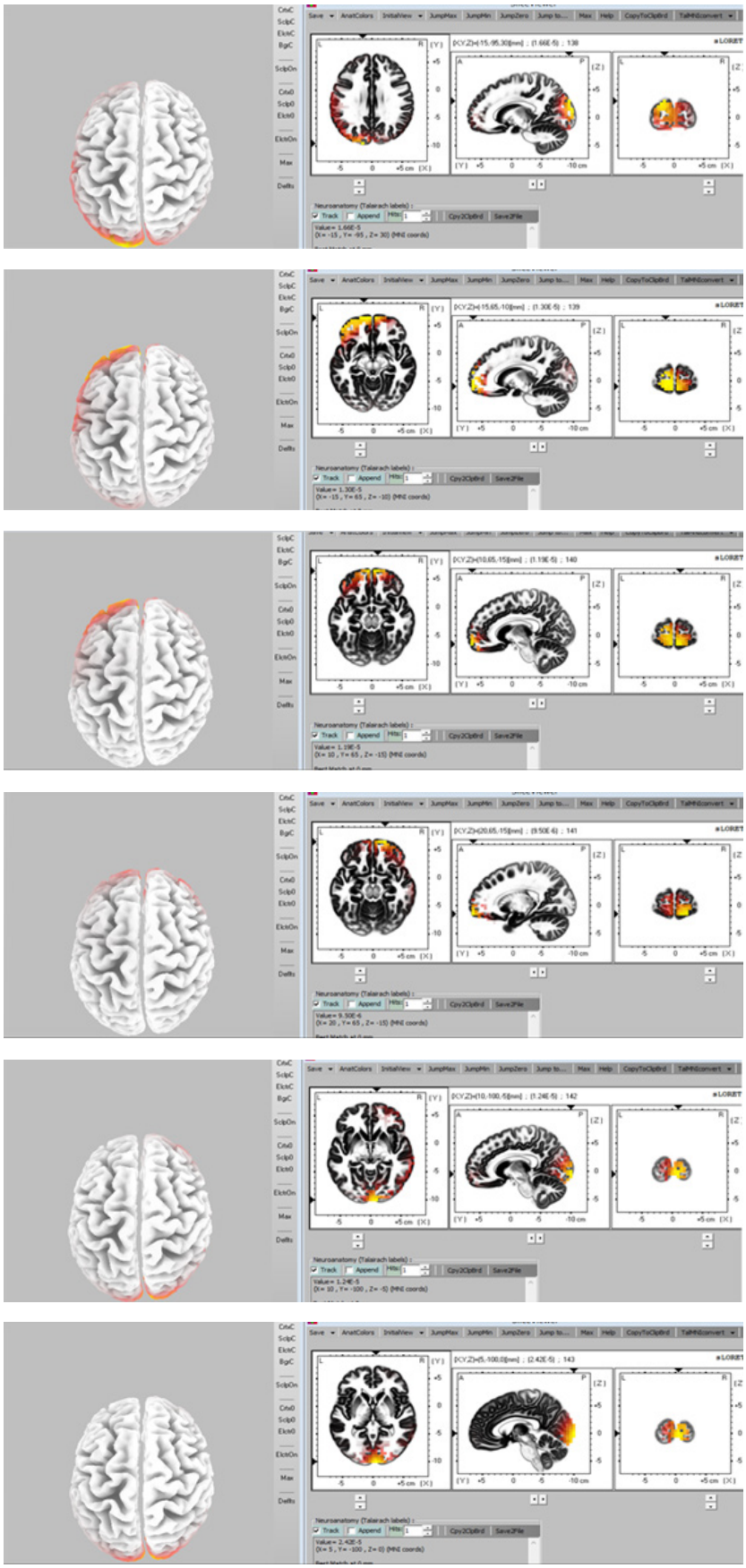
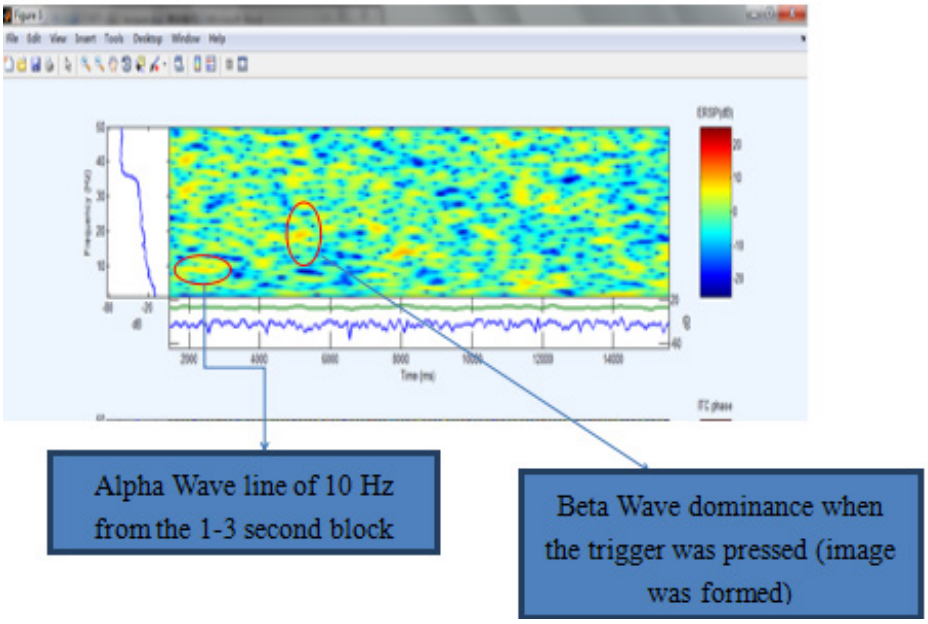
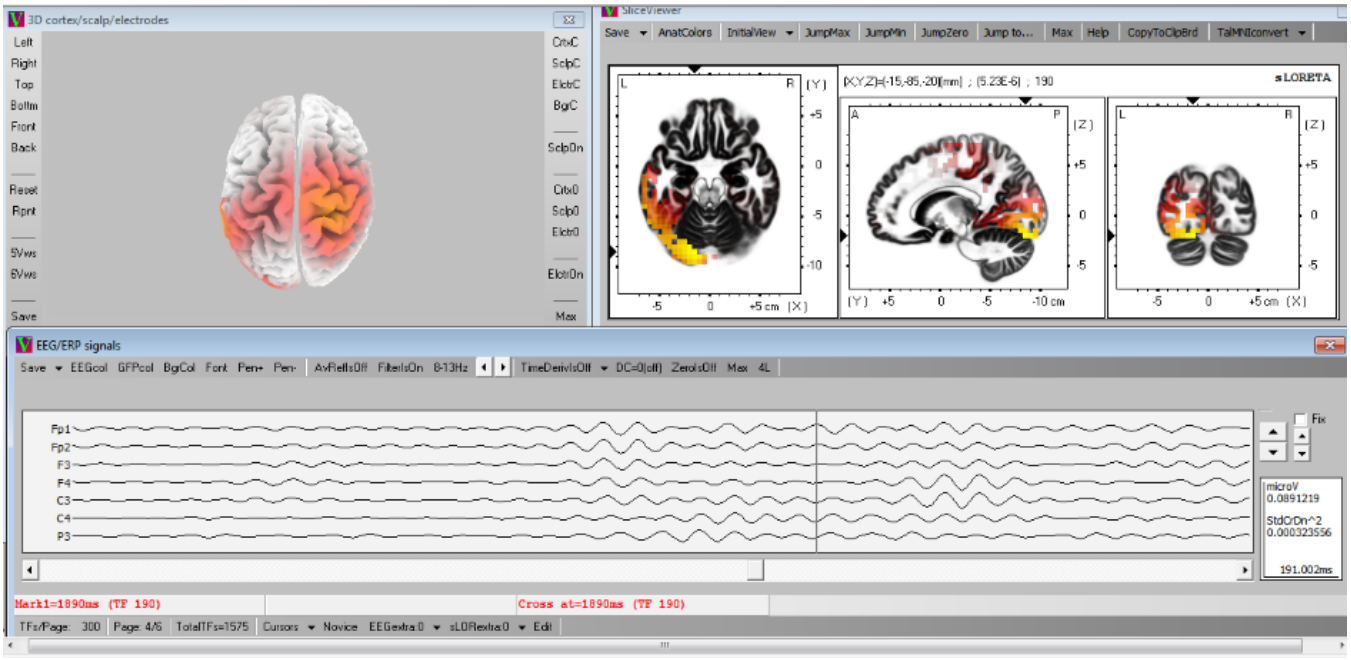


Figure 7: Performance of Beta rhythm activities (From upside to down side, the plots present Beta rhythm activities 80ms, 70ms, 60ms and 50ms before the subject realized the image formation.)





(a)



(b)

which also indicates that the Beta activity's spatial shift's speed is at least 15 to 18 times faster than that of Delta oscillations. This signifies that the activated brain regions during Beta oscillations are relatively higher in interactivity with each other than that of Delta oscillations. The spatial shift of Beta and Delta oscillations observed from these sLORETA plots show back-and-forth tracking arrows between frontal lobe and visual cortex, which can also be regarded as certain communication between these two brain regions, more frequently for beta wave than that for delta wave activity. This gets further asserted by our standard deviation analysis of functional connectivity computed via sliding window approach as the maximum and medium (median) standard deviation is the lowest for beta wave activity among all other waves (Figure 5). Hence, mostly Beta wave activity related neuron oscillations are correlated in the most consistent manner than the other waves. This strongly suggests that beta wave activity is involved with locking neurons together thereby playing a critical utility role for information query.

between its activated brain regions whereas Beta wave activity, with least deviation, represents most consistently correlated activity. Considering the fact that all subjects were asked to generate various kinds of images in their mind, we can draw a conclusion that prefrontal lobe participate in the processing of image formation with visual cortex under both low frequencies (Delta oscillations, 0-4Hz) and Beta oscillations (13-30Hz) as seen from our sLORETA plots.

Cortex maps plotted by sLORETA indicate that Theta oscillations do not show a retention at the frontal lobe within the 300ms before the subjects generate images in their mind. However, a detectable annular movement trace can be observed, as displayed in Figure 8, which is transmitting along the edge of the cortex map from its top view. Considering the sampling rate, this trace with Theta wave was continuous during the processing of image formation unlike the sudden shifts of activated regions as seen with the Delta and Beta oscillations. This circulating movement of activated Theta rhythm's regions would be either clockwise or counter clockwise. The cycle of annular movement varies for different subjects, as well as for different tasks of image formation for the same subject. For most datasets that we recorded from the six subjects, these cycles were between 60ms and 150ms long in duration.

Since all subjects were completing the tasks with their eyes closed while EEG signals were being recorded, Alpha waves dominated and presented a relatively higher intensity than other waves for most of the time, which can be finalized from the time-frequency plots of EEG signal on O1 channel (left side occipital region) in Figure 9(a). For some EEG datasets, the image formation in subjects' mind also excites Mu waves, which share the same frequency range of Alpha rhythm, but are detected from the motor cortex, rather than occipital region. Figure 9(b) is an example of Mu waves on motor cortex. We believe that this domination of alpha wave and mu wave is also the reason why we see its correlation is overall consistent and not highly erratic (Figure 5). Due to the activity of Alpha waves and Mu waves, the mirror neural system must be suppressed by the corresponding executive network. A global view of Alpha rhythm activity during the analyzed 300ms presents an anomalous movement trace. However, compared with Theta and Beta rhythm, the performance characteristics of Alpha rhythm activity regions mostly falls in between them, which means that for some cases, an obvious annular movement trace along the cortex edge from a top view is observed; for the other cases, oscillations of brain activity under Alpha rhythm between frontal lobe and visual cortex are present; or even both phenomenon are detected in the same dataset at different time period.

On an overall level, we observe from connectivity analysis (Figure 5) that channel-pairs with consistent correlation, i.e. less varied correlation or minimum standard deviation in correlation, tend to remain least varied for the entire task for Alpha followed by Beta, Delta and then Theta oscillations. Whereas, for pairs of channels that are highly erratic in phase with each other, i.e. with maximum or medium standard deviation in correlation, tend to be least varied for Beta followed by Alpha, Theta and then Delta oscillations. The precise nature behind this pattern observation remains elusive. It was not clear why Alpha oscillation's pair of nodes remain least varied for minimum standard deviation and not for maximum and median standard deviation group. Here, we quantified task-evoked correlation's variation for each wave and we see that overall, Alpha and Beta oscillations tend to be at the lower end of variation for correlated pairs. We believe this could be for the following reason. Alpha wave is dominating as subject's eyes were closed as seen from frequency plots in Figure 9(b) also and Beta wave appears to be involved with locking neurons and information query as described earlier in Figure 7.

Individual studies of Delta, Theta, Alpha and Beta oscillations during the 300ms period before subjects realized image formation in their mind demonstrate totally different performance of these rhythms. This also indicates that the rhythms provide different functions for the processing of self-contemplating image formation with its corresponding executive network. Self-contemplating image formation involves multiple brain regions' activation, and these activations of brain is dynamic during the period from the formation in the mind to the time when subjects realize this image formation. The higher frequency the rhythm has, the higher the speed of spatial shifts, i.e. movement of activated brain regions, was observed.

Finally, for each EEG dataset, we generated 30 time progressive plots representing the 300ms time to trace the highest brain wave intensity positions. Figures 10 to 12 show accumulated traces of the 30 points in sequence. In Figure 10, the activity of forming an image of someone they love, which produced heavy activity towards the front cortex of the brain and some activity in the visual cortex. This persisted in multiple subjects with the Alpha brain wave. Alpha wave displayed both characteristics of wrapping around the brain, clockwise or counter clockwise motion and persisting activity in visual cortex and frontal lobe. With Figure 11, the subject is completing a simple mathematical calculation. Note that Beta wave moves at a much faster pace, seeing as it is at 13 – 30Hz. Beta displays characteristics of rapid movement between the frontal lobe and the visual cortex. The activity of a mathematical cognition does show a higher concentration in the frontal lobe. And we see a much more even distribution between the left and right brain with Beta wave, however, Alpha becomes more likely to favor the left or right side of the brain. Figure 12 displays a subject's performance while thinking about some part of their body. The Theta wave, instead of confining towards a specific region of the brain, seems to have a wraparound pattern, either clockwise or counterclockwise. Other subjects also show a favoritism of the right side of the brain. Delta wave moved at a much slower pace, as we had anticipated since it is defined between 0 – 4Hz, concentrating a majority of its time by moving between the frontal lobe, further towards the positive Y axis, and the visual cortex, further back towards the negative Y axis. However, it travels at around 45 degree angle to the X and Y axis. The above described brainwave activity characteristics reproducibly happened to every subject under test.

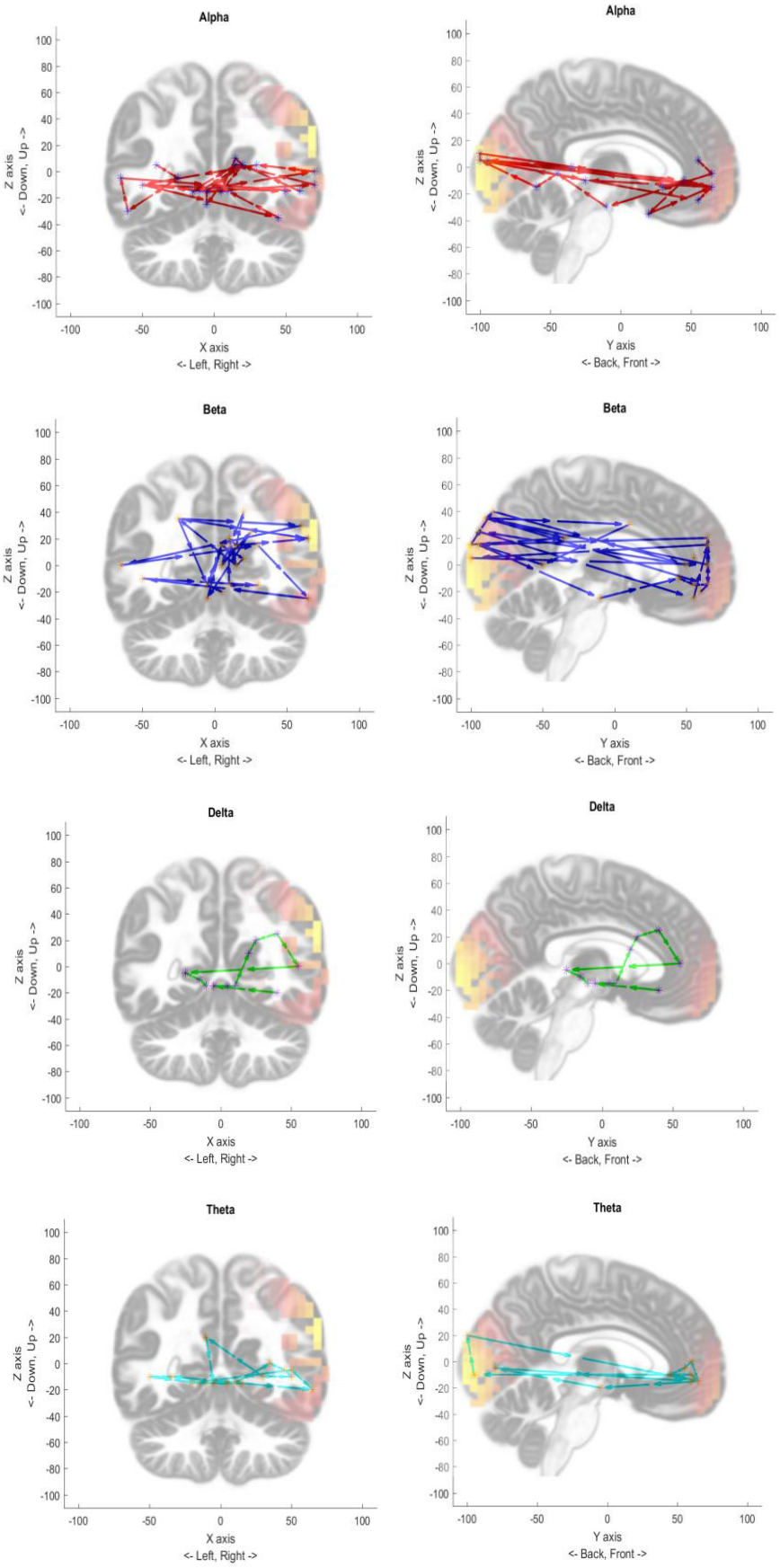


Figure 10: Performance of Alpha, Beta, Delta, and Theta wave across the brain while subject thought about the faces of the people that they loved.

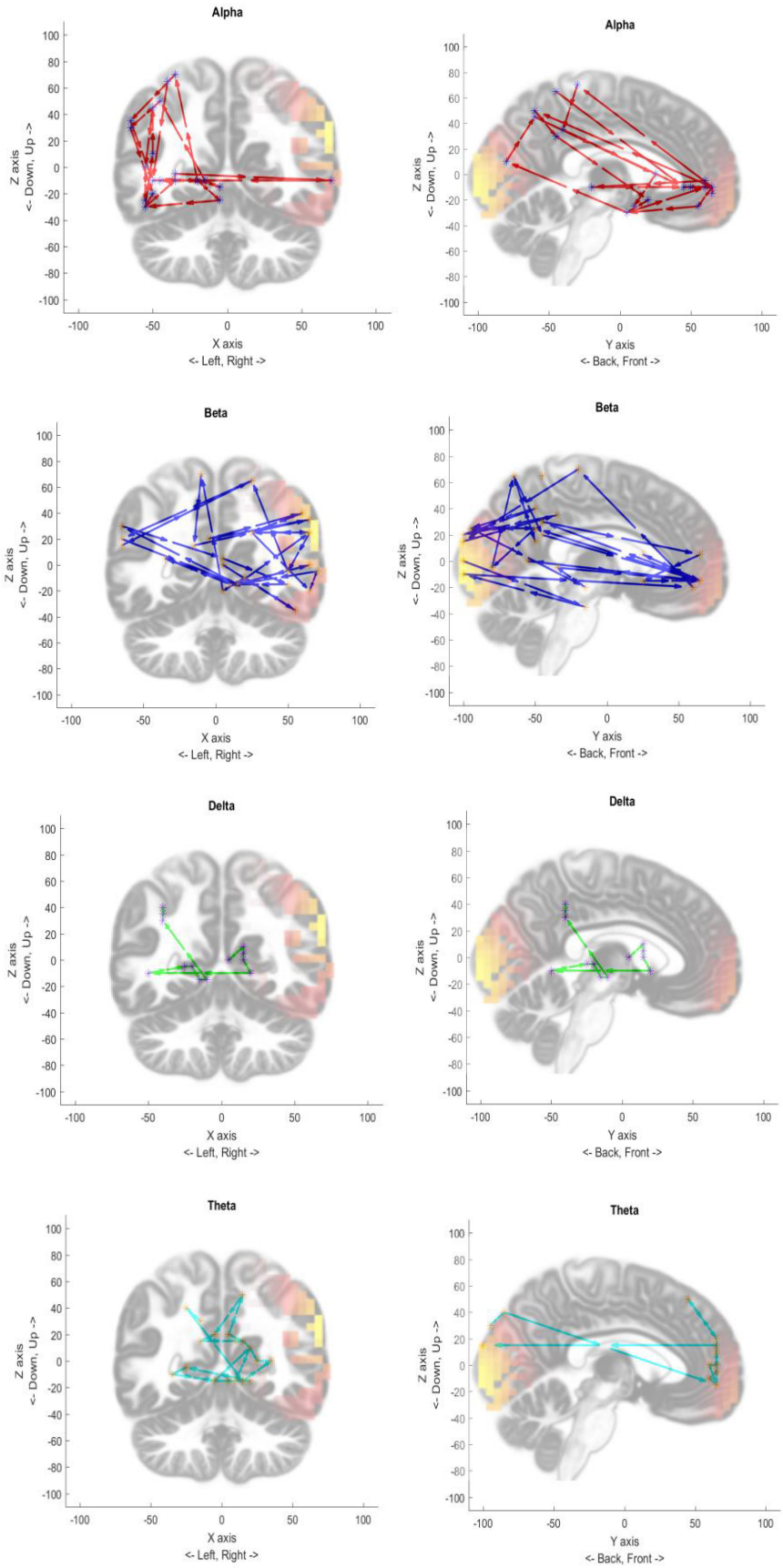


Figure 11: Performance of Alpha, Beta, Delta, and Theta wave across the brain finish some simple mathematical calculation, in this case 15*18.

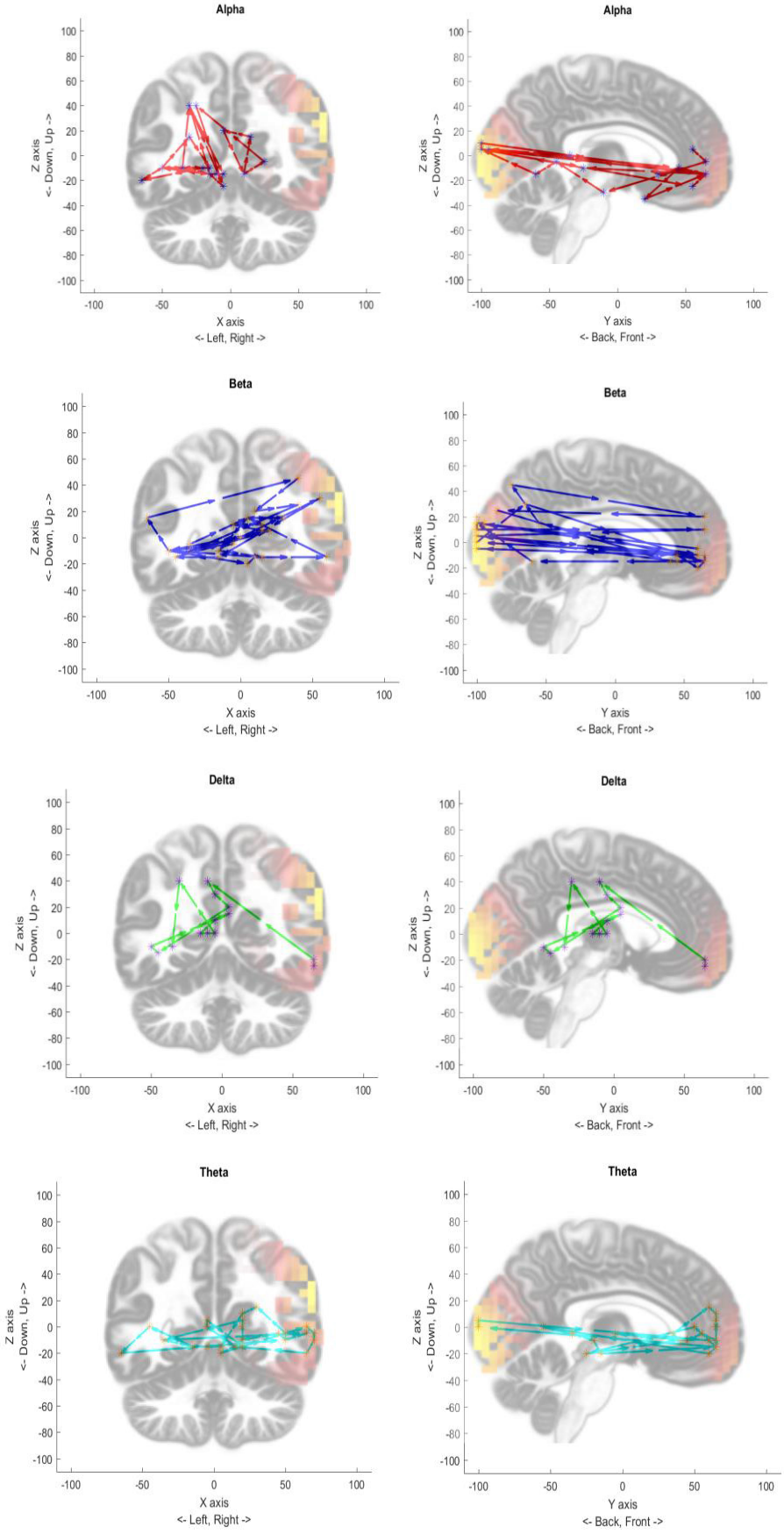


Figure 12: Performance of Alpha, Beta, Delta, and Theta wave across the brain while the subject thought about some parts of their body, in this case the left hand.

4. Conclusions

In this work we proposed a new method to monitor dynamic activities of brain functions at different frequency band without using averaging techniques like even repeats. In our example experiments, to associate subject behaviors with 3-D tracing characteristics, we found that each wave band had its own identifiable behavior that reproducibly generated notable characteristics along the 3 categories including emotion, math calculation, and motor functions. Our data analysis indicates that beta waves play key important roles in linking different brain areas together for information inquiry. Theta and low Alpha are more likely playing the role of information control, integration, and image formation. Conclusions derived from these tracing results were further confirmed with sliding window spatial correlation analysis.

Acknowledgment

The research is supported by the intramural research program of NIH, NIDA, NSF grant ECCS-1631820, NIH grants MH112180, MH108148, MH103222, and a Brain and Behavior Research Foundation grant.

References

[1] A. Wudenhe, Q. Meng, and F. S. Choa, "Three-dimensional eeg signal tracking for reproducible brain activity monitoring". In *Signal Processing in Medicine and Biology Symposium (SPMB), Philadelphia, USA, 2016*.

[2] T. W. Picton, S. Bentin, P. Berg, E. Donchin, S. A. Hillyard, R. Johnson, G. A. Miller, W. Ritter, D. S. Ruchkin, M. D. Rugg and M. J. Taylor, "Guidelines for using human event-related potentials to study cognition: recording standards and publication criteria". *Psychophysiology*, *37*(2), pp.127-152, 2000.

[3] Q. Meng, F. S. Choa, E. Hong, Z. Wang, and M. Islam, " Control channels in the brain and their influence on brain executive functions". In *SPIE Sensing Technology+ Applications* (pp. 910716-910716). International Society for Optics and Photonics, Baltimore, USA, 2014.

[4] T. J. Buschman, E. L. Denovellis, C. Diogo, D. Bullock and E. K. Miller, "Synchronous oscillatory neural ensembles for rules in the prefrontal cortex". *Neuron*, *76*(4), pp.838-846, 2012.

[5] A. Damasio, *Self comes to mind: Constructing the conscious brain*, Vintage, 2010.

[6] R. D. Pascual-Marqui, "Standardized low-resolution brain electromagnetic tomography (sLORETA): technical details". *Methods Find Exp Clin Pharmacol*, *24*(Suppl D), pp.5-12, 2002.

[7] J. Bosch-Bayard, P. Valdes-Sosa, T. Virues-Alba, E. Aubert-Vazquez, E. R. John, T. Harmony, J. Riera-Diaz and N. Trujillo-Barreto, "3D statistical parametric mapping of EEG source spectra by means of variable resolution electromagnetic tomography (VARETA)". *Clinical Electroencephalography*, 32(2), pp.47-61, 2001.

[8] W. Miltner, C. Braun, R. Johnson, G. V. Simpson, and D. S. Ruchkin, D.S, "A test of brain electrical source analysis (BESA): a simulation study". *Electroencephalography and clinical neurophysiology*, *91*(4), pp.295-310, 1994.

[9] R. D. Pascual-Marqui, "Review of methods for solving the EEG inverse problem". *International journal of bioelectromagnetism*, *1*(1), pp.75-86, 1999.

[10] M. J. Herrmann, J. Römmler, A. C. Ehlis, A. Heidrich, and A. J. Fallgatter, "Source localization (LORETA) of the error-related-negativity (ERN/Ne) and positivity (Pe)". *Cognitive brain research*, *20*(2), pp.294-299, 2004.

[11] S. Mientus, J. Gallinat, Y. Wuebben, R. D. Pascual-Marqui, C. Mulert, K. Frick, H. Dorn, W. M. Herrmann and G. Winterer, "Cortical hypoactivation during resting EEG in schizophrenics but not in depressives and schizotypal subjects as revealed by low resolution electromagnetic tomography (LORETA)". *Psychiatry Research: Neuroimaging*, *116*(1), pp.95-111, 2002.

[12] B. Saletu, P. Anderer and G. M. Saletu-Zyhlarz, "EEG topography and tomography (LORETA) in diagnosis and pharmacotherapy of depression". *Clinical EEG and neuroscience*, *41*(4), pp.203-210, 2010.

[13] R. W. Thatcher, D. M. North and C. J. Biver, "Diffusion spectral imaging modules correlate with EEG LORETA neuroimaging modules". *Human brain mapping*, *33*(5), pp.1062-1075, 2012.

[14] T. Rosburg, P. Trautner, C. E. Elger and M. Kurthen, "Attention effects on sensory gating—intracranial and scalp recordings". *Neuroimage*, *48*(3), pp.554-563, 2009.

[15] R. Agarwal, J. Gotman, D. Flanagan and B. Rosenblatt, "Automatic EEG analysis during long-term monitoring in the ICU". *Electroencephalography and clinical Neurophysiology*, *107*(1), pp.44-58, 1998.

[16] V. Krajča, S. Petránek, I. Patáková and A. Värri, "Automatic identification of significant graphoelements in multichannel EEG recordings by adaptive segmentation and fuzzy clustering". *International journal of bio-medical computing*, *28*(1-2), pp.71-89, 1991.

[17] D. Michael and J. Houchin, "Automatic EEG analysis: a segmentation procedure based on the autocorrelation function". *Electroencephalography and clinical neurophysiology*, *46*(2), pp.232-235, 1979.

[18] E. A. Allen, E. Damaraju, S. M. Plis, E. B. Erhardt, T. Eichele and V. D. Calhoun, "Tracking whole-brain connectivity dynamics in the resting state". *Cerebral cortex*, *24*(3), pp.663-676, 2014.

[19] R. M. Hutchison, T. Womelsdorf, E. A. Allen, P. A. Bandettini, V. D. Calhoun, M. Corbetta, S. Della Penna, J. H. Duyn, G. H. Glover, J. Gonzalez-Castillo and D. A. Handwerker, "Dynamic functional connectivity: promise, issues, and interpretations". *Neuroimage*, *80*, pp.360-378, 2013.

[20] V. Gallese, C. Keysers and G. Rizzolatti, "A unifying view of the basis of social cognition". *Trends in cognitive sciences*, *8*(9), pp.396-403, 2004.

[21] M. Lotze and U. Halsband, "Motor imagery". *Journal of Physiology-Paris*, *99*(4), pp.386-395, 2006.

[22] J. A. Pineda, "The functional significance of mu rhythms: translating "seeing" and "hearing" into "doing"". *Brain Research Reviews*, *50*(1), pp.57-68, 2005.

[23] G. Rizzolatti and L. Craighero, "The mirror-neuron system". *Annu. Rev. Neurosci.*, *27*, pp.169-192, 2004.

[24] N. Uchida, A. Kepecs and Z. F. Mainen, "Seeing at a glance, smelling in a whiff: rapid forms of perceptual decision making". *Nature reviews. Neuroscience*, *7*(6), p.485, 2006.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1647-1653 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Optimizing the front-end power solution for automotive ADAS systems

Beniamin Dragoi*

Valeo Vision System, Tuam Galway, Ireland

ARTICLE INFO

Article history:
Received: 31 May, 2017
Accepted: 14 August, 2017
Online: 24 August, 2017

Keywords:
Front-end DC-DC converter
ADAS automotive systems step-up step-down converter Non-inverting buck-boost Sepic Zeta

ABSTRACT

In this paper the front-end power solution for automotive ADAS electronic systems is analyzed. The increased complexity of electronic systems in ADAS automotive brings new challenges for hardware design, especially with regard to improving the efficiency of the switched-mode power supply while maintaining the overall cost reasonably low and keeping the occupied area on the PCB as small as possible. Four topologies of Non-inverting step-up/step-down converter are analyzed. A real-life case is considered and 4 DC-DC converters are designed and simulated to fulfill the requirements. Finally a comparison between the circuits is undertaken. Results are analyzed from the point of view of efficiency, area and price and a decision is reached for the most optimized converter.

1. Introduction

This paper is an extension of work originally presented in 2016 at the 12th IEEE International Symposium on Electronics and Telecommunications (ISETC) [1].

Advanced driver assistance systems (ADAS) are probably the fastest-growing electronic systems in the automotive industry. Every year more and more features are added to the cars to make them safer and to help drivers. A brief list of the kind of feature available today is as follows: pedestrian detection, traffic sign recognition, blind spot monitor, lane departure warning, forward collision warning, driver monitor, adaptive cruise control and emergency braking control, and the list can continue. This extraordinary increase in available ADAS brings to reality something that once was just a dream or a nice theme for Sci-Fi movies: autonomous car. Researches and trials for self-driving cars have a long history of about 100 years. But only in the last 10 years have notable results been achieved. ADAS are basically the foundation of the self-driving vehicle. Today there are already cars on the road with self-driving features with the most advanced publicly available systems being at Level 2 of autonomy (based on SAE Society of Automotive Engineers International's J3016 document definition [2]). Level 3 cars are expected to be on the road in 2018.

To support all these features for ADAS and autonomous driving, electronic systems in cars are very complex. To be able to detect the environment many sensors of different types must be available and the information collected must be processed extremely fast. Systems for ADAS and autonomous driving utilize today the following: cameras (4-12), ultrasonic sensors (12-20), radar (short and long range) and LiDAR. Each of these sensors has its own specific capability and by combining the information from them the system is more capable of reading the world around the car.

However the next challenge is to process the huge amount of data received from all these sensors. In a sense self-driving cars will be extremely powerful computers on wheels.

Extremely powerful multi-core processors like Renesas R-Car, Nvidia Parker, EyeQ4 from Mobyleye, etc. are available. Electronic control units (ECU) that can handle and process such large amounts of data are usually multi-processor systems with 1-4 processors and 1 – 2 microcontrollers (MCUs). Such ECU contains also all the required digital interfaces and decoding blocks necessary to capture data from external sensors. Power consumption and power dissipation for these ECUs is important and becomes critical in the automotive context. High requirements for robustness and reliability combined with the need for a long life (10 years) usually demands passive cooling (i.e. with no fans) for the vehicle's electronics. This is not easy to achieve when the ambient temperature can reach 85°C. These are some of the reasons that compel hardware design engineers to focus on reducing the power consumption and power dissipation of its devices. Besides selecting modern ICs (advanced technology) one can reduce the dissipated power by increasing the efficiency of local power supplies.

The following pages cover the topic of selecting and designing the optimised front-end power supply for ADAS electronics and examine several crucial aspects for automotive use:

- Efficiency, with the overall goal of reducing power loss;

*Corresponding Author Name: Beniamin Dragoi, Valeo Vision System, Tuam Galway, Ireland | Email: beniamin.dragoi@valeo.com

- Board occupied area, considering that the PCB is one of the most expensive parts in an ECU;
- Price reduction, which is the goal when the technical and quality aspects are fulfilled.

2. Power supplies in automotive

Most of the cars in the world today use 12-volt automotive electrical systems. Typically a car battery operates with a voltage in the range of 9V to 16V. Therefore, all electronic circuits powered by the battery power line should be able to account for this input-voltage variation. For modern electronics with most ICs powered at 3.3V or below (2.5V, 1.8V, 1.2V, 1.1V and lower) and only a few ICs that still require 5V or higher (CAN, Flexray, LIN transceivers), this input voltage range (9-16V) seems to be perfect. However, if for the battery the normal voltage is 12V (engine off) and 14.4V (engine on), this also can fall to 7V (warm / hot start) or as low as 3.2V (during cold start) or increase to as high as 34V (load dump).

Figure 1 and Figure 2 show typical warm and cold start profiles, while Figure 3 presents the typical load dump [3]

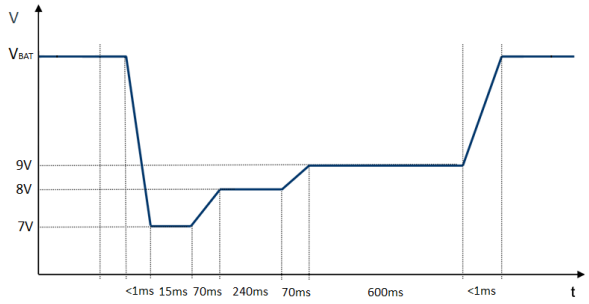


Figure 1. Example of a warm start profile

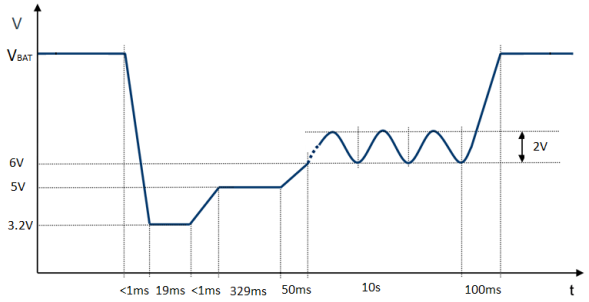


Figure 2. Example of a cold start profile

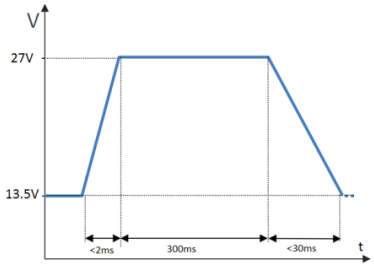


Figure 3. Example of a load dump profile

Voltage levels and timings presented in these figures are typical but they can vary from one car manufacturer to another. They are also dependent on the location of the ECU in the car. Standard profiles for these events are described by ISO 7637-2.

Some electronic systems in the car (engine ECU, safety systems, navigation, infotainment, dashboards, etc) must be live during a cold-crank event and also during a warm start with the requirement to be in Functional Status A: *the system must fulfill all functions during and after exposure to the event.*

In conclusion many electronic circuits for automotive must be designed to be powered from batteries with a voltage range of 3V – 34V. This wide voltage range is challenging from at least two aspects:

- Minimum input voltage is around 3V and is below some very common voltage rails needed for ICs like 3.3V and 5V. This leads to the need for a way to boost the input voltage to the required levels.
- Maximum input voltage is much higher than the usual rails for the processor's cores (1V, 0.8V or even lower). It is difficult to directly step down a high voltage like 34V, to a level like 0.8V because the duty cycle is very small. The minimum duty-cycle is limited by the converter's minimum ON time and its switching frequency. For such a low duty cycle it is highly likely that the step-down converter enters into "pulse-skipping" mode with the risk of increased output voltage ripple and a shift in frequency. Both effects might be unacceptable and result in a high voltage noise that is too high and the risk of failing EMC (electromagnetic) emissions.

To solve both above-mentioned aspects one can use the power supply topology presented in Figure 4. Below can be found a typical block diagram for a power supply system based on an intermediate voltage rail. This is probably the most used power supply topology for complex ECUs in automotive.

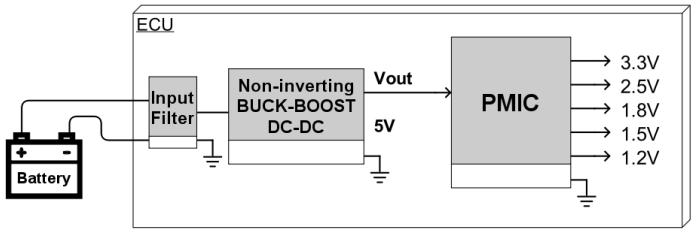


Figure 4. Typical block diagram for an automotive power supply

For this case where the intermediate regulated power rail is $V_{OUT}=5V$, if the cold start profile goes below 5V, a front-end step-up/step-down DC-DC is required. In Figure 4 this converter is mentioned as being Non-inverting BUCK-BOOST because the input voltage range can be below or above the output voltage (5V) and the output voltage has the same polarity as the input one. 5V as output of the front end power supply is not a magic number, but it is a suitable value for at least 3 reasons:

- A 5V rail is still needed in the automotive electronics for circuits like CAN, and Flexray;
- Most of the available power management ICs (PMIC) have a maximum input voltage of 5.5V;
- A lower value for the output of the front-end supply brings the converter into the zone of very low duty cycles and the associated risks and issues.

3. Non-inverting Step-up/Step-down converter topologies

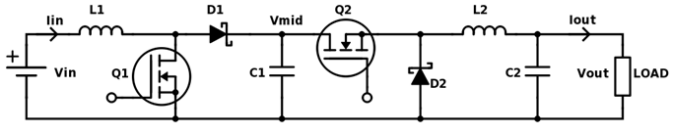
In the literature, there are few DC-DC converter topologies that can produce the *non-inverting step-up/step-down* function [4].

Some of them, such as BOOST-BUCK and SEPIC are well-known, whereas others such as 4-switch BUCK-BOOST and ZETA are not so popular. One solution is to use two-stage conversion: one BOOST is followed by one BUCK (BOOST-BUCK). Another approach is the single-stage buck-boost converters such as: 4-switch BUCK-BOOST, SEPIC or ZETA.

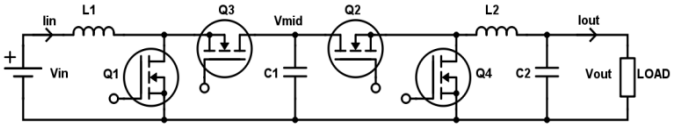
3.1. BOOST-BUCK converter

Figure 5 shows the *Boost-Buck* converter with two implementations: a) asynchronous and b) synchronous. This converter is basically a cascade of two converters, a BOOST formed by L1, Q1, Q3 (or D1) and C1 followed by a BUCK made by Q2, Q4 (or D2), L2 and C2. The purpose of the Boost converter is to provide a minimum intermediate voltage V_{mid} during cold or warm start battery voltage sags. V_{mid} is set to provide enough input voltage for the Buck converter. As long as $V_{in} > V_{mid}$ the Boost converter is disabled and the Buck converter is powered through L1 and Q3 (or D1). For this case Q3 or D1 are ON and Q1=OFF. During this phase V_{mid} is not regulated and is: $V_{mid}=V_{in}-V(L1)-V_{DS}(Q3)$ for the synchronous converter or $V_{min}=V_{in}-V(L1)-V_F(D1)$ for asynchronous one.

When $V_{in} \leq V_{mid}$ the boost converter kicks in and maintains V_{mid} as a regulated voltage.



a) Asynchronous

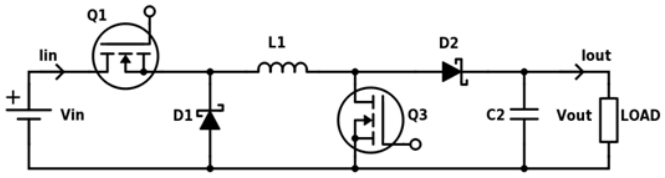


b) Synchronous

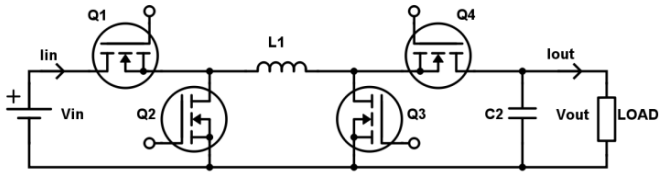
Figure 5. BOOST-BUCK converter

3.2. 4 switch BUCK-BOOST converter

Another topology capable of producing a positive regulated output voltage for V_{in} below or above V_{out} is the synchronous *4-switch Buck-Boost* – Figure 6 a) and b).



a) Asynchronous



b) Synchronous

Figure 6. 4-Switch BUCK-BOOST converter

Basically, this is a simplification of one BUCK converter followed by one BOOST converter; both converters share the same inductor L1 [6]. There are 3 modes of operation based on the relationship between V_{in} and V_{out}:

1. V_{in} much higher than V_{out} → Buck mode

Q3 is constant OFF and Q4 (or D2) is constant ON. This is the preferred mode of operation because the efficiency is good. This is basically a buck converter's efficiency reduced by the Q4 loss in constant conduction ($I_{out}*R_{ds(on)}$). However, for the asynchronous version, Q4 is replaced by a diode D2 and there could be important power loss due to the forward voltage of the diode being continuously ON ($I_{out}*V_{F(D2)}$). If the output voltage V_{out} selected is low enough (e.g. 5V), this mode of operation is the mode of operation for most of the time, normal operation, warm start and load dump. So it is important to minimize the power loss produced by the inactive Boost section.

2. V_{in} much lower than V_{out} → Boost mode

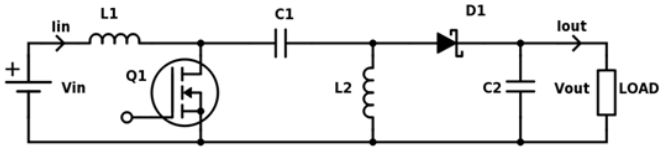
Q1 is constant ON and Q2 (or D1) is constant OFF. If V_{out}=5V, according to Figure 2, the converter will work in this mode only for about 20ms (lowest section of the cold start profile).

3. V_{in} close to V_{out} → Buck-boost mode.

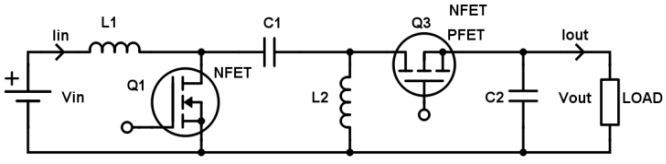
In this mode all 4 transistors are controlled by the driving controller to ensure a proper transition from Buck to Boost or vice-versa. The converter will work in this mode only for few hundreds of milliseconds, when the battery is recovering from the cold crank.

3.3. SEPIC converter

The most common single-stage converter with non-inverting buck-boost behaviour is SEPIC. Asynchronous and synchronous versions are shown in Figure 7. This topology uses only two switches (transistor of diode) [7]. The most known SEPIC configuration is the asynchronous one but here the synchronous SEPIC converter is considered as well [8]. The synchronous SEPIC convertor can be implemented with Q3 NFET (N-channel Field Effect Transistor) of PFET (P-channel Field Effect Transistor) depending on the available driving controller. Overall efficiency of the converter is reduced because two inductors are used in operation and due to high currents and voltages for the switch elements.



a) Asynchronous

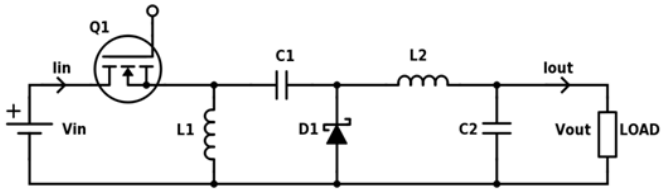


b) Synchronous

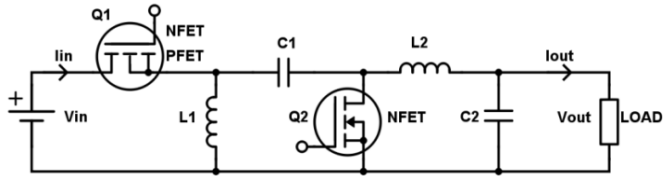
Figure 7. SEPIC converter

3.4. ZETA converter

Zeta is a very similar topology to SEPIC; it is named "inverse SEPIC" and has the topology shown in Figure 8. It has two inductors and two active switches [4]. One of the active switches can be a diode (D1) or a FET (Q2). Like SEPIC the two inductors can be coupled or independent.



a) Asynchronous



b) Synchronous

Figure 8. ZETA converter

ZETA topology is better for lower output voltage ripple due to continuous output current in L2. This reduces the required output capacitance to maintain a low output voltage ripple. However due to the input discontinuous current (through Q1) ZETA shows higher input voltage ripple. ZETA convertor can be implemented with Q1 NFET of PFET, depending on the available controller.

3.5. Topology comparison

As previously shown, all 4 types of non-inverting step-up/step-down converters considered here have their asynchronous and synchronous variants. To avail of increased efficiency this work considers only the synchronous type of converters here. This one reduces the power loss associated with diodes in conductions and proves to have higher efficiency for higher current load [5].

Some advantages and disadvantages have been already mentioned but a more detailed comparison is undertaken in Table 1. This table contains two sections, one for advantages for each converter and one for disadvantages [9, 10].

Each type of SMPS presented here has its own advantages and disadvantages, so there is no unique answer to the question: '*what non-inverting step-up/step-down is the best?'* The final decision in selecting one or another topology belongs to the design engineer based on his application's requirements. The next section contains a comparison for all these 4 types of converter for a real life case, comparisons that helps the designer to decide what the most optimized converter for an application is.

4. Case study and simulations

A real life case for a front-end power supply was selected to show the differences between these 4 converters. Design requirements are listed in Table 2.

To be able to compare the board occupied area and price for all converters it was decided to use real components, available on the market, components that are automotive qualified. For the same reasons it was decided to use controllers from a single supplier - Linear Technology, now part of Analog Devices [12]. All four designs were simulated using the same Spice simulator – Ltspice.

Table 1. Comparison of advantages and disadvantages for 4 non-inverting step-up / step-down converters

BOOST-BUCK	4 switch BUCK-BOOST	SEPIC	ZETA
Advantages			
Continuous, non-pulsating input and output currents due to input and output inductors (L1 and L2). This reduces ripple voltage (input and output) and improves the EMI noise, and can simplify the input filter design and size.	Single inductor design. Lower price and smaller board occupied area.	Only two active switching components are required	Only two active switching components are required
Easy to design and compensate. Both converters, Buck and Boost are mature circuits and well known, easy to design and easy to compensate for stability	Switch rating = V_{in} for input section, V_{out} for output section => Mosfet transistors with lower V_{DS} can be used, so with better (lower) $R_{ds(on)}$	Low input voltage ripple	Low output voltage ripple
Can be designed using two completely separated controllers / converters	High efficiency	Inherent short-circuit protection at output due to DC-block capacitor	Inherent short-circuit protection at output due to DC-block capacitor
Disadvantages			
Two inductors are required: price and area. Inductors are expensive and big components.	Input and output are both noisier due to discontinuous currents *Iin* and *Iout*	High output ripple (similar to Boost)	High input voltage ripple (similar to Buck)
4 switches are needed - price	4 switches are needed	Two inductors are required	Two inductors are required
Low efficiency – during normal Vbat range converter works in Buck mode but it has a power loss in L1 and Q3	The controller for the entire converter is more complex and contains two driver sections	Coupling capacitor C1 is needed. In automotive use two series capacitors might be need to protect against MLCC failure (failure mode is *short)*	Coupling capacitor is needed but the voltage across it is lower (V_{out}). Series capacitors might not be needed
Controller needed is complex (two driver sections or two controllers)	Special consideration for crossing point between *Buck-mode* and *Buck-mode* is needed.	Frequency compensation is more difficult to be implemented	Frequency compensation is more difficult to be implemented. Zeta is not well understood.

Table 2. Design Requirements for the front-end SMPS

V_{IN}	4-35V
f_{sw}	400KHz
V_{OUT_ripple}	<50mV
V_{OUT}	5V
I_{OUT}	5A

The switching frequency is set at 400 kHz as a trade-off between several factors: *efficiency, area, cost* and *EMC risks.*

It is known that lower switching frequency improves efficiency due to lower switching losses, but also increases the occupied area (expensive) [11]. Big and expensive inductors and big and expensive capacitors are needed to reduce the input and output voltage ripple. So, higher efficiency at low frequency comes with higher cost [12]. On the other hand, these DC-DC converters must comply with very strict requirements in terms of conducted and radiated EMC emissions. It is very important to avoid generating noise in the radio AM band (535 kHz to 1705 kHz) so the switching frequency should be below or above AM band. Switching frequencies above AM band suffer from low efficiency and might force the controller into pulse-skipping mode during load-dump. So, the switching frequency is selected just below AM band.

All four designs use as a starting point the application notes available from the supplier. The target is to design a stable power supply capable of fulfilling the requirements with the minimum of components. However, to have a fair comparison passive components from the same family/series are used and where possible even the same components are used. This approach helps to obtain a fair comparison in terms of occupied area and price. In Table 3 a BOM comparison is shown as a result of the design of these for SMPS.

Next controllers are used in these designs:

- Boost-Buck - LTC7812
- 4-switch Buck-Boost - LTC3789
- SEPIC - LT8710. Un-coupled inductors are used.
- ZETA - LT3840 Un-coupled inductors are used.

Table 3. BOM comparison

	BOOST-BUCK	**4-switch BUCK-BOOST**	**SEPIC**	**ZETA**
Controller	Dual: Boost + Buck	4 Switch Buck-Boost	BOOST	BUCK
L1	2.2uH / 14A	2.2uH / 14A	2.2uH /14A	2.2uH /14A
L2	3.3uH / 8A	-	3.3uH / 8A	3.3uH / 8A
Q1	NFET – 40V, 49A, 9.3mΩ		NFET – 40V, 49A, 9.3mΩ	
Q2			-	NFET – 40V 49A, 9.3mΩ
Q3			PFET, 60V 64A, 14mΩ	-
Q4			-	-
C_{IN} – cer.	1x2x 10uF 50V	1x2x 10uF 50V	1x2x 10uF 50V	1x2x 10uF 50V
C_{IN} – Bulk	150uF / 35V	150uF / 35V	47uF / 35V	150uF / 35V
C_{Mid} – cer.	3x 2x 10uF 50V	-	-	-
C_{Min} – Bulk	270uF / 35V	-	-	-
C_{OUT} – cer.	2x 22uF / 10V	3x 22uF / 10V	4x 22uF /10V	1x 22uF /10V
C_{OUT} –Bulk	220uF / 6.3V	220uF / 6.3V	220uF / 6.3V	220uF / 6.3V
Coupling Cap- Cer. (Include RC damping)	-	-	2x 2x 33uF/ 35V	4x 22uF / 10V
Output Ferrite Bead	-	Z=60Ω @ 100MHz	Z=60Ω @ 100MHz	-

4.1. Simulation results

Due to lack of space only the schematics for *4-switch Buck-Boost* converter (Figure 9) and *ZETA* converter (Figure 10) are provided here.

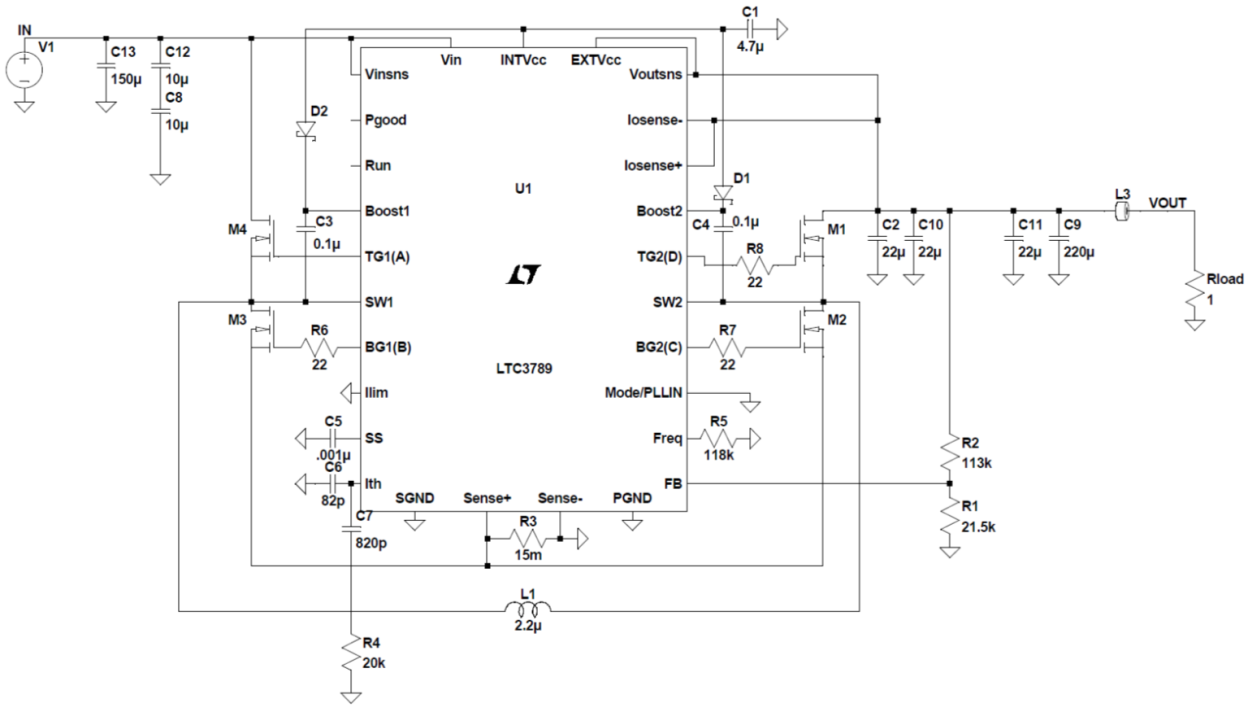


Figure 9. 4-switch Buck-Boost converter schematic

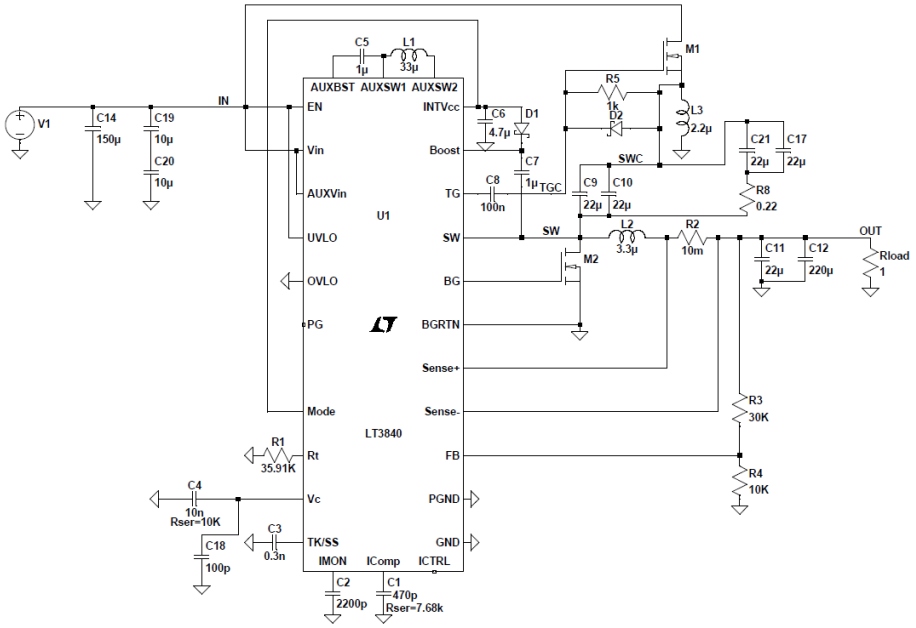


Figure 10. ZETA converter schematic

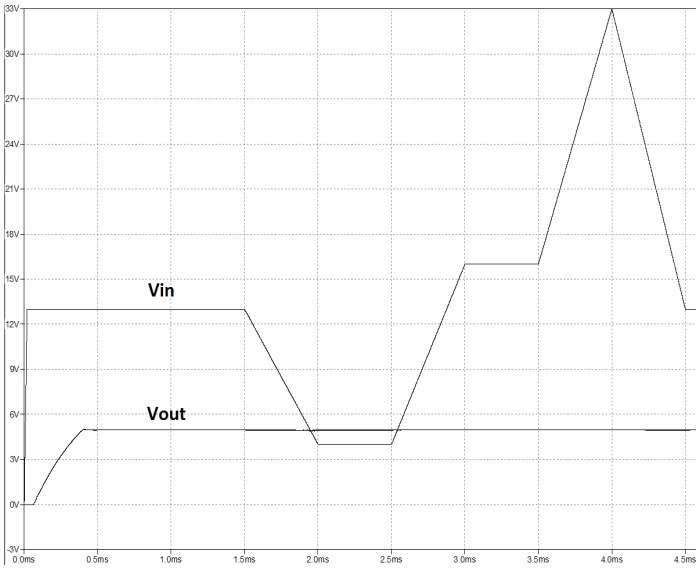


Figure 11. Simulations results for *4-switch Buck-Boost*

Figure 11 shows the transient simulation result intended to verify the behavior of the converter when the input voltage V_{in} varies from 0V to 33V. The response to Cold-crank and Load-Dump events is verified in the same simulation. To speed up the simulation, all the timings from these two profiles are compressed and reduced. However, the converter reacts very well to these faster transitions too.

All simulation results for all 4 considered converters are virtually identical with those shown in Figure 11.

Based on the simulation performed and their results it can be concluded that all 4 designed converters are suitable to be used for the case being considered. All converters can handle the automotive cold-crank, warm-start and load-dump events and provide a regulated output voltage.

4.2. Comparison for efficiency, are and cost

Because all 4 converters are suitable to be used as a front-end power supply for ADAS ECUs (from electrical point of view), one can optimize the power solution for automotive applications by using several criteria:

- *Efficiency* comparison is shown in Figure 12. Ideally, one would like to select the converter that has the best efficiency over the entire input voltage range for a given output current and system requirements; however, the converter stays a brief time in cold-crank so the efficiency in this phase might be not so important for some applications. The left side of Figure 12 shows the efficiency comparison during cold-crank profile for a weak battery while the right side shows the efficiency for each converter when the battery is healthy.

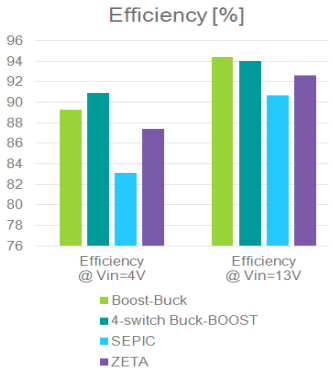


Figure 12. Efficiency comparison

- *Occupied area, with and without controller* – Figure 13. The entire PCB occupied area by a converter is important and it is shown in the left section of Figure 13. Controller size / package might vary from one supplier to another while the external components remain almost the same so a comparison of the space requirement without the controller is presented in the right side of Figure 13. For comparison of the occupied area the sum of areas of all components was considered plus an additional 0.5mm keep-out area on each side for each component;

- *Price with and without controller* - Figure 14. Prices used in the comparison are obtained from the supplier website (Linear Tech) for SMPS controller and all the other parts are from component distributor Mouser. Prices are expressed in € and are for 1k pieces.

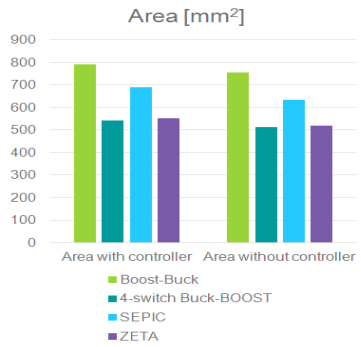


Figure 13. Occupied area comparison

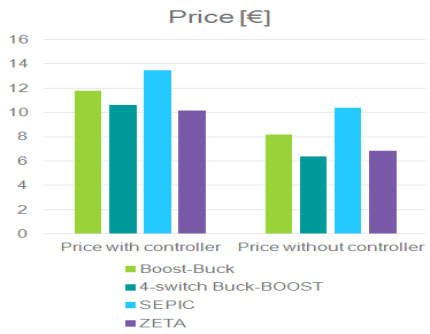


Figure 14. Price comparison

A summary of these comparisons is available in Table 4.

Table 4. Summary of comparisons

	BOOST-BUCK V_{mid} =7V	4-switch BUCK-BOOST	SEPIC	ZETA
Efficiency @ V_{in}=4V	89.3%	90.9%	83.1%	87.4%
Efficiency @ V_{in}=13V	94.4 %	94.0%	90.7%	92.6%
V_{OUT_ripple} [mV_{pkpk}]	22.4	42.6	46.6	27.6
Area [mm²] with controller	791	541	689	553
Price [€] with controller	11.8	10.6	13.5	10.17
Area [mm²] without controller	755	511	633	518
Price [€] without controller	8.2	6.36	10.4	6.83

5. Conclusions

Four synchronous converters suitable to perform the non-inverting step-up/step-down function for powering an ADAS electronic system are considered in this work. A short theoretical analysis is presented and a list of advantages and disadvantages of all topologies is undertaken and summarized, from the perspective of automotive applications. One must know all the advantages and disadvantages of all suitable DC-DC converters before being able to obtain an optimized solution. A method of optimization for automotive ADAS front-end power supply is to consider the efficiency, board occupied area and cost along with the quality parameters of an SMPS (not studied in this work) (as input and output voltage ripple, loop response, stability and transient response).

A real life case was considered and according to the simulation results obtained and the comparisons made for efficiency, occupied area and cost, it can be concluded that the *4-switch Buck-Boost* converter is the optimal one for the real-life case which was considered. This converter meets all the requirements specified in Table 2 and has the smallest occupied area (calculated with or without controller), has a low price and exhibits high efficiency. The SEPIC converter (with un-coupled inductors) proves to be the worst solution: lowest efficiency, highest cost (due to high voltage coupling capacitors) and almost the biggest PCB occupied area. An interesting solution is ZETA that is very competitive in terms of price and space. Also ZETA shows decent efficiency. However precautions must be taken to properly compensate it in frequency.

Future work is planned to analyze the suitability and performance of new/improved converter topology as multi-phase converters (buck-boost, SEPIC) and multi-switch SEPIC/ZETA [13].

Conflict of Interest

The author declares no conflict of interest.

References

[1] B. Dragoi, "On selecting a front-end DC-DC converter for automotive applications", Electronics and Telecommunications (ISETC), 2016 12th IEEE International Symposium on, 27-28 Oct. 2016

[2] Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles, SAE international, http://standards.sae.org/j3016_201609/

[3] International Standard, "Road vehicles — Electrical disturbances from conduction and coupling. Part 2", ISO 7637-2, 2004.

[4] M. Wens, M. Steyaert, "Design and Implementation of Fully-Integrated Inductive DC-DC Converters in Standard CMOS", ISBN 978-94-007-1435-9, pp.90-99, 2011.

[5] R. Nowakowski, N. Tang, "Efficiency of synchronous versus nonsynchronous buck converters", Texas Instrument, Analog Applications Journal, Fouth Quarter, 2009, pp.15-18.

[6] R.W. Erickson, D. Maksimovic, "Fundamentals of Power Electronics", Second Edition, Kluwer Academic Publisher, eBook ISBN: 0-306-48048-4, 2004, pp.134-137.

[7] D. Zhang, "Designing A SEPIC Converter", Texas Instruments, AN-1484, SNVA168D, April 2008.

[8] Shashikumari J. and Ramya N, "Design of Synchronous SEPIC and Synchronous Zeta Converter for Stand-Alone Photovoltaic System" IJRSI, vol. III, Issue V, May 2016, pp. 529-551, ISSN 2321 – 2705

[9] J. Betten. "Synchronous Zeta Converter Outperforms The SEPIC", How2Power Today, Issue May 2014

[10] V. Choudhary, "Selecting the right buck-boost converter for wide-VIN rails", Electronic Products and Technology, www.ept.ca, 18.06.2016.

[11] A.I. Pressman, K. Billings, T. Morey, "Switching Power supply Design", Third Edition, McGraw Hill, ISBN 978-0-07-148272-1, 2009, pp. 20-21

[12] R. Nowakowski, B. King , "Choosing the optimum switching frequency of your DC/DC converter" , 25.10.2006, http://www.eetimes.com/document.asp?doc_id=1272335

[13] M-S. Song, Y-D Son, K-H Lee, "Non-isolated Bidirectional Soft-switching SEPIC/ZETA Converter with Reduced Ripple Currents", Journal of Power Electronics, Vol. 14, No. 4, pp. 649-660, July 2014, ISSN(Print): 1598-2092 / ISSN(Online): 2093-4718

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1654-1659 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Emotional state recognition in speech signal

Krystian Kapala*, Dawid Krawczyk, Stefan Brachmanski

Wroclaw University of Science and Technology, Department of Acoustics and Multimedia, Faculty of Electronics, Wybrzeze Wyspianskiego 27, 50-372 Wrocław, Poland

ARTICLE INFO

Article history:
Received: 05 April, 2017
Accepted: 24 June, 2017
Online: 24 August, 2017

Keywords:
Speech signal processing
Emotion recognition

ABSTRACT

The matters regarding speech signal processing and analyzing in terms of emotional states recognition were presented in this paper. An experiment was conducted to perform both objective and subjective emotional states recognition tests for Polish language.

1. Introduction

The aim of the experiment was to test the effectiveness of fixed seven emotional states recognition. Tests were carried out with use of Polish language speech recordings from two acoustic databases. Samples consisted of professional actor and amateur speakers voice recordings. During studies, an objective, computer method was used composed of speech signal parameters extraction, selection of parameter vectors, and classification tests. The investigations were extended by carrying out subjective emotional state recognition tests performed by group of respondents. In various cases [1] emotional states investigations are implemented, e.g. in automated systems, analysing customer satisfaction level in telecommunication companies hotline systems. The experiments described in the paper should be regarded as pilot studies. However, their subsequent development could contribute to emergence of an automatic system, designed for Polish language, capable of recognizing emotions in speech signal.

2. Acoustic Databases Acquisition

Emotions are one of the most important elements of human life. We can define them as short and usually strong mental arousal induced under the influence of a particular stimulus, often leading to physical reaction. For example, a person in a state of anger, joy or fear usually responds by speaking louder, faster and with increased energy in the higher frequency band.

A popular way of classification is the division of emotions into basic and complex. Basic emotions are up to a dozen [2], and they result from the natural human reactions, while the complex states are to be their appropriate combinations. Despite the lack of consensus on a single set of basic emotions, there is one popular theory named "The big six" by Paul Ekman [3]. Accordingly, basic emotional states are: joy, anger, disgust, surprise, fear and sadness. They can be recognized from facial expressions, gestures and voice, even by people from different cultures, speaking different languages. The studies of emotional states in the speech, in some cases, use only 3 states [4]: positive, negative and neutral. However, due to its versatility and easy recognition "big six" was chosen for conducted experiment.

Emotional state databases contain 3 types of recordings:

- Spontaneous speech recordings.
- Forced emotional states recordings.
- Simulated emotions recordings.

Spontaneous speech databases include recordings of emotions induced in a natural way. To create this type of database usually recordings from multiple sources are used. These include conversations with emergency dispatchers [5], utterances of the participants of TV game shows, journalist relations from dramatic events and the interaction with a robot. Second type of databases include collections of enforced emotions recordings. Each emotional state in speakers voice needs to be induced. It is possible, e.g. through, presentation of movies or recordings stimulating appropriate emotions. Another, slightly more controversial method relies on putting speaker in a stressful situation. The advantage of this type of database is to obtain naturally excited emotions.

The result of this work was development of two acoustic databases (MD – amateur and TS – professional speaker). The

*Corresponding Author: Krystian Kapala, Wroclaw University of Science and Technology, Department of Acoustics and Multimedia, Wybrzeze Wyspianskiego 27, 50-372 Wrocław, Poland | Email: krystian.kapala@pwr.edu.pl

structure of each was divided into training and testing sets. The first (MD) training set contained 543 recordings of women and 556 men. Test set consisted of 509 women recordings and 510 men. In total MD database was composed of 2118 recordings. TS base contained only one speaker recordings. Training and test sets consisted of adequately 219 and 243 recordings, in total - 462.

Table 1. Comparison of databases

	MD	TS
Place of recordings	WUT recording studio	Amateur recording studio
Microphone	T-Bone SCT 700 - condenser	Shure SM 58 - dynamic
Mixer	Yamaha 03D	Behringer Eurorack mx - 802a
Sound Card	SoundMax Integrated Digital Audio	Sound Blaster Audigy 2 NX
Sampling frequency	44.1 kHz	44.1 kHz
Resolution	16 bits	24 bits
Language	Polish	Polish
Speakers	Amateurs - men and women	Professional actor
Speakers number	6 male, 7 female	1 male
Utterances	10 sentences	10 nearly identical sentences
Simulated emotions	6 + neutral state	6 + neutral state
Number of recordings	At least 1	At least 6

Recording sessions for each database were carried out in different conditions as Table 1 shows. A set of 10 sentences was selected, and put in the following order (translations in brackets).

- •Jutro pójdziemy do kina.
  - (Tomorrow we are going to the cinema.)
- •Musimy się spotkać.
  - (We have to meet.)
- •Najlepsze miejsca są już zajęte.
  - (The best seats are already occupied.)
- •Powinnaś zadzwonić wieczorem.
  - (You should have called yesterday.)
- •To na pewno się uda.
  - (I am sure it will succeed.)
- •Ona koniecznie chce wygrać.
  - (She really wants to win.)
- •Nie pij tyle kawy.
  - (Do not drink that much coffee.)
- •Zasuń za sobą krzesło.
  - (Tuck in your chair.)
- •Dlaczego on jeszcze nie wrócił.
  - (Why has not he returned yet.)
- Niech się pan zastanowi.
  - (Think about it, sir.)

Table 2. Phonems appearance frequencies in databases

Phonem	Frequency [%]		Phonem	Frequency [%]	
	TS	MD		TS	MD
e	11.67	12.43	ʧ	1.67	2.16
a	10	9.73	d	1.67	1.62
o	8.33	8.11	l	1.67	1.62
n	6.11	5.95	w	1.11	1.08
j	5.00	5.41	g	1.11	1.08
v	4.44	4.32	ʦ	1.11	1.08
i	3.89	3.78	x	1.11	1.08
p	3.89	3.78	õ	1.11	1.08
u	3.89	3.78	b	0.56	0.54
ɲ	3.89	3.78	ʤ	0.56	0.54
t	3.33	3.24	c	0.56	0.54
s	3.33	3.24	ʣ	0.56	0.54
ɨ	2.78	2.70	ɛ̃	0	0
m	2.78	2.70	ɳ	0	0
ɕ	2.78	2.70	ɻ	0	0
z	2.78	2.70	ʤ	0	0
k	2.22	2.16	ʑ	0	0
r	2.22	2.16	f	0	0
tɕ	2.22	2.16	ʒ	0	0
ʃ	1.67	2.16	#	0	0

Databases of simulated emotional states include recordings of either amateur or professional speakers. This type is the easiest to obtain and analyze. Recording conditions may be adjusted. The list of utterances, with suggested emotional states and number of repetitions are prepared before the recordings. Typically they contain recordings of single sentences, sometimes longer statements [5]. Usually, the content of spoken sentences is not emotionally characterized (e.g. interrogative sentences are rarely used, for possible suggestion of surprise). The biggest drawback is the difference between the natural and simulated emotions, which is impossible to determine. In addition an amateur speaker may simulate either expressively or faintly, which may also affect the results of subsequent analysis. There are many databases of simulated emotional states recorded in multiple languages e.g.. Spanish [6], German [7], Danish [8] or Polish [9].

3. Parameters and vector selection

Recognition of emotions in utterances requires parameterization of the speech signal. After parameters extraction

investigated acoustic material is represented by a vector of extracted features. Unfortunately, there is no universal set of parameters allowing to achieve optimum subsequent analysis, therefore usually a large number of parameters is extracted, which then is used for determining the most representative sets of features - selection.

The following parameters were selected and extracted for further analysis:

- Pitch frequency,
- The first four formant frequencies,
- Intensity of the speech signal,
- LPC coefficients.

Parameters extraction was conducted with Praat and jAudio software. Parameters were characterized with values such as minimum, maximum, mean, median, standard deviation, range and mean absolute slope. Extracted features were exported to a text file and each sample was assigned to a specific emotional state. Parameters selection was conducted in Weka software [10] with usage of ChiSquaredAttributeEval algorithm.

4. Classification

As first trials showed algorithms using support vector machines [11-13] were the most effective, thus both SVC were selected to carry out the final tests. Table 3 presents the effectiveness of various algorithms used for database classification.

Most often training sets contain linearly inseperable data. In these situations algorithms that allow for the existence of components beyond the margin are used e.g. C-SVC. When forming the hyperplane, requirements reduction is determined with special value. Parameter C (cost) is responsible for the number of vectors inside the margin. Its value is determined experimentally in order to obtain maximum efficiency of classification process. Increasing the parameter C causes an effect similar to the linear classification. Small values allow to adjust the hyperplane to the training set.

Tests were performed for the seven cases of various TS (professional actor) and MD (amateur speakers) training and test sets combinations (Table 4).

The algorithm using υ-SVC classification is very similar to C-SVC. An additional parameter is υ coefficient with a value between <0; 1>. This parameter specifies the number of possible violations found in the case of linear classification and the number of support vectors. With increasing υ value the amount of vectors within the margin increases. When υ coefficient has a value of 0, linear classification is performed.

LibSVM library imported to Weka software enabled conducting classification processes with usage of support vector classifiers such as C-SVC and υ-SVC. During tests υ-SVC algorithm showed better effectiveness, thus further classification results were described only for this method.

Table 3. Recognition results for selected algorithms

Used classifier	Effectiveness [%]
Fast decision tree [14]	55.82
K-star	56.04
NN [15]	56.62
KNN5[16]	59.08
KNN10	57.31
Naive Bayes [16]	38.57
Multi layer perceptron	52.60
C-SVC (polynomial kernel)	56.33
C-SVC (radial basis kernel)	63.69
υ-SVC (polynomial kernel)	56.92
υ-SVC (radial basis kernel)	64.38

Table 4. Test and training sets combinations

No	Test set	Training set	Test symbol
1	MD	MD	MD>MD
2	TS	TS	TS>TS
3	MD	TS+MD	TS+MD>MD
4	TS	TS+MD	TS+MD>TS
5	MD	TS	TS>MD
6	TS	MD	MD>TS
7	TS+MD	TS+MD	TS+MD>TS+MD

Table 5. Summary of selected parameters

Parameter \ Set	MD	TS	MD + TS
Pitch frequency F0	average, maximum, minimum, range, median	average, maximum, minimum, range, median, standard deviation, mean absolute slope	average, maximum, minimum, range, median, standard deviation, mean absolute slope
Speech signal intensity	average, maximum, minimum, range, median, standard deviation	average, maximum, minimum, range, median, standard deviation	average, maximum, range, median, standard deviation
Formant F1	average, median	bandwidth mean	average, median
Formant F2	average, median, bandwidth mean	average, range, median, bandwidth mean	average, median, standard deviation, mean absolute slope
Formant F3	-	average, median, bandwidth mean	bandwidth mean
Formant F4	average, median, bandwidth mean	average, minimum, range, median, standard deviation	average, median
LPC parameters	(2, 3, 4, 8, 9, 10 order mean values)	(2, 3, 4, 6 order mean values)	(2, 3, 4, 8, 9, 10, 11 order mean values)

5. Results presentation

Recognition efficiency results for each emotional state and all 7 test were presented in confusion matrixes, respectively Tables 6-12.

Conducted classification tests brought varied results. The lowest effectiveness was achieved for different database training and test set cases. In case of MD test and TS training set, overall effectiveness was 21.6%. Fear and surprise were recognized with the best score - above 40%. State of anger was recognized as fear 75 times out of 157 samples. The opposite situation (MD training, TS test set) improved results to 29.6%. Anger and neutral emotional states were recognized in 54.1 and 57.6% of cases.

Table 6. Confusion matrix (MD>MD)

	anger	neutral	joy	sadness	fear	disgust	surprise	Effectiveness [%]
anger	**106**	6	14	4	13	10	4	67.5
neutral	2	**103**	1	24	1	23	1	66.5
joy	15	8	**91**	0	10	12	11	61.9
sadness	1	26	3	**99**	2	18	0	66.4
fear	11	2	14	8	**77**	8	7	60.6
disgust	10	9	7	27	5	**81**	5	56.3
surprise	7	5	5	5	12	7	**99**	70.7

Table 7. Confusion matrix (TS>TS)

	anger	neutral	joy	sadness	fear	disgust	surprise	Effectiveness [%]
anger	**33**	0	3	0	1	0	0	89.2
neutral	0	**28**	0	0	2	2	1	84.8
joy	6	0	**26**	0	1	0	2	74.3
sadness	0	0	0	**27**	3	4	2	75.0
fear	2	0	3	0	**24**	4	0	72.7
disgust	0	1	4	2	1	**23**	3	67.6
surprise	2	1	4	0	3	3	**22**	62.9

Table 8. Confusion matrix (TS+MD>MD)

	anger	neutral	joy	sadness	fear	disgust	surprise	Effectiveness [%]
anger	**116**	3	13	1	9	9	6	73,9
neutral	1	**84**	6	20	2	42	0	54,2
joy	16	5	**95**	0	9	12	10	64,6
sadness	3	23	4	**84**	1	30	4	56,4
fear	11	2	21	3	**77**	4	9	60,6
disgust	13	6	12	15	1	**87**	10	60,4
surprise	8	3	8	5	10	8	**98**	70

Table 9. Confusion matrix (TS+MD>TS)

	anger	neutral	joy	sadness	fear	disgust	surprise	Effectiveness [%]
anger	**27**	0	8	0	2	0	0	73.0
neutral	1	**27**	3	2	0	0	0	81.8
joy	10	0	**16**	0	6	0	3	45.7
sadness	0	17	1	**12**	0	2	4	33.3
fear	6	5	7	6	**6**	3	0	18.2
disgust	1	12	11	2	1	**3**	4	8.8
surprise	1	3	5	4	4	3	**15**	42.9

Table 10. Confusion matrix (TS>MD)

	anger	neutral	joy	sadness	fear	disgust	surprise	Effectiveness [%]
anger	**9**	0	41	0	75	10	22	5.7
neutral	0	**1**	1	29	48	36	40	6.0
joy	13	2	**38**	1	37	26	30	25.9
sadness	0	2	14	**36**	23	19	55	24.2
fear	11	0	34	5	**56**	5	16	44.1
disgust	5	1	17	31	43	**19**	28	13.2
surprise	4	0	37	1	31	6	**61**	43.6

Maximum efficiency results as expected, were achieved while using training and test sets from the same databases. In case of MD database overall performance was 64.4%. The best result for single state concerned surprise - 70.7%, the worst one, disgust 56.3%. In the classification conducted on sets from the database TS total effectiveness of 75.3% was obtained. Most likely this was a result of single person recordings within database - professional actor using similar expression of emotions in each repeated sentence. All the emotions were identified with efficiency higher than 60%, achieving the best results for anger - 89.2% and neutral state - 84.8%.

Important tests were also carried out for combined training sets of both databases (TS+MD). The conducted classifications highest score (62,9%) was obtained for the test set derived from the MD database. The lowest result was (49,6%) observed in case of TS database test set. Anger state recognition results in both cases were very similar, whereas in case of neutral state for TS derived test set was higher by nearly 30%. The rest of the emotional states were recognized with greater efficiency when testing MD derived set. It should be noted that the recognition results for TS test set in case of single emotional states were varied - the respective values ranged between 8.8 and 81.8%. An experiment was also conducted on combined collection of training and test sets of both databases. The overall result was 57.4%. The results for individual emotional states recognition were similar to the ones obtained when testing MD database itself, most likely due to the impact of large number of samples from the

MD database. They were not strongly diversified and the results for all emotional states exceeded 50%. Adding TS training set deteriorated 4 emotional states recognition effectiveness by approximately 10%, leaving the rest with similar results.

Table 11. Confusion matrix (MD>TS)

	anger	neutral	joy	sadness	fear	disgust	surprise	Effectiveness [%]
anger	**20**	1	7	0	7	0	2	54.1
neutral	2	**19**	1	9	0	2	0	57.6
joy	13	0	**6**	0	7	0	9	17.1
sadness	0	19	0	**10**	0	4	3	27.8
fear	8	4	5	1	**2**	13	0	6.1
disgust	3	14	9	2	0	**2**	4	5.9
surprise	3	1	5	6	4	3	**13**	37.1

Table 12. Confusion matrix (TS+MD>TS+MD)

	anger	neutral	joy	sadness	fear	disgust	surprise	Effectiveness [%]
anger	**122**	3	39	2	12	11	5	62.9
neutral	2	**99**	10	27	3	46	1	52.7
joy	21	5	**116**	0	16	13	11	63.7
sadness	5	25	9	**107**	1	35	3	57.8
fear	14	4	35	7	**83**	9	8	51.9
disgust	14	11	24	18	5	**94**	12	52.8
surprise	10	4	17	10	16	15	**103**	58.9

6. Empirical studies

Empirical studies were done in two stages. The first step involved preparing the appropriate test material from both acoustic databases. Step two involved empirical research and analysis of results.

The task of the person taking part in the experiment was to assign the proper emotional state to each sample. The selection was made from only seven states (6 basic and neutral one). Preliminary studies have shown that the test material meets the criteria of objectivity. After listening to each sound sample, the listener was supposed to determine what emotional content it was carrying. The listener answered by clicking on the appropriate emotional state button. If in doubt, the sample could be replayed. A picture representing questionnaire appearance was added below (Figure 1)

The first step was to determine the number of people to take part in the experiment so that the results could be representative. The selected group consisted of 10 people (2 female, 8 male participants). Age of group members varied between 26 and 32.

Survey was composed of two parts. The first one referred to recordings from the amateur speakers database. Table 13 presents the recognition effectiveness for each individual participant. It turns out that the accuracy of their emotional states recognition in given experiment ranged between 55 and 70 percent. These results could be the beginning of an interesting discussion connected with their interpretation. Psychological reasons were omitted in this article. They will be addressed during subsequent studies.

Figure 1. Questionnaire used in subjective tests

Table 13. Test participants results

Number of samples	Participant ID	Right answers	Effectiveness [%]
266	1	146	54,9
267	2	151	56,6
266	3	154	57,9
267	4	157	58,8
266	5	162	60,9
267	6	166	62,2
267	7	172	64,4
267	8	180	67,4
267	9	180	67,4
267	10	188	70,4
In total		**In total:**	**Average:**
2667		1656	62,1

A confusion matrix was presented (Table 14), which also show the effectiveness of recognizing individual emotional states. The best recognized emotion for ameteur speakers among the participants was „surprise", which reached almost 75 percent of the correct interpretations. The worst state to determine was "disgust". This state was correctly identified in 40 percent of cases.

Table 14. Confusion matrix (amateurs)

Amateur speakers database	anger	neutral	joy	sadness	fear	disgust	surprise	Effectiveness [%]
anger	**236**	39	14	8	14	38	10	65.7
neutral	5	**257**	3	62	7	32	1	70.0
joy	14	62	**210**	1	25	16	20	60.3
sadness	2	45	3	**231**	20	46	4	65.8
fear	18	33	13	27	**176**	28	26	54.8
disgust	22	77	2	71	11	**149**	40	40.0
surprise	8	5	10	5	30	28	**253**	74.6

In the second part of the survey, the respondents were presented with sound samples of professional actor voice (confusion matrix in Table 15). The best interpreted state was "anger" recognized correctly in all cases. The least effectiveness was achieved for fear, where efficiency was only 20%. In fact fear was even more times interpreted as anger which could indicate on quite specific, different actor expression.

Table 15. Confusion matrix (professional)

Professional speakers database	anger	neutral	joy	sadness	fear	disgust	surprise	effectiveness [%]
anger	**29**	0	0	0	0	0	0	100
neutral	1	**28**	0	0	0	1	0	93,3
joy	1	1	**29**	0	0	0	0	93,5
sadness	2	8	0	**12**	2	6	0	40
fear	9	5	0	4	**6**	4	2	20
disgust	7	5	0	0	2	**15**	1	50
surprise	0	0	4	0	0	1	**25**	83,3

The differences in test results obtained from both parts of empirical studies seem to be quite obvious, given the various way of expression. Actor utterances could have been significantly different from the amateur's in pace or dynamics of the voice. Another important fact is the use of only one speaker in comparison with 13 people (both genders) in amateur database. Table 16 shows the results of emotion recognition for each recorded speaker. Attention was drawn to the fact that some non-professional speakers have gained better interpretation results of their emotions.

Table 16. Speakers individual results

Amateurs:			
Speaker ID (M- Male, F-Female):	**Number of samples**	**Right answers**	**Effectiveness [%]**
M4	198	96	48,5
M7	171	89	52
M5	207	108	52,2
K6	201	113	56,2
K4	117	67	57,3
M2	203	120	59,1
M1	150	89	59,3
M6	206	123	59,7
K5	210	139	66,2
K3	186	125	67,2
K1	188	129	68,6
K2	209	145	69,4
M3	211	169	80,1
Professional:			
M1	144	210	68,6

7. Empirical studies

The results obtained during the investigations both for empirical and computer methods are comparable, which could suggest that there is a possibility to create and further develop an effective automatic tool capable of recognizing emotions from speech signal. In experiment both amateur and professional speakers were used. The results show that unlikely to initial assumptions including professional speaker in investigations did not improve the recognition effectiveness – using his recordings in training sets in objective method gave good result in case of test set composed only of his own recordings. Additionally during empirical studies some amateur speakers were recognized with better effectiveness.

References

[1] L. Vidrascu, L. Devillers., Detection of real-life emotions in call centers. in Proc. Eurospeech 2005, Lizbona, 2005.

[2] Cowie R., Describing the Emotional States Expressed in Speech. In Proceedings of the ISCA, Belfast, 2000, pp. 11-18.

[3] Ekman P.: Emotion in the human face. Pergamon Press, 1972.

[4] Rong J., Chen Y. P., Chowdhury M., Li G., Acoustic Features Extraction for Emotion Recognition. Computer and Information Science, 2007. ICIS 2007. 6th IEEE/ACIS International Conference, 11-13 July 2007, pp. 419-424.

[5] Ambruš D., Collecting and Recording of an Emotional Speech Database. Technical Report, Faculty of Electrical Engineering and Computer Science, Institute of Electronics, University of Maribor.

[6] Montero J. M., Gutierrez-Arriola J., Colas J., Enriquez E., Pardo J. M., Analysis and modeling of emotional speech in Spanish. In Proc. ICPhS'99, pp. 957 - 960, San Francisco 1999.

[7] Burkhardt F., Paeschke A., Rolfes M., Sendlmeier W., Weiss B., A Database of German Emotional Speech. Proc. Interspeech 2005.

[8] Engberg I. S., Hansen A. V., Documentation of the Danish Emotional Speech Database. Internal AAU report, Center for Person Kommunikation, Department of Communication Technology, Institute of Electronic Systems, Aalborg University, Denmark, September 1996.

[9] J. Cichosz, K. Ślot, Low-Dimensional Feature Space Derivation for Emotion Recognition. ICSES 2006, Łódź, 2006.

[10] Witten I. H., Frank E., Data Mining: Practical machine learning tools and techniques. 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

[11] Vapnik V. N., Statistical Learning Theory. Wiley, 1998.

[12] Kwon O., Chan K., Hao J., Lee T., Emotion Recognition by Speech Signals. Eurospeech, Geneva, Switzerland, Sep. 01-03, 2003.

[13] Zhou J., Wang G., Yang Y., Chen P., Speech emotion recognition based on rough set and SVM. Cognitive Informatics, 2006. ICCI 2006. 5th IEEE

[14] Cichosz J., Ślot K, Emotion Recognition in Speech Signal Using Emotion-extracting Binary Decision Trees. Affective Computing and Intelligent Interfaces – ACII 2007, Lisbon.

[15] Kang B., Han Ch., Lee S., Youn D., Lee Ch., Speaker Dependent Emotion Recognition Using Speech Signals. In Proc. ICSLP, 2000, pp. 383-386.

[16] Wang Y., Guan L, An Investigation of Speech-Based Human Emotion Recognition. Multimedia Signal Processing, 2004 IEEE 6th Workshop, 29 Sept.-1 Oct. 2004, pp. 15- 18.

[17] Kapala K., Krawczyk D. Brachmanski S., Tests of selected emotional states recognition., 20th. IEEE SPA Conference , Poznan, 2016.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1660-1664 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Factors Affecting the Adsorption of Trivalent Chromium Ions by Activated Carbon Prepared from Waste Rubber Tyres

Sylvia E. Benjamin[*,1], Muhammad Ashfaq Sajjid[2]

[1]*Forman Christian College - A chartered University, Gulberg II, Lahore 54600, Pakistan*

[2]*Obeikan Investment Group, 2nd Industrial City, Riyadh 17952, Saudi Arabia*

A R T I C L E I N F O

Article history:
Received: 16 June, 2017
Accepted: 08 August, 2017
Online: 24 August, 2017

Keywords:
Adsorption
Chromium (III)
Tyre Activated Carbon (TAC)

A B S T R A C T

Economic gains are generally the outcome of industrialization and consequently urbanization. However, positive fiscal index generates a negative impact on natural environment sources heaving pollutant burden on soil, air and water. Industries throw tones of contaminated water into soil and water bodies without proper treatment and create a potential threat for both living and non-living species. Chromium in trivalent state (Cr^{3+}) is added in water bodies and soil through waste water from tanneries, cooling water systems, chemical and pulp and paper industries. The present research work aims at the preparation of an inexpensive activated carbon prepared from non- degradable waste scrap rubber tyres. The carbon produced from scrap rubber tyres was activated by 5% solution of $BaCl_2$ and 0.4 N solution of HCl and verified by ethylene blue solution. The adsorption capacity of the Tyre activated carbon (TAC) was investigated for different parameters i.e., initial chromium (III) ion concentration, activated carbon dosage, contact/ stirring time and pH. The adsorption capacity of TAC depends on the initial metal ion concentration and the TAC dose. pH of the chromium solution effects the adsorption capacity of TAC due to the formation of tetra hydroxochromate(III) complexes,. The results show that TAC offers a cost effective reclamation process for the removal of Cr^{3+} from effluent waters.

1. Introduction

Industrialization and urbanization at a fast rate, especially in the developed countries, is todays leading need due to rapid population growth and thus directly linked with advancements in economics. High economic gains has its beneficial as well as adverse effects. Industrial progress has intensified pollution problems worldwide. Land, air and water pollution is building up due to inefficient execution of environmental laws. One of the major concerns is water pollution caused by effluents from various industries. The present paper is an extension of the original work that presented the removal of Copper ions (Cu^{2+}) from industrial waste waters by TAC [1].

Economic booms have lifted up many people from agonizing poverty through industrialization. However, this has resulted in serious environmental degradation (China Water Quality Management Policy and Institutional Considerations, World Bank, 2006). Industries use tones of water for various processes and also release large quantities of waste water (loaded with spent chemicals and metallic residues) in the nearby lakes, rivers and canals without suitable treatment. Many areas of the world have now lost access to safe drinking water. It is estimated that 2000 children under the age of five die daily. 90 % of such deaths are linked with diarrheal diseases due to intake of contaminated water and improper sanitation and hygiene conditions. A press release by UNICEF (New York, 22 March 2013) states that 783 million people do not have access to safe drinking water (comprising of China on the top with 119 million, then India, Nigeria and lastly Pakistan at the bottom with 15 million).

Release of chemicals in aquatic environment brings changes in the structure and functional properties of water species [2] and sometimes results in their total extinction. A major concern of today are the heavy metals which are toxic above certain concentration [3-4] and henceforth, a potential threat for the consumers. Severe ecological problems are arising due to the contamination of surface and ground water. Some plant species

[*]Corresponding Author: Sylvia E. Benjamin, Forman Christian College - A chartered University, Gulberg II, Lahore 54600, Pakistan, +92 42 33342997934, Email: s_e_benjamin@yahoo.com

are able to withstand low concentrations of certain metals but generally face eradication at high levels [5]. Mostly heavy metals are carcinogenic and get incorporated in the food chain. Biomagnification and bioaccumulation [6] of heavy metals in the living beings has further aggravated the situation.

Chromium metal possesses three stable valence states i.e., 0, +3 and +6. Dietary Cr^{3+} is an essential trace mineral requirement of human body for the control of glucose metabolism by insulin [7]. is proved to be a potential carcinogen [8-9]. In [10]-[11], the authors have reported that even Cr^{3+}, when taken as supplement, can accumulate in tissues and can bring about structural changes in DNA. A major release of Cr^{3+} is through tanneries and cooling water systems, and from refractory, chemical and metallurgical industrial operations. Cr^{3+} incorporation into the ecosystem and delivery to humans via e.g., vegetables [12] and seepage to the ground water has adversely affected human lives. It is therefore, immensely important to address this pertinent issue which needs strict implementation of Cr^{3+} removal systems from waste waters.

Several methods for the removal harmful ions in water are effective e.g., chemical precipitation, electrolysis, membrane technology, ion exchange etc. and are in common practice by the industries. Adsorption of ions onto a suitable adsorbent is a low cost, effective and easy method for the separation of undesired chemicals. Adsorption [13-14] i.e., physiosoprtion or chemisorption can adequately remove matter in solid, gas or liquid form. Several waste materials are reported to be good adsorbents [15-16]. Activated carbon is especially favored due to its inert nature [17-18]. In this study, activated carbon derived from scrap rubber tyres is used as an adsorbent for the removal of Cr^{3+} from aqueous solutions. The major aim was to test the performance of TAC as the function of adsorbent dose, Cr^{3+} concentration in solution, pH and the contact time. This will be a cost-effective utilization of non-degradable scrap rubber tyres for water purification.

2. Materials and Methods

2.1. Preparation of Activated Carbon from Scrap Rubber Tyre

Scrap rubber tyres were collected from disposal sites. These were cut into small pieces, washed several times with distilled water and dried. Pieces were then crushed and grinded in mortar and pestle. 600gms of the ground rubber tyres was taken in a round bottomed flask fitted with a water condenser. Dry distillated of tyres was done at 450 °C. Three main fractions were obtained i.e., chars, oils plus water mixture and obnoxious gases. The residue chars were heated in a muffle furnace at 800-900 °C for 30 seconds and cooled to room temperature in desiccators. It was then soaked in 5% $BaCl_2$ solution for 12 hours for chemical activation and then filtered. The carbon was firstly washed with HCl (0.4 N) and then several times with distilled water to remove the oxides and soluble impurities. The final residue sample was dried in an oven overnight at 105°C. The dried TAC was sieved and 200 mesh size was selected for Cr^{3+} removal studies.

2.2. Confirmatory Test for the adsorption capacity of TAC.

The efficiency of TAC was tested by Methylene Blue test. 0.2gm of TAC was immersed in 10ml of 1% Methylene blue solution in a 50ml beaker. Same test was performed with commercially available Merck activated carbon for comparison. The absorbance was measured at λ_{max} of 660nm.

The adsorption capacity. AC (%) of TAC was calculated using following relationship:

$$AC\ (\%) = \frac{Ci - Ct}{Ci} x\ 100$$

Where

$Cí$ = initial concentration of the simulated solution

Ct = The concentration of solution after treatment.

The adsorption capacity (or the removal of Cr^{3+} - %) of TAC was checked on Varian spectrophotometer using an acetylene-air flame and lamp current of 7 mA for Cr^{3+} determination.

2.3. Preparation of Chromium(III) solutions

Chromium sulphate ($Cr_2(SO_4)_3.12H_2O$) salt was dried in an oven overnight at 50 °C. Stock solution of 1000 ppm was prepared by dissolving 7.70 gms of salt in de-ionized in a measuring flask of 1L. Further dilutions of 50 ppm, 100 ppm & 200 ppm solutions were made from stock solution.

The following tests were performed to check the adsorption capacity of the prepared TAC.

2.4. Effect of TAC dosage on adsorption of chromium (III) ions

Five volumetric flasks of 50 ml capacity were taken. 25 ml of 200 ppm solution of Cr^{+3} ions was poured in each flask. 0.1 gm, 0.2 gm, 0.3 gm, 0.4 gm and 0.5 gm of TAC was added. The solutions were stirred for 3 hours on magnetic stirrer at 600 rpm and then filtered. The concentration of Cr^{3+} left in solution after adsorption tests was checked by Atomic absorption spectrophotometry (AAS).

2.5. Effect of chromium (III) ions concentration on adsorption by the TAC

25 ml of 10 ppm, 50 ppm, 100 ppm, 150 ppm and 200 ppm Cr^{3+} solutions were taken in five measuring flasks. 0.1 gm of TAC was added. The solutions were stirred for 3 hrs on magnetic stirrer at 600 rpm and then filtered. The amount of Cr^{3+} remaining in solution in each flask was measured by AAS.

2.6. Effect of contact time on the adsorption capacity of TAC

25 ml of Cr^{3+} solution (200 ppm) was taken in five separate 50 ml volumetric flasks (x5). TAC (0.1 gm) was added to each flask and the solutions were stirred for different time intervals i.e., 1, 2, 3, 4 and 5 hours on magnetic stirrer at 600 rpm. Cr^{3+} remaining in solution after each time interval was checked by AAS.

2.7. Effect of solution pH on the adsorption capacity of TAC.

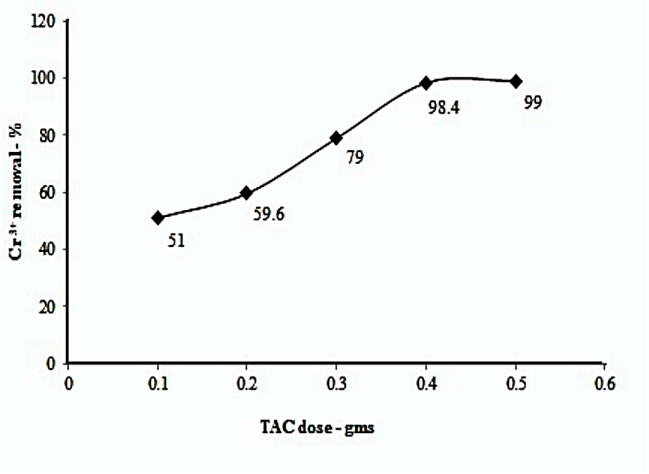
Cr^{3+} solutions (25 ml) were taken in four separate 50 ml volumetric flasks. The pH of solutions was maintained at pH 4, pH 5, pH 7 and pH 10. TAC (0.1gm) was added to each flask and the solutions were stirred for 3 hours at 600 rpm. After the adsorption test each solution was filtered and the Cr^{3+} concentration was checked by AAS.

3. Results & Discussion

Water pollution by the industries have risen to colossal disasters in many parts of the world. The developing countries are facing tremendous ill health effects due to water contamination especially by heavy metals [19-20]. Many living species are eradicated leaving environmental issues that need remediation. This has led the scientists to find new methods and materials for survival of living beings. Adsorption of harmful chemicals and heavy metals is a cheap method for the removal of contaminants. In recent years, several adsorbents and particularly bio adsorbents [21-23] have been investigated to combat pollution arising as a consequence of anthropogenic activities. Rubber tyres are non-biodegradable. These have no use except illegally burned to produce obnoxious fumes or dumped in tones at waste sites. We have made use of waste rubber tyres and converted it to activated carbon for the reclamation of waste water.

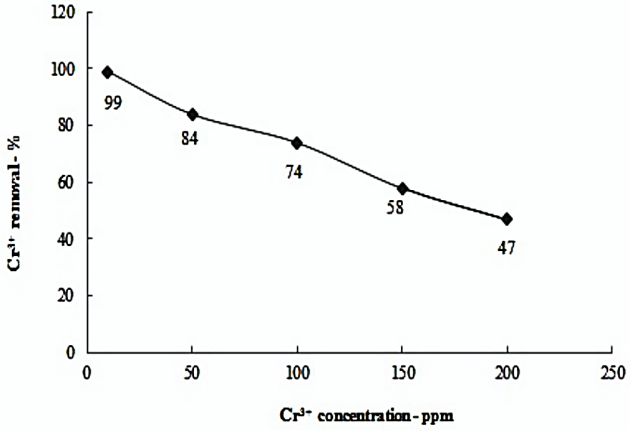
The efficiency or the adsorption capacity of the prepared TAC was checked by the Methylene blue test and compared with commercial Merck Activated Carbon (powdered) at λ_{max} of 660 nm. The blank solution showed an absorbance of 2.707. The dye absorbance test showed that TAC had absorbance of 0.053 and the commercial Merck Activated Carbon had absorbance of 0.035. Thus, our prepared TAC showed good absorbance capacity for the dye.

It is envisaged that as the amount of adsorbent is increased, the removal of metal ions should also increase systematically. TAC required to effectively adsorb Cr^{3+} was checked by varying TAC dose for 200 ppm concentration. The results are presented in Figure 1.



The adsorption percentage of Cr^{3+} (200 ppm) increases with the amount of adsorbent dose. Maximum adsorption of 99 % was observed with 0.5 gm TAC of 200 mesh size at pH 6.5 (nearly neutral) after a contact period of 3 hours held at a stirring speed of 600 rpm. This important parameter has been studied by several researchers before [24-25]. In our earlier work [1], TAC proved to be an effective adsorbent for the removal of Copper (II) ions from industrial waste water. Sharma and Goyal [26] have found strong effect of adsorbent dose on removal of Cr^{3+} from tannery effluents at pH-4 by Aspergillus sp biomass. Recently, Jamil et al [27] have reported similar dose effect of Pakistani coal derived activated carbon on removal of Cr^{3+} from simulated solutions. The exchangeable sites for the metal ion adsorption increase with the increase in TAC dose and hence, the water is cleaned from the contaminant ions. The distribution coefficient for the TAC dose tests varies in the range of 10.4 to 198. The Gibbs free energy (ΔG)changes from -5.8 KJ/mole through -13.4 KJ/mole telling that Cr^{3+} removal by adsorption is a favorable process. We selected TAC (mesh-200) dose of 0.1gms and Cr^{3+} concentration of 200 ppm for further experimentation so that variations can be studied more thoroughly.

The effect of initial metal ion concentration is shown in Figure 2.



The removal of metal ions from aqueous solutions by adsorption is highly dependent on the initial metal ion concentration in solution [28-29]. The adsorption efficiency of adsorbents is applicable in a specific range of concentrations as the active site may be filled above certain concentrations. Figure 2 shows that the adsorption decreased when the concentration of the metal ions increased. Maximum adsorption of 99% was observed at low concentration of Cr^{3+} ions i.e., 10 ppm and linearly declined as the initial Cr^{3+} concentration increased. Adsorption reduced to 47% with 0.1gms of TAC at 6.2 pH at contact time of 3 hours. TAC amount is the limiting factor. The exchangeable sites with a low dose amount of 0.1gms became saturated with Cr^{3+} ions at 10 ppm concentration. The site coverage is reached and equilibrium is established. At higher initial concentrations of Cr^{3+} ions, the adsorption percentage decreases due to low availability of active sites. We observed the surface charge interaction phenomenon in absence of diffusion.

The contact time of adsorbent with the metal ion solution is an important factor that effects the adsorption capacity [30, 31]. Cr^{3+} absorption by 0.1 gm TAC dose as the function contact time is shown in Figure 3.

The results show that as the contact time of TAC with Cr^{3+} solution is increased, the removal of Cr^{3+} also increases. Under constant stirring conditions, maximum adsorption of Cr^{3+} (50%) was observed in 3 hours. After 3 hours, no significant change in Cr^{3+} removal is observed. Thus, 3 hours was found sufficient for achieving equilibrium.

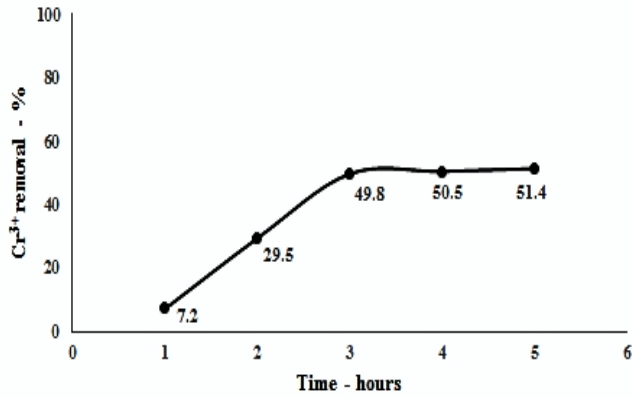


Figure.3. Effect of contact time on Cr (III) adsorption (initial Cr^{3+} concentration: 200 ppm, TAC dose: 0.1g, pH - 6, stirring speed: 600rpm)

The effect of pH on adsorption process is presented in Figure 4.

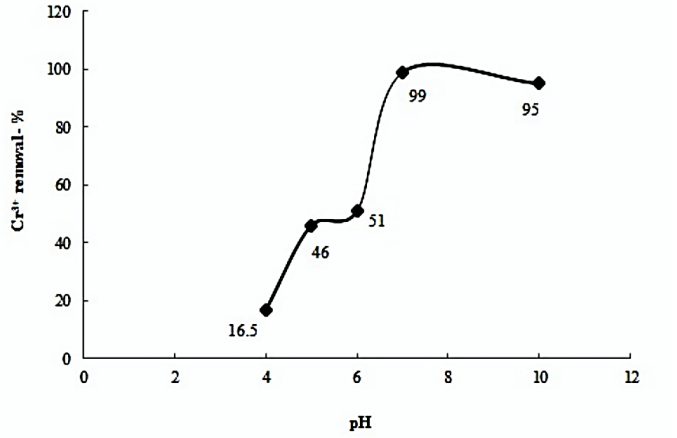


Figure.4. Effect of solution pH on Cr (III) adsorption (initial Cr^{3+} concentration 200ppm, contact time: 3 hours, activated carbon dose: 0.1g, stirring speed: 600rpm)

The removal of metal ions i.e., the absorptive capacity of adsorbents is dependent upon optimal pH level of the metal ions solution (32-33). In acidic medium (pH-4), Cr^{3+} is present as mononuclear hydrated ion. Figure 3 shows that the adsorption efficiency of TAC for Cr^{3+} is very low i.e., only 16.5%. At acidic pH, the hydronium (H^+) ions affiliation to the adsorbents surface is responsible for TACs' poor adsorption efficiency. The adsorption efficiency of TAC for Cr^{3+} increases with an increase in pH till pH-5 to 46 %. Deprotonation of metal binding sites on TAC also occurs when the solution pH increases. The Cr^{3+} ions are attracted to the vacant sites and hence, higher adsorption is observed. Decrease in positive charge at TAC removes repulsion and hence, improves the attraction of Cr^{3+} ions for active adsorption. Lyubchik et al. [34] have identified carboxylic groups on surface of activated carbons for enhancement in removal of Cr^{3+} at initial solution pH between 2 till 4. Weak change in adsorption occurs between pH 5 through pH 6. At this pH, Cr^{3+} precipitates as hydrated $Cr(OH)_3$. As the pH rises, solubility of chromium (III) increases. Apparently, adsorption of the Cr^{3+} ions is maximum at pH 7 i.e., 99% which is close to 95% adsorption at pH -10. At alkaline pH, adsorption efficiency is disturbed due to the precipitation of Cr^{3+} as tetra hydroxochromate(III) complexes, $Cr(OH)_4^-$ [35]. Cu^{2+} [1] go into hydroxide formation at high pH. Precipitation is not as effective as adsorption as the separation of precipitates from water is cumbersome. The adsorption efficiency may increase at acidic pH with that activated carbon which has negatively charged surface. Dai, Ren and Tao [36] used solution pH for distinction between Cr (VI) and Cr^{3+} adsorption on chemically modified Chitosan. The adsorption efficiency of Cr(VI) was maximum i.e., 97% at pH 4.0, whereas the adsorption efficiency of Cr(III) was 5%. We have also observed very weak adsorption of Cr^{3+} ions i.e., 16.5 % at pH-4.

4. Conclusions

TAC is an effective low cost adsorbent for the removal of Cr^{3+} and Cu^{2+}ions. It offers treatment of impaired water at reduced energy and time consumption. The inert nature of TAC secures its reliability versus biomaterials which degrade over long time exposures. TAC almost completely removes Cr^{3+} ions from solutions when the metal ions are present in low concentration at pH 5.5. In higher alkaline medium, tetra hydroxo-Cr(III) complexes formation increases.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

The authors are indebted to thank Prof. Dr. P.H. Armacost, Rector, Forman Christian College – A chartered University, Lahore, Pakistan for providing the necessary laboratory facilities and Mr. Rashid Masih, Laboratory Supervisor, Department of Chemistry, Forman Christian College – A chartered University, Lahore, Pakistan, for his constant help in doing the atomic absorption work.

References

[1] S.E. Benjamin & Muhammad A. Sajjid, "Purification of Copper ion contaminated Industrial water by activated carbon prepared from scrap rubber tyres", Proc. 4th Int. Conf. on Energy, Environment and Sustainable Development (EESD) 2016, Nov. 1-3, 2016, Mehran Univ. of Eng. and Tech., Pakistan. http://eesd.muet.edu.pk/ sites/default/files/imagecache/ EESD_2016 paper_284.pdf

[2] M. ShahidulIslam, M. Tanaka, "Impacts of pollution on coastal and marine ecosystems including coastal and marine fisheries and approach for management: a review and synthesis", Marine Poll. Bull., **48**, 624–649, 2004.

[3] I. M. Madany, A.A.A. Wahab, Z. Alawi, "Trace metals concentrations in marine organisms from the coastal areas of Bahrain, Arabian Gulf", Water, Air, and Soil Poll., **91**(3), 233–248, 1996.

[4] A. M. Freije, "Heavy metal, trace element and petroleum hydrocarbon pollution in the Arabian Gulf: Review", J. of the Assoc. of Arab Univ. for Basic & Appl. Sciences (JAAUBAS), **17**, 90–100, 2015. http://dx.doi.org/10.1016/j.jaubas.2014.02.001.

[5] J. Antonovics, A.D. Bradshaw, R.G. Turner., "Heavy Metal Tolerance in Plants" Adv. in Ecological Res., 7, 1–85, 1971.

[6] N.S. EL-Shenawy, N. Loutfy, M.F.M. Soliman, M. M. Tadros, A. A. Abd El-Azeez, "Metals bioaccumulation in two edible bivalves and health risk assessment", Environ Monit Assess, 188**,** 139, (2016). doi:10.1007/s10661-016-5145-2.

[7] T. H. Lim, T. Sargent, N. Kusubov, "Kinetics of trace element chromium(III) in the human body", Am J Physiol., **244**(4), R445-54, 1983.

[8] D. L. Carlisle, D. E. Pritchard, J. Singh, S. R. Patierno, "Chromium(VI) Induces Dependent Apoptosis in Diploid Human Lung and Mouse Dermal Fibroblasts", Molecular carcinogenesis, **28**(2), 111 -118, 2000, DOI: 10.1002/1098-2744(200006)28:2< 111:AID-MC7>3.0.CO;2-Y.

[9] S. S. Wise, J. H.C. Schuler, S.P. Katsifis, J. P. Wise Sr, "Barium chromate is cytotoxic and genotoxic to human lung cells", Environ. Mutagenesis, **42** (4), 274–278, 2003.
[10] D. M. Stearns, J. J. Belbruno, K. E. Wetterhahn, "Prediction of chromium(III) accumulation in humans from chromium dietary supplements", FASEB J, **9** (15),1650-7, 1995.
[11] D. A. Eastmond, J. T. Macgregor, R. S._Slesinski, "Trivalent chromium: assessing the genotoxic risk of an essential trace element and widely used human and animal nutritional supplement", Crit Rev Toxicol., 38 (3), 173-90, 2008. Doi:10.1080/10408440701845401.
[12] A. Nishat, "Flora damage assessment due to Chromium in the vicinal areas of tanneries", M.Phil Thesis, Forman Christian College – A chartered University, Pakistan, 2011.
[13] X. Chu, S. Liu, S. Zhou, Y. Zhao, "Adsorption and separation of carbon dioxide and methane in new zeolites using the Grand Canonical Monte Carlo method, Adsorption", 22, 891-899, 2016. doi:10.1007/s10450-016-9793-x.
[14] W.S.W. Ngah, M. A. K. M. Hanafiah, "Removal of heavy metal ions from wastewater by chemically modified plant wastes as adsorbents: A review", Biores. Technol., **99**, 3935–3948, 2008.
[15] C. Namasivayam, D. Kavitha, "Removal of Congo Red from water by adsorption onto activated carbon prepared from coir pith, an agricultural solid waste", Dyes and Pigments, **54**, 47–58, 2002.
[16] A. Demirbas, "Heavy metal adsorption onto agro-based waste materials: A review", J. of Hazard. Mat., 157 (2–3), 220–229, 2008.
[17] M. Karnib, A. Kabbani, H. Holail, Z. Olama, "Heavy Metals removal using Activated Carbon, Silica and Silica Activated Carbon Composite", Energy Procedia, **50,** 113-120, 2014.
[18] T. S. Anirudhan, S. S. Sreekumari, "Adsorptive removal of heavy metal ions from industrial effluents using activated carbon derived from waste coconut buttons", J. of Environ. Sciences, **23** (12), 1989-1998, 2011.
[19] J. O. Duruibe, M. O. C. Ogwuegbu, J. N. Egwurugwu, "Heavy metal pollution and human biotoxic effects", Inter. J. of Physical Sci., **2** (5), 112-118, 2007. http://www.academicjournals.org/IJPS ISSN 1992 - 1950.
[20] V. M. Chaitali, J. Dhote, "Review Of Heavy Metals in Drinking Water and their Effect on Human Health", Inter. J. of Innovative Res. in Sc., Engr. and Techol. (IJIRSET), **2** (7), 2992-2996, 2013.
[21] S.N.A. Abas, M.H.S. Ismail, M. L. Kamal, S. Izhar, "Adsorption Process of Heavy Metals by Low-Cost Adsorbent: A Review", World Appl. Sci. J., **28** (11), 1518-1530, 2013. Doi:10.5829/idosi.wasj.2013.28.11.1874.
[22] M. Jain, V.K. Garg, K. Kadirvelu, "Equilibrium and kinetic studies for sequestration of Cr(VI) from simulated wastewater using sunflower waste biomass", J of Hazard Mater**, 171** (1–3), 328-334, 2009.
[23] S. Gupta, B.V. Babu, "Removal of toxic metal Cr(VI) from aqueous solutions using sawdust as adsorbent: Equilibrium, kinetics and regeneration studies", Chem Engr J, **150**, 352-365, 2009.
[24] A. Sarı, M. Tuzen, M. Soylak, "Adsorption of Pb(II) and Cr(III) from aqueous solution on Celtek clay", J. of Hazard Mater, 144 (1–2), 41–46, 2007.
[25] T. R. Chaudhury, K. M. Pathan, M. Amin, M. Ali, S. B. Quraishi, A. I. Mustafa, "Adsorption of Cr (III) from aqueous solution by ground nut shell", J. of Environ. Sci and Water Resources, **1** (6), 144 - 150, 2012.
[26] I. Sharma, D. Goyal, "Adsorption Kinetics: Bioremoval of trivalent Chromium from Tannery Effluents by Aspergillus sp Biomass, Res. J. of Environ. Sci,, 4, 1-12, 2009.
[27] A. Jamil, S. Umar, M. Salman, J. M. Anzano, "Removal of chromium (III) by using coal as adsorbent", J. of Hazard Mater, 171 (1-3), 797-801. 2009. Doi: 0.1016/j.jhazmat.2009.06.076.
[28] L. J. Yu, S.S. Shukla, K.L. Dorris, A. Shukla, J. L. Margrave, "Adsorption of chromium from aqueous solutions by maple sawdust", J. of Hazard Mater, 100 (1–3), 53–63, 2003.
[29] A, Özer, D. Özer, A. Özer, "The adsorption of copper (II) ions on to dehydrated wheat bran (DWB): Determination of the equilibrium and thermodynamic parameters" Process Biochem, **39**, 2183– 2191, 2004.
[30] B. Meroufel, O. Benali, M. Benyahia, Y. Benmoussa, M.A. Zenasni, "Adsorptive removal of anionic dye from aqueous solutions by Algerian kaolin: Characteristics, isotherm, kinetic and thermodynamic studies", J. Mater. Environ. Sci. **4** (3): 482-491, (2013).
[31] M. Salem, MSM H. Sharawy, M. E. Ossman, "Kinetics modeling and Adsorption isotherm studies for Cr(III) removal using Boehmite Nano-powder", IJCBS, **3**: 9-18, 2013.
[32] D. Das, M. K. Sureshkumar, K. J. Radhakrishnan, "Adsorptive removal of Cr(III) from aqueous solution using tripolyphosphate cross-linked chitosan beads", Radioanal Nucl Chem, **289**, 275, 2011. Doi:10.1007/s10967-011-1074-2.
[33] J. V. Flores-Cano, R. Leyva-Ramos, F. Carrasco-Marin, "Adsorption mechanism of Chromium(III) from water solution on bone char: effect of operating conditions", Adsorption, **22**, 297, 2016. Doi:10.1007/s10450-016-9771-3.
[34] S. B. Lyubchik, I. I.Perepichka, O. L. Galushko, A. I. Lyubchik, E. S. Lygina, I. M. Fonseca, "Optimization of the Conditions for the Cr (III) Adsorption on Activated Carbon", Adsorption, **11**, 581, 2005. Doi:10.1007/s10450-005-5616-1.
[35] N. Torapava, A. Radkevich, D. Davydov, A.Titov, I. Persson, "Composition and Structure of Polynuclear Chromium(III) Hydroxo Complexes", *Inorg. Chem., **48** (21):* 10383–10388, 2009. DOI: 10.1021/ic901539g
[36] J. Dai, F. L. Ren, C. Y. Tao, "Adsorption of Cr(VI) and Speciation of Cr(VI) and Cr(III) in Aqueous Solutions Using Chemically Modified Chitosan", Int. J. Environ. Res. Public Health, **9**, 1757-1770, 2012. Doi:10.3390/ijerph9051757.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1665-1669 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

A prospective design of a social assistive electronic system for the elderly

Arsénio Reis[*] [1], Hugo Paredes[1], Isabel Barroso[2], Maria João Monteiro[2], Vitor Rodrigues[2], Salik Ram Khanal[1], João Barroso[1]

[1]*INESC TEC and Universidade de Trás-os-Montes e Alto Douro, Vila Real, Portugal*

[2]*CIDESD and Universidade de Trás-os-Montes e Alto Douro, Vila Real, Portugal*

ARTICLE INFO

Article history:
Received: 21 June, 2017
Accepted: 28 July, 2017
Online: 24 August, 2017

Keywords:
Wellbeing
Elderly
Ambient Assisted Living
Human Computer Interaction

ABSTRACT

Aging is a natural process that progressively introduces limitations in a person's life, which can have dramatic effects on the person's lifestyle and wellbeing and in most cases is related to the strength of the person's social bonds with the family and friends group. Therefore, it is important to maintain these bonds in the various stages of life, especially in the later life, when the individual has health limitations and may be living in a care centre. In this work, we developed an exploratory approach to the usage of ICT systems in order to autonomy assist the elderly in maintaining their social connections and relationships with family and friends. It doesn't substitutes the human care, but it should assist and encourage the elderly to preserve and develop their social activities and the relationships with their family and friends social group. We propose the design of a software system, capable of running in autonomous devices, such as, robots and other consumer appliances, enabling them to recognize and interact with the users, according to their state of mind and the specific current moment context. The interactions should be simple and focused on keeping the users engaged and informed about the current life events of their family and friends, and create the opportunities for the users to interact with the participants of those events by creating interaction proposals. On a technical level, the system should have knowledge about the user and be able to acquire and update context information from social media, video cameras, email, etc, regarding the user and the persons from his family and friends groups, in order to develop meaningful interactions with the user. An adaptive user interface would present and manage the interactions, acting as a mediator between the user and his family and friends group.

1. Introduction

This paper is an extension of a work originally presented in the 2017 International Conference on Technology and Innovation on Sports, Health and Wellbeing [1].

In Portugal, it was estimated that by 2016 there should be a ratio of three elderly people per youth and an average longevity of 81 years. In the last 2011 Census, this trend was confirmed, with a percentage of 19% of the population at or over 65 years old [2]. This increase longevity is a trend of the modern societies, from the second half of the twentieth century, which are facing a substantial growth of the aging population, conducting to new and relevant economic, social and political challenges

One of these challenges regards the sustainability of the health and social security systems, which have to deal with old and dependent people, living alone. In fact, about 60% of the elderlies is living alone or in the exclusive company of other, also elderly people. The elderly dependency ratio has been increasing gradually and, in 2012, it was at 28.8%, which corresponds to more than a quarter of the active population, and positions the elderly people in a category of individuals associated with isolation, loneliness, disease, poverty and even social exclusion [3][4].

In this context, the older people group is a social and economic vulnerable group, suffering from losses at the psychic and physical levels, which inevitably introduces modifications on their social relations and increases the risk of isolation and loneliness. On this regard, the World Assembly on Ageing under the theme "Society for all ages" refers to the importance of the full realization of the right of all people to enjoy the maximum physical and mental health [5].

To address these modifications and changes that occur in the family and in the individual's social interaction patterns, it is necessary to research alternative solutions for new models of

[*]Corresponding Author: Arsénio Reis, INESC TEC and Universidade de Trás-os-Montes e Alto Douro, Vila Real, Portugal | Email: ars@utad.pt

aging, in which the individual retains emotional, physical, mental and social stability. This approach requires a new strategy for healthy aging, promoting intergenerational cooperation, community cohesion and social inclusion. During the aging process, the elderly individuals should maintain healthy social relationships in order to minimize psychological decline, which is particularly serious in age related institutionalized persons [6]. Becoming elder shouldn't introduce major changes in the individual's social role, otherwise, there will be great difficulties and the person, will struggle to adapt to the changes. Instead, the person should be active and engaged in stimulating activities [7].

Aging and the evolutions of the elderly's social roles is setting the tone for the development of specific devices and services that will support the user in the later phases of the aging process, particularly in regard to quality of life, maintaining social and family interactions, reducing the risk of loneliness and loss of social status in institutionalized elderly. Without detracting from the importance of the care, as provided by professionals and the value of humanized personal relationships, there is the need for additional aid, and the solutions based on social robotics are becoming a specialized technological branch, which can support the interactive dimension of the care process [8] [9]..

2. Assisting social activity

The process of an individual becoming older or aging encompassed physical, psychological and social changes. It is one of the highest risk factors known for most human diseases and it will introduce limitations to the individual life style [10]. The reduction in physical and social activity will rise these risk factors and degrades the aging process, causing or increasing suffering to the individual and to his family and friends [11]. Most times, the individual well-being is related to the strength of the social bonds with the family and friends group, so it is important to keep this bonds in the various stages of life, especially in the later life, when the individual has health limitations and is living in a community centre [12].

In a study regarding the older people active aging perception, the authors found that the most common perception were related to: maintaining physical health and functioning (43%), leisure and social activities (34%), mental functioning and activity (18%) and social relationships and contacts (15%) [13]. This study reveals that the maintenance of the social and mental activity accounts for the 57% of the perceive aging and is an important issue regarding successful aging and the Later Life or eight stage of life [14].

On another recent study, conducted on several Portuguese care centers, the users (elderlies) were interviewed regarding their motivation to use information and communication technologies (ICT), including artificial intelligent personal assistants as companions and assistive mediators. It was concluded that in this group, elderly people living in a care center, there is a strong will and open mind to try to use new technologies and new forms of interaction. In fact, this population is composed mostly by old people (more than 70 years old), suffering from some type of chronic disease and loneliness related problems (95%). The health issues also manifest as disabilities (75%), mostly causing mobility problems (53%) and sight problems (40%). The adoption of social robots or assistants would be welcome as well as other assistive technologies [15]. These conclusions were surprising, in the sense that the literature suggests that the usage of ICT by older adults is complex and extends beyond the dichotomy between 'successful users' and 'unsuccessful non-users' [16]. On closer look, the successful adoption of ICT solutions by older people is related to attitudes, experience of use, and perceived benefits [17], with many studies emphasizing the potential benefits of the use of ICT for elderly people, particularly by allowing them to stay connected to what is happening around them and to interact with others [18][19][20]. This interaction possibility is particularly important, if not dramatic, for people living in a care centre, which might explain the motivation to use ICT solutions.

The Ambient Assisted Living embraces the concept of helping senior citizens to improve their quality of life and expand their participation in the community by using systems to overcome their special needs, based on ubiquitous sensing, individual-environment interaction, context awareness, and learning information systems [15]. The European overview report of Ambient Assisted Living (AAL) researches this concept and aims to find out an efficient solution to help these elderly people independently living [21]. The usage of ICT has been researched as a tool to support the therapeutic process, in which the main actors are the physicians [22]. The usage by the patients themselves is a new approach and requires the patient ability to properly use a system [23].

3. Autonomous assistance in social activities

This prospective proposal has the objective of design and implement a system to monitor and support users with age related limitations, in order to promote their daily well-being. Most times, these users are living in an institution and have health limitations affecting their social activity. They also have limitations on using modern information and communication based equipment. This systems aims to overcome those limitations by promoting and reinforcing the bonds between the user and the elements of his family and friends group, with the usage of information and communication technologies.

The software system should enable autonomous equipment to recognize and interact with the users, according to their state of mind and the specific momentary context. The main features should provide the ability to identify the users and have a set of interactions to inform the users about the everyday life of their family and friends group. It should promote interactions with them in order to mitigate the tendency for social isolation and its negative consequences.

Due to the user's limitations, the system's user interface should be designed by following the adaptive interfaces paradigm [24], in accordance with the user profile and the context. It should use simple interactions, based on natural language [25], in order to be user friendly to the expected age group.

The system is meant to be installed in autonomous equipment, such as autonomous robots or consumer appliances, and should be tested in elderly care institutions, such as social support centres, where users are admitted as residents. Specific testing plans should be designed as intervention plans suited to test the equipment and the software in order to evaluate their suitability and feedback the system development project [26].

This assistive system should forge a new type of relationship with the users by maximizing their social activity and maintenance of the relationship with their family and friends social group, which is often degraded due to the user's limitations and institutionalization.

4. Methods

The teams working on the system development should have two distinct backgrounds: software engineering; and health care. The work should be done in short, interactive, cycles, in which the healthcare team should validate the requirements; verify the usability of the developed features; and test the features with real users. The software engineering team should develop the system using the state of the art technologies, following a user centric development method in order to meet all the usage requirements, related to the user's limitations [27][28][29][30].

The development and implementation of this system should have two main phases. A first phase, focused on software development and a second phase, focused on the field tests

In the first phase it should be designed and developed the system software, with the necessary features in order to accomplish the main objectives. The main areas of software development will be related to:

- Image acquisition and analysis, for the users identification and state of mind recognition;
- Application design with databases and social media APIs, for the development of the user activities;
- Natural langue and adaptive user interfaces, for the development of the system's user interface;
- Remote equipment management, for the planning and execution of the equipment deployment during the field tests.

The software development should be executed using agile methods in an incremental process. In this phase, the healthcare professionals should be part of the team in order to create the requirements and specifications as well as to monitor the development results.

On a second phase, the focus should be on the field tests and the monitoring of the system usage by the real users. Some development will still be necessary in order to adjust and develop some features, according to the users monitoring feedback. In this phase, a strong effort should be done to monitor the users in different usage contexts.

The autonomous equipment should be adapted for usage in the testing institutions, including the necessary adjustments for the remote management of the equipment.

5. Requirements and features

The system should provide a set of basic functionalities, features or requirements, related to the interaction with the users, namely: user identification; state of mind assessment; current context; personal information retrieval; a set of activity proposals.

The user identification and state of mind evaluation will be based on image acquisition and analysis, after which the system should retrieve the user profile and decide on an activity proposal. The activity proposal could be as simple as a personalized greeting or a more complex interaction, according to the current context and the user state of mind.

The context, personal information and state of mind, should include information about: gender, age, special needs, personal preferences, time of day, time of year, physical location, state of mind (sadness, joy, etc.), etc.

The interactions with the user should be developed as a set of activities that can be proposed and executed with the user. Figure 1 and Figure 2 have sequence diagrams illustrating two interactions. Some activities to be developed are: (1) Basic greeting; (2) Email management; (3) Social Network management; (4) Social and family events management; and (5) Social games.

- The basic greeting activity is a simple greeting interaction, in which the system greets the user according to the state of mind and context of the moment.
- The email management activity should provide a personal email box to each user and inform the user about the current messages status. In particular dates, such as family anniversaries, the system should offer the user a chance to send a congratulations message.
- The social media management is a more sophisticated activity in which the system monitors the user group of family and friends, regarding their activity in the social networks, e.g., facebook, twitter, Instagram, etc., and informs the user about relevant updates.
- The social games activity has the objective of adapt and create a set of social games, e.g., cards, trivia, etc., that the user can play in the system.

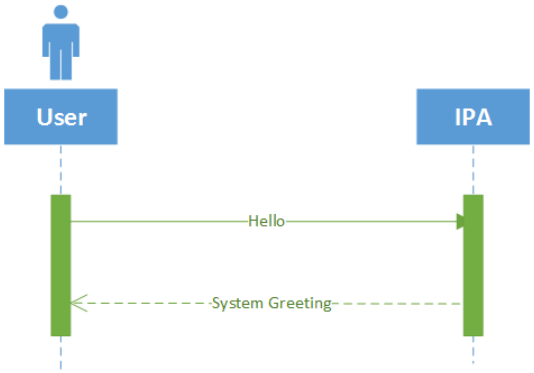


Figure 1 - Sequence diagram of basic greeting activity

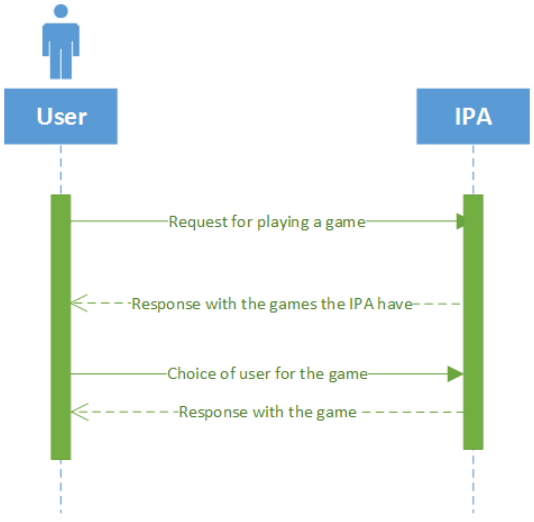


Figure 2 - Sequence diagram for social game activity.

The email and social media management are specially targeted to strengthen the user bonds with the elements of his family and friends. The user should be kept up to date with the course of lives of his friends and family and the systems should provide

opportunities for the user to interact with those person on the social media networks. In Figure 3, it is described the basic workflow of an interaction with a user. As previously described, the system engages with the user and based on the available data about the user and the current context, it proposes an activity in order to forward an interaction.

The set of features should be extensively tested with real users in order to evaluate their tangible contribution to the main objective and their usability, considering the user interface and the user's limitations. So, a series of interventions plans should be created in order to test the system and the equipment in several real scenarios, together with partner institutions. For each intervention, a specific plan should be created, considering the specific conditions such as: the user's characteristics; the context of the hosting institution; the technical resources available.

6. The current consumer electronic assistants

Currently there are on the market several electronic assistants, e.g., Amazon Alexa, Google Assistant, Apple Siri, Microsoft Cortana, that are able to assist the user in some daily tasks. Some of this companies also provide hardware appliances uniquely designed to house their assistant software. In TABLE I are summarized the main features of this assistants.

Table I – Electronic assistants main features

Interactions	Google Assistant	Amazon Alexa	Apple Siri	Microsoft Cortana
Create a shopping list	yes	yes	no	no
Check Calendar	yes	yes	yes	yes
Manage Calendar	no	yes	yes	yes
Check date and time	yes	yes	yes	yes
Set an Alarm	yes	yes	yes	yes
Check Weather	yes	yes	yes	yes
Check Traffic	yes	yes	yes	yes
Play news	yes	yes	no	no
Play music	yes	yes	yes	yes
Search Online	yes	yes	yes	yes
Perform Calculations	yes	yes	yes	yes
Call a Uber	no	yes	yes	no
Control home temperature and lights	yes	yes	yes	no
Lock/Unlock Doors	no	no	yes	no
Play games	yes	yes	yes	yes
Make a call	yes	yes	yes	yes
Read books	no	yes	no	no
Shopping on amazon	no	yes	no	no
Order pizza	no	yes	yes	yes
Manage SMS	yes	yes	yes	yes
Read emails	no	no	yes	yes
Send emails	no	no	yes	no
Send posts in Facebook	yes	yes	yes	no
Send posts in Twitter	yes	yes	yes	
Receive Facebook notifications	no	no	no	no
Receive Twitter notifications	no	no	no	no

It is a very promising set of interactions, but on some preliminary tests we concluded that there is a gap between what the assistants can do and what they effectively do when confronted with a user expecting to execute an interaction. We cannot categorically state that the assistants can't execute the actions, but effectively it would be very difficult for an elderly to use them in order to maintain some sort of social interaction with their friends and families. The preliminary testing suggest that these assistants lack the empathy and common sense engagement necessary to interact with an elderly. In future work we expect to measure the quality of the interaction and assess the effectiveness of the interaction.

7. Conclusion

In Figure 3 it is presented the general system workflow, regarding an autonomous user interaction. Although it is a simple workflow, the strict software implementation depends of a diverse set of concepts and technologies. The field tests and implementation, with real users in real institutions, are a main care endeavour and its constant feedback is determinant to the success of the system development.

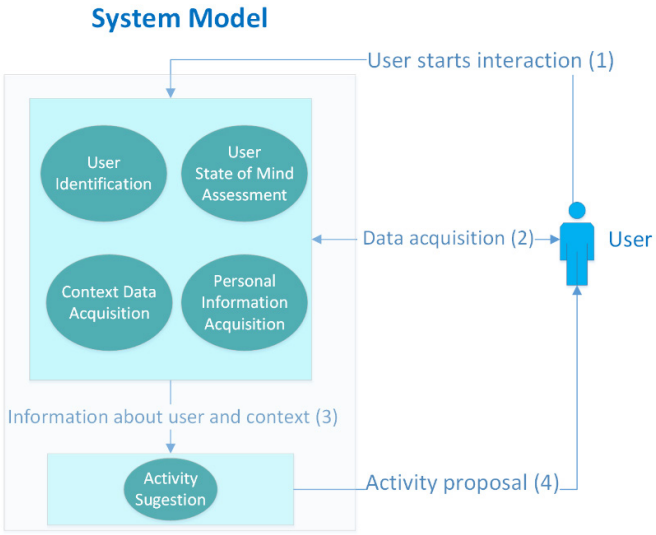


Figure 3 – System workflow

The success of this proposal is dependent of two main fiends of study: software engineering and elderly care. In such a sensible context, of elderly users with age related limitations, the subtle details of the system, in particular the user interface usability, are paramount and an excellent multidisciplinary team management is necessary. In TABLE I is summarized the features and the related research areas that the software system combines.

Table II - Features and research areas

Features	Research areas
User identification	Image processing (face recognition)
State of mind assessment	Image processing (emotion detection)
Context acquisition	On-line services APIs Localization services Image proecessing (Object detection and recognition)
Data retrieval	Databases
Social media related activities	Social media APIs
Social games	Gaming and Serious games
Natural language user interface	Adaptive interfaces Natural language

The success of the features implementation must be assured by a user centric development with the continuous participation of the health care team elements. They should also act on behalf of the users in order to assure a user participated development.

This system proposal is intended for usage with autonomous systems such as robot devices, but it can be adapted and incorporated in other user appliances, such as settop tv boxes or even mobile phones. We didn't refer to a system architecture, but in a real system, with an extensive catalogue of user interactions and activities proposals, it should be considered the usage of a cloud backend implementation and a user device front end.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

This work was supported by Project "NanoSTIMA: Macro-to-Nano Human Sensing: Towards Integrated Multimodal Health Monitoring and Analytics/NORTE-01-0145-FEDER-000016" financed by the North Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, and through the European Regional Development Fund (ERDF).

References

[1] Arsenio Reis, Hugo Paredes, Isabel Barroso, Maria Joao Monteiro, Vitor Rodrigues, Salik Ram Khanal, Joao Barroso: Autonomous systems to support social activity of elderly people a prospective approach to a system design. 2016 1st International Conference on Technology and Innovation in Sports, Health and Wellbeing (TISHW); 12/2016, DOI:10.1109/TISHW.2016.7847773

[2] Instituto Nacional de Estatística (2011). Censos 2011– Resultados Provisórios. http://censos.ine.pt/xportal/xmain?xpid=CENSOS&xpgid=ine_censos_publicacao_det&contexto=pu&PUBLICACOESpub_boui=122073978&PUBLICACOESmodo=2&selTab=tab1&pcensos=61969554

[3] Instituto Nacional de Estatística (2012). O Envelhecimento em Portugal. Situação demográfica e socio-económica recente das pessoas idosas, Serviço de Estudos sobre a População do Departamento de Estatísticas Censitárias e da População.

[4] Fernandes, A. (2001). Velhice, solidariedades familiares e política social: itinerário de pesquisa em torno do aumento da esperança de vida. Sociologia, Problemas e Práticas [online], n.36, pp.39-52.

[5] Conselho Económico e Social (2013). Parecer de iniciativa sobre as consequências económicas, sociais e organizacionais decorrentes do envelhecimento da população. Lisboa:

[6] Palmer, A. D.; Newsom, J. & Rook, K. (2016). How does difficulty communicating affect the social relationships of older adults? An exploration using data from a national survey. Journal of Communication Disorders, 62:131-146.

[7] Cabral, M. (coord.) (2013). Processos de Envelhecimento em Portugal. Lisboa: Fundação Francisco Manuel dos Santos.

[8] Reza K., Sima S., & Mei-Tai Chu (2014). Socially Assistive Robots in Elderly Care: A Mixed-Method Systematic Literature Review. Intl. Journal of Human–Computer Interaction, 30: 369–393.

[9] Tânia Rocha, Hugo Fernandes, Arsénio Reis, Hugo Paredes, João Barroso: Assistive Platforms for the Visual Impaired: Bridging the Gap with the General Public. Recent Advances in Information Systems and Technologies, 03/2017: pages 602-608; , ISBN: 978-3-319-56537-8, DOI:10.1007/978-3-319-56538-5_61

[10] Dillin, Andrew, Daniel E. Gottschling, and Thomas Nyström. "The good and the bad of being connected: the integrons of aging." Current opinion in cell biology26 (2014): 107-112.

[11] Rowe, J. W., & Kahn, R. L. (1987). Human aging: usual and successful.Science, 237(4811), 143-149.

[12] Rook, K. S. (1984). The negative side of social interaction: impact on psychological well-being. Journal of personality and social psychology, 46(5), 1097.

[13] Bowling, A. (2008). Enhancing later life: how older people perceive active ageing?. Aging and Mental Health, 12(3), 293-301.

[14] Fisher, B. J. (1995). Successful aging, life satisfaction, and generativity in later life. The International Journal of Aging and Human Development, 41(3), 239-250.

[15] Sun, H., De Florio, V., Gui, N., & Blondia, C. (2009, April). Promises and challenges of ambient assisted living systems. In Information Technology: New Generations, 2009. ITNG'09. Sixth International Conference on (pp. 1201-1207). Ieee.

[16] Selwyn, N., Gorard, S., Furlong, J., & Madden, L. (2003). Older adults' use of information and communications technology in everyday life. Ageing and society, 23(05), 561-582.

[17] Arsénio Reis, Dennis Paulino, Hugo Paredes, João Barroso: Using Intelligent Personal Assistants to Strengthen the Elderlies' Social Bonds. Universal Access in Human–Computer Interaction. Human and Technological Environments, 01/2017: pages 593-602; , ISBN: 978-3-319-58699-1, DOI:10.1007/978-3-319-58700-4_48

[18] Hernández-Encuentra, E., Pousada, M., & Gómez-Zúñiga, B. (2009). ICT and older people: Beyond usability. Educational Gerontology, 35(3), 226-245.

[19] Weatherall, J. W. A. (2000). A grounded theory analysis of older adults and information technology. Educational Gerontology, 26(4), 371-386.

[20] Magnusson , L. , Hanson , E. , Birto , L. ,Berthold , H. , Chambers , M. , & Daly ,T. (2002). Supporting family carers through the use of information and communication technology—the EU project ACTION . International Journal of Nursing Studies , 39 , 369 – 381 .

[21] Steg, H., & Strese, H. (2005). Ambient assisted living–european overview report.

[22] Reis, A., Lains, J., Paredes, H., Filipe, V., Abrantes, C., Ferreira, F., . . . Barroso, J. (2016). Developing a System for Post-Stroke Rehabilitation: An Exergames Approach. In M. Antona & C. Stephanidis (Eds.), Universal Access in Human-Computer Interaction. Users and Context Diversity: 10th International Conference, UAHCI 2016, Held as Part of HCI International 2016, Toronto, ON, Canada, July 17-22, 2016, Proceedings, Part III (pp. 403-413). Cham: Springer International Publishing.

[23] Paredes, H., Reis, A., Marinho, S., Lains, J., Filipe, V., & Barroso, J. (2016). A study of ICT active monitoring adoption in stroke rehabilitation - Proceedings of the 3rd IPLeiria's International Health Congress. BMC Health Services Research, 16(3), 111-242. doi: 10.1186/s12913-016-1423-5.

[24] Sa-nga-ngam , P. , & Kurniawan , S. (2006). An investigation into older people's browsing activities . Psychology Journal , 4 (3), 245 – 265 .

[25] Bobillier Chaumon, M. E., Michel, C., Tarpin Bernard, F., & Croisile, B. (2014). Can ICT improve the quality of life of elderly adults living in residential home care units? From actual impacts to hidden artefacts. Behaviour & Information Technology, 33(6), 574-590.

[26] Arsénio Reis, Isabel Barroso, Maria João Monteiro, Salik Khanal, Vitor Rodrigues, Vitor Filipe, Hugo Paredes, João Barroso: Designing Autonomous Systems Interactions with Elderly People. Universal Access in Human–Computer Interaction. Human and Technological Environments, 01/2017: pages 603-611; , ISBN: 978-3-319-58699-1, DOI:10.1007/978-3-319-58700-4_49

[27] Jameson, A. (2009). Adaptive interfaces and agents. Human-Computer Interaction: Design Issues, Solutions, and Applications, 105, 105-130.

[28] Liddy, E. D., Paik, W., McKenna, M. E., Weiner, M. L., Edmund, S. Y., Diamond, T. G., ... & Snyder, D. L. (2000). U.S. Patent No. 6,026,388. Washington, DC: U.S. Patent and Trademark Office.

[29] Gulliksen, J., Göransson, B., Boivie, I., Blomkvist, S., Persson, J., & Cajander, Å. (2003). Key principles for user-centred systems design. Behaviour and Information Technology, 22(6), 397-409.

[30] Beck, K., Beedle, M., Van Bennekum, A., Cockburn, A., Cunningham, W., Fowler, M., ... & Kern, J. (2001). Manifesto for agile software development.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 2, No. 3, 1670-1676 (2017)
www.astesj.com
Special Issue on Recent Advances in Engineering Systems

ASTES Journal
ISSN: 2415-6698

Theoretical developments for interpreting kernel spectral clustering from alternative viewpoints

Diego Peluffo-Ordóñez [*1], Paul Rosero-Montalvo[1,2], Ana Umaquinga-Criollo[1],Luis Suárez-Zambrano[1], Hernan Domínguez-Limaico[1], Omar Oña-Rocha[1], Stefany Flores-Armas[1], Edgar Maya-Olalla[1]

[1] *Universidad Técnica del Norte, Facultad de Ingeniería en Ciencias Aplicadas, 100150, Ecuador*
[2] *Instituto Tecnológico Superior 17 de Julio, Yachay, Ecuador*

ARTICLEINFO

Article history:
Received: 03 April, 2017
Accepted: 23 July, 2017
Online: 27 August, 2017

Keywords:
Kernel
Spectral Clustering
Support Vector Machines

ABSTRACT

To perform an exploration process over complex structured data within unsupervised settings, the so-called kernel spectral clustering (KSC) is one of the most recommended and appealing approaches, given its versatility and elegant formulation. In this work, we explore the relationship between (KSC) and other well-known approaches, namely normalized cut clustering and kernel k-means. To do so, we first deduce a generic KSC model from a primal-dual formulation based on least-squares support-vector machines (LS-SVM). For experiments, KSC as well as other consider methods are assessed on image segmentation tasks to prove their usability.

1 Introduction

In general, for classifying or grouping a set of objects (represented as data points) into subsets holding similar objects, the field of machine learning - specifically, the pattern recognition- provides two great alternatives being essentially different from each other: Supervised- and Unsupervised-learning-based approaches. The former ones normally establish a model from beforehand known information on data normally provided by an expert, while the latter ones form the groups by following a natural clustering criterion based on a (traditionally heuristic) procedure of data exploration [1]. Therefore, unsupervised clustering techniques are preferred when object labelling is either unavailable or unfeasible. In literature, we can find tens of clustering techniques, which are based on different principles and criteria (such as: distances, densities, data topology, and divergences, among others) [2]. Some remarkable, emerging applications are inbalanced data analysis [3] and time-varying data analysis [4]. Particularly, spectral clustering (SC) is a suitable technique to deal with grouping problems involving hardly separable clusters. Many SC approaches have been proposed, among them: Normalized-cut-based clustering (NCC), which, applied as explained in [5], heuristically and iteratively estimates binary cluster indicators [5] or approximates the solution in a one-iteration fashion by solving a quadratic programming problem [6]. Kernel k-means (KKM) that can be formulated using eigenvectors [7]. Kernel spectral clustering (KSC), which uses a latent variable model and a least-squares-support-vector-machine (LS-SVM) formulation [8]. This work has a particular focus on KSC, being one of the most modern approaches. It has been widely used in numerous applications such as time-varying data [9,10], electricity load forecasting [11], prediction of industrial machine maintenance [12], and among others. Also, some improvements and extensions have been proposed [12–14].

The aim of this work is to demonstrate the relationship between KSC and other approaches, namely NCC and KKM. To do so, elegant mathematical developments are performed. Starting from either the primal or dual formulation of KSC, we show clearly the links with the other considered methods. Experimentally, in order to assess the clustering performance, we explore the benefit of each considered method on image segmentation. In this connection, images extracted from the free access Berkeley Segmentation Data Set [15] are used. As a meaningful result of this work, Also, we provide mathematical and experimental evidence of the usability of combining together a LS-SVM formulation and a generic latent variable model for clustering purposes.

The rest of this paper is organized as follows: Section 2 outlines the primal-dual formulation for KSC starting with a LS-SVM formulation regarding a variable model, which naturally yields an eigenvector-

[*] Diego Peluffo, , Contact No (+593)991250815 & dhpeluffo@utn.edu.ec

https://dx.doi.org/10.25046/aj0203208

based solution. Sections 3 and 4 explore and discusses on the links of KSC with NCC and KKM, respectively. Some experimental results are shown in Section 5. Section 6 presents some additional remarks on improved versions of KSC and its relationship with spectral dimensionality reduction. Finally, section 7 draws some final and concluding remarks.

2 Kernel Spectral clustering

Accounting for notation and future statements, let us consider the following definitions: Define a set of N objects or samples represented by d-dimensional feature vectors. Likewise, consider a data matrix holding all the feature vectors, so that $\mathbf{X} \in \mathbb{R}^{N\times d}$: $\mathbf{X} = [\mathbf{x}_1^\top, \ldots, \mathbf{x}_N^\top]^\top$, where $\mathbf{x}_i \in \mathbb{R}^d$ is the i-th d dimensional feature vector or data point. KSC is aiming to split $\mathbf{X}$ into K disjoint subsets, being K the number of desired groups.

2.1 Latent variable model and problem formulation

In the following, the clustering model is described. Let $\mathbf{e}^{(l)} \in \mathbb{R}^N$ be the l-th projection vector, which is assumed in the following latent variable form:

$$\mathbf{e}^{(l)} = \boldsymbol{\Phi}\mathbf{w}^{(l)} + b_l \mathbf{1}_N, \tag{1}$$

where $\mathbf{w}^{(l)} \in \mathbb{R}^{d_h}$ is the l-th weighting vector, b_l is a bias term, n_e is the number of considered latent variables, notation $\mathbf{1}_N$ stands for a N dimensional all-ones vector, and the matrix $\boldsymbol{\Phi} = [\phi(\mathbf{x}_1)^\top, \ldots, \phi(\mathbf{x}_N)^\top]^\top$, $\boldsymbol{\Phi} \in \mathbb{R}^{N\times d_h}$, is a high dimensional representation of data. The function $\phi(\cdot)$ maps data from the original dimension to a higher one d_h, i.e., $\phi(\cdot) : \mathbb{R}^d \to \mathbb{R}^{d_h}$. Therefore, $\mathbf{e}^{(l)}$ represents the latent variables from a set of n_e binary cluster indicators obtained with $\text{sign}(\mathbf{e}^{(l)})$, which are to be further encoded to obtain the K resultant groups.

From the least-squares SVM formulation of equation (1), the following optimization problem can be stated:

$$\max_{\mathbf{e}^{(l)},\mathbf{w}^{(l)},\mathbf{b}^{(l)}} \frac{1}{2N}\sum_{l=1}^{n_e} \gamma_l \mathbf{e}^{(l)\top}\mathbf{V}\mathbf{e}^{(l)} - \frac{1}{2}\sum_{l=1}^{n_e} \mathbf{w}^{(l)\top}\mathbf{w}^{(l)} \tag{2a}$$

$$\text{s.t. } \mathbf{e}^{(l)} = \boldsymbol{\Phi}^\top \mathbf{w}^{(l)} + b_l \mathbf{1}_N, \tag{2b}$$

where $\gamma_l \in \mathbb{R}^+$ is the l-th regularization parameter and $\mathbf{V} \in \mathbb{R}^{N\times N}$ is a diagonal matrix representing the weight of projections.

2.2 Matrix problem formulation

For the sake of simplicity, we can express the primal formulation (2) in matrix terms, as follows:

$$\max_{\mathbf{E},\mathbf{W},\mathbf{b}} \quad \frac{1}{2N}\text{tr}(\mathbf{E}^\top\mathbf{V}\mathbf{E}\Gamma) - \frac{1}{2}\text{tr}(\mathbf{W}^\top\mathbf{W}) \tag{3a}$$

$$\text{s.t. } \mathbf{E} = \boldsymbol{\Phi}\mathbf{W} + \mathbf{1}_N \otimes \mathbf{b}^\top, \tag{3b}$$

where $\mathbf{b} = [b_1, \ldots, b_{n_e}]$, $\mathbf{b} \in \mathbb{R}^{n_e}$, $\Gamma = \text{Diag}([\gamma_1, \ldots, \gamma_{n_e}])$, $\mathbf{W} = [\mathbf{w}^{(1)}, \cdots, \mathbf{w}^{(n_e)}]$, $\mathbf{W} \in \mathbb{R}^{d_h\times n_e}$, and $\mathbf{E} = [\mathbf{e}^{(1)}, \cdots, \mathbf{e}^{(n_e)}]$, $\mathbf{E} \in \mathbb{R}^{N\times n_e}$. Notations $\text{tr}(\cdot)$ and $\otimes$ denote the trace and the Kronecker product, respectively. By minimizing the previous cost function, the goals of minimizing the weighting variance of $\mathbf{E}$ and maximizing the variance of $\mathbf{W}$ are reached simultaneously. Let $\Sigma_\mathbf{E}$ be the weighting covariance matrix of $\mathbf{E}$ and $\Sigma_\mathbf{W}$ be the covariance matrix of $\mathbf{W}$. Since matrix $\mathbf{V}$ is diagonal, we have that $\text{tr}((\mathbf{V}^{1/2}\mathbf{E})^\top\mathbf{V}^{1/2}\mathbf{E}) = \text{tr}(\Sigma_\mathbf{E})$. In other words, $\Sigma_\mathbf{E}$ is the covariance matrix of weighted projections, i.e., the projections scaled by square root of matrix $\mathbf{V}$. As well, $\text{tr}(\mathbf{W}^\top\mathbf{W}) = \text{tr}(\Sigma_\mathbf{W})$. Then, KSC can be seen as a kernel, weighted principal component analysis (KWPCA) approach [8].

2.3 Solving KSC by using a dual formulation

To solve the KSC problem, we form the corresponding Lagrangian of the problem from equation (2) as follows:

$$\begin{aligned}\mathcal{L}(\mathbf{E},\mathbf{W},\Gamma,\mathbf{A}) = &\frac{1}{2N}\text{tr}(\Gamma\mathbf{E}^\top\mathbf{V}\mathbf{E}) - \frac{1}{2}\text{tr}(\mathbf{W}^\top\mathbf{W}) \\ &- \text{tr}(\mathbf{A}^\top(\mathbf{E} - \boldsymbol{\Phi}\mathbf{W} - \mathbf{1}_N \otimes \mathbf{b}^\top)),\end{aligned} \tag{4}$$

where matrix $\mathbf{A} \in \mathbb{R}^{N\times n_e}$ holds the Lagrange multiplier vectors $\mathbf{A} = [\alpha^{(1)}, \cdots, \alpha^{(n_e)}]$, and $\alpha^{(l)} \in \mathbb{R}^N$ is the l-th vector of Lagrange multipliers.

Solving the partial derivatives on $\mathcal{L}(\mathbf{E},\mathbf{W},\Gamma,\mathbf{A})$ to determine the Karush-Kuhn-Tucker conditions, we obtain:

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\mathbf{E}} = 0 \quad &\Rightarrow \mathbf{E} = N\mathbf{V}^{-1}\mathbf{A}\Gamma^{-1}, \\ \frac{\partial\mathcal{L}}{\partial\mathbf{W}} = 0 \quad &\Rightarrow \mathbf{W} = \boldsymbol{\Phi}^\top\mathbf{A}, \\ \frac{\partial\mathcal{L}}{\partial\mathbf{A}} = 0 \quad &\Rightarrow \mathbf{E} = \boldsymbol{\Phi}\mathbf{W}, \\ \frac{\partial\mathcal{L}}{\partial\mathbf{b}} = 0 \quad &\Rightarrow \mathbf{b}^\top\mathbf{1}_N = 0.\end{aligned}$$

Therefore, by eliminating the primal variables from initial problem (2) and assuming a kernel trick such that $\boldsymbol{\Phi}\boldsymbol{\Phi}^\top = \boldsymbol{\Omega}$, being $\boldsymbol{\Omega} \in \mathbb{R}^{N\times N}$ a given kernel matrix, the following eigenvector-based dual solution is obtained:

$$\mathbf{A}\Lambda = \mathbf{V}(\mathbf{I}_N + (\mathbf{1}_N \otimes \mathbf{b}^\top)(\boldsymbol{\Omega}\Lambda)^{-1})\boldsymbol{\Omega}\mathbf{A}, \tag{5}$$

where $\Lambda = \text{Diag}(\lambda)$, $\Lambda \in \mathbb{R}^{N\times N}$, $\lambda \in \mathbb{R}^N$ is the vector of eigenvalues with $\lambda_l = N/\gamma_l$, $\lambda_l \in \mathbb{R}^+$.

Also, taking into account that the kernel matrix represents the similarity matrix of a graph with K connected components as well as $\mathbf{V} = \mathbf{D}^{-1}$ where $\mathbf{D} \in \mathbb{R}^{N\times N}$ is the degree matrix defined as $\mathbf{D} = \text{Diag}(\boldsymbol{\Omega}\mathbf{1}_N)$; then the $K-1$ eigenvectors contained in $\mathbf{A}$, associated to the largest eigenvalues, are piecewise constant and become indicators of the corresponding connected parts of the graph. Therefore, value n_e is fixed

to be $K-1$ [8]. With the aim of achieving a dual formulation, but satisfying the condition $\mathbf{b}^\top \mathbf{1}_N = 0$ by centering vector $\mathbf{b}$ (i.e. with zero mean), the bias term should be chosen in the form

$$b_l = -1/(\mathbf{1}_N^\top \mathbf{V}\mathbf{1}_N)\mathbf{1}_N^\top \mathbf{V}\boldsymbol{\Omega}\alpha^{(l)}. \quad (6)$$

Thus, the solution of problem of equation (3) is reduced to the following eigenvector-related problem:

$$\mathbf{A}\boldsymbol{\Lambda} = \mathbf{V}\mathbf{H}\boldsymbol{\Omega}\mathbf{A}, \quad (7)$$

where matrix $\mathbf{H} \in \mathbb{R}^{N\times N}$ is the centering matrix that is defined as

$$\mathbf{H} = \mathbf{I}_N - \frac{1}{\mathbf{1}_N^\top \mathbf{V}\mathbf{1}_N}\mathbf{1}_N\mathbf{1}_N^\top, \mathbf{V},$$

where $\mathbf{I}_N$ denotes a N-dimensional identity matrix and, $\boldsymbol{\Omega} = [\Omega_{ij}]$, $\boldsymbol{\Omega} \in \mathbb{R}^{N\times N}$, being $\Omega_{ij} = \mathcal{K}(\mathbf{x}_i, \mathbf{x}_j), i, j \in [N]$. Notation $\mathcal{K}(\cdot,\cdot) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ stands for the kernel function. As a result, the set of projections can be calculated as follows:

$$\mathbf{E} = \boldsymbol{\Omega}\mathbf{A} + \mathbf{1}_N \otimes \mathbf{b}^\top. \quad (8)$$

Once projections are calculated, we proceed to carry out the cluster assignment by following an encoding procedure applied on projections. Because each cluster is represented by a single point in the $K-1$-dimensional eigenspace, such that those single points are always in different orthants due also to the KKT conditions, we can encode the eigenvectors considering that two points are in the same cluster if they are in the same orthant in the corresponding eigenspace [8]. Then, a code book can be obtained from the rows of the matrix containing the $K-1$ binarized leading eigenvectors in the columns, by using $\text{sign}(\mathbf{e}^{(l)})$. Then, matrix $\widetilde{\mathbf{E}} = \text{sgn}(\mathbf{E})$ is the code book being each row a codeword.

2.4 Out-of-sample extension

KSC can be extended to out-of-samples analysis without re-clustering the whole data to determine the assignment cluster membership for new testing data [8]. In particular, defining $\mathbf{z} \in \mathbb{R}^{n_e}$ as the projection vector of a testing data point $\mathbf{x}_{\text{test}}$, and by taking into consideration the training clustering model, the testing projections can be computed as:

$$\mathbf{z} = \mathbf{A}^\top \boldsymbol{\Omega}_{\text{test}} + \mathbf{b}, \quad (9)$$

where $\boldsymbol{\Omega}_{\text{test}} \in \mathbb{R}^{n_e}$ is the kernel vector such that

$$\boldsymbol{\Omega}_{\text{test}} = [\Omega_{\text{test}_1}, \ldots, \Omega_{\text{test}_N}]^\top,$$

and $\Omega_{\text{test}_i} = \mathcal{K}(\mathbf{x}_i, \mathbf{x}_{\text{test}})$. Once, the test projection vector $\mathbf{z}$ is computed, a decoding stage is carried out that consists of comparing the binarized projections with respect to the codewords in the code book $\widetilde{\mathbf{E}}$ and assigning cluster membership based on the minimal Hamming distance [8].

2.5 KSC algorithm

Following the pseudo-code (Algorithm 1) to perform KSC is shown.

Algorithm 1 Kernel spectral clustering: $[q_{\text{train}}, q_{\text{test}}] = \text{KSC}(\mathbf{X}, \mathcal{K}(\cdot,\cdot), K)$

1: Input: K, $\mathbf{X}$, $\mathcal{K}(\cdot,\cdot)$
2: Form the kernel matrix $\boldsymbol{\Omega}$ such that $\Omega_{ij} = \mathcal{K}(\mathbf{y}_i, \mathbf{x}_j)$
3: Determine $\mathbf{E}$ through (8)
4: Form the training codebook by binarizing $\widetilde{\mathbf{E}} = \text{sgn}(\mathbf{E})$
5: Assign the output training labels q_{train} according to similar codewords
6: Compute the training codewords for testing
7: Assign the output testing labels q_{test} according to the minimal Hamming distance when comparing with training codewords
8: Output: $q_{\text{train}}, q_{\text{test}}$

3 Links between KSC and NCC

This section deals with the relationship between KSC and NCC, starting from the formulation of the NC problem until reaching a weighting principal component analysis (WPCA) formulation in a finite domain.

3.1 Multi-cluster spectral clustering (MCSC) from two point of view

In [5], the so-called Multi-cluster spectral clustering (MCSC) is introduced, which is based on the well-known k-way normalized cut-based formulation given by:

$$\max \frac{1}{K}\frac{\text{tr}(\mathbf{M}^\top \boldsymbol{\Omega}\mathbf{M})}{\text{tr}(\mathbf{M}^\top \mathbf{D}\mathbf{M})} = \max_{\mathbf{m}^{(k)}} \frac{1}{K}\frac{\sum_{k=1}^K \mathbf{m}^{(k)\top}\widehat{\boldsymbol{\Omega}}\mathbf{m}^{(k)}}{\sum_{k=1}^K \mathbf{m}^\top \mathbf{m}} \quad (10a)$$

$$\text{s.t. } \mathbf{M} \in \{0,1\}^{N\times K}, \quad \mathbf{M}\mathbf{1}_K = \mathbf{1}_N. \quad (10b)$$

Expressions (10a) and (10b) are the formulation of the NC optimization problem, named (*NCPM*). Previous formulation can also be expressed as follows. Let $\widehat{\boldsymbol{\Omega}} = \mathbf{D}^{-1/2}\boldsymbol{\Omega}\mathbf{D}^{-1/2}$ be a normalized kernel matrix and $\mathbf{L} = \mathbf{D}^{1/2}\mathbf{M}$ be a binary matrix normalized by the square root of the kernel degree. Then, a new *NCPM* version can be expressed as:

$$\max_{\mathbf{L}} \frac{1}{K}\frac{\text{tr}(\mathbf{L}^\top \widehat{\boldsymbol{\Omega}}\mathbf{L})}{\text{tr}(\mathbf{L}^\top \mathbf{L})} = \max_{\ell^{(k)}} \frac{1}{K}\frac{\sum_{k=1}^K \ell^{(k)\top}\widehat{\boldsymbol{\Omega}}\ell^{(k)}}{\sum_{k=1}^K \ell^{(k)\top}\ell^{(k)}} \quad (11a)$$

$$\text{s.t. } \mathbf{D}^{-1/2}\mathbf{L} \in \{0,1\}^{N\times K}, \quad \mathbf{D}^{-1/2}\mathbf{L}\mathbf{1}_K = \mathbf{1}_N, \quad (11b)$$

where $\ell^{(k)}$ is the column k of $\mathbf{L}$.

Solution of former problem has been addressed in [5, 16] by introducing a relaxed version, in which numerator is maximized subject to denominator is constant, so

$$\max_{\mathbf{L}} \frac{1}{K}\text{tr}(\mathbf{L}^\top \widehat{\boldsymbol{\Omega}}\mathbf{L}) \qquad \text{s.t. tr}(\mathbf{L}^\top \mathbf{L}) = \text{const.} \quad (12)$$

Indeed, authors assume the particular case $\mathbf{L}^\top\mathbf{L} = \mathbf{I}_K$, i.e. letting $\mathbf{L}$ be an orthonormal matrix. Then, solution correspond to any K-dimensional basis of normalized matrix eigenvectors. Despite that in [16] it is presented an one-iteration solution for *NCPM* with suboptimal results avoiding the calculation of SVD per iteration, the omitting of the effect of denominator $\text{tr}(\mathbf{L}^\top\mathbf{L})$ by assuming orthogonality causes that the solution cannot be guaranteed to be a global optimum. In addition, this kind of formulation provide non-stable solutions due to the heuristic search carried out to determine an optimal rotation matrix [16].

3.2 Solving the problem by a difference: Empirical feature map

Recalling original problem 3.1, we introduce another way to solve the *NCPM* formulation via a minimization problem where the aims for maximizing $\text{tr}(\mathbf{L}^\top\widehat{\mathbf{\Omega}}\mathbf{L})$ and minimizing $\text{tr}(\mathbf{L}^\top\mathbf{L})$ can be accomplished simultaneously, so:

$$\max_{\mathbf{L}} \text{tr}(\mathbf{L}^\top\widehat{\mathbf{\Omega}}\mathbf{L}\,\text{Diag}(\gamma)) - \text{tr}(\mathbf{L}^\top\mathbf{L}) \tag{13}$$

where $\gamma = (\gamma_1,\ldots,\gamma_N)^\top$ is a vector containing the regularization parameters.

Let us assume $\mathbf{\Omega} = \mathbf{\Psi}\mathbf{\Psi}^\top$ where $\mathbf{\Psi}$ is a $N \times N$ dimensional auxiliary matrix, and consider the following equality:

$$\begin{aligned}\text{tr}(\widehat{\mathbf{\Omega}}) &= \text{tr}(\mathbf{D}^{-1/2}\mathbf{\Omega}\mathbf{D}^{-1/2}) = \text{tr}(\mathbf{D}^{-1}\mathbf{\Omega}) \\ &= \text{tr}(\mathbf{D}^{-1}\mathbf{\Psi}\mathbf{\Psi}^\top) = \text{tr}(\mathbf{\Psi}^\top\mathbf{D}^{-1}\mathbf{\Psi}),\end{aligned}$$

then

$$\mathbf{D}^{-1/2}\mathbf{\Omega}\mathbf{D}^{-1/2} = \mathbf{\Psi}^\top\mathbf{D}^{-1}\mathbf{\Psi}.$$

Previous formulation is possible since kernel matrix Ω is symmetric. Now, let us define $\mathbf{h}^{(k)} \in \mathbb{R}^N = \mathbf{\Psi}^\top\ell^{(k)}$ as the k-th projection and $\mathbf{H} = (\mathbf{h}^{(1)},\cdots,\mathbf{h}^{(K)})$ as the projections matrix. Then, formulation given by (13) can be expressed as follows:

$$\max_{\mathbf{h}^{(k)},\ell^{(k)},\gamma_k} \frac{1}{2K}\sum_{k=1}^{K}\gamma_k\mathbf{h}^{(k)\top}\mathbf{V}\mathbf{h}^{(k)} - \frac{1}{2}\sum_{k=1}^{K}\ell^{(k)\top}\ell^{(k)} \tag{14a}$$

$$\text{such that } \mathbf{h}^{(k)} = \mathbf{\Psi}\ell^{(k)}, \tag{14b}$$

where matrix $\mathbf{V} \in \mathbb{R}^{N\times N}$ can be chosen as:

- $\mathbf{I}_N$: We can normalize matrix $\mathbf{\Omega}$ in such way for all i condition $\sum_j^N \omega_{ij} = 1$ is satisfied and therefore we would obtain a degree matrix equaling the identity matrix. Then, $\sum\mathbf{h}^{(l)\top}\mathbf{h}^{(l)} = \text{tr}(\mathbf{H}^\top\mathbf{H})$, which corresponds to a PCA-based formulation.
- Diag($\mathbf{v}$): With $\mathbf{v} \in \mathbb{R}^N$ such that $\mathbf{v}^\top\mathbf{v} = 1$, we have a WPCA approach.
- $\mathbf{D}^{-1}$: Given the equality $\mathbf{V} = \mathbf{D}^{-1}$, optimization problem can be solved by means of a procedure based on random walks; being the case of interest in this study.

3.2.1 Gaussian processes

In terms of Gaussian processes, variable $\mathbf{\Psi}$ represents a mapping matrix such that $\mathbf{\Psi} = (\psi(\mathbf{x}_1),\ldots,\psi(\mathbf{x}_N))$ and where $\psi(\cdot) : \mathbb{R}^d \to \mathbb{R}^N)$ is mapping function, which provides a new N-dimensional data representation where resultant clusters are assumed to be more separable. Also, matrix $\mathbf{\Omega}$ is to be chosen as a Gaussian kernel [17]. Therefore, according to optimization problem given by (14), term $\mathbf{h}^{(k)}$ is to be the k-th projection of normalized binary indicators as $\mathbf{h}^{(k)} = \mathbf{\Psi}\ell^{(k)}$.

3.2.2 Eigen-solution

We present a solution for 14, which after solving the KKT conditions on its corresponding Lagrangian, an eigenvectors problem is yielded. Then, we first solve the Lagrangian of problem (14) so:

$$\mathcal{L}(\mathbf{h},\ell,\gamma,\alpha) = \frac{1}{2K}\mathbf{h}^\top\mathbf{V}\mathbf{h} - \frac{1}{2}\ell^\top\ell - \alpha^\top(\mathbf{h} - \mathbf{\Psi}\mathbf{w}), \tag{15}$$

where α is a N-dimensional vector containing the Lagrange multipliers.

Solving the partial derivatives to determine the KKT conditions, we have:

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\mathbf{h}} = 0 \quad &\Rightarrow \mathbf{h} = \frac{K}{\gamma}\mathbf{D}\alpha, \\ \frac{\partial\mathcal{L}}{\partial\ell} = 0 \quad &\Rightarrow \ell = \mathbf{\Psi}^\top\alpha, \\ \frac{\partial\mathcal{L}}{\partial\alpha} = 0 \quad &\Rightarrow \mathbf{h} = \mathbf{\Psi}\ell.\end{aligned}$$

Eliminating the primal variables, we obtain the following eigenvector problem:

$$\lambda\alpha = \mathbf{D}^{-1}\mathbf{\Omega}\alpha, \tag{16}$$

where $\lambda = N/\gamma$. Then, matrix $\mathbf{\Delta}_K = (\alpha^{(1)},\cdots,\alpha^{(K)})$ can be computed as the eigenvectors associated with the first K longest eigenvalues of $\mathbf{D}^{-1}\mathbf{\Omega}$.

Finally, projections matrix $\mathbf{H}$ is in the form

$$\mathbf{H} = \mathbf{\Psi}\mathbf{L} = \mathbf{\Psi}\mathbf{D}^{1/2}\mathbf{M} = \mathbf{\Omega}\mathbf{\Delta}_K, \tag{17}$$

and therefore $\mathbf{M} = \mathbf{\Psi}^{-1}\mathbf{D}^{-1/2}\mathbf{\Omega}\mathbf{\Delta}_K$, where $\mathbf{\Psi}$ can be obtained from a Cholesky decomposition.

Then, within a finite domain, both solution and formulation of NCC can be expressed similarly as done in KSC. So it is demonstrated the relationship between a kernel-based model and Gaussian processes.

4 Links between KSC and KKM

Kernel K-means method (KKM) is a generalization of standard K-means that can be seen as a spectral relaxation when introducing a mapping function in the objective function formulation [18]. As mentioned throughout this paper, spectral clustering approaches usually are performed on a lower-dimensional space, keeping the pairwise relationships among nodes. Then, it often leads to a relaxed NP-problems where continuous solutions are obtained by a eigen-decomposition. Such an eigen-decomposition is regarding the normalized similarity matrix (Laplacian, as well). In a Kernel K-means framework, eigenvectors are considered as geometric coordinates and then K-means methods is applied over the eigen-space to get the resultant clusters [19, 20].

Previous instance is as follows: Suppose that we have a gray scale matrix $m \times n$ pixels in size. Characterizing each image pixel with d features -e. g., color spaces, morphological descriptors- it is yielded as a result a data matrix in the form $\mathbf{X} \in \mathbb{R}^{N \times d}$, where $N = mn$. Afterwards, the eigenvectors $\mathbf{V}\mathbb{R}^{N \times N}$ of a normalized kernel matrix $\mathbf{P} \in \mathbb{R}^{N \times N}$ such that $\mathbf{P} = \mathbf{D}^{-1}\mathbf{\Omega}$, being $\mathbf{\Omega}$ the kernel matrix and $\mathbf{D} \in \mathbb{R}^{N \times N}$ the corresponding degree matrix. Then, we proceed to cluster $\mathbf{V}$ into K groups using K-means algorithm: $\mathbf{q} = \text{kmeans}(\mathbf{V}, K)$, being $\mathbf{q} \in \mathbb{R}^N$ the output cluster indicator such that $q_i \in [K]$. The segmented image is then a $m \times n$ sized matrix holding regions in accordance with $\mathbf{q}$.

Briefly put, one simple way to perform a KKM procedure is applying k-means over the eigen-space. In Equation (7), the dual formulation is regarding the matrix $\mathbf{S} = \mathbf{VH\Omega}$ where weighting matrix can be chosen as $\mathbf{V} = \mathbf{D}^{-1}$ and $\mathbf{D}$ is the degree of the data-related graph. Since $\mathbf{H}$ causes a centering effect, matrix $\mathbf{S}$ is the same as $\mathbf{P}$ when kernel matrix $\mathbf{\Omega}$ is centered. In other words, KKM can be seen as a KSC formulation with an incomplete latent variable model being a non-centered one (with no bias term).

5 Results and discussion

In order to show how considered methods work, we conduct some experiments to test their clustering ability on segmenting images. To do so, the segmentation performance is quantified by a supervised index noted as Probabilistic Rand Index (PR), explained in [21], such that $PR \in [0, 1]$, being 1 when regions are properly segmented. Images are drawn from the free access Berkeley Segmentation Data Set [15]. To represent each image as a data matrix, we characterize the images by color spaces (RGB, YCbCr, LAbB, LUV) and the xy position of each pixel. At the end, data matrix $\mathbf{X}$ gathers N pixels represented by d characteristics (variables). To run the experiment, we resize the images at 20% of the original size due to memory usage restrictions. All the methods are performed with a given number of clusters K manually set as shown in shown in Fig. 5 and using the scaled exponential similarity matrix as described in [19], setting the number of neighbors to be 9.

To test all the methods in a fair scenario, kernel-based methods (KSC and KKM) use $\mathbf{\Omega}$ as kernel matrix, whereas such a matrix is the affinity matrix for NCC. As well, to perform the clustering procedure, the number of clusters is the same for all the considered methods. As can be readily appreciated, KSC overcome the rest of studied clustering methods. This fact can be attributed to the KSC formulation, which involves a whole latent variable model being in turn incorporated within a LS-SVM framework. Indeed, just like principal component analysis (PCA), KSC optimizes an energy term. Differently, such an energy term is regarding a latent variable instead of directly the input data matrix. Concretely, a latent variable model is used, which is linear and formulated in terms of projections of the input data. The versatility of KSC relies on the kernel matrix required during the optimization procedure of its cost function. Such a matrix holds pairwise similarities, then KSC can be seen as data-driven approach that not only consider the nature of data but yields a true clustering model. It is important to quote that -depending on the difficulty of the segmentation task- data matrices representing images yield features spaces, which may present hardly separable classes. Then, we have demonstrated the benefit of the KSC approach that uses a model along with a LS-SVM formulation -everything within a primal-dual scheme. Other studies have also proven the usability and versatility of this kind of approaches [8, 22].

6 Additional remarks

As explained in [23], KSC performance can be enhanced in terms of cluster separability by optimally projecting original input data and performing the clustering procedure over the projected space. Given the unsupervised nature, spectral clustering becomes very often a parametric approach, involving then a stage of selection/tunning of collection of initial parameters to avoid any local-optimum solution. Typically, the initial parameters are the kernel or similarity matrix and the number of groups. Nonetheless, in some problems when data are represented in a high-dimensional space and/or data-sets are non-linearly separable, a proper feature extraction may be an advisable alternative. In particular, a projection generated by a proper feature extraction procedure may provide a new feature space wherein the clustering procedure can reach more accurate cluster indicators. In other words, data projection accomplishes a new representation space, where the clustering can be improved, in terms of a given mapping criterion, rather than performing the clustering procedure directly over the original input data.

The work developed in [23] introduces a matrix projection focusing on a better analysis of the structure of data that is devised for a KSC. Since data projection can be seen as a feature extraction process,

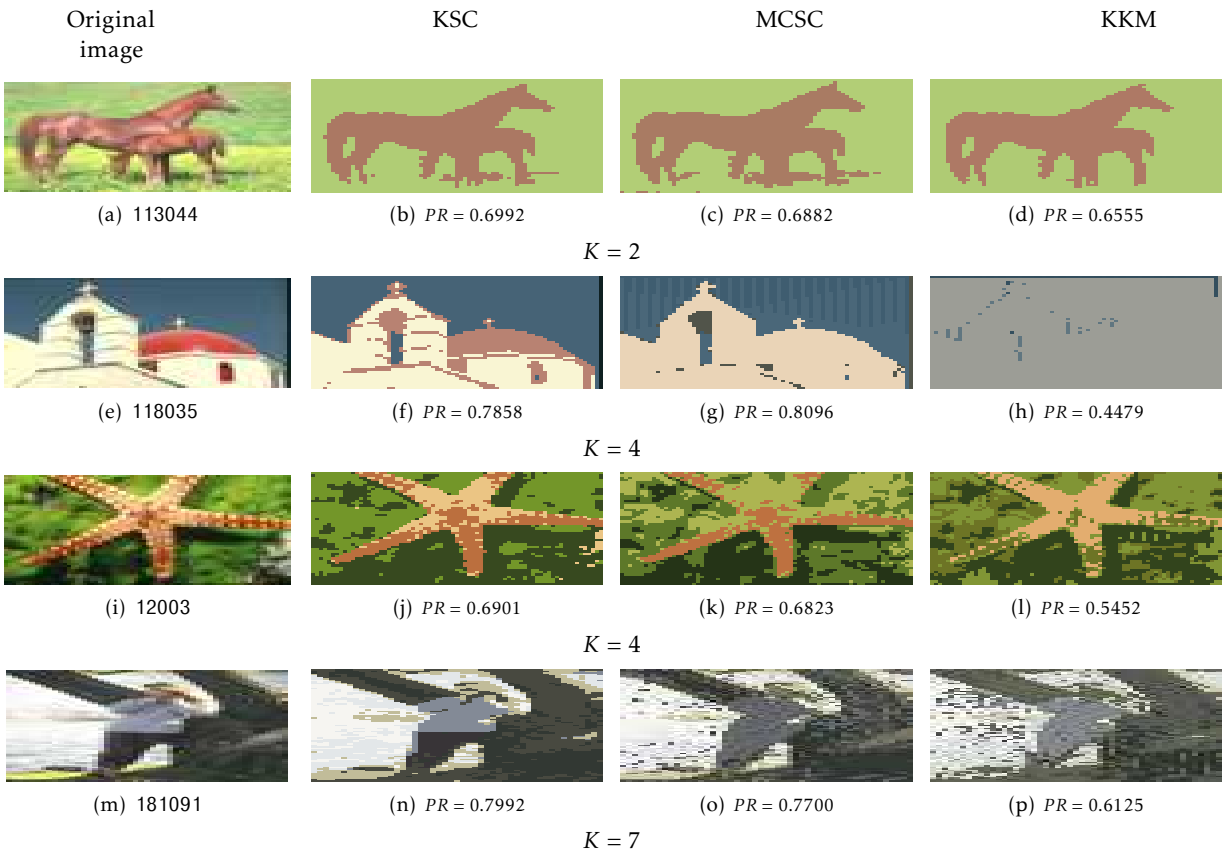


Figure 1: Clustering performance on image segmentation reached by all the considered methods. It is noticeable that KSC overcome the remaining methods. Images are data that traditionally involve highly non-separable clusters. Therefore, the benefit of using a whole latent variable model within a LS-SVM formulation is verified.

we propose the M-inner product-based data projection, in which the similarity matrix is also considered within the projection framework, similarly as discussed in [24]. There are two main reasons for using data projection to improve the performance of kernel spectral clustering: firstly, the data global structure is taken into account during the projection process and, secondly, the kernel method exploits the information of local structures.

Another study [25] explores the links of KSC with spectral dimensionality reduction from a kernel viewpoint. Particularly, the proposed formulation is LS-SVM in terms of a generic latent variable model involving the projected input data matrix. In order to state a kernel-based formulation, such a projection maps data onto a unknown high-dimensional space. Again, the solution of the optimization problem is addressed through a primal-dual scheme. Finally, once latent variables and parameters are determined, the resultant model outputs a versatile projected matrix able to represent data in a low-dimensional space. To do so, since the optimization is posed under a maximization criterion and dual version has a quadratic from, the eigenvectors associated with the largest eigenvalues can be chosen as a solution. Therefore, the generalized kernel model may represente a weighted version of kernel principal component analysis.

7 Conclusions

This works explores a widely-recommended method for unsupervised data classification, namely kernel spectral clustering (KSC). From elegant developments, the relationship between KSC and two other well-known spectral clustering approaches (normalized cut clustering and kernel k-means) is demonstrated. As well, the benefit of KSC-like approaches is mathematically and experimentally proved. The goodness of KSC relies on the nature of its formulation, which is based on a latent variable model incorporated into a least-square-support-vector-machine framework. Additionally, some key aspects and hints to improve KSC performance as well as its ability to represent dimensionality reduction approaches are briefly outlined and discussed.

As a future work, a generalized clustering framework is to be designed so that a wide range of spectral approaches can be represented. Doing so, the task of selecting and/or testing a spectral clustering method would become easier and fairer.

References

[1] F. Schwenker and E. Trentin, "Pattern classification and clustering: a review of partially supervised learning approaches," *Pattern Recognition Letters*, vol. 37, pp. 4–14, 2014.

[2] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern recognition letters*, vol. 31, no. 8, pp. 651–666, 2010.

[3] S. Belarouci and M. A. Chikh, "Medical imbalanced data classification," *Advances in Science, Technology and Engineering Systems Journal*, no. 3, pp. 116–124.

[4] H. Chebi, D. Acheli, and M. Kesraoui, "Dynamic detection of abnormalities in video analysis of crowd behavior with DBSCAN and neural networks," *Advances in Science, Technology and Engineering Systems Journal*, no. 5, pp. 56–63.

[5] Y. S. X. and S. Jianbo, "Multiclass spectral clustering," in *ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision*. Washington, DC, USA: IEEE Computer Society, 2003, p. 313.

[6] D. H. Peluffo-Ordóñez, C. Castro-Hoyos, C. D. Acosta-Medina, and G. Castellanos-Domínguez, "Quadratic problem formulation with linear constraints for normalized cut clustering," in *Iberoamerican Congress on Pattern Recognition*. Springer, 2014, pp. 408–415.

[7] I. S. Dhillon, Y. Guan, and B. Kulis, "Kernel k-means: spectral clustering and normalized cuts," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2004, pp. 551–556.

[8] C. Alzate and J. A. K. Suykens, "Multiway spectral clustering with out-of-sample extensions through weighted kernel pca," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 2, pp. 335–347, 2010.

[9] D. H. Peluffo-Ordónez, S. García-Vega, A. M. Alvarez-Meza, and C. G. Castellanos-Domínguez, "Kernel spectral clustering for dynamic data," in *Iberoamerican Congress on Pattern Recognition*. Springer, 2013, pp. 238–245.

[10] R. Langone, C. Alzate, and J. A. Suykens, "Kernel spectral clustering with memory effect," *Physica A: Statistical Mechanics and its Applications*, 2013.

[11] C. Alzate and M. Sinn, "Improved electricity load forecasting via kernel spectral clustering of smart meters," in *2013 IEEE 13th International Conference on Data Mining*. IEEE, 2013, pp. 943–948.

[12] R. Langone, R. Mall, C. Alzate, and J. A. Suykens, "Kernel spectral clustering and applications," in *Unsupervised Learning Algorithms*. Springer, 2016, pp. 135–161.

[13] D. H. Peluffo-Ordóñez, C. Alzate, J. A. K. Suykens, and G. Castellanos-Domínguez, "Optimal data projection for kernel spectral clustering," in *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2014, pp. 553–558.

[14] S. Mehrkanoon, C. Alzate, R. Mall, R. Langone, and J. A. Suykens, "Multiclass semisupervised learning based upon kernel spectral clustering," *IEEE transactions on neural networks and learning systems*, vol. 26, no. 4, pp. 720–733, 2015.

[15] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. 8th Int'l Conf. Computer Vision*, vol. 2, July 2001, pp. 416–423.

[16] D. Peluffo, C. D. Acosta, and G. Castellanos, "An improved multi-class spectral clustering based on normalized cuts," *XVII Iberoamerican Congress on Pattern Recognition - CIARP*, 2012.

[17] C. Rasmussen and C. Williams, *Gaussian processes for machine learning*. MIT press Cambridge, MA, 2006, vol. 1.

[18] H. Zha, C. Ding, M. Gu, X. He, and H. Simon, "Spectral relaxation for k-means clustering," *Advances in neural information processing systems*, vol. 14, pp. 1057–1064, 2001.

[19] L. Zelnik-manor and P. Perona, "Self-tuning spectral clustering," in *Advances in Neural Information Processing Systems 17*. MIT Press, 2004, pp. 1601–1608.

[20] I. S. Dhillon, Y. Guan, and B. Kulis, "Kernel k-means, spectral clustering and normalized cuts," 2004.

[21] R. Unnikrishnan, C. Pantofaru, and M. Hebert, "Toward objective evaluation of image segmentation algorithms," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 6, pp. 929–944, 2007.

[22] D. H. Peluffo-Ordóñez, J. A. Lee, and M. Verleysen, "Generalized kernel framework for unsupervised spectral methods of dimensionality reduction," in *Computational Intelligence and Data Mining (CIDM), 2014 IEEE Symposium on*. IEEE, 2014, pp. 171–177.

[23] D. H. Peluffo-Ordóñez, C. Alzate, J. A. Suykens, and G. Castellanos-Domínguez, "Optimal data projection for kernel spectral clustering." in *European Symposium on Artificial Neural Networks - ESANN*, 2014.

[24] J. Rodríguez-Sotelo, D. Peluffo-Ordóñez, D. Cuesta-Frau, and G. Castellanos-Domínguez, "Unsupervised feature relevance analysis applied to improve ecg heartbeat clustering," *Computer Methods and Programs in Biomedicine*, 2012.

[25] X. Blanco-Valencia, M. Becerra, A. Castro-Ospina, M. Ortega-Adarme, D. Viveros-Melo, J. Alvarado-Pérez, and D. Peluffo-Ordóñez, "Kernel-based framework for spectral dimensionality reduction and clustering formulation: A theoretical study," *ADCAIJ: Advances in Distributed Computing and Articial Intelligence Journal*, 2017.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1677-1692 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Modeling Double Subjectivity for Gaining Programmable Insights: Framing the Case of Uber

Loretta Henderson Cheeks[1*], Ashraf Gaffar[2], Mable Johnson Moore[3]

[1]*School of Computing, Informatics, and Decision Systems Engineering (CIDSE), Arizona State University, Tempe, AZ, 85287, U.S.A.*

[2]*School of Computing, Informatics, and Decision Systems Engineering (CIDSE), Arizona State University, Mesa, AZ, 85212, U.S.A.*

[3]*Chief Information Officer, Savannah State University, Savannah, GA, 31404, U.S.A.*

ARTICLE INFO

Article history:
Received: 31 May, 2017
Accepted: 10 August, 2017
Online: 04 September, 2017

Keywords:
Subjectivity
Artificial intelligence
Frames
News source sentiment
Disruptive technology
Social influence
Online news
Amplification
Machine learning
Natural language processing
Uber

ABSTRACT

The Internet is the premier platform that enable the emergence of new technologies. Online news is unstructured narrative text that embeds facts, frames, and amplification that can influence society attitudes about technology adoption. Online news sources are carriers of voluminous amounts of news for reaching significantly large audience and have no geographical or time boundaries. The interplay of complex and dynamical forces among authors and readers allow for progressive emergent and latent properties to exhibit. Our concept of "Double subjectivity" provides a new paradigm for exploring complementary programmable insights of deeply buried meanings in a system. The ability to understand internal embeddedness in a large collection of related articles are beyond the reach of existing computational tools, and are hence left to human readers with unscalable results. This paper uncovers the potential to utilize advanced machine learning in a new way to automate the understanding of implicit structures and their associated latent meanings to give an early human-level insight into emergent technologies, with a concrete example of "Uber". This paper establishes the new concept of double subjectivity as an instrument for large-scale machining of unstructured text and introduces a social influence model for the discovery of distinct pathways into emerging technology, and hence an insight. The programmable insight reveals early spatial and temporal opinion shift monitoring in complex networks in a structured way for computational treatment and visualization.

1. Introduction

This paper is an extension of work originally presented in 2017 IEEE 11th International Conference on Semantic Computing (ICSC) [1]. The Internet is the premier platform for the proliferation of voluminous online news and has been compared to a huge social and psychological laboratory [2]. It has enabled a globally and connected world which is greatly transforming computational approaches for exploring online social interactions for gaining early insight into emergent technology. News sources propagate online news that function within a complex and dynamic system via the Internet. This system comprises informational flows, interactions, dynamical forces, and emerging properties, which lead to asymmetries and have a tendency to influence society. Online news is a form of narrative text that embeds facts and strategic communication frames for shaping and transforming group standards, values, attitudes, and beliefs about technology adoption. The embedded facts represent absolute, relevant truths characterized by their rigid (or fixed) nature. These absolute truths are shared among news sources. However, purely factual articles would be hard to write and read. Adding tone, writing style, attitude, opinions, related context, and non-factual comments and arguments often make text more subjective but easier to read. The exercise of news framing—and the amplification of subjectivity that it communicates—is where multidimensional aspects of the content emerge and where expressivity happens for transformational and maximal influence. Framing is the use of strategic devices for presenting prominent aspects and perspectives about an issue using certain words as well as stereotyped images and sentences for the purpose of conveying hidden meanings about an issue [3].

*Corresponding Author: Loretta Henderson Cheeks, School of Computing, Informatics, and Decision Systems Engineering (CIDSE), Arizona State University, Tempe, AZ, 85287, USA | Email: loretta.cheeks@asu.edu

1.1. What is Insight?

News sources play a critical role in the activation of frames about technology society adopt, and the amplification of subjectivity in online news drives a wedge between evidence and beliefs [4]. In addition, the combination of facts, frames, and amplified subjectivity represents instruments that offer a space of possibilities for deeper exploration of the cognitive aspects of narrative text. Natural Language Processing (NLP) made significant progress using Artificial Intelligence (AI) to parse narrative text by applying analytical techniques. However, less success has been made with computational treatment when it comes to semantic levels. Semantic parsers work well only at a phrase level, but it becomes challenging to semantically parse several sentences together. Some approaches can look for keywords to determine the sentiment of a text, provide a summary, or a basic language-to-language translation, but they remain by far limited in understanding the deep latent meanings which contribute to the richness and depth of ideas within a complete text or of a related set of articles at the same level humans do. We define this as an "Insight" which is "the ability to understand the latent meanings within an article". The definition of Insight can be generalized to a group of related articles, and refers to "the ability to synthesize the relationship between a group of articles as well as how the contents of latter articles were affected or influenced by those in earlier articles of the same group." The need to understand this insight becomes particularly important with online news, as the number of related articles tend to be generally larger than printed media due to the far reach and hence the significantly larger and more diverse readership of the Internet. This "understanding" process is typically highly cognitive, and is best done by humans. However, with the large number of online news, the need for software tools becomes essential for gaining programmable insight at a very large scale. We define programmable insight as "the ability to apply computational models and methods and hence software tools to leverage human cognitive ability to gain insight into very large narrative text".

1.2. Challenges of Processing Subjective Text

When attempting to identify potential pathways to address the complexities and dynamics that exist in online news text, frames are instruments for experimenting and learning about the production of language and forces at work for gaining programmable insight that may lead to new instruments for exploring cognitive computation. Here, pathways mean the potential routes that one may take to explore patterns about a subject. The authors recognize the need for an integration of multiple models and perspectives that advance what is known about hidden meanings and latent structures in text. Modeling narrative text allows software tools to provide help in automating human efforts using computational leverage. However, providing a deep model that represents hidden meaning in text is relatively hard. Several NLP methods in artificial intelligence use word matching and probability functions to determine the correctness and possible meaning of words with the help of language corpus containing trillions of words. This "brute force" and highly analytical top-down approach could be computationally heavy, and does not provide an insight beyond the phrase level. An acknowledgment of the multidimensional aspects—degrees of freedom—of the narrative text in news calls for computational research to fill knowledge gaps and to transform the current understanding of alternative pathways to learning that leverage cognitive science research.

Leveraging the scholarly work of mathematicians, the fathers of the digital computer developed the initial logical machine using a rather formal grammar structure. The first level of grammatical expressivity the computer machine could understand is propositional logic—that is, knowledge of facts or truths. Considering the "form" of the English language (i.e., subject, verb, object), the logical machine could make valid arguments using as few as two premises and simple symbols, thereby drawing a logical conclusion. Propositional logic became the basic component of symbolic logic. It became the enabler for the logical machine to reason using simple English grammatical structure. This breakthrough led to machines being able to reason as either true (1) or false (0), using simple binary symbols and boolean algebra to express "and," "or," "not," "implies/then," and "if and only if."

By switching out fixed sentence structure for symbols, as was done in mathematics for centuries, the logical machine could express a primitive level of truth. The computer could express with "0" or "1" with this simple syntactical machine code. Although this advanced in the logical machine was very impressive in its day, this level of formalism with its strict grammatical structure falls short of reaching the goal of AI, as it gave no insight into the topic proposed in the sentence structure.

Propositional logic usage in the logical machine had no properties nor a sense of relationships, all of which were embedded in the internal structure of the text. Rather, its sole focus was on pure factual, blocks of sentence structures as isolated objects for expressing what logicians for centuries thought to be real. Real objects were point masses and forces; as such, properties leading to an understanding about the relationships and internal structures of text were dismissed as subjective. Discounting relational properties as the inferior status of "secondary" properties resulted in excessive adherence to the theoretical approach of attaining machine intelligence. However, researchers soon came to realize that if there is any future for theory, it must be applied. This meant that the internal structure in text must delve into deeper levels of expressivity where hidden meanings and relationships could be explored for advancing machine intelligence. Since this time, as shown in Fig. 1, numerous levels of machine logic have been established to allow for the variety of properties to reason about relationships, individuals, identities, and quantifiers—thereby, giving way to make more elaborate assertions, inferences, and semantic expressions of phrases. Fig. 1 shows the progressive climb up the stack from the formal space where computer machines focused on propositional logic, formal grammar, and, thereafter, phrases.

With the pendulum swinging from machine-to-human logical expression—where the programmer of the computer machine had to learn the psychology of the machine—to human-to-machine logic for expressing rational human thoughts, higher-level programming languages were developed, such as Fortran, Prolog, and Lisp. With the foundations for reasoning and with higher levels of expressivity at the researcher's grasp, attention turned to another field of research, AI.

Here, at the phrases layer (NLP using AI) is where much of the computer science research today continues to focus. This has led to great breakthroughs for learning AI, NLP, and fuzzy logic; however, this approach shed very little insight into cognitive computation.

This research advances beyond the phrases layer and bridges the computational and cognitive worlds using the instrument of

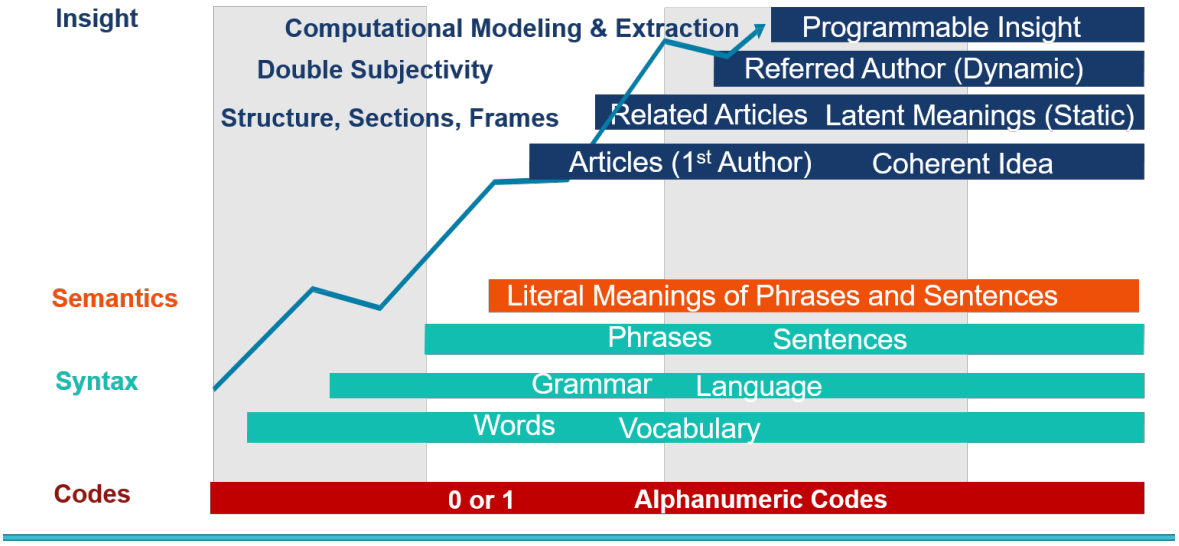


Fig. 1. A view of the end-to-end bridge to cognitive computation model. Here, the stack, from bottom to top, represents the evolution of computational processing from grammar to insight, as shown on the right, and the value or outcome that results as shown on the left. For instance, computer processing of 0 and 1 bits were the catalyst for primitive storage, transmission, and processing logic.

frames that exist in narrative text for exploring programmable insight that may lead to deeper learning for machine intelligence.

In exposing the discovery of new knowledge about text, transformations (or shifts) in which dynamic forces reach a critical point make clear the new structures, pathways, and influence patterns to be accepted or new questions or alternatives in the scope of an issue to be revised. The ability to seek out novel, uncharted relationships that are implicit from a body of unstructured narrative text—yet not explicitly stated within that text—has obvious scientific value. This process is called "novel knowledge discovery." It enables the use of the current state of knowledge for an emergence of possible new relationships and pathways that have yet to be studied.

1.3. Structured Expert Support (SES) Models

Fully structured data render itself easily to computation. A sales receipts generated by point-of-sale (POS) software in most retail stores includes a fully-structured list of items code, text, sale price as well as subtotal, total, date and time of each receipt. Terabytes of such data follow a well-known model and would be easy to process with software tools and provides solid foundation of business intelligence (BI) and other advanced insight, data science, and consumer-behavior pattern discovery. Narrative text, on the other hand, is highly unstructured and hence represent a nightmare to computational algorithms trying to extract deep knowledge buried within. One of the main reasons is the lack of a representative model that this data follow. While humans can understand and even enjoy the contents of such text, the lack of concrete models prohibits the computational treatment of such text. In the proposed approach, the authors use the framing theory—as will be shown later—to build representative models of narrative text. While these models need to be built by hand using human experts (thus the name expert support), they are done in a structured way to normalize human cognitive ability into nominal structure (hence the qualifier Structured Expert Support) as shown in Fig 2. Fig. 2 also is a graphical depiction of the process for discovering news frames. The formal problem definition of framing is presented as follows:

Let U_t be a universe of online news documents at time t. Let $\{d_i\}, 1 \leq i \leq n$, denote a finite subset of documents from the universe of documents U_t, where n is the total number of documents being considered. Let $S_i = \{P_i, C_i\}$ be the set of document properties, where P_i is the source that produced the article and that propagates a central organizing idea (i.e., framing) and C_i is the set of terms (i.e., features or content descriptors) from the news article where common words have been removed—all of which supply a context and suggest what the issue is through the use of selection, emphasis, exclusion, cues, and elaboration [5, 6]. For issues of interest, the process leads to the discovery of standard frames dominantly used in U.S. news coverage [7]: human interest, conflict, economic, managerial, and science. Let f_1, f_2, f_3, f_4, and f_5 correspond to each of these frames, where each f_i is associated with a pair $\{x_i, y_i\}$; $\boldsymbol{x_i}$ is the feature vector, and $\boldsymbol{y_i}$ is the vector with the corresponding weights. Discover five dominate news frame signatures.

Humans would manually extract certain keywords present in a text, and build representative data structures (vectors) representing hidden meanings within the text. On the one hand, these vectors will be a strong representation of the hidden frames in a text, and -on the other hand- they will be highly structured and hence readily computational. As will be discussed later, the proposed SES relies heavily on the theory of framing as the modeling concept and the bridging mechanism between cognitive contents and computational capability of narrative text.

This manuscript includes a model for expressing the multidimensional aspects of news framing—and the amplification that it conveys—for transformational and maximal influence. The model is a pathway to explore programmable insight to expose a deep level of expressivity into the hidden structures in text—what is referred to in this paper as "double subjectivity," a term introduced here [8]. Double subjectivity is the term to describe the secondary subjectivity that results after receiving indirect informal cognitive data in online news. The double subjectivity method expands Axelrod's [9] and Klienberg's [10-12] research on cultural dynamics and social influence for constructing an issues

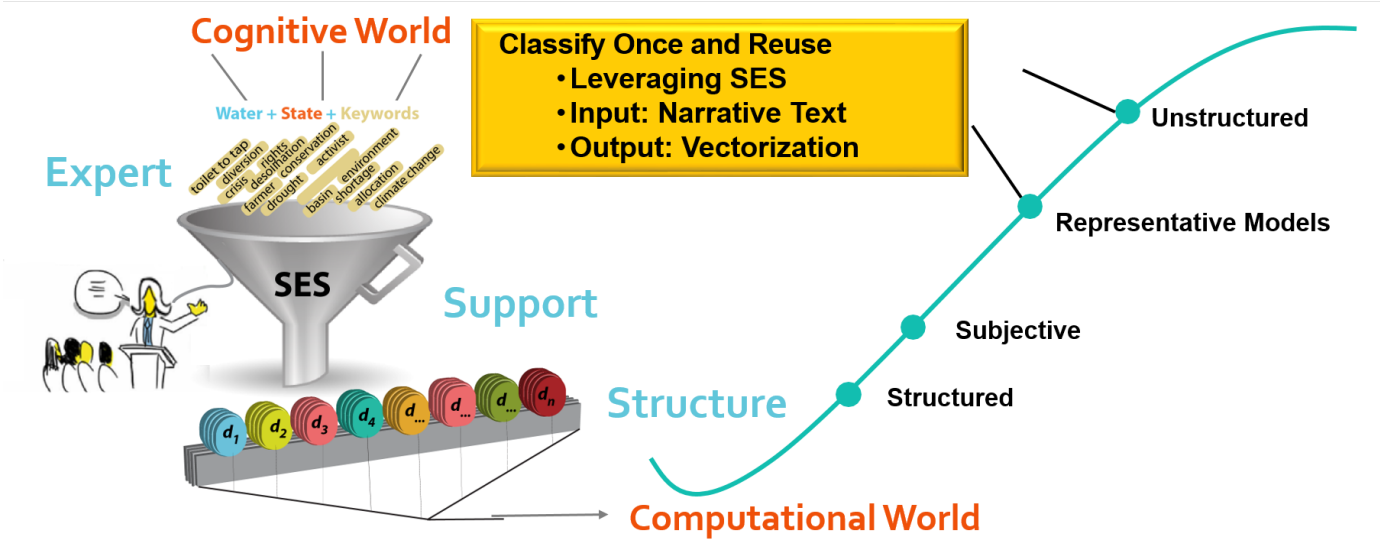


Fig. 2. Depiction of SES significance in the development of a representative model in support of the discovery of frames that exist in unstructured narrative text.

network representation and designing a visual language for exploring spatio-temporal dynamics. The issues network is exemplar of a complex dynamical system whereby migration paths may be visualized. The migration path is defined as the movement of frame choices in articles across a network to establish a new bridge for information exchange. The way the issues network grow (gain node connections) and shrink (loss node connections) is discussed in a subsequent section in this manuscript. The aim of the authors is to demonstrate news framing as a step toward deep learning about the production of language that may inform scholars about pathways that offer new approaches that advance cognitive computation for improved intelligent behavior in machines.

To the best of the authors knowledge, in computer science scholarly research, examining double subjectivity using online news and framing effects for gaining programmable insight has not been addressed previously for the class of problems discussed in this paper. This research inquiry will give insight into the following questions: What is the migration path of online news frames as it evolves over time? What is the effect of double subjectivity? What are the conditions observed that allow for the evaluation of shifts in influence or amplification intensity?

The following sections will describe the conceptual framework, related works, background on the issues of application, methods, possibility space transitions, observations of the models in action, and the conclusion.

2. Conceptual Framework

In the last decade, advances in fields such as AI, text summarization, and NLP have enabled gains in understanding text syntactic structure and limited gains in text semantics by offering algorithmic and theory-based methodologies. However, these fields have struggled in efforts to unlock hidden meanings and abstractions found in unstructured text such as jokes, poems, or art that are readily expressed through cognitive structures. This is due in large part to barriers that exist when using machine learning algorithms that are still limited in their understanding of complex unstructured text where factors such as semantic context, culture, and complex relationships must be considered for making sense of the narrative text.

Humans have the unique and superior ability to get beyond the syntax of data in order to make meaning. However, a great deal of cognitive science research has focused on cognitive structures called “frames” (sometimes “schemas”) that exist in the recesses of the brain [13]. The framing of online news articles by news sources consequently has the effect of evoking a word and phrase in the unconscious cognitive structure in the mind of the reader. These structures are built over time through repetition. The repetitive framing of critical issues will strengthen the circuits about the issue in a reader’s brain. When online news sources repeat frames often, they become “normalized” unconscious activation points in the brains of readers. This behavior by online news sources is equivalent to placing ideas in a container, referred here as an article. These ideas are sent through communication channels that have tremendous influence on the perception and real value of knowledge. When neighboring news sources take the ideas out of the container by way of adoption or copying, the cumulative effect is double subjectivity. A property of double subjectivity is spatio-temporal dynamics, which has the capacity to display a larger context of influence. As a number of related articles grows, the cumulative effect of subjectivity could grow and take on certain direction or shape, influenced by the article-to-article changes or the amplification or suppression of frames. This dynamic behavior may be best shown through a network visualization. A network visualization is a mechanism that has been used to leverage people’s ability to acquire knowledge for revealing hidden meanings, social influence, and migration paths of information and interactions.

The automation of unstructured text can be generally categorized under two main areas: computational data and cognitive data.

2.1. Computational Data

Computational data—represented by mathematical formulas, propositional logic, and formal (computer) languages—have little room for subjective interpretation and are typically concise, context-free, and fully predictable. For instance, a correctly developed source code can compile and run on a computing

machine and will always produce the same output for a specific input. The combination of a formal set of instructions resulting from the successful compilation and run of source code can transition a computer through a set of well-defined states, depending exclusively on the compiled code and input data. This leaves no room for interpretation or subjectivity. In other words, the meaning and interpretations must be fully represented in the source code. A good programming language must, therefore, have sufficient expressiveness to fully represent a feasible solution in a specific problem space. If a source code arrives at an undefined state, it will simply freeze or crash; it is unable to make an independent decision. Some areas of computer science such as fuzzy logic and AI attempt to address these inherent limitations by providing a default or recovery system. Some languages also provide exception handling constructs and error recovery mechanisms. However, these approaches remain rigid mechanical methods, and are not the focus of the work conveyed here. This expresses the "formal space" as depicted in Fig. 4.

Although the computational language of source code is a good medium for communicating with computers, natural languages that embed cognitive structures are the mediums of communication among humans, and natural languages are informal and ambiguous in comparison to the computational language of computer code. In Fig. 4, the "informal space" is shown to describe the embeddedness of informal cognitive data structures that humans are superior over machines in conveying which are contained in narrative text. Cognitive data can be further divided into two subcategories, depending on the source of origination: direct and indirect.

2.2. Direct Cognitive Data

Direct cognitive data are the result of a person's encounter with an event and the perception, in part, of its details firsthand. Human perceptions of events lead to personal interpretations of events. Interpretation depends greatly on the viewer's perception, as different people can arrive at different interpretations of the same event. Perception is often associated with the viewer's state of mind, prejudice, background, and beliefs. Two persons witnessing the same event or being involved in the same situation could assign different interpretations and end with separate conclusions. Although the external situation is the same (the perception of which is uniquely defined in this paper as the absolute baseline), it could be internalized differently by different people. Such a case is a first-level subjectivity. Even when different people are asked to report an incident or a situation (the same absolute baseline), different accounts are often the result, indicating first-level subjectivity. Many factors further affect this subjectivity:

- Subjectivity: Humans have different expressive abilities and styles in reporting events using natural languages, leading to increased subjectivity.
- Ambiguity: The ambiguity of natural languages allows for different interpretations of spoken or written sentences.
- Complexity: Complex, extended situations, such as movie showings, are harder to perceive and interpret compared to simple or atomic events like the occurrence of a rocket launch.
- Expressiveness: Unlike a computational language, a natural language does not fully express the contents in its syntax. Natural languages report about something, rather than fully describe an action, artifact, or event in a specific way.
- Context: Natural languages are highly context dependent. The same sentence could have different meanings in different contexts.
- Modality: Natural languages are often combined with other modalities like para-verbal communication, body language, and face impressions, which can influence the meanings or interpretation of spoken words.

The aforementioned factors explain why—even for different people directly witnessing the same situation—substantial subjectivity can be present in reporting or discussing an event that was mutually witnessed.

People often try to reduce first-level subjectivity. Television reporters try to abide by certain rules to maximize neutrality and leave it up to the viewers to build their own opinions freely. The scientific community tries to show the stated facts through rigorous research and scientific experiments in attempts to eliminate or reduce subjectivity. However, subjectivity is not so simple to discard. Consider newspaper and television reporters who are known for their opinions.

2.3. Indirect Cognitive Data

The Internet and World Wide Web have led to an explosion of indirect cognitive data although events can still be directly perceived and subjectively interpreted. Because of technology advancements, it is possible to receive a much larger number of events indirectly. Because these events are not directly witnessed, an absolute baseline is not available. Because knowledge of an event is often based on the written or spoken reports of others, the baseline will be relative to the reporter's subjectivity or simply "float."

With the acceptance of indirect reports and their interpretation, the receiver adds his or her own subjectivity. Again, no longer is the absolute baseline available. It is replaced with a reporter-relative baseline or a floating baseline. If the reporter's subjectivity is strongly framed and influenced towards certain beliefs, the receiver of the account already starts with a baseline that is completely different from the absolute baseline. This new baseline could strongly affect the receiver's beliefs. Double subjectivity is the term introduced in this paper to describe the secondary subjectivity that results after receiving indirect informal data. Fig. 3 illustrates the cumulative influence effects of double subjectivity.

The illustrated formula depicts a generic function for double subjectivity in online news. Here, the assumption is that there is an observer-dependent function $f(x)$, such that the level of subjectivity L_i is the relative baseline of the situation or event dependent on firsthand account or cumulative effect. First-level subjectivity it is expressed as,

$$L_1 = f_1(L_0), \text{ and} \quad L_2 = f_2(L_0) \tag{1}$$

whereas, double subjectivity is expressed as,

$$L_1 = f_1(L_0), \text{ and} \quad L_2 = f_2(L_1) \tag{2}$$

f is an amplification intensity function with cumulative influence effects B_i that depends on the person observing, then reporting a situation or event.

An analogy drawn from photocopying illustrates double subjectivity in that, ideally, a photograph copy should be a close replica of the copied photograph. Yet, in actuality, the resulting photograph is the copied photo with its amplification plus the distortion amplification gained from the copying process or machine.

Online news is exemplary of double subjectivity in that its assemblage consists of subjective properties such as context, interpretation, multimodal interaction, background, compensation, and assumptions. These properties are critical building blocks that characterize narrative text for its a) signification (i.e., the semantic content signifying an association or concept given a word, sentence, or phrase) and b) significance (i.e., the relevance, rank, importance, or capacity to make a difference). Unfortunately, there remain substantial gaps in advancing what is known about the production of language in narrative text where double subjectivity exists. Therefore, exploring deeper levels of expressivity with an emphasis on double subjectivity may inform a new direction in research.

Discovering double subjectivity in narrative text poses a difficult problem, as it seeks to give insight and meaning to open-ended statements and indirect data. More complex structures within the narrative text provide an expansive landscape for advancing effective tools and crafting applications to automate many language-related tasks—for example, document

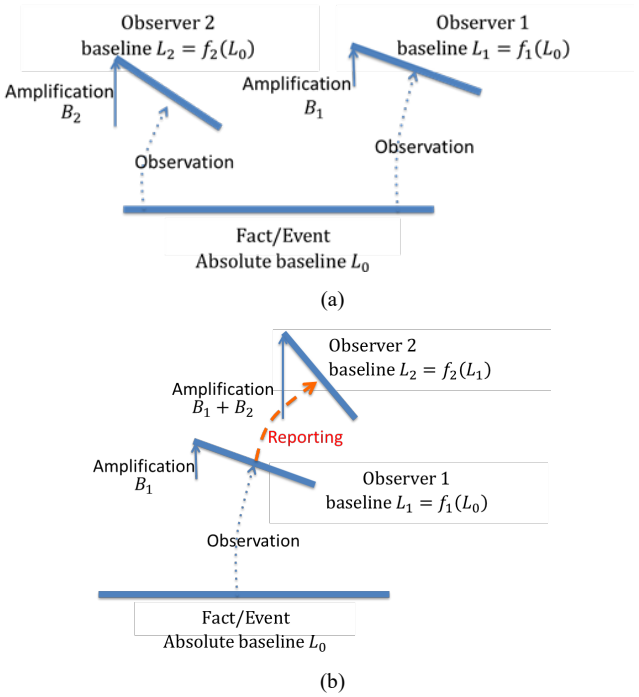


Fig. 3. Two representative illustrations of the double subjectivity framework. (a) There is an event (Observation) about an issue that two observers see: Observer 1 and Observer 2. Although it is the same event, their perceptions are different. Their amplification may be different, but independent. (b) This graph illustrates the introduction of an indirect account of an event, whereby the reporter does not receive the information firsthand. Rather, Observer 2 adopts the report of Observer 1, which sets into motion the cumulative effect of double subjectivity.

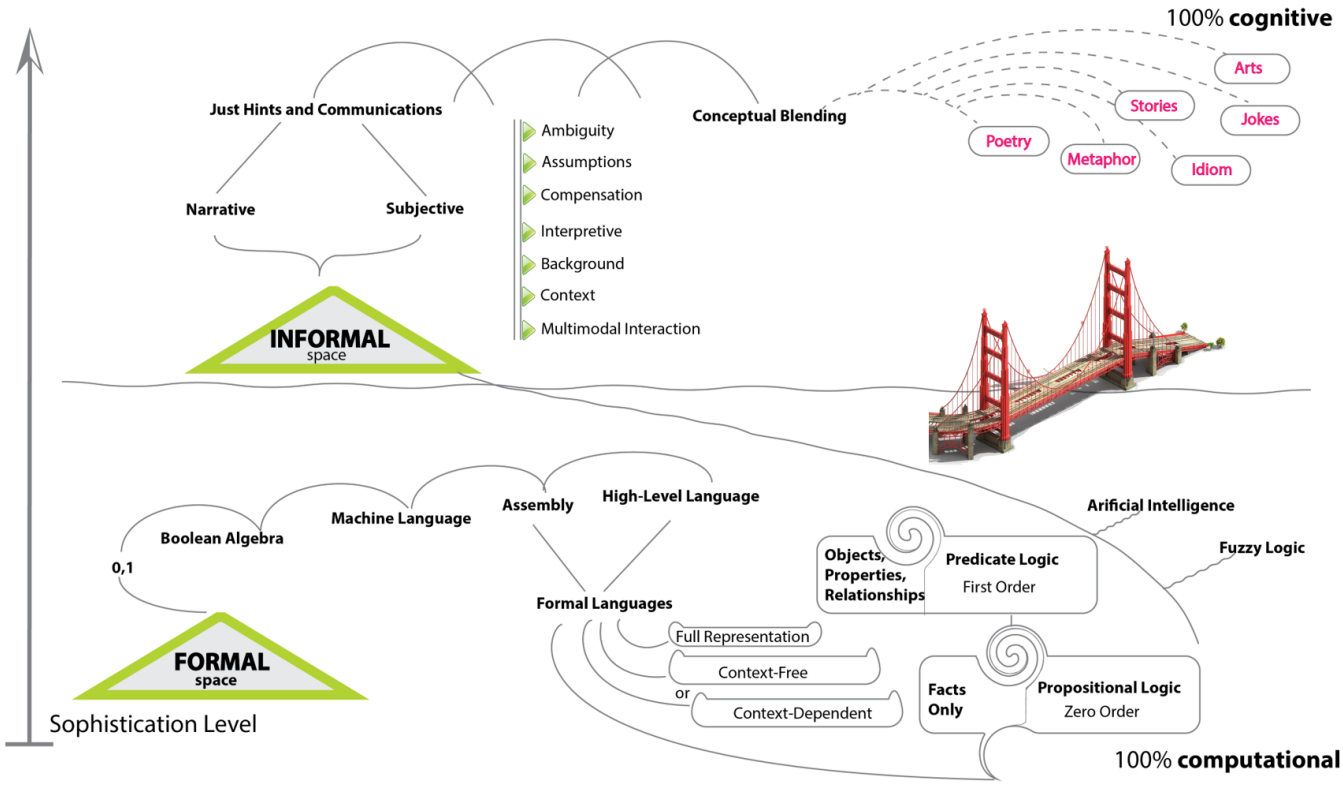


Fig. 4. A depiction of the evolution of first-level subjectivity and double subjective potentials.

summarization, automated text generation, and many others. To show the gains in logic and reasoning, Fig. 4 depicts advances made using current tools and new frontiers that double subjectivity may act as a bridge, such as poems, metaphors, idioms, jokes, art, and storytelling. This is to be achieved by exploring levels of narrative text.

3. Broader Impact of This Work

Media attention and framing can influence society concern about a specific topic. To observe the potential pathway of cognitive, informal, natural language, the authors focus on Uber as a case study. Uber intersects the economic and technology domain with characteristics of emergence, complex and dynamical forces, and the amplification of frames. Uber uses data as foundation for transforming the transportation ecosystem. It has played a key role in shaping the shared economy that heavily rely on social influence and online news frames. The framing of an issue has an impact on the ultimate course of the issue. The ways that Uber use framing has a direct impact on dynamical forces, such as, accessibility, mobility, shared value, or policy outcomes. Uber is a good case because it is context specific, complex, and characterized by multiple interacting forces. Newspapers in the U.S. regularly produce in-depth articles about automation of machines, emergence of platforms, and the influence of the crowd that are published online. However, the general public often has little experience with understanding the underpinning dynamical forces at work. As an example, amplified framing is more likely to influence uninformed respondents [14] or respondents with reduced exposure to or interest in an issue [15].

Therefore, consumers attitudes and beliefs about emergence and behavior shifts are likely influenced by the way reporters frame this issue.

This research can generally apply to any large, unstructured dataset that is affected (or amplified) by framing. Therefore, it can be extended to other applications such as health, economics or political discourse. It may lend multiple benefits to analyses of other domains, such as politics or presidential elections: 1) a tool for measuring amplification intensity over time and in real time, 2) a framework for exploring the social influence attributed to the double subjectivity with its primary property being the unique cumulative effect afforded in an online news network, 3) a model for exploring the dynamic forces at work when shifts occur in attitudes and beliefs, and 4) a real-time visualization of these dynamic forces.

4. Related Works

In the literature reviewed, no computational treatments were identified to address a) the concept of double subjectivity for automatic emergence of knowledge of narrative text or b) a model for exploring the interactions and implications of news frame choices. However, several works were identified that have some relevance for the research topic.

4.1. Text Summarization, Text Data Mining, and Information Retrieval

In computer science, scholarly research on latent meanings in the association between terms and documents to reveal relationships is found in literature related to text summarization, text data mining, and information retrieval. The earliest paper on text summarization identified for the current report is a description of work at IBM in the 1950s [16]. In his work, the authors proposed that the frequency of a particular word provides a useful measure of its significance, and he identified the concept "term frequency" (TF), which states that it is possible to identify the significant terms based solely on the term's calculated frequency within that document [16]. Term frequency relates to average information, or entropy, of a term or group of terms in the ranking of their relationship to each other.

Outstanding contributions to text data mining in the 1990s and beyond, with the advent of machine learning, include text representation and models construction [17-20]; data dimensions reduction re-search in feature extraction [21, 22]; and deep semantic mining [23].

The research on online news frames reported in this paper is motivated by framing theory, which focuses on understanding the latent meanings of observable messages in their contexts [15] and can provide important insight into how the presentation or framing of an issue affects the choices people make. Other disciplines have focused on framing. In linguistics research, similar approaches are also described as "latent semantic analysis" (LSA) [24]. Furthermore, social network analysis (SNA) focuses on the importance of relationships among interacting units [25].

4.2. Opinions and Social Influence

Whereas text data mining techniques and methods have given insight into first-level subjectivity, dynamic systems modeling provides insight into the double subjectivity that exists in narrative text. Models of social influence, cultural dynamics, and information diffusion (i.e., topic modeling, sentiment analysis, and opinion mining) are active areas of research [9, 1, 12, 13].

It has been posited that social impact theory explains the impact of a social group on an agent as being dependent on the prominence of the social sources, their proximity, and source group mass (i.e., the number in the group) [26]. Political opinions and Axelrod's cultural dynamics model behave similarly to these models that capture the interplay between selection and influence [27-30].

4.3. Networks

The ability to visualize vast amounts of data clearly creates a capacity for gaining insights about critical issues affecting society. At the most basic level, a network is a collection of points joined together in pairs [31]. Using Graph Theory, the data points are called nodes (or vertices) and the pairs connections are called edges (on links). Graph theory provides a strong mathematical foundation that helps us in visualizing the information geometric (special) aspects like the relationships be-tween different news sources, as well as it's progress and evolution over time (temporal aspects). Here, in this manuscript the terms nodes and edges are used. Networks is a powerful way to represent patterns of connections or interactions between the parts of a system [31]. The issues network is best expressed using a net-work to understand the underlying global and local structure of the network, the connections and interactions between nodes, the clusters of highly connected nodes, influence patterns, and migration paths.

5. Contextual Case

Founded in San Francisco, California in 2009, Uber Technologies describes itself as a transportation network company which develops, markets and operates the Uber car transportation and food delivery mobile apps [32]. In Ubers' short history, the

company that connects riders with drivers using the Uber mobile app has become one of the fastest growing companies in the United States and worldwide. With approximately $200K in seed capital from founders Travis Kalanick and Garrett Camp, UBER entered the market with an improved automated car service that offers better pricing over traditional taxi providers. This improved car service could be easily accessed on demand by consumers using a smartphone. It is on the basis of consumers shifts in values, attitudes, and behaviors from ownership to the shared economy—ride sharing—and the advances in technology that make access to product and services a realization that afforded Uber to become the prominent company it is today. In contrast to traditional taxi providers who own assets (e.g., cars, shuttles, or limousine), Uber use drivers cars for delivering convenience and efficiency. This change have inspired the shared economy to reframe the way in which businesses operate for reaching large numbers of people across cities in the United States and worldwide. As a result, the company has met many legal complications, court injunctions, complaints from taxi providers of unfair competition, and complaints from public utilities commissions about Uber's lack of compliance to safety rules that are adhered to by traditional taxi providers. These challenges did not impede Uber's influence and growth with its $68 billion paper valuation. In just 6 years, it surpass the valuation of 100 year-old companies like General Motors and Ford, as well as "traditional" short-term transportation companies like Hertz and Avis [33].

With Uber's recognition of consumer behavior shifts and advances in technology as emergent properties within complex and dynamical systems, the company set out to chart an alternative way for delivering cars and convenience to consumers, particularly through the use of social influence and opinions formation in online media and narrative text. Harnessing, the power of social influence, Uber evinces their reliance on consumers, lobbyist, and elites as influencers in online media as they a) created a team for community management, social-media outreach, and the creation of articles about Uber in online media and b) enlist celebrities, VIPs, and elite investors (i.e., Morgan Stanley, Saudi Arabia's Public Investment Fund, Goldman Sachs, and Fidelity Investments are among Uber's billion dollar investors). The emergence of consumers shifts, technology advances, and the interplay of complex and dynamical forces provides a useful context in which to explore Uber's approach to transform the transportation ecosystem through the use of social influence for gaining deeper understanding of the cumulative effects of double subjectivity and the expressive function that programmable insight offers.

Uber's mission to "bring safe reliable transportation to everyone everywhere" [34] is accomplished using its unique software platform and infrastructure that enable the company to meet consumers transportation needs with data rather than traditional methods, such as steel, glass, rubber, and salespeople. Uber's reliance on data to drive social influence and opinions formation, places a demand for the company to hire an enormous number of data scientist and data engineers to gain command of the vast amounts of data that derives from interactions between consumers, drivers, network traffic, pricing algorithms, and online media. These interactions leave traces of patterns about behaviors that indicate the interworking of forces that may go unnoticed due to the limitations of existing tools available in machine learning and NLP.

While text data mining techniques and methods have proven to give insight to first-order subjectivity (i.e., words, grammar, phrases, and sentences—syntax and semantics), complex and dynamical systems modeling offers insight into the cumulative effects of double subjectivity that exist in narrative text [8]. Uber's data scientist use tools that allows for treatment of disassembled factors and/or forces, however Uber data analytics could benefit from an account for the composite of data over space and time as a whole, not just its parts. An assemblage treatment of data will give Uber the needed insight for understanding its effective use of social influence and data-driven techniques and technology that enables them to reframe transportation and logistics globally.

Uber's data management and analytics can be improved on using the double subjectivity model for gaining programmable insights where spatio-temporal dynamics is the expected outcome as described in this manuscript. The double subjectivity model could benefit Uber by a) offering a visualization tool for managing data more efficiently, b) delivering accurate results of the various interacting factors and forces in a shorter time period that considers the spatio-temporal dynamics of incoming data, and c) a human level visualization tool. This could benefit Uber's data scientists from doing data analytics individually on dissembled data factors and forces. According to [8], the cumulative effects of the double subjectivity model has the ability to apply computational models and methods and hence software tools to leverage human cognitive ability to gain insight into very large narrative text. As a result, the double subjectivity model graphs will improve the accuracy of solving abstract engineering and data problems and will give Uber a tool for getting a handle on their effective use of social influence for shifting values, attitudes, and behaviors.

5.1 Forces of Emergence, Change, and Conversion

Elucidating the underpinning forces at work in complex interactions and social influence—the capacity to affect or be affected—is essential for an adequate explanation of emergence, change, and conversion. There were a number of forces that converge to make Uber an attractor and help situate Uber as market leaders. The timing of Uber's innovative mobile app coincides with shifts in consumers values and attitudes about ownership versus the shared economy. These shifts stem from forces of emergence, such as ubiquitous computing and mobility, economic pressures and recessions, the desire for improved conveniences, access to transportation in cities and communities where consumers were underserved or poorly served, a new awakening of limited resources, and Uber's use of online media for propagating a message about their elite consumers. "Coverage of elites in articles is a property of particular interest in this study, as this factor has an effect on the bias intensity that news sources adopt over time [8]." According to Druckman [35, 36], the way elites frame an issue is a driving force for shaping public opinion. An example should help to clarify the implications the interplay of forces within the complex interactions and social influence that gave rise to Uber's emergence. Elites are individuals in power relationships or positions of authority.

Imagine Uber decides to enter a city, it follows a number of aggressive steps to quickly penetrate the market. Uber enters a city secretly and uses the local media to influence consumers by letting others know that it has entered the market (i.e., free advertising). Likewise, a promotion offering first-time consumers free rides while marketing to recruit new drivers aggressively in the new city starts immediately. Uber ignores regulators or fights them with an army of lawyers. The causal effects of Uber's entrance into a city

is felt immediately as evidence by price reduction in transportation fares, an initial increase to drivers commissions, and large sign-up bonuses for on boarding drivers. Uber braces itself for the challenges that it must overcome when entering a city, particularly regulatory constraints and competitors complaints, therefore the company invest a great amount of their cash base (in billions) in lobbyist who act as carriers of influence in the online media and within policy makers circles. Leveraging consumers sentiment about the shared economy—ride sharing and a less expensive alternative to car ownership, Uber enlist consumers as influencers and carriers of their messaging in online narrative text as they report their support of more efficient and less expensive transportation services. According to [37], "Uber employees whose sole job in the 'playbook' is to gain political influence start pushing for regulations that legalize Uber's operations. In almost all cases, this has worked and Uber is given the chance to gain a firm foothold in the city." Economic forces caused by the recession of 2007 put into motion the need for individuals to seek alternative income sources as many Americans were experiencing unemployment and had much less disposable income. Uber drivers may work part-time and full-time with the freedom to set their own schedule. This income on demand is a perk for individuals needing to supplement their income or those who desire non-traditional work flexibility. What all this means is that Uber's consumers, drivers, lobbyist, competitors, and other stakeholders adopt and copy Uber's narrative of sharing, efficiency, and convenience which reporters use in online news. The double subjectivity model captures cumulative effects of copying and gives insight to complex and dynamical forces whereby gradual shifts in amplification-intensity in online news is made visible.

5.2 Uber's Future — Double Subjectivity as a Key to Programmable Insight

Uber's decision to rely on data for social influence affords the company to expand to new lines of business. For example, Uber recently expands to on-demand trucking, b) car-pooling, c) mass transit, and d) meal delivery services. A consequence of Uber's expansion is the need to manage a vast amount of generated data on a daily basis—primarily narrative text—,this the human cannot scale through the hiring of more data scientist. Unfortunately, the existing tools machine learning and NLP has available to analyze these large quantities of data are designed for structured text which are inadequate for handling the Uber's data. Furthermore, existing tools must go beyond identifiable variables involved in the spatio-temporal production of human-generated data to an understanding of the entire assemblage. This involves factoring in the interworking of complex and dynamical forces coupled with cultural dynamics and social influence such that effective adjustments may be made that will transform Uber's generated data into information advantages.

Modeling double subjectivity provides a human readable tool in real-time whereby gradual cumulative effects are visible that considers identifiable variables and interactions between diverse forces. This tool may give Uber programmable insight into social influence effects and shifts in values, attitudes, and behaviors. A tool for gaining programming insight into how the myriad human and non-human variables and forces affect one another in associations for shifting values, attitudes, and behaviors should reveal more about Uber's long-term capacity to grow using data as its driving force for market penetration.

6. Methods

6.1. Dataset Description

The news data for the study were collected from Google News,[1] a newsfeed aggregator. Each article was selected based on four characteristics: a) its publication by an online news source, b) the prominence of the news source that generated the articles, c) the frequency of news publications by sources over time, and d) the salient terms contained in the article. All articles were published on the Internet between January 2015 and March 2017.

Articles were gathered by limiting the search for keywords associated with Uber and transportation consequences. After the collection of the articles, the articles were stored in a database for undergoing data preprocessing that included removing errors and inconsistencies to improve data quality.

6.2. Issues Networks for Online News

The first step in the construction of an issues network for news frames is to define the nodes. The issues network is a collection of news sources and article nodes of interest joined together in pairs by edge connections. In this research, the news articles being published by online news sources and the news sources that generate the articles served as the nodes. Though the proposed model and network extends to most bodies of unstructured narrative text, where frame amplification may exist, this manuscript focus on the case of Uber for illustrating the way complex and dynamical forces effect shifts in attitudes, values, and behaviors. It is the reasoning of the authors that news sources play a critical role in the propagation of amplification about issues, and amplification in online news drives a wedge between evidence and beliefs [4].

Uber is transforming the way data is used for shaping on-demand transportation service. As such, one can expect that the number of articles being produced on the issue is an indicator of its prominence and the framing choices of the news sources. The framing of Uber case is key to shaping attitudes and values. Such framing also indicates areas of amplification. Connections are made between news sources through the capture of framing adoption over time as it happens during migration. Migration pathways can be understood as assemblages and adopters of online articles frames. This adoption is a result of the framing effects that happens when online news sources repeat a narrative and its neighboring nodes copy what is reported, thereby this news narrative increases prominence in the network and causes growth and clustering of the mass. Undoubtedly, reciprocal relationships (i.e., news sources using the same frame choice) can be valuable, but such relationships are a challenge to capture using the article content. This is largely due to news sources not crediting other news sources in reports. The resulting network aim is to capture structures that were created and maintained through repeated patterns of framing and to capture the migration path of those structures over time.

6.3. Graphs and Problem Formalization

Framing of critical socio-environmental issues in online news is a complex and dynamic social network, which can be studied as

[1] http://news.google.com/

a graph, as it is here. Graph analysis has become important in understanding the dynamic process in the production of news frames. Graphs allow for the visualization of the social network, showing interdependency of actors (or nodes) in terms of the social relationships such as friendship, kinship, or financial exchange [38, 39]. Graphs represent objects where order and disorder coexist. Graphs serve well for showing social interactions, influence, migration paths, and framing effects. For instance, a graph may expose which news sources hold central positions that function as points of prominence, control, and stability and may show edges that act as highways for lead relationships and exchanges of news frames. These intertwined dynamics, coupled with the vast amounts of online news being produced daily, make graphs an important tool for visualization and analysis of information flows pertaining to critical issues.

The overarching research problem is expressed through the construction of an issues network, and the model of a news frame issues network is presented in this paper as a graph $G = ({S_g}^t, {A_h}^t, {E_r}^t)$ consisting of a) S_g that denotes a set of source nodes containing $g = \{s_1, s_2, \cdots, s_g\}$ elements and b) A_h as a set of article nodes containing $h = \{a_1, a_2, \cdots, a_h\}$ elements that operate within the time window t. In general, t indicates one discrete time step of a reporting period. Let $S_g = \{C_i, F_i\}$ be the set of source properties, where a) C_i is the set of terms (i.e., sub-issue keywords or content descriptors) and b) F_i is the dominant frame choice (identity) that the news source uses. Each relation R_r has a corresponding set of edge connections, ${E_r}^t$, directed/undirected as edge elements. The issues network subscript $r = 3$, corresponds to the total number of relations. The edge set represents the communication channels between node pairs. A_h are children of S_g; therefore, they may inherit the properties of the parent news source, as shown in Fig. 5.

The relations (or rules denoted as R_1, R_2, R_3) for E_r edge connections are as follows:

$\boldsymbol{R_1}$. A & A (articles to articles). A non-directional edge connection is constructed for E_1 when $A_h(a) = \{a: (a_1, a_2, \cdots, a_h) \in E_1\}$ is the set of neighbors of $\boldsymbol{a_h}$ such that $\boldsymbol{a_h}$ represents articles with a similar sub-issue.

$\boldsymbol{R_2}$. A & S (articles produced by the same source). A directional edge connection is constructed for E_2 when $S_g \in E_2 \rightarrow A_h(a) = \{a: (a_1, a_2, \cdots, a_h) \in E_2\}$ is the source (or producer) of the set of neighbors of $\boldsymbol{a_h}$ articles, where $\boldsymbol{a_h}$ represents articles with similar sub-issues.

$\boldsymbol{R_3}$. S & F (sources with sub-issue to sub-issue with the same frame choice). A non-directional edge connection is constructed for E_3 when $S_g(s) = \{s: (s_1, s_2, \cdots, s_g) \in E_3\}$ is the set of neighbors of $\boldsymbol{s_g}$ such that $\boldsymbol{s_g}$ shares the same dominant frame choice.

Problem: Given $G = ({S_g}^t, {A_h}^t, {E_r}^t)$ about I issue, construct a graph for determining how frames are produced in online news media over time.

6.4. Model Setup

To model the news frame issues network, the authors of this paper propose formulating a new model for information feedback based on a social system. The new framework is within the tradition of previous frameworks. These are the previous models [27, 9, 11, 40, 41] for studying the evolution of personal position, social influence multi-dimensions, and forces that cause cultural shifts. Axelrod's model [9] is built on two simple assumptions: a) selection, a phenomenon in which people are more likely to interact with those who are more similar to them, and/or to be more receptive to influence from those who are similar and b) influence, a term that describes the case in which interactions tend to cause similarity among interacting actors. The model proposed in the present study promotes the additional emphasis of amplification that emerges as double subjectivity in narrative text. Although the behavior of the proposed model is similar to that in Axelrod's [9], when deciding how to interact in the network, the proposed model differs in that insight is given into the endogenous and exogenous forces that may be at play in dynamic systems. Therefore, the double subjectivity in the proposed social influence model integrates cultural dynamics extensions [9, 10-12, 42, 43] with the cusp catastrophe model [41]. The cusp catastrophe model is useful for describing nonlinear relationships such as those found in narrative text. In the proposed model, the particular focus is on the different forces at work in the amplified slant employed in news frames as it allows for the degrees of freedom whereby alternative pathways may be realized. The cusp catastrophe model allows for understanding the different forces at work when quantifying framing amplification and the migration path.

Conceptually, the switch from linear to nonlinear relationships involves taking into account not only a news source's capacity to select (or affect), but also another news source's capacity to be influenced (or affected). Thus, an important aspect of selection and influence is the characterization of dynamical systems not just by their properties, but also by their capacities.

Consider this contextual example. When a news event occurs that impacts society, it thereby becomes an issue. It is partly defined by its properties, such as the absolute baseline facts, as well as being in a certain state, like the event having taken place on a certain day. This same news framing, however, has the capacity to intensify its amplification that may have the causal effect of a revolt or eruption by the people, an exacerbation of

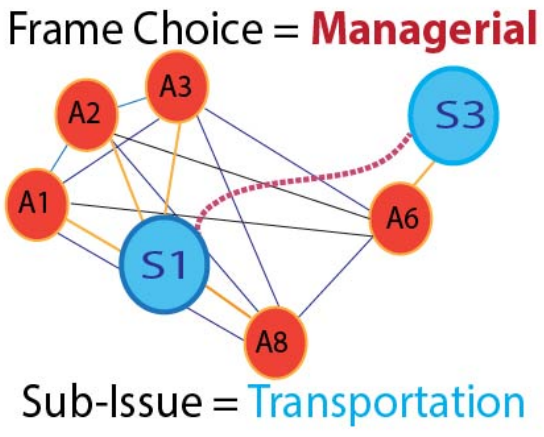


Fig. 5. A subset of the news frame issues network with two news source nodes with connecting articles using the same sub-issue and the same frame choice. The blue circles around notation S_g show online news sources node that produced the article and propagates a central organizing idea (i.e., framing), orange circles around A_h show an online news articles node, black lines show edge connections among neighboring articles on the similar issue, golden lines show edge connections of news sources who produce articles covering similar issue, and the red line shows a managerial news frame.

sentiment about a group of people, or the viral sensation of the news frame. This can happen through interactions with news sources that have the capacity to exercise strong amplification and to simultaneously weaken the agency of the people, group, or opposing view.

Properties are always reality based, since at any given point in time the facts—absolute baseline—are either true or false. Hence, facts are logical, formal, and rigid, but the causal capacity of news when exercising double subjectivity has a cumulative effect, which acts as a fabrication of what is real without being actual (or absolute). Unlike the limitations that facts in online news offer, amplification intensity within the construct of double subjectivity allows for the freedom to change dimensions through the parameterization of different forces—frame identity, justice, and amplification intensity—and distributions of sudden changes. The end effect is the perception of reality about an issue in online news that may offer new structure formations (or pathways) that lead to alternative views for shaping attitudes and beliefs. Stated differently, that which was only potential becomes a reality when changes in the different forces reach a certain critical threshold and a sudden jump happens. At this point, changes in beliefs can be tracked with the proposed model, even if the news framing amplification mechanism morphs the facts into something far from the truth.

The cusp catastrophe model calls this possibility space "the response surface space and control space," which represents all possible states in the system. In this research, the relationship among frame identity, justice (or a sense of fairness), and amplification intensity are acknowledged through the cusp catastrophe differential calculus formula, which allows the measurement of the rapidity or slowness with which forces can change. In the geometric approach to the calculus, each degree of freedom becomes one dimension of a possibility space. The space of possible states in the system allows for the differential relations between them to determine a certain distribution of stable states around points or loops of attractions. They have a fractal dimension (intermediate between one and two) and are referred to as "chaotic attractors" [40]. Thom's cusp catastrophe model [41] shows specific transitions of the forces, which have the tendency to cycle through the same set of states over and over [44].

The integration of cultural dynamics and the cusp catastrophe model for defining the new double subjectivity social influence model may shed insight into deep levels of expressivity in narrative text where amplification offers alternative pathways and possibilities. This remains an emerging research topic for understanding the production of language that may help in synthesizing vast amounts of unstructured text and learning online social behaviors.

The authors of this study explore the process that news sources use when producing online news articles. The main factors considered in this production are the facts, amplification intensity, and signals that derive from neighboring news sources. The signals that derive from news sources are likened to the reporter-relative or a floating baseline as mentioned previously. In the process of time, the news sources' absolute baselines will undergo adjustments as the amplification intensity changes.

6.5. Double Subjectivity Social Influence Model

The proposed double subjectivity social influence model integrates cultural dynamics extensions [9, 10-12] with the cusp catastrophe models [41, 44] for modeling double subjectivity in narrative text. The cusp catastrophe model is useful for describing nonlinear relationships such as those found in narrative text. The cusp catastrophe model allows for understanding the forces at work when quantifying amplification and the migration path.

Online news sources tend to produce articles about an issue I of interest to society, such as Uber emergence, changes, and sentiment that fosters shared value and cost savings. Additionally, news sources may receive signals β from other news sources on the importance of the issue, based on observed increase to the number of articles other neighboring news sources produce. They receive facts f about an issue. Each news source, then, selects a amplification strategy α using as the basis of its calculation the amplification intensity. More specifically, each news source S_g with interest in I issue at time t generates news articles

$$I(\mathrm{g}) = \alpha_g + \beta_g(f) \quad (1)$$

where α_g and β_g are the coefficients of the news source while f are the facts received. Furthermore, it is supposed that α_g corresponds to the endogenous propensity of the news source nodes to express its own amplification (or double subjectivity), while β_g represents the exogenous force in the network. In the first interaction, set $\alpha_g = 0$, denotes amplification intensity; this value will be calculated upon interacting with neighboring nodes and should change to give shape to the news source's absolute baseline. Time evolves in discrete steps t = 0, 1, 2, . . ., and $S_g(t)$ denotes the mass count on S_g at time t. The maximization of $\{S_g\}$ gives the sub-issue most important, M_i—a count of the largest cluster (the mass) of nodes reporting on an issue (case). This initialization of α_g and β_g provides access to new nodes to enter the overall conversation about Uber by linking to the most prominent mass, thereby, leveraging facts and the neighboring nodes of the relative baseline. This is similar to the operation observed in the hidden Markov model (HMM), whereby the news sources' absolute baseline is hidden with adjustments made over time. In contrast, the relative baseline is likened to the observations one can make in the HMM for getting a sense of the state [45].

Problem: Given $\{S_g\}$ interactions within graph G_t over time, determine the migration path and the nonlinear, causal linkages about $I(\mathrm{g})$.

At the start of the process, each news source $g \in S$ has a nonnegative node mass associated with it, corresponding to the fraction of the news sources that initially reports (i.e., through signaling) on similar sub-issues about Uber or transportation sub-issue. The dynamical system allows for each news source to switch frame choice, $F(e)$, and sub-issue interest, thereby enabling random selection of news source interaction. Also, news sources are susceptible to being influenced when they interact with neighboring news sources that share similar amplification intensity. The full state space may be calculated by counting the news sources as expressed as the mass vector $S_g(t)$.

6.6. Defining Amplification Intensity

The quantification of amplification intensity is situated in terms of the mathematical catastrophe theory [41]. Catastrophe theory is a branch of nonlinear dynamic systems theory that originated with the work of the mathematician Rene´ Thom [41] to help explain biological morphogenesis as one of the great

mysteries confronting mathematical biology. A key property in catastrophe theory is that the system under study is driven toward an equilibrium through its use of gradient descent or potential function for seemingly automatic guidance (i.e., through the law of attraction) occurring in the system, which is important in research on social influence, particularly when considering the property of convergence and stable states. The cusp model is the most well-known and the simplest model of catastrophe theory, positing that nonlinear transition within a system from one state to another is guided by two controlling variables: asymmetry and the bifurcation factor.

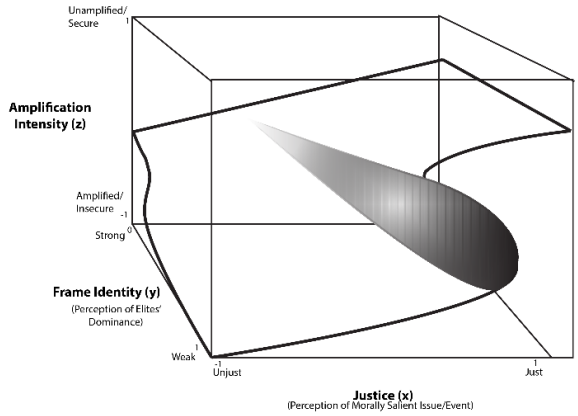


Fig. 6. The image depicts the cusp catastrophe model for measuring the shift in news sources amplification intensity.

Consider, for example, the social and economic inequality movement Occupy Wall Street. What began as small grassroots pressure points to address pay inequality among fast food workers and Walmart employees has transitioned to giving voice to 99% of the U.S. population and moving the political conversation in the U.S. election. This transition, from a grassroots protest in Manhattan's Zuccotti Park to making inequality and the wealth gap the core of the political race, is a catastrophe in accordance with catastrophe theory.

6.7. Model Variables Defined

Frames are a characteristic of human cognition structures, that operate simultaneously in human minds and texts. These latent structures of subjective interpretation are often produced and transmitted through the lens of power relations, hence elites [35].

The strategic use of managerial, economic, or conflict frames as expressivity in language embedded within online news, particularly when reporting on critical issues that impact society, have grave cost and consequences, which are far reaching. In the end, perceptions, beliefs, and attitudes by society become polarized or conflicted about critical issues even in the same domain, rather than society coming to a common consensus.

According to Lakoff [13], deep frame structures—where moral foundations exist—must be in operation in the mind during the production of language for justice about critical issues to prevail.

These emerging insights is justification for using justice in relation to frame identify as forces for understanding shifts in amplification intensity that lead to programmable insights about online news sources reporting on critical issues.

Fig. 6 graphically displays the possibility space of cusp model, an extension of the cusp catastrophe base model.

Let $M = \{Managerial, Economic, Human\ Interest\}$ and $S = \{Human\ Interest, Science\}$ be the aggregate count of the sets of frame choices obtained from the edge connection. S_g is assigned the frame identity corresponding to its F_i property at time t.

Definition: Amplification is represented by , a projection on the behavior surface for predicting the migration path that shows the gradual shift in amplification intensity of the online news source when shaping the frame narrative about the issue under study. This factor is a measure of the strength of amplification—that is, "weak amplification" is considered safe and is represented by *S(ecure)*, and "strong amplification" is considered insecure or harmful and is represented by *I(nsecure)*.

Definition: Asymmetry control (or normal) factor is represented by $\boldsymbol{x}$. This factor receives the label *Justice*, denoted on a scale that ranges from $J(ust)\ to\ U(njust)$. Justice is a measure of the news source's perception of forces about fairness where a multi-dimensional formulation is a desirable consideration. A utility-like function, we call the "aggregate" variables comprising of accessibility, mobility, disposable funds, low cost for ride share, and ubiquitous wireless technology are used to summarize forces that give one a sense of Justice.

Definition: Splitting factor or bifurcation factor is represented by $\mathbf{y}$. This factor receives the label *Frame Identity*, denoted as $S(trong)$ or $W(eak)$. This is the perceived dominance by elites when reporting the news.

The authors, here, argue that Frame Identity and the perception of Justice are key forces for predicting the amplification intensity of news sources when producing articles about emergent technology.

Frame Identity is one form of amplification; it is shaped by strategic devices for presenting prominent aspects and perspectives about an issue using a strong slant for the purpose of conveying latent meanings about an issue [3, 35, 46]. The position of the authors is that *Justice* is of equal importance, as it captures the perceived sense of fairness about an issue. A discretized approach to opinion research has been conducted in scholarly research [44] and applications of cusp catastrophe [47, 48].

7. Visual Language

Traces of visualization for communicating ideas predates to thousands of years. Recently, attention has turned to the usage of visualizations for audience influence and persuasion. Most importantly, with the influx of big data on the Internet, visualizations have become the premier tool for exploratory data analysis that enables the formation of hypothesis about a network.

Visualization is a powerful human-oriented perception and comprehension tool. It provides special usability benefits, particularly when the goal is to monitor and analyze vast amounts of multi-dimensional data over time. Visualizations can provide insights into the spatial and temporal informational flows and dynamics, they allow for inference and discovery.

The properties of visual, spatial, and time are essential to building human intelligence. As this research introduces the double subjectivity social influence model for exploring the interplay of the dynamical system about critical issues. In this case, Uber.

In this research, a visual language that comprises alphabets for expressing nodes and grammar for expressing relational edge

connections is presented. It is used for communicating double subjectivity cumulative effects for understanding the spatial-temporal dynamics underlying the network of collective influence of framing in online news and as a factor in the production of language in text narratives.

Table 1. Visual language notations, symbols, and definitions

NOTATION	SYMBOL
S_g	
Thick blue circles around the notation S_g denote online news sources node.	
A_h	
Thin orange circles around the notation A_h denote an online news articles node.	
E_1	
Solid black lines correspond to edge connections among neighboring articles covering the similar issue.	
E_2	
Solid lines with colors {red, green, purple} correspond to edge connections of news sources who produce articles covering similar issue within domains.	
E_3	
Dotted lines with colors {red, blue, green, purple, orange} correspond to news frame choice (or strategy).	

Fig. 7 is an illustration given to show scenarios where news sources belief converges to a state of consensus, as shown in Fig. 7(a)., and conflict, as shown in Fig. 7(b). For this illustration, one may use a Euclidean distance measure for showing the degree of change among news sources.

The overarching research inquiry may be explored and is expressed through the construction of an online news frame issues network. Table 1 is the visual language for expressing and exploring the double subjectivity social influence model using the issues network for understanding the network structure and underpinning dynamical forces.

Given behavior of news sources when reporting on issues that impact society, like Uber, are complex, the interactions are not static, even when measured for very short intervals. The double subjectivity social influence model, make visible dynamics that may be easily missed in big data platforms, such as the Internet. The instruments identified in Table 2 may be used as probes for gaining programmable insights into the network and the forces at work.

This research offers a method for time stepwise transitions when navigating the dynamic online news issues network. The double subjectivity social influence model provides a mechanism for control over belief learning using the values of α_g, while β_g offers news sources inside signaling mechanism by leveraging historical knowledge about neighboring nodes.

Table 2. Instruments for measuring

NOTATION	DESCRIPTION
G_t	Graph composed of nodes and edges with respect to time t
S	Total number of online news source nodes in a graph
A	Total number of online news articles nodes in a graph
E_a	Total number of article to article edges in a graph
E_s	Total number of article to source edges in a graph
E_f	Total number edges of sources employing the same frame in a graph
$S_g(t)$	Number of online news source nodes in a graph at time t
$A_h\ (t)$	Number of online news articles nodes in a graph at time t
$A\ (e)$	Number of article edges in a graph at time t
$S\ (e)$	Number of source edges in a graph at time t
$F\ (e)$	Number of frame edges in a graph at time t
b	Amplification Intensity Constant
$f(b)$	Amplification Intensity Function

8. Possibility Space for Transitions and Steady States

Nonlinear relationships represent a variety of possibilities and pathways of which the linear case is a limiting one. The authors propose to express the possibility space for the study using three variables as shown in Table 3. The possibility space represent simultaneous actuality and potential structures formed by differential interacting forces and distributions of singularities. The term singularities means an non-ordinary event that triggers a shift, such as, the repetitive framing of the need for car ownership in a way that bring into question its utility given its economic cost and utilization. The possibility space described here, at a minimum, is three-dimensions as we consider frame identity and an aggregate of justice as forces affecting the amplification intensity, which may cause shifts in attitudes. The

state of online news framing on the issue at a particular time is actual, while all the other available states are potential, waiting to be triggered into actuality by a force.

Table 3. Possibility space index

FRAME IDENTITY (SPLITTING FACTOR)		
INDEX-1	**DESCRIPTION**	**VALUES**
S	S(trong)	0
W	W(eak)	1

JUSTICE (NORMAL FACTOR)		
INDEX-2	**DESCRIPTION**	**VALUE**
J	J(ust)	1
U	U(njust)	-1
0	N(eutral)	0

AMPLIFICTION INTENSITY		
INDEX-3	**DESCRIPTION**	**VALUE**
S	S(ecure)	1
I	I(nsecure)	-1
0	N(eutral)	0

Note. Explanation of Index meanings. Index 1 represents the control variable frame identity, which is the perceived dominance by elites when reporting the news. Index 2 represents the sense of justice in the handling of the issue. Index 3 represents the dependent variable amplification intensity, which has the effect of signaling a secure (or safe) or insecure (or harmful) issue.

Index-1, Index-2, and Index-3 are combined to denote state transitions for updating shrinkages and growth. For instance, SJS denotes strong frame identity (S), a sense of justness (J), and a sense of security or being safe (S). SJS represents an actual state.

News sources' selection and influence occur through signaling news message amplification intensity. Table 4 indicates where news sources change amplification intensity based on interactions and it is an illustration of the gains and losses that may occur over time. Each column shows location of growth in the number of recipients switching from one news source to another, as a result of observed amplification intensity. The shrinkage and growth at a single site are proportional to the number of recipients and the number of senders signaling.

9. Discussion on Influence Dynamics and Migration Path

In this manuscript, models of cultural dynamics are integrated with the cusp catastrophe model to explore double subjectivity through quantifying the amplification intensity. The analysis shows that this measure (i.e., amplification intensity) has potential to change the network structure, as prominent mass count is not the only law of attraction at work. Figure 6 is a representation of the outcome of the model socio-temporal dynamics.

Although the proposed model is capable of scaling to full dataset capacity, a small subset of the network is used for this preliminary evaluation. Given online news sources receive signals that provide a relative baseline—leveraging other news sources' opinions, interpretations, and perceptions—it is most likely that a news source will choose interactions with similar sources. A chance exists for a news source to interact with another news source based on an increase in its amplification intensity although no connection exists. However, the chance is higher that news sources will interact only with connected sources; thus, breaking out of the cluster is caused by catastrophic shifts in amplification.

Table 4. Migration possibility space.

	Signaling content (state of news sender)						
Recipient	**SUI**	**S00**	**SJS**	**WUI**	**W0I**	**W0S**	**WJS**
SUI			S00			WUI	W0I
S00	SUI		SJS				
SJS	S00			WOS	WJS		
WUI	SUI		W0I				W0I
W0I	S00						
W0S			S00				
WJS	W0S		W0S	WOS			

With the amplification intensity function, a variant of the cusp catastrophe model shows possibility space in behavior when making decisions about trust and distrust. Any space on the surface response represents the state of the observed news source. As the news source perceives that the issue will not lead to fair consequences, the cusp triggers them to start attributing blame toward the population causing harm. For instance, the most recent report of Uber previous CEO distracting behavior. The contextualization of this case with Uber shows that a news source with ties to nodes that employ the frame choice of perceived unfairness (unjust) will notice that the sources more aligned with elites are becoming increasingly bullish on the value of the emergent technology issue that promises to deliver improved efficiencies and conveniences. These news sources who initially framed the issue as unjust attract other news sources (i.e., homophily) because their narrative is spreading. When feelings of unfairness intersect with the dominance of the elites, the news sources start scanning for those to whom to attribute blame, thereby entering a space of distrust and strong amplification. This kind of model is associated with hysteresis, making it hard to shift between surface response planes, as the migration path will be different.

10. Result of Research

Uber is a major international company with millions of customers. While the company uses advanced technology to offer its services, its main reliance is on influencing public opinions to improve its acceptance, reputation, and hence maintain a sustainable growth. Uber identified the need to keep a close eye on what is being said about the company in online news and different media. The current technology largely depends on humans—and traditional analytical tools—to follow-up and analyze Uber-related media, which can be expensive and limited in scale.

This manuscript presents the programmable insight approach that allows the new software to manage and analyze very large number of digital data. Thereby, providing immediate insight into different Uber social influence as characterized through opinions formation, network growth or shrinkage, amplification shifts, and news influence on consumer behaviors. This approach has the

potential to give Uber a large scale, real-time overview of digital data social influence and causal effects of complex dynamics, such as their sources, growth rate, or effects of frame choice.

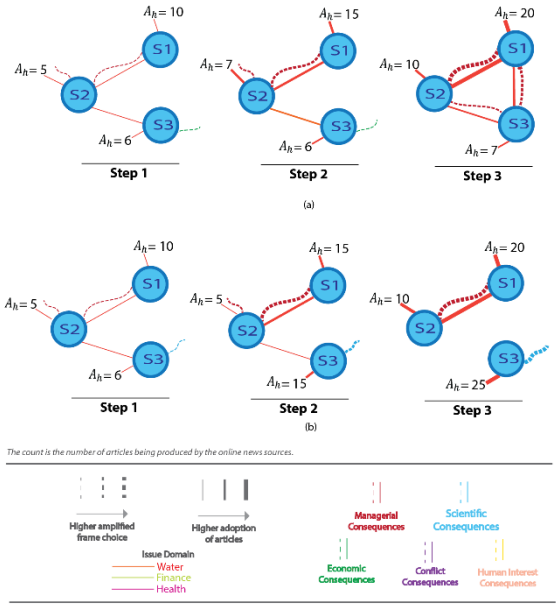


Fig. 7. A visual representation of the model. This is an illustrative scenario of news source interactions where Double Subjectivity has spatial-temporal dynamics and social influence with causal effect. Diagram (a) shows interactions that results in consensus. Whereas, (b) shows the formation of an alternative path that may be expressed as conflict in beliefs or attitudes about Uber. This depiction shows the five standard communications frames used in this research and three domains where vast amounts of narrative unstructured text exist online.

Over time, the programmable insight software will track the progress and trend of online data, providing Uber executives with a true map of ongoing opinions. Uber will be able to enhance positive trends as well as observe negative opinions early on, track their influence, and manage any related problems early on.

Results will be displayed visually using a graph with nodes representing news sources and articles produced, and edge relationships representing the frame type and strength of the associated influence. A human executive will be able to read the graph and observe any new node appearing on it, as well as influence level measures; represented by the number of links going out of the node and their thickness. This node or edge metric will immediately inform executives the emergence of new social influence and opinion sources (the node), as well as their significance (the number and thickness of edges).

While these metrics can be displayed in an excel sheet, it renders itself much easier to the graph theory, and a very large amount of information can be comprehended by a human person much faster.

The performance gain is threefold:

1. *Spatial*: The ability to observe, analyze, and provide "Insight" on a very large number of digital data emanating from many sources in real-time. Traditionally, this would require an army of human analysts using simple tools to work on each document, which could be very time consuming.
2. *Temporal*: The software will be able to continuously observe a large number of online news and other digital data over an extended period of time, and displaying the temporal progress of opinions, showing not only all existing ones, but the history of each news source, and how much it has been influencing others. This will emphasize the dynamics of opinion, and will help focus on significant ones while ignoring others with less or diminishing influence.
3. *Cost*: Running the programmable insight software will cost less than the efforts of human analysts.

11. Conclusion

In this manuscript, a new concept was presented that allows for an adjustment of the absolute baseline of a news source (or agent) that takes into account one's amplification for expressing double subjectivity through news frames. Further, offered in this research is a) the first known formal news frame issues network in computer science, b) a model for learning the migration paths and patterns about issues, and c) the first known formulation of a amplification intensity function using the cusp catastrophe model for showing the possibility space and gradual shifts in views that are sensitive to news framing amplification. Preliminary experiments suggest that the integration of cultural dynamic models with a cusp catastrophe model is promising for exploring double subjectivity for revealing latent relationships found in online news articles. Future work will involve testing other hypotheses and unfolding the potential function to explore conditions of convergence. Also, future work will give treatment to other hypotheses associated with amplification that leads to trust or distrust, as a result of perceived positive consequences.

Conflict of Interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

[1] L. H. Cheeks, A. Gaffar, Exploring Double Subjectivity Through News Frames in Online News Sources: A Network Approach in IEEE 11th International Conference on Semantic Computing (ICSC), 2017.

[2] L. Skitka & E. Sargis, The internet as psychological laboratory. Annu Rev Psychol 57: 529–555, 2006.

[3] R. M. Entman, "Framing: Towards clarification of a fractured paradigm", Journal of Communication, 43(4), 51-58, 1993.

[4] R. K.Garrett, B. E. Weeks, R. L. Neo, Driving a Wedge Between Evidence and Beliefs: How Online Ideological News Exposure Promotes Political Misperceptions, Journal of Computer-Mediated Communication, 4 August 2016.

[5] J. Tankard, The empirical approach to the study of media framing. In S. Reese, O. Gandy, & A. Grant (Eds.), Framing public life (pp. 95–106), 2001, Mahwah, NJ: Erlbaum.

[6] W. J. Severin, J. W. Tankard, Communication theories: origins, methods, and uses in the mass media, 4th Edition, Longman Publishing Group, New York, 1997.

[7] S. K. An, K. K. Gower, "How do the news media frame crises? A content analysis of crisis news coverage", Public Relations Review 35, 107–112, 2009.

[8] L. H. Cheeks, T. L. Stepien, D. M. Wald, and A. Gaffar, Discovering news frames: An approach for exploring text, content, and concepts in online news sources, International Journal of Multimedia Data Engineering and Management (IJMDEM), 7, pp. 45–62, 2016.

[9] R Axelrod. The dissemination of culture - A model with local convergence and global polarization. Journal of Confict Resolution, 41(2):203{226, APR 1997. ISSN 0022-0027. doi: f10.1177/0022002797041002001g.

[10] D. Kempe, J. Kleinberg, S. Oren. and A. Slivkins, Selection and influence in cultural dynamics. Network Science, 4(01), pp.1-27, 2016.

[11] D. Bindela, J. Kleinberg, & S. Orenb, How bad is forming your own opinion?, Games and Economic Behavior, 92, 248–265, 2015.

[12] D. Kempe, J. Kleinberg, and ´ E. Tardos, "Maximizing the spread of influence through a social network," in Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp. 137–146, 2003.

[13] G. Lakoff, Why it Matters How We Frame the Environment, Environmental Communication Vol. 4, No. 1, pp. 7081, March 2010.

[14] G.F. Bishop, R.W. Oldendick, A.J.Tuchfarber,. Effects of presenting one versus two sides of an issue in survey questions. The Public Opinion Quarterly 46, 69–85, 1982.

[15] D. A. Scheufele, "Framing as a theory of media effects", The Journal of Communication, 49(1), 103-122, 1999.

[16] H. P. Luhn. "The Automatic Creation of Literature Abstracts," IBM Journal of Research and Development, 2, No. 2, 159, April 1958.

[17] R. Steinheiser, and C. Clifton, "Data Mining on Text." 2012 IEEE 36th Annual Computer Software and Applications Conference IEEE Computer Society, p. 630, 1998.

[18] A. Tan, N. Zhong and L. Zhou, "Text Mining: The state of the art and the challenges." Proceedings of the PAKDD Workshop on Knowledge Discovery and Data Mining, 65—70, Beijing, China 2000.

[19] S. Tan, X. Cheng, B. Wang, H. Xu, M. M. Ghanem, Y. Guo, "Using dragpushing to refine centroid text classifiers", In Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '05). ACM, New York, NY, USA, 653-654. 2005.

[20] M. Tsytsarau & T. Palpanas, Survey on mining subjective data on the web, Data Mining and Knowledge Discovery, 24(3), 478-514, 2012.

[21] H. Karanikas, C. Tjortjis, and B. Theodoulidis. "An Approach to Text Mining using Information Extraction", Proc. Workshop Knowledge Management Theory Applications (KMTA 00), 2000.

[22] Q. Hu, D. Yu, Y. Duan, & W. Bao, A novel weighting formula and feature selection for text classification based on rough set theory, Natural Language Processing and Knowledge Engineering, 2003, Proceedings. 2003 International Conference on. IEEE, 2003.

[23] B. Liu, Sentiment analysis and opinion mining, Synthesis lectures on human language technologies, 5(1), 1-167, 2012.

[24] S. Dumais, G.W. Furnas, T.K. Landauer, S. Deerwester, Using latent semantic analysis to improve information retrieval. In Proceedings of CHI'88: Conference on Human Factors in Computing, New York: ACM, 281-285, 1988.

[25] C. Fellbaum, "A Semantic Network of English: The Mother of all WordNets," in: Computers and the Humanities 32: 209-220. 1998.

[26] B. Latane′. The psychology of social impact. American psychologist, 36(4):343, 1981.

[27] Eli Ben-Naim, Paul Krapivsky, and Sidney Redner. Bifurcations and patterns in compromise processes. Physica D, 183, 2003.

[28] C. Castellano, S. Fortunato, and V. Loreto, Statistical physics of social dynamics. Reviews of Modern Physics, 81:591–646, 2009.

[29] G. Deffuant, D. Neau, F. Amblard, and G. Weisbuch, Mixing beliefs among interacting agents, Advances in Complex Systems, 3:87–98, 2000.

[30] R. Hegselmann and U. Krause, Opinion dynamics and bounded confidence: Models, analysis and simulation, Journal of Artificial Societies and Social Simulation, 5(3), 2002.

[31] Newman, M. E. J., Networks an Introduction. Oxford University Press, 2011.

[32] UBER, Retrieved from https://uber.com, May 2017.

[33] G. Petropoulos, Uber and the economic impact of sharing economy platforms, retrieved from http://bruegel.org/2016/02/uber-and-the-economic-impact-of-sharing-economy-platforms/, 2016.

[34] L. Bradford, What You Need to Know About Getting a Tech Job at Uber. Retrieved from https://www.thebalance.com/what-you-need-to-know-about-getting-a-tech-job-at-uber-4036276, 2016.

[35] D. Chong & J. Druckman, Framing public opinion in competitive democracies. American Political Science Review, 101(4), 637-655, 2007a.

[36] D. Chong & J. Druckman, A theory of framing and opinion formation in competitive and elite environments. Journal of Communication, 57, 99-118, 2007b.

[37] A. Henshall). The Aggressive Processes Uber is using for Global Expansion. Retrieved from http://www.business2community.com/strategy/aggressive-processes-uber-using-global-expansion-01827929#G3L6C8oArOWPii3q.97, 2017.

[38] K. C. Chen, M. Chiang, H. V. Poor, "From technological networks to social networks," published in IEEE J. Sel. Areas Commun.., 2013.

[39] Y. F. Chou, R. G. Cheng, and H. H. Huang, "Performance evaluation of referral selection algorithms in human centric communication networks,"in Proc. 2012 International Symposium on Wireless Personal Multimedia Communications, pp. 638–642, 2012.

[40] I. Stewart, Does God Play Dice?: The Mathematics of Chaos, Basil Blackwell (Oxford), pp 84–94, 1989.

[41] R. Thom. Structural Stability and Morphogenesis, 1972. D. H. Fowler (trans.). Reading, MA: W.A. Benjamin.

[42] T. A. DiPrete and G. M. Eirich, Cumulative Advantage as a Mechanism for Inequality: A Review of Theoretical and Empirical Developments, Annual Review of Sociology, Vol. 32: 271-297, Volume publication date August 2006.

[43] M. McPherson, L. Smith-Lovin, and J. M. Cook, Birds of a feather: Homophily in social networks, Annual Review of Sociology, 27:415–444, 2001.

[44] P. Sobkowicz, Discrete model of opinion changes using knowledge and emotions as control variables. PloS one, 7(9), e44489, 2012.

[45] L. R. Rabinerand B. H. Juang,"An introduction to hidden Markovmodels," IEEE ASSP Mag., vol. 3, no. l, pp. 4-16, 1986.

[46] B. K. Jaworski, Media Framing Of U.S. Health Care Reform: A New Era Or Reinforcing Dominant Ideologies Of Health And The Health Care System?, 2012.

[47] T. F. Homer-Dixon, Environment, Scarcity, and Violence. Princeton University Press, 1999.

[48] Homer-Dixon, T. F., The upside of down: Catastrophe, creativity, and the renewal of civilization. Washington: Island Press, 2006.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1693-1710 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

1-D Wavelet Signal Analysis of the Actuators Nonlinearities Impact on the Healthy Control Systems Performance

Nicolae Tudoroiu[*1], Sorin Mihai Radu [1], Elena-Roxana Tudoroiu [2], Wilhelm Kecs [2], Nicolae Ilias[3], Liana Elefterie[4]

[1]*Mechanical and Electrical Engineering Faculty, University of Petrosani, 332006, Romania*

[2]*Science Faculty, Mathematics and Informatics, University of Petrosani, 332006, Romania*

[3]*Mechanical and Electrical Engineering Faculty, University of Petrosani, 332006, Romania*

[4]*Spiru Haret University, Faculty of Economics Sciences, Constanta, Romania*

ARTICLE INFO

Article history:
Received: 01 June, 2017
Accepted: 11 July, 2017
Online: 04 September, 2017

Keywords:
1-D wavelet analysis
wavelet transform
fault detection and isolation
neutralization process
HVAC control systems

ABSTRACT

The objective of this paper is to investigate the use of the 1-D wavelet analysis to extract several patterns from signals data sets collected from healthy and faulty input-output signals of control systems as a preliminary step in real-time implementation of fault detection diagnosis and isolation strategies. The 1-D wavelet analysis proved that is a useful tool for signals processing, design and analysis based on wavelet transforms found in a wide range of control systems industrial applications. Based on the fact that in the real life there is a great similitude between the phenomena, we are motivated to extend the applicability of these techniques to solve similar applications from control systems field, such is done in our research work. Their efficiency will be demonstrated on a case study mainly chosen to evaluate the impact of the uncertainties and the nonlinearities of the sensors and actuators on the overall performance of the control systems. The proposed techniques are able to extract in frequency domain some pattern features (signatures) of interest directly from the signals data set collected by data acquisition equipment from the control system.

1. Introduction

This paper attention is focused now on the most likely actuators prone to failures during operation for a particular case study such as for example, an electro-pneumatic control valves integrated in the forward path of a neutralization wastewater plant control structure. These control valves dose the input flows of the acid and base reactants and are controlled by a proportional-integral (PI) controller in order to keep the pH-level of the neutralized solution at its target value. A pH-probe measures pH-actual value of the neutralized solution inside the reactor during the transient and steady state neutralization process, transmitting its feedback to the controller with time delay. Typically, the cause roots of the majority faults in neutralization control processes are the result of unexpected control valves failures during the frequent opening and closing operations, due to the backlash, dead-band, leakage, and blocking. In the literature several works are dedicated to identify some of the control valve critical failures, as a fuzzy classification solution for fault diagnosis of valve actuators in 2003 well documented in [1]. Also, in 2005 are proposed methods to detect and to diagnose faults in HVAC control systems, including backlash, based on frequency and spectral analysis such as in [2]. Furthermore, in 2006 a graph method that describes each fault by three levels of knowledge is suggested in [3] by using a structural analysis as a powerful tool for early determination of the possibility to detect and isolate the faults. The results evaluated on the DAMADICS control valve benchmark model reveal "how to determine which faults in the benchmark need further modeling to get desired isolation properties of the diagnosis system". In 2007 was proposed a nonparametric statistical method in order to diagnose at least four valve failure issues, among them the backlash, dead-band, leakage, and blocking, as is mentioned in [4]. Moreover, in 2007 was proposed a new method for detection and estimation of

[*]Corresponding Author: Nicolae Tudoroiu, John Abbott College, 2127 Lakeshore Road, Sainte-Anne-de-Bellevue, QC, H9X 3L9, Canada
Email: ntudoroiu@gmail.com

backlash in control loops such in [5]. The detection procedure is automated and based on normal operating data, without measuring the backlash output. In addition, the proposed detection procedure provides an estimate of the dead-band caused by the backlash, in order to offer all information needed to compensate for the backlash. The authors from [4] revealed also in 2012 an interesting statistic record that confirms again that "among the most frequents "loop illnesses", the valve has one major disorder: around of 30% of all valves has any degree of damage, being responsible for increase significantly loop variability. Two of the most frequent valve injuries are stiction and backlash". The stiction causes limit-cycles in the loop that increases its variability, while the backlash increases the loop variability, reducing the control performance and inserting limit-cycles only for integrator plants, similar to stiction failure [4].

The distinction between stiction and backlash that generate limit cycles in the case of integrator plants is done based on a data-based method developed in [4]. Well, it sounds that the closed-loop patterns features provide enough information to verify when either stiction or backlash is available [4]. Also in 2012 a new nonlinear control valve model was developed in [6] and its effectiveness in simulating valve stiction using MATLAB/SIMULINK software package was demonstrated. In this research defective measurement and control loops equipment, in particular a pH and acid concentration loops with a control valve as actuator are under investigation. More precisely, we assume that the neutralization control process is subjected to known deterministic disturbances, and its pH level and acid concentration are controlled separately by two negative feedback closed-loops by using a state feedback control law strategy with two PI separate controllers integrated in the closed-loop control structures. As a most suitable real time implementation and analysis tools are the Wavelet and Signal Processing Toolboxes provided by the one of the most powerful and well spread tools from the market such as the MATLAB/SIMULINK software package. This paper is organized as follows: in section 2 a brief control valve nonlinearities –terminology is introduced. In section 3 is made a brief description of the wavelet. In section 4 is introduced the neutralization process as a case study proposed in this paper. In section 5 are presented the main results obtained in MATLAB SIMULINK in time domain related to process modeling and closed-loop control. In sections 6 and 7 are presented a multi signal 1-D wavelet analysis used finally as a basics detection tool of the actuators faults in control systems. The paper ends with the concluding remarks.

2. The Control Valve Nonlinearities -Terminology

To understand better the nonlinearity control valve effects during its operation in order to control the input acid reactant flow on the overall performance of the neutralization control process in open-loop, as well as in closed loop, the key concepts that inspire the following discussion of static and dynamic friction in control valves need first to be defined. Furthermore, the reader gets a better insight of the stiction (static friction) and backlash mechanisms and a more formal definition of all possible valve actuator nonlinearities well documented in [7]. This section reviews the American National Standard Institution's (ANSI) formal definition of terms related to the control valve nonlinearity effects, in a similar way as is presented in [7]:

1. **Backlash:** ''In process instrumentation, it is a relative movement between interacting mechanical parts, resulting from looseness, when the motion is reversed''.
2. **Hysteresis:** ''Hysteresis is that property of the element evidenced by the dependence of the value of the output, for a given excursion of the input, upon the history of prior excursions and the direction of the current traverse. It is usually determined by subtracting the value of dead-band from the maximum measured separation between upscale-going and downscale-going indications of the measured variable (during a full-range traverse, unless otherwise specified) after transient have decayed".

This concept is illustrated in Figure 1 (a) and (c)[7] for a pneumatic control valve with the layout shown in Figure 2 [8].

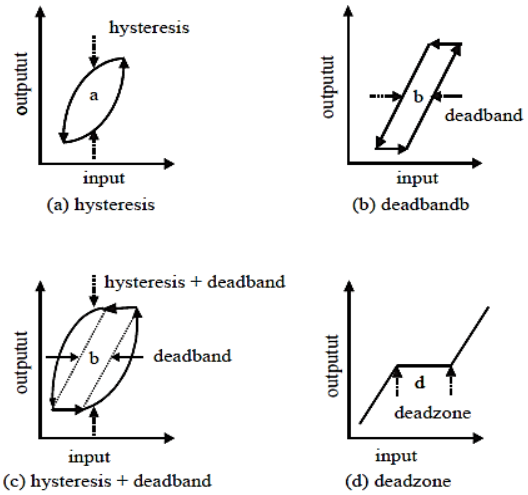


Figure 1: The hysteresis, dead-band, and dead-zone valve nonlinearities (screenshot view [7], according to ANSI/ISA-S51.1-1979).

3. **Dead-band:** ''In process instrumentation, it is the range through which an input signal may be varied, upon reversal of direction, without initiating an observable change in output signal'', as is shown in Figure 1(b).

The dead-band is not a bijection relationship; it is characterized by an input-output phase lag, expressed in percent of its span [7]. Moreover, a combination of the effects of dead-band and the hysteresis may be met as a new nonlinearity in a control valve actuator, as is shown in Figure 1(c). "Some reversal of output may be expected for any small reversal of input. This distinguishes hysteresis from dead-band'' [7].

4. **Dead-zone:** ''It is a predetermined range of input through which the output remains unchanged, irrespective of the direction of change of the input signal'', as is shown in Figure 1(d).

Unlike dead-band, the dead-zone doesn't produce an input - output phase leg, so it is expressed by a bijective input-output relationship.

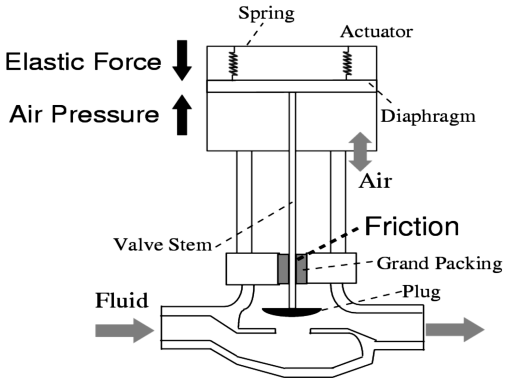


Figure 2: The structure of the pressure control valve (snapshot view [8])

These definitions reveal that the term ‘‘backlash’’ specifically applies to the slack or looseness of the mechanical part when the motion changes its direction [7]. In addition, the above ANSI (ISA-S51.1-1979) definitions illustrated clearly in Figure 1 prove without doubt that hysteresis and dead-band nonlinearities have distinct effects on a control valve. More precisely, the dead-band can be quantified in terms of x-axis input signal span, while hysteresis refers to a separation in the y-axis measured (output) response [7]. The combined effects of the static and dynamic effects of the previous nonlinearities generate one of the most complex control valve nonlinearity known under the name of control valve stiction (static friction). Its input–output characteristics consist of four components: dead-band, stick-band, slip jump and the moving phase of a stiction faulty valve as is shown in Figure 3 [7].

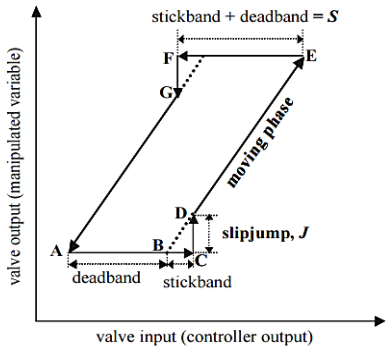


Figure 3: The characteristics of a sticky faulty control valve (Screenshot view [7])

The behavior of a sticky faulty valve is completely described in [7] as ‘‘a property of an element such that its smooth movement in response to a varying input is preceded by a sudden abrupt jump called the slip-jump. Slip-jump is expressed as a percentage of the output span. Its origin in a mechanical system is static friction which exceeds the friction during smooth movement’’. In this description the dead-band and stick-band represent the behavior of the non-moving control valve, while its input keeps changing. The slip jump releases suddenly the potential energy stored in the control valve chambers due to its stem high static friction in order to balance the kinetic energy as the valve starts to move. More precisely, the magnitude of the slip jump is one of the main causes in generating the limit cyclic behavior by the control valve stiction [7]. Once the valve slips, it continues to move until it sticks again (e.g., in the point E from Figure 3), and during all this moving-phase was proved that the dynamic friction may be much lower than the static friction [7]. The definition of the stiction well illustrated in Figure 3 can be considered as a rigorous description of the effects of friction in a control valve. Closing, these definitions will be very useful in the following sections to characterize the pattern features of the impact of several actuators nonlinearities in industrial control systems practice; under consideration will be the both continuous and discrete time state-space models of healthy and faulty open or feedback closed-loop control systems configurations. The healthy and faulty control systems responses from time domain will be converted in frequency domain by using different types of wavelet transforms that will be defined in the next section. Based on wavelet analysis in frequency domain will be extracted the matching pattern features that characterize each control valve nonlinearity under investigation.

3. Wavelet Signals Processing Technique Approach

The signal processing approach is one of the most used techniques used in practice for signal analysis. It is focused on particular signal characteristics, dealing also with the effects of the “white” or “colored” noises that contaminate the useful signals, by using statistical methods, such as the signals correlation and autocorrelation functions, covariance and cross-covariance, power spectral density, likelihood, or the autoregressive-moving-average (ARMA) models [9]. Frequency analysis is suitable for detecting the signals that contain particular meaningful frequency information, very useful to be used for example to detect, diagnose and isolate the healthy or faulty behavior of the control systems as response to the effects of a particular actuator nonlinearity or sensor faults, known in the control system literature as fault detection and isolation (FDI) techniques [10]. In this paper work are extracted only some of the pattern features of the healthy or faulty responses corresponding to control valve (actuator) nonlinearities under investigation. The wavelet transform is useful to detect the abrupt changes (signals pattern features) in the faulty actuators or sensors. It provides a useful set of time-scale or time-frequency domain tools and techniques to operate on a large range of signals [9]. The wavelet transform carries out a special form of analysis by shifting the original signal from the time domain into the time–frequency [11]. According to [11] “the idea behind the wavelet transform is to define a set of basic functions that allow an efficient, informative and useful representation of signals”, and first preliminary attempts to demonstrate the new theory based on the wavelet analysis started with the modeling of “a certain signal by a combination of translations and dilations of a simple, oscillatory function of finite duration, named wavelet “, and this technique is referred to as a continuous wavelet transform (CWT) [9]. In [11] also is mention that “Having emerged from advancement in time–frequency localization from the short–time Fourier analysis, the wavelet theory provides facilities for a flexible analysis as wavelets figuratively “zoom” into a frequency range. Moreover, the “wavelet methods constitute the underpinning of a new comprehension of time–frequency analysis”. Based on CWT many other wavelet analysis techniques have been developed in the literature [9], among them, due to paper space limitation, only one of these it will be considered. Depending on the intended purpose, if low-frequency information is required then the wavelet analysis deals with long time intervals, otherwise it will deal with short time intervals when high-frequency information is preferred. In the following we will give some

definitions that can be of real interest for the readers in order to have a good insight on all these concepts related to wavelet transforms as well as their manipulation in signals analysis and processing.

CWT Definition 1: The continuous-time wavelet transform of one-dimensional real-valued function, $f(t) \in L^2(R)$-Hilbert space, measurable and square-integrable, i.e. $\int_{-\infty}^{+\infty} |f(t)|^2 dt < \infty$, with respect to a wavelet function $\psi(t)$ is defined as the sum over all time of the signal multiplied by scaled, shifted versions of function $\psi(t)$ [9]:

$$Wf(s,u) = \frac{1}{\sqrt{s}} \int_{-\infty}^{+\infty} f(t) \psi^*(\frac{t-u}{s}) dt \quad (1)$$

where $\psi^*(t)$ represents the complex-conjugate of the "*mother wavelet*" $\psi(t)$, a real or complex-valued continuous time function that must satisfy the following two conditions:

$\psi(t)$ is a function with zero-average:

$$\int_{-\infty}^{+\infty} \psi(t) dt = 0 \quad (2)$$

$\psi(t) \in L^2(R)$, i.e. is a squared-integrable function, i.e. $\int_{-\infty}^{+\infty} |\psi(t)|^2 dt < \infty$.

A "mother wavelet" is a waveform for which the most energy is restricted to a finite duration [9], with the mention that there are an infinite number of the functions that candidate to be considered as a "mother wavelet".

In Eq. (2) the variable s is a "scale" or "dilation" variable that performs a stretching or compressing action on the "mother wavelet" while the variable u is referred as a "time shifting" or "translation" that delays or hastens the signal start.

More precisely, the function $\psi^*(\frac{t-u}{s})$ means that the mother wavelet $\psi(t)$ is shifted over time with u units of time and s times dilated. Therefore the wavelet analysis is a powerful tool that provides a time-scale view of the signal under investigation.

Denoting by

$$\overline{\psi}_s(t) = \frac{1}{\sqrt{s}} \psi(-\frac{1}{s}) \quad (3)$$

then the CWT of the continuous function $f(t) \in L^2(R)$ can be expressed by a the following convolution product (*):

$$Wf(s,u) = f * \overline{\psi}_s(u) \quad (4)$$

Also, it is worth to emphasize that the result of applying CWT to a signal under investigation is a wavelet coefficient vector that is a function of scale and translation g(s, u). An interesting interpretation related to a wavelet coefficient vector is done in [9], stating that "each coefficient represents how closely correlated a scaled wavelet is with the portion of the signal which is determined by translation". The CWT coefficients are nothing else than time-scale view of the analyzed signal, and so the CWT is an important analysis tool capable to "offers insight into both time and frequency domain signal properties" [9]. The results of this interpretation lead to the following useful observations [9] that will be considered for developing the proposed wavelet signals processing and analysis strategy:

The higher scales correspond to the "most" stretched wavelets, furthermore "the more stretched the wavelet, the longer the portion of the signal with which is compared, and thus the coarser the signal patterns features measured by the wavelet coefficients".

The "…coarser features called "approximations" providing basic shapes and properties of the original signal under investigation correspond to low frequency components, whereas the low scale components capture the high frequency information, called "details"…".

The main drawback of CWT is its computationally inefficiency due to the calculation of the CWT coefficients at every single scale with an impressive amount of work generating a huge bunch of data that must be stored in a considerable computer memory space. An alternative to the CWT is the discrete wavelet transform DWT, much more efficient and of high accuracy. The DWT is based on the wavelet analysis at particular scales and translations that are power of two, such as 2, 4, 6, 16, and so on [9], [10]. In [9] is stated that "the approximations" of the signals under investigation "provide basic trends and characteristics of the original signals, whereas the details provide the flavor of the signal", and the result of the applying DWT on the analyzed original signal is the so called wavelet decomposition around two key coefficient vectors, one of them is an "approximation" coefficient vector Ca, and other one is a "detail" coefficient vector Cd, representing "approximations" and "details" of the original signals under consideration. If the decomposition is repeated on the approximations in each stage, then the multiple stage DWT will break down the original analyzed signal into many successively lower resolution components, as is shown in [9], section 5.1.2, pp.79. According to [9] "at each stage, the approximation coefficient Ca represents the signal trend, and the detailed coefficient Cd includes the information on noise or nuance". The inverse process opposite to decomposition is the signal reconstruction by using an inverse discrete wavelet transform (IDWT). More details about sample wavelet definitions known as Haar, Mexican Hat, Morlet and Daubechies wavelets are well documented in [10]. Using the MATLAB/SIMULINK Wavelet and Processing Toolboxes will be implemented in real time the proposed multifractal and multisignal 1-D wavelet analysis strategy following the guidelines from [13], [14].

4. Case Study: pH Neutralization Plant

For simulation purpose the experimental setup for the pH neutralization process and the dynamic model coefficients are the same as in [15]. The tank layout is shown in Figure 3.

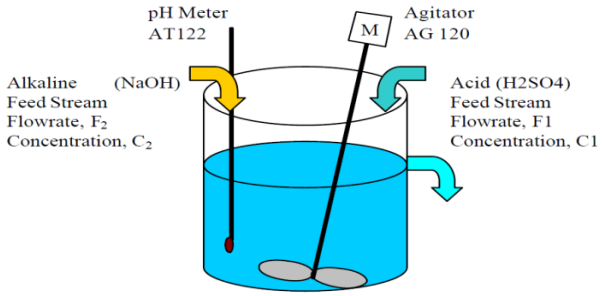


Figure 4: The reactor tank layout (screen shot view [11], pp. 41, Section 4)

Of technical viewpoint as the piping and instrumentation diagram layout, the tank and the overall architecture we preferred the simplified representation shown in Figure 4 to Figure 7 reproduced from [11] for their simplicity and better insight. They are related to the same pH neutralization process, only the coefficients of the models make the difference due to the changes in reactors geometry, flow rates and concentrations of acid and alkaline reactants.

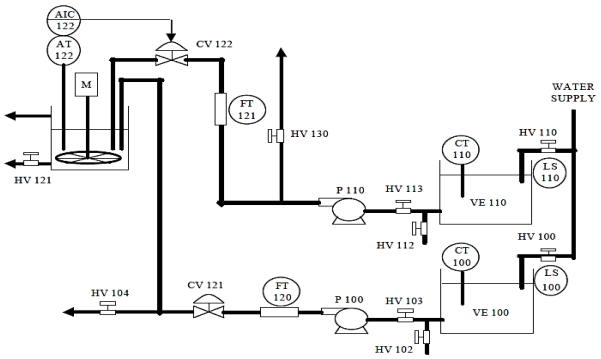


Figure 5: Piping and Instrumentation Diagram of pH neutralization plant (screen shot view [11], pp. 37, Section 3)

Figure 6: The picture of the neutralization pilot plant (snapshot view [11], pp. 38, Section 3)

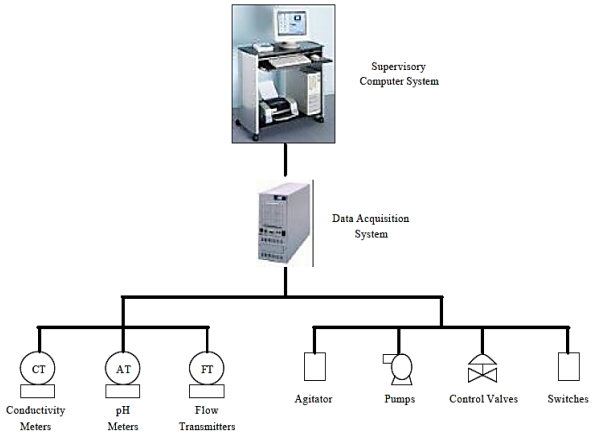


Figure 7: Overall architecture of the pH neutralization pilot plant (screen shot view [11], pp. 39, Section 3)

In Figure 8 is shown the block scheme of the both control loops, i.e. pH concentration value, and level of the solution inside the neutralization reactor.

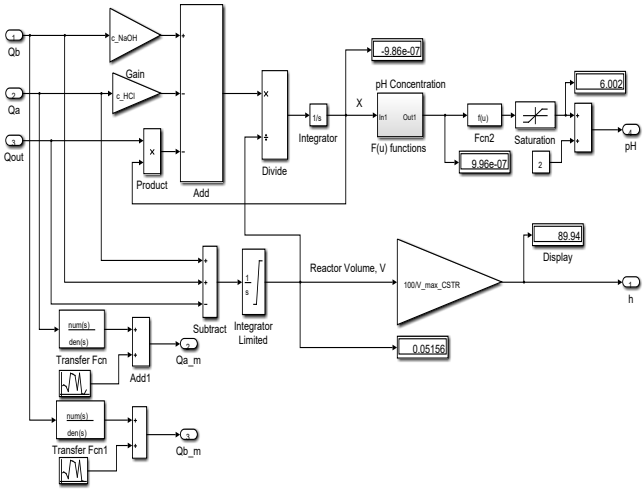


Figure 8: The SIMULINK model of the solution level control inside of the neutralized reactor and pH control closed-loops [15]

In Figure 9 is shown the SIMULINK model of the pH-controller with a back-propagation windup reset to avoid the saturation of the integrator due to the saturation nonlinearity of the control valve actuator. In Figure 10 and Figure11 is shown in detail the PI level controller with a windup reset from the same raison as for PI pH-controller.

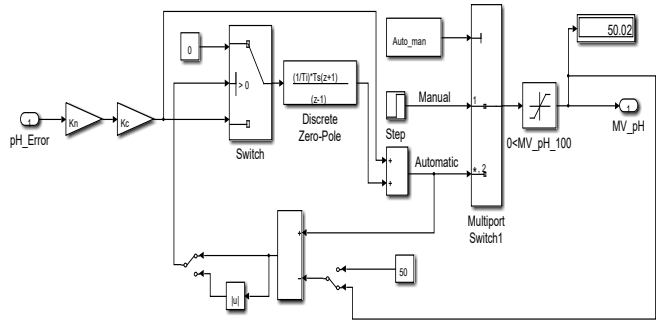


Figure 9: The SIMULINK model of pH controller with windup integrator [15]

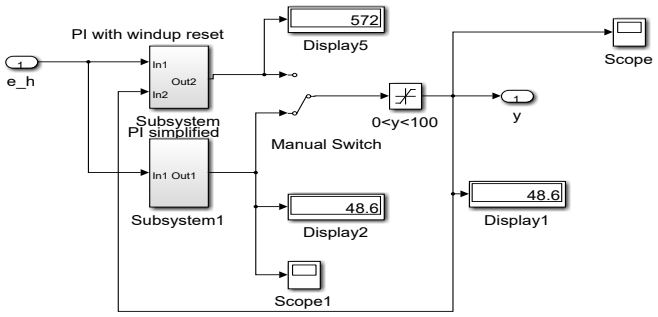


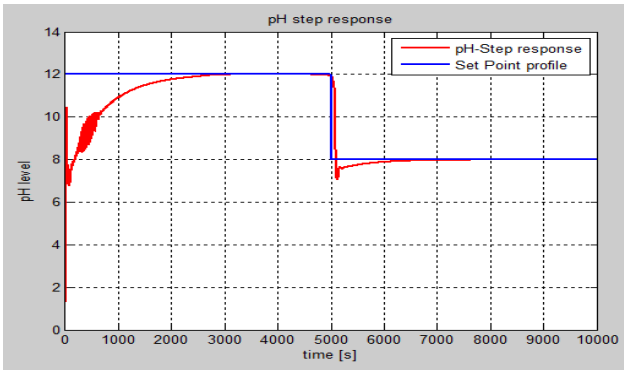
Figure 10: The simplified SIMULINK model of the Level PI Controller Block [15]

5. MATLAB and SIMULINK Simulations Results in Time Domain

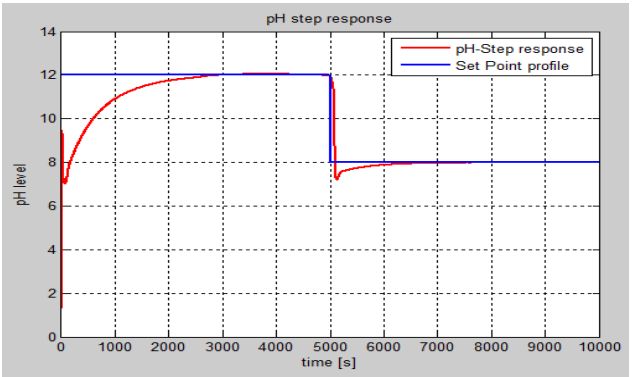
In this section we will present the simulation results of the control system performance in closed-loop, in time domain.

A. Healthy Control System with Pump and Control Valve Actuator Enabled

The pH step response is shown in Figure 11, for which the input profile changes the initial step input from pH12 after 5000 seconds (1h23min) to standardized pH8, the total simulation time taking 10000 seconds (2h46min).



(a)



(b)

Figure 11: The step response of the pH concentration of the neutralized solution inside the reactor.

Legend: a. Control valve enabled

b. Pump enabled

From performance analysis prospective the simulation results reveal a very accurate performance for the both PI controllers with windup reset. All these simulation results are carried out in a MATLAB SIMULINK simulation environment based on the SIMULINK model of the overall control system structure with the both feedback closed-loops, shown in Figure 12, similar as is developed in [16]. The step response of the level of neutralized solution inside the reactor controlled by a PI controller with windup reset is shown in Figure 12. In Figure 13 is presented the step response of the neutralized solution level inside the reactor. The controller efforts for pH and level are shown in Figure 14 and Figure 15.

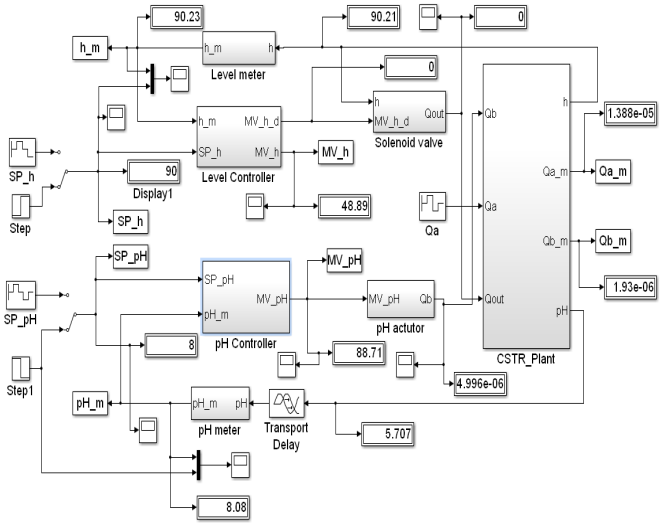


Figure 12: The SIMULINK model of the overall closed-loop control system [15]

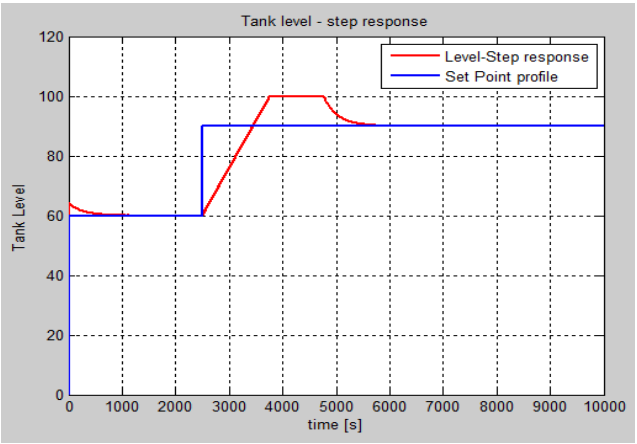
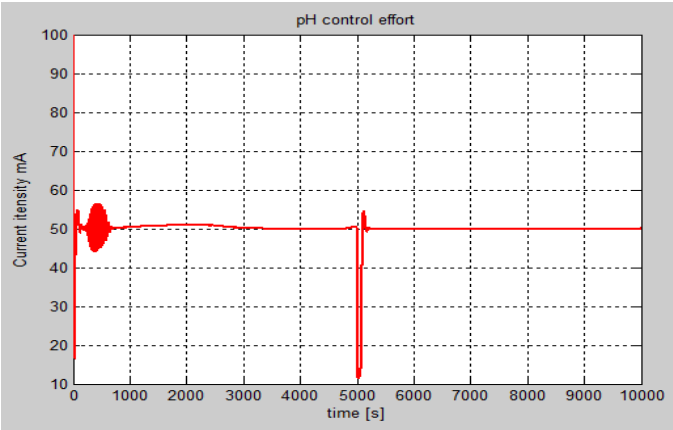


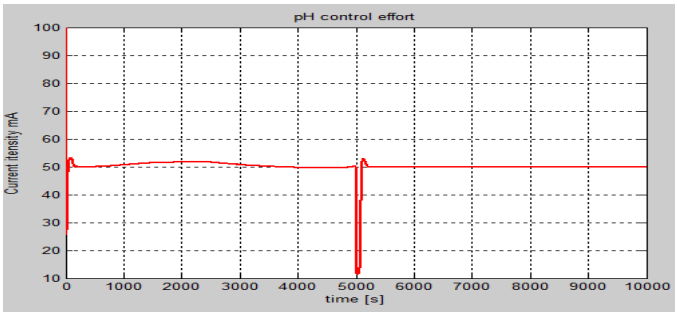
Figure 13: The step response of the neutralized solution level inside the reactor.

Let us now to consider the presence of a time delay in the pH control loop, i.e. pH sensor transmits the pH measured value to the input of the pH controller with certain amount of transport delay, let say for example 10 seconds, same value as is the time constant of the pH controller, you can see in Figure 16 that the pH control loop starts to oscillate at beginning of the simulations reaching the target steady-state for a while (2000 seconds) since after the first switching moment in setting point the closed-loop reaches the limit of stability around the second target setting

point, thus the controller becomes now inadequately. In this case is required a simple adaptive Smith-Predictor control structure for controlling time-delay systems, such is developed in [16].



(a)



(b)

Figure 14: The SIMULINK simulations of the pH controller effort in closed-loop

Legend: a. Control valve enabled

b. Pump enabled

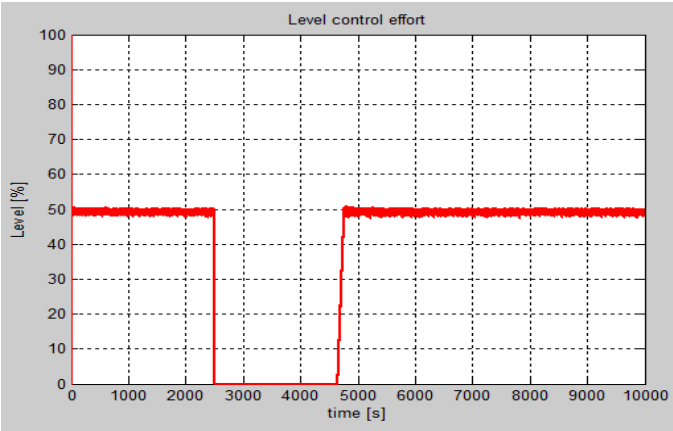


Figure 15: The SIMULINK simulations of the controller level effort in closed-loop

A considerable control effort of pH controller can be seen in this case in Figure 17. During these simulations no interferences between the two controls loops were noticed, thus the level control loop still remains very accurate.

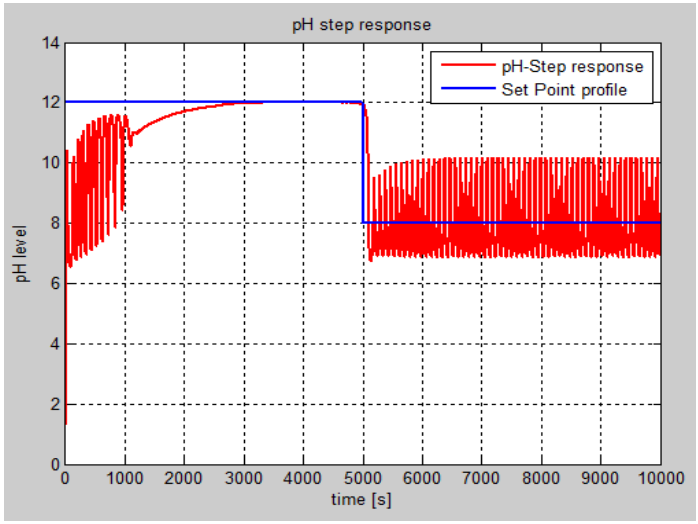


Figure16: The SIMULINK simulations of the pH concentration in closed-loop with 10 seconds time delay in signal transmission

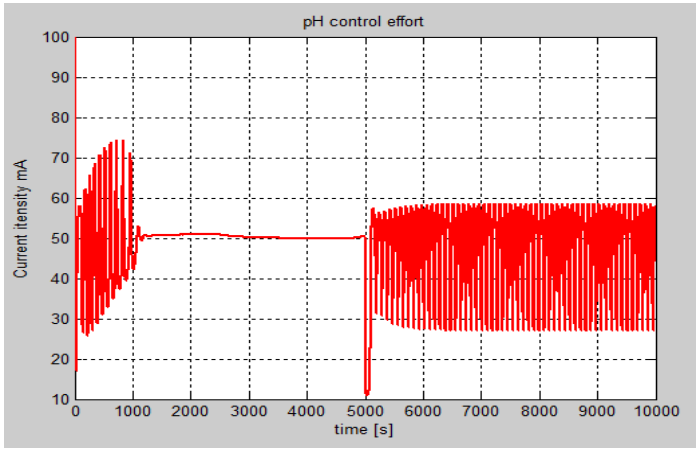


Figure 17: The SIMULINK simulations of pH controller effort in the presence of 10 seconds time delay in signal transmission

B. Faulty Control System with Control Valve Actuator Enabled

In this subsection we analyze the impact of control valve actuator nonlinearities on the overall control system performance.

1. Backlash control valve actuator nonlinearity shown in Figure 18 with dead-band width = 0.7, SIMULINK model.

In Figure 19 is shown the impact of the control valve backlash nonlinearity on the pH closed-loop control for the same set point input profile setup and the same tuning values for the pH controller parameters as for healthy control system structure. In this case the overall performance of pH control closed-loop degrades significantly. This is the first pattern feature extracted from measured pH output signal in time domain that will be analyzed in frequency domain by using a fractal analysis based on the wavelet transforms. The control effort is considerable also in this case, as is shown in Figure 20.

2. Coulomb and Viscous friction nonlinearity in control valve actuator shown in Figure 21 with the coefficient of viscous friction (gain) = 0.1, the Coulomb friction value (offset) = 1, SIMULINK model.

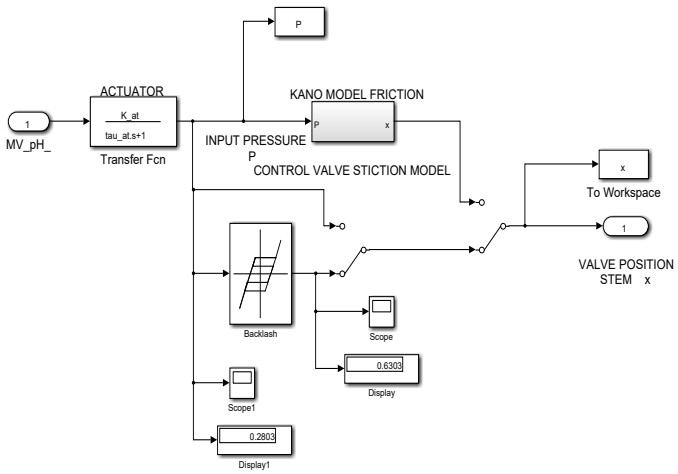


Figure 18: SIMULINK model of the control valve actuator with backlash nonlinearity (dead-band width = 0.7)

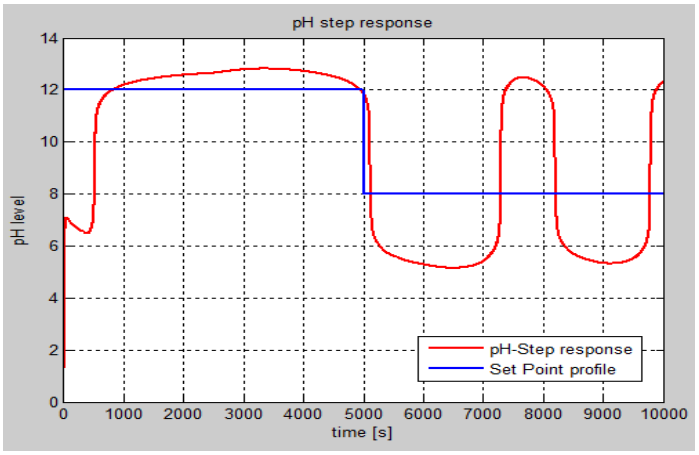


Figure 19: The impact of backlash control valve actuator nonlinearity on the pH evolution in closed-loop (dead-band width = 0.7, SIMULINK model).

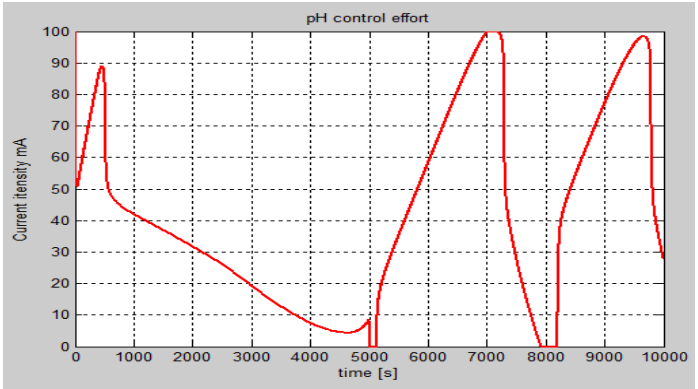


Figure 20: The pH control effort in the presence of backlash nonlinearity in the control valve actuator integrated in the pH control closed-loop (dead-band width = 0.7).

3. Dead-zone nonlinearity in control valve actuator shown in Figure 23 with a symmetric dead-zone width = 1, SIMULINK model.

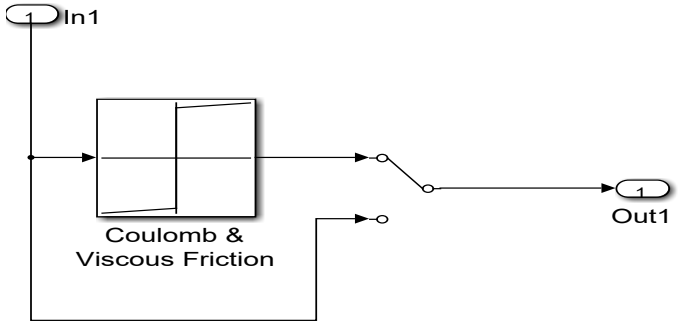


Figure 21: SIMULINK model of the control valve actuator with Coulomb and Viscous friction nonlinearity, SIMULINK model

Similar, compared to backlash nonlinearity case the overall performance of pH control closed-loop degrades drastically, as is shown in Figure 22. This is the second pattern feature extracted from measured pH output signal in time domain that will be analyzed in frequency domain by using a fractal analysis based on the wavelet transforms.

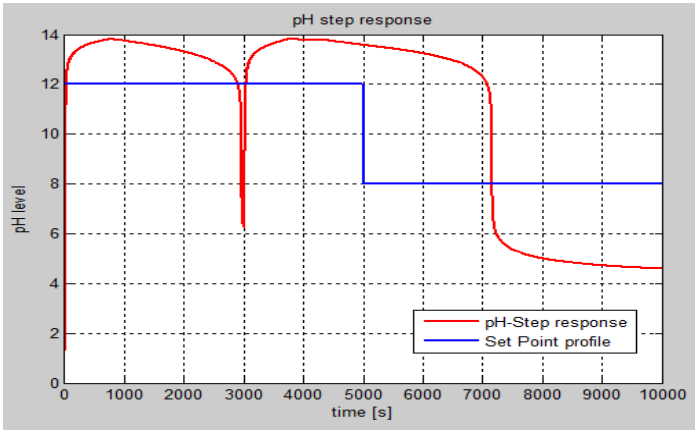


Figure 22: The impact of Coulomb and viscous friction nonlinearity of control valve actuator on the pH evolution in closed-loop (Coefficient of viscous friction (gain) = 0.1, the Coulomb friction value (offset) = 1, SIMULINK model).

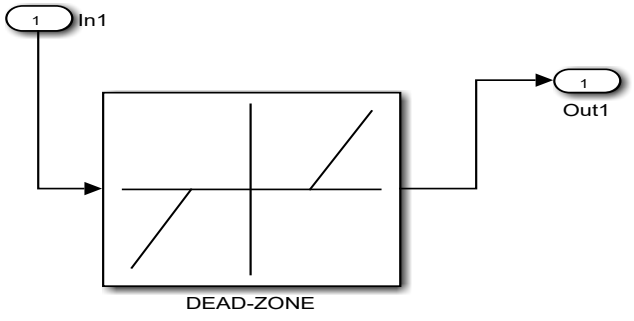


Figure 23 SIMULINK model of the control valve actuator with Dead-zone nonlinearity, SIMULINK model (dead-zone width =1)

The impact of dead-zone nonlinearity on the overall performance is shown in Figure 24. This is the third pattern feature extracted from measured pH output signal in time domain that will be analyzed in frequency domain by using a fractal analysis based on the wavelet transforms.

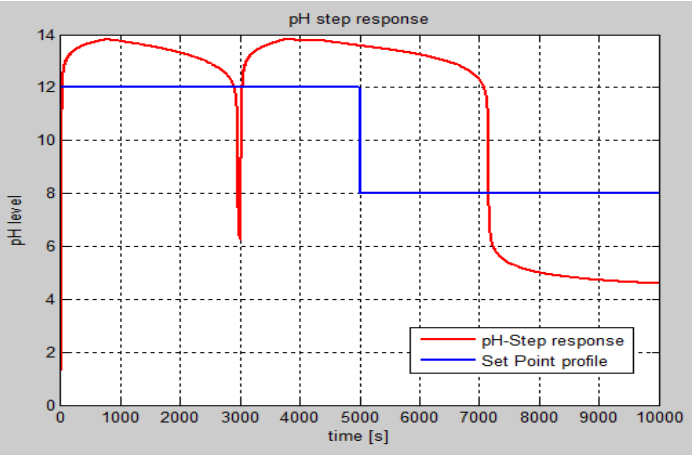


Figure 24: The impact of Dead-zone nonlinearity of control valve actuator on the pH evolution in closed-loop (dead-zone width = 1, SIMULINK model).

4. ***Saturation nonlinearity*** in control valve actuator shown in Figure 25 with superior limit = 0.5 [Pa], and the inferior limit = 0[Pa], SIMULINK model.

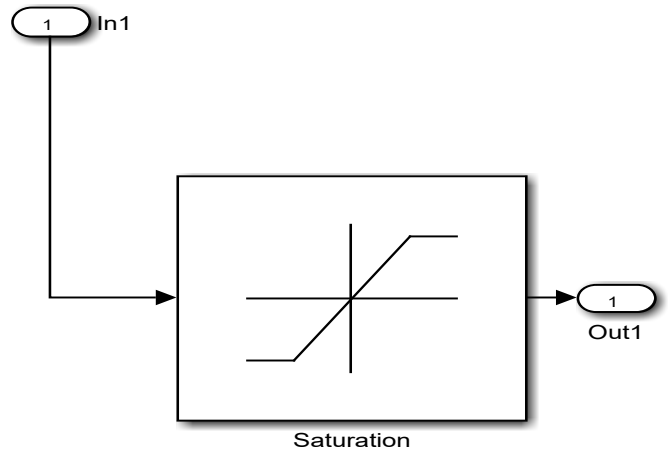


Figure 25: SIMULINK model of the control valve actuator with saturation nonlinearity, SIMULINK model

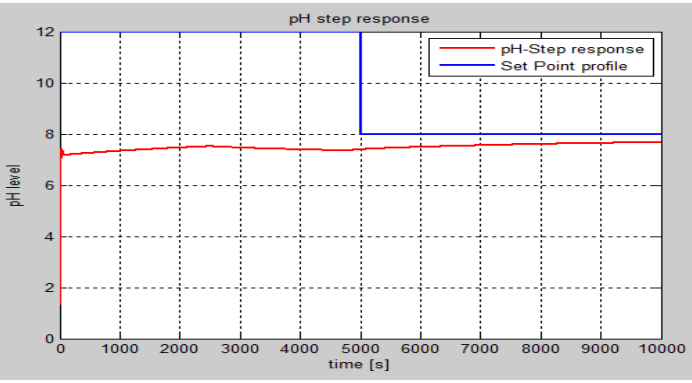


Figure 26: The impact of saturation nonlinearity of control valve actuator on the pH evolution in closed-loop (dead-zone width = 1, SIMULINK model).

6. Multisignal 1-D Wavelet Analysis - MATLAB Simulations Results

1-D Multisignal Definition 2: A 1-D multisignal is a set of 1-D signals of same length stored as a matrix organized rowwise (or columnwise) [14]. In the proposed case study all the signals are stored in a matrix with two rowwises, in a first row is stored the healthy signal and in second one is stored the corresponding faulty signal representing the impact of each actuator nonlinearity on the overall performance of healthy control system. The purpose of this section is to make a wavelet analysis of the 1-D multisignal set containing all the nonlinearities described in the previous section. How you will see the wavelet analysis is a precious analysis tool to denoise, compress and cluster different representations or their simplified versions. In first step a deeply analyze is made for all the signals collected in the previous section from the closed-loop control system, and then we will find some representations and also simplified versions for these signals by reconstructing their approximations at given levels, denoising and compressing them.

Denoising and compressing are two of the main applications of wavelets, often used as a preprocessing step before clustering [14]. The last step performs several clustering strategies and compares them. It allows summarizing a large set of signals using sparse wavelet representations. In Figure 27 you get an overall image about the impact of all control valve actuator nonlinearities on the closed-loop performance of the neutralization control system discussed in previous section.

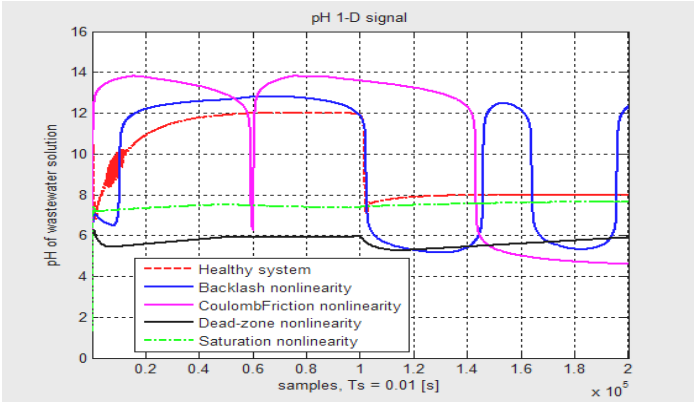


Figure 27: The impact of the control valve nonlinearities on the closed-loop performance of pH control system

The same overall image you can get as more attractive in a 3-D representation of all these nonlinearities including also the healthy control system, as a reference for the severity of each impact, is offered in Figure 28, The 3-D representation is preferred in addition to highlight the periodicity of the multisignal.

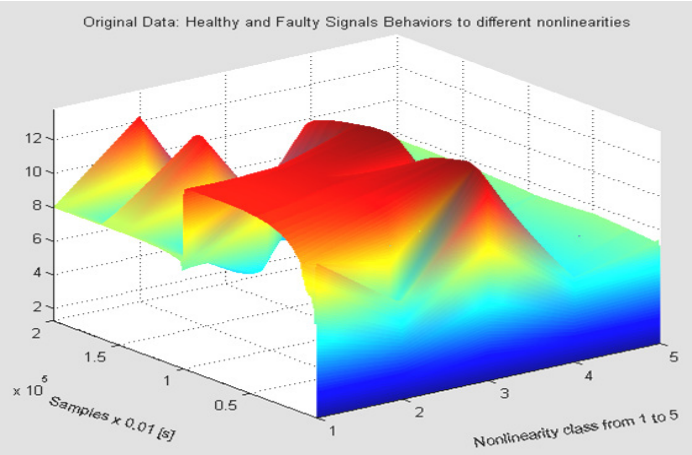


Figure 28: The 3-D representation of the impact of several control valve nonlinearities, on the closed-loop control system performance.

In 3-D representation each nonlinearity impact can be seen more clearly. In Figure 29 is shown the row decomposition of the multisignal related to the following fields of the generated structure in MATLAB R2013a:

```
dirDec: 'r' (row decomposition)
         Level of decomposition: 7
         wname: 'sym4'(near symmetric wavelet, as is shown
                       in Figure 29)
         dwtFilters: [1×1 struct]
         dwtEXTM: 'sym'
         dwtShift: 0
         dataSize: [35 1440]
            ca: [35×18 double]
            cd: {1×7 cell}
```

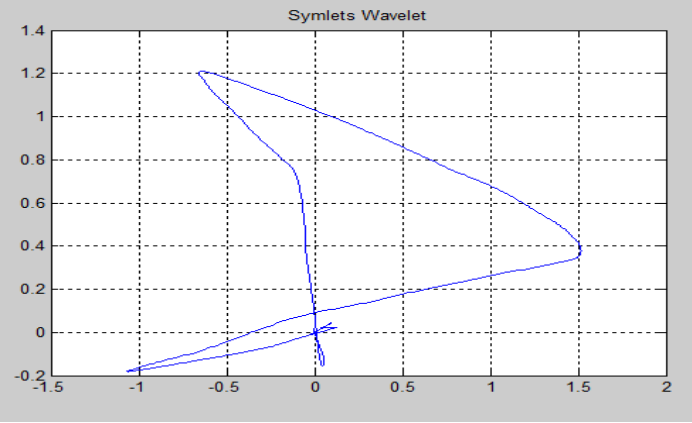


Figure 29: The symlet wavelet ("sym4') representation

MATLAB Wavelets Toolbox offers you more facilities to get information about all these wavelets functions using the MATLAB command waveinfo('name'), where 'name' is the assigned short name for the wavelets families such as 'sym', 'morl', 'haar', 'db', and so on. The reconstruction of the approximations at level 7 for each row signal is shown in Figure 30. The signals reconstructed are also compared in the same graph to the original signals. Furthermore, to get a better insight about the control valve nonlinearities under investigation in Figure 31 the signals are split in two groups, in the first group the healthy signal, the backlash and Coulomb viscous friction, and in second group the last two nonlinearities, control valve dead-zone and saturation.

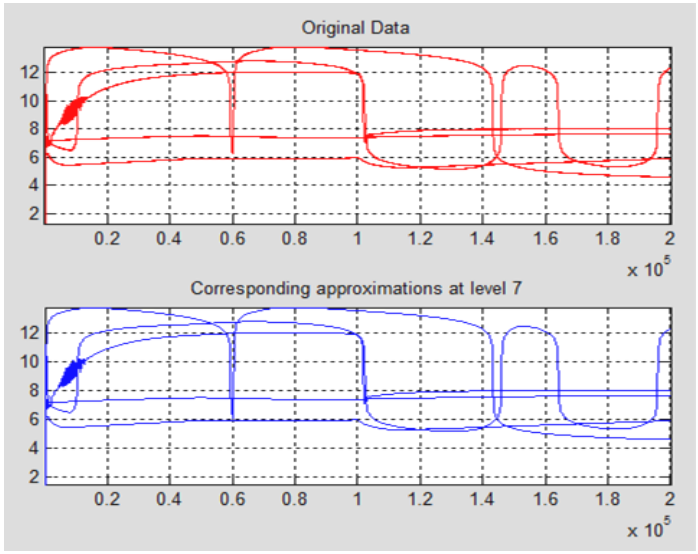


Figure 30: The original signals and the reconstruction of the approximations at level 7 for each row.

The top plot shows all the original signals and the bottom one shows all the corresponding approximations at level 7. As it can be seen, the general shape is captured by the approximations at level 7, but sometimes some interesting features are lost, as for example, the bumps at the beginning and at the end of the signals could disappear.

In order to perform a more subtle simplification of the multisignal preserving these bumps a denoise operation of the multisignal is useful. The denoising procedure is performed in MATLAB R2013a following three steps, namely [15]:

- **Decomposition**: First is required to select a wavelet function and to set up the level of its decomposition N, and then compute the wavelet decompositions of the signals at level N.
- **Thresholding**: For each level from 1 to N and for each signal, a threshold is selected and thresholding is applied to the detail coefficients.
- **Reconstruction**: Compute wavelet reconstructions using the original approximation coefficients of level N and the modified detail coefficients of levels from 1 to N.

If a slightly change will be made at the level of decomposition by changing N from 7 to N = 5 in Figure 32 can be seen similar results as in Figure 30, but now the quality of the results is better since the bumps at the beginning and at the end of the signals are well recovered. Conversely, in the same figure the residuals look like a noise except for some remaining bumps due to the signals. Furthermore, the magnitude of these remaining bumps is of a small order.

In Figure 33 are shown the results of the signals compression operation that follows the same three steps as in denoising case, a slightly difference in terms of the procedure is found only in step 2 that where two compression approaches are available [15]:

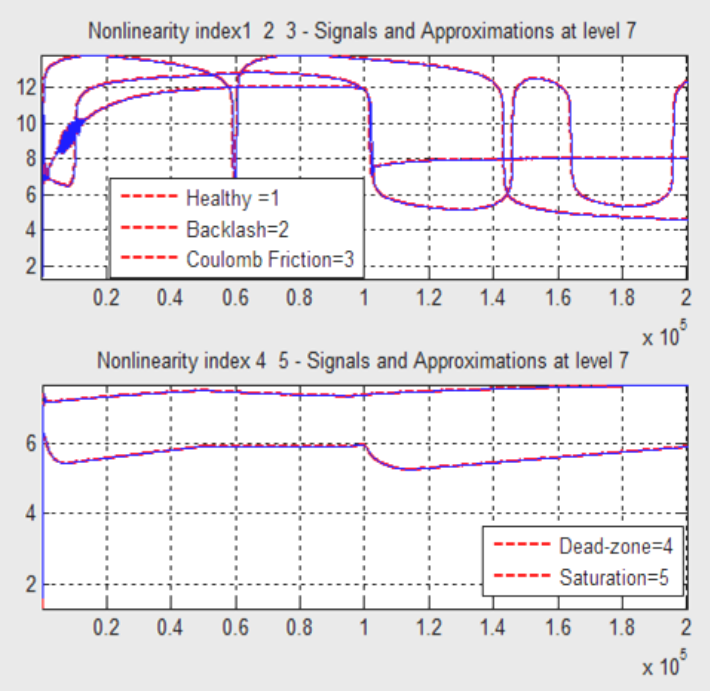


Figure 31: The signals and approximations at level 7 of row decomposition split in two groups

In the first approach the wavelet signals are expanded and keep the largest absolute value coefficients; in this case, a global threshold, a compression performance, or a relative square norm recovery performance can be set [15]. Therefore, for this approach only a single signal-dependent parameter needs to be selected. In the second approach a determined level-dependent threshold is chosen visually. The compression performance can be evaluated by calculating the corresponding densities of nonzero elements, as is shown in Figure 34.

For all the control valve nonlinearities under investigation that causes faulty signals, as well as for healthy signal, the percentage of required coefficients to recover 99% of the energy is the same, approximately 0.75%. This is one of the stronger wavelets features to prove a high capacity to concentrate all signal energy in few coefficients [15].

The last step is the clustering of the wavelets signals that offers a convenient procedure to summarize a large set of signals using sparse wavelet representations (with small number of elements). In Figure 35 the wavelets signals are clustered by 3. The first one is related to the Coulomb viscous friction nonlinearity of the control valve, the second cluster is related to the healthy and baclash nonlinearity, and the third cluster is related to dead-zone and saturation nonlinearities.

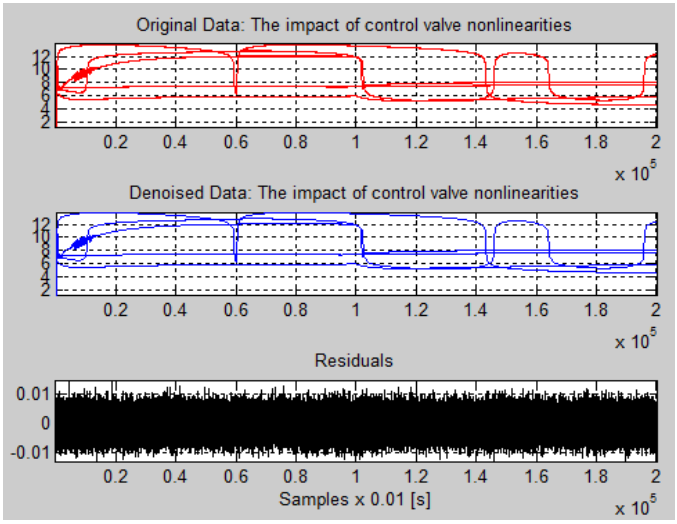


Figure 32: The original signals, their approximations at level 5 of row decomposition, and the residuals.

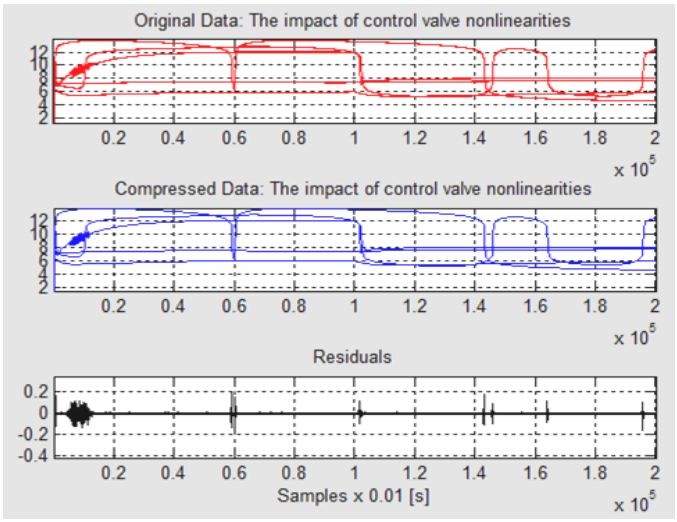


Figure 33: The original and compressed signals and the corresponding residuals

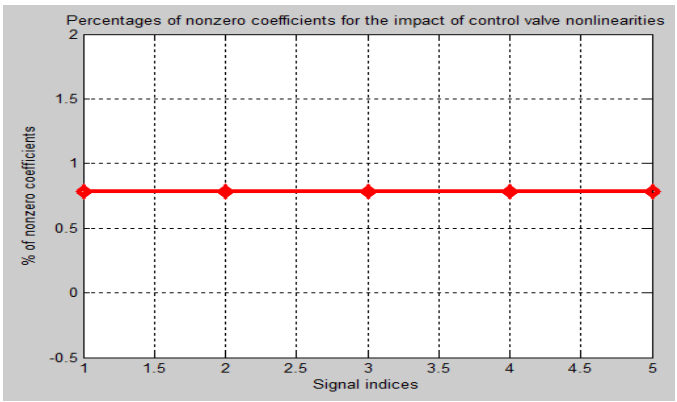


Figure 34: The percentage of nonzero coefficients for all the control valve nonlinearities wavelet signals

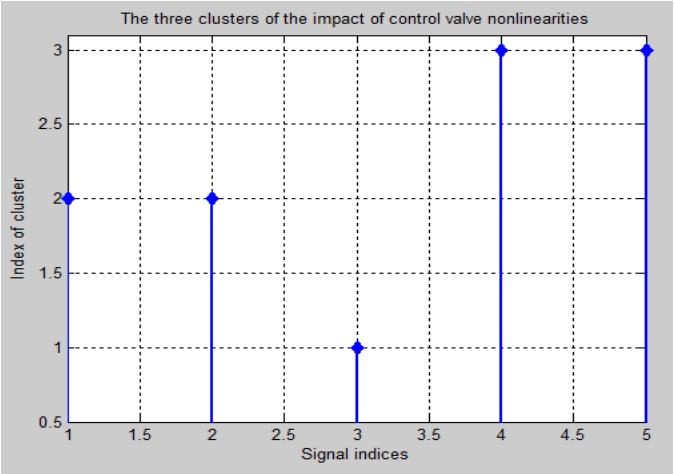


Figure 35: The three clusters of the wavelets signals correspondin to the healthy and control valve nonlinearities

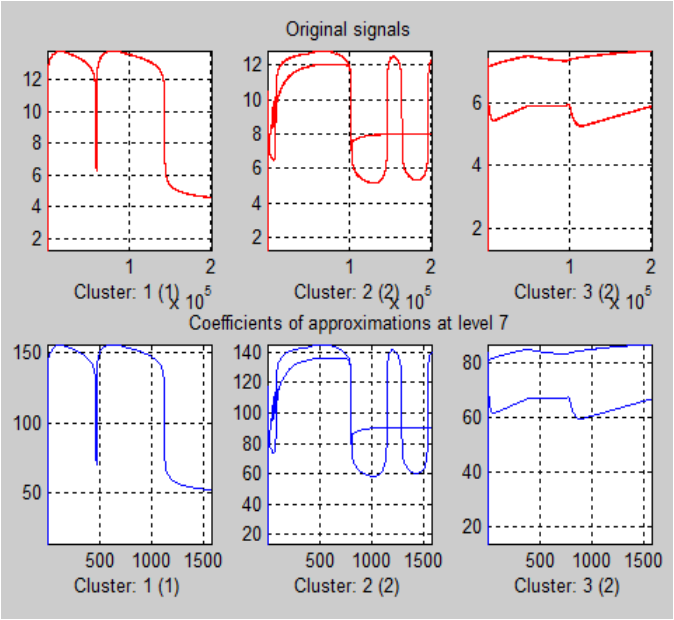


Figure 36: The partition of each cluster in healthy and faulty control valve signals

7. 1-D Wavelet Analysis as a Detection Tool of the Actuators Faults in Control Systems

In this section we investigate how to use the 1-D wavelet analysis to detect changes in the variance of a process control.

Changes in signal variance provide precious information anytime when something fundamental has changed about the data-generating mechanism that is basically the main idea of implementation for Fault Detection Diagnosis and Isolation (FDDI) strategies capable for detection, diagnosis and isolation of the faults that occur in actuators and sensors in a lot of control systems applications. In this section we investigate the effectiveness of using the 1-D wavelet analysis to detect some anomalies caused by malfunctioning of the equipment or different control system parts, especially actuators and sensors more likely prone to the errors. Encouraged by the experience and the preliminary results obtained in the control systems field, such as modeling and process identification, state estimation and FDI control strategies, a new FDI approach based on signal processing analysis to improve the accuracy, robustness and implementation design of these techniques is a big challenge. For simulation purpose we analyze only the impact on the overall closed-loop control performance of two control valve actuator nonlinearities, namely the backlash and Coulomb viscous friction, met frequently in the control industrial applications, as is shown in the Figures 37 and 38. In the Figure 37 is shown the SIMULINK model of the backlash nonlinearity that occurs often in control valve actuator, with the injection mechanism of the related fault at the instant $t_{fault\_injection}$ = 7000 seconds. The set point input profile changes also at the instant t_{SP1}=5000 seconds, from pH12 to pH8. Similar, in the Figure 38 is shown the SIMULINK model of the Coulomb viscous friction control valve nonlinearity and the injection mechanism of the related fault. Similar, in Figure 39 is presented the mechanism of generating the values for the backlash width using the *clock* block and *if-then* with separate triggered action block.

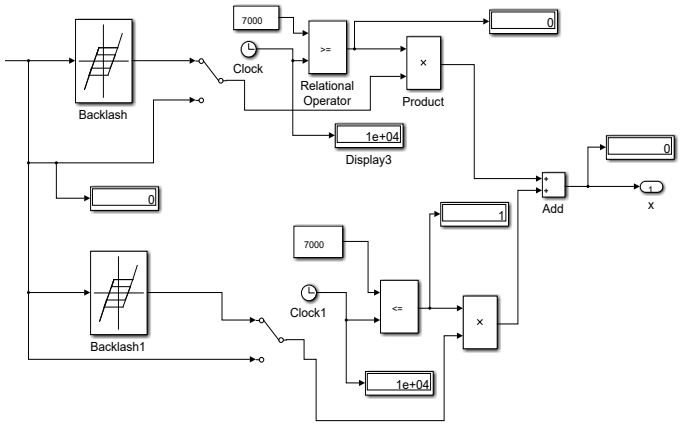


Figure 37: The SIMULINK model of backlash nonlinearity in the faulty control valve and the injection mechanism

In the Figures 40 and 41 are shown the injection instant of the both faults (backlash and Coulomb viscous friction), and also the faulty behavior of the control system after this instant.

Precious information about the time detection and the severity of the faults in each case is provided by the residuals defined as a difference between the healthy signal output (free fault) of the control system and faulty signal output, as is shown in Figure 42 and 43.

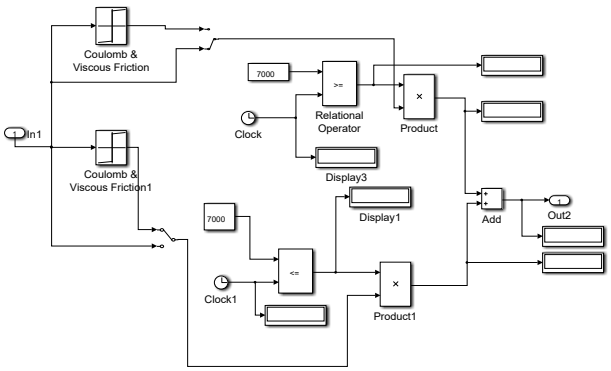


Figure 38: The SIMULINK model of Coulomb viscous friction nonlinearity in the faulty control valve and the injection mechanism.

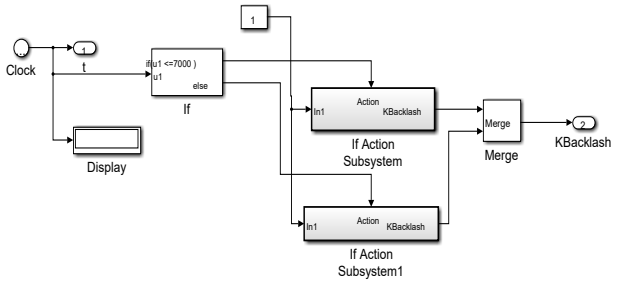


Figure 39: The SIMULINK model of generating the width values for backlash nonlinearity in the faulty control valve required in the injection mechanism.

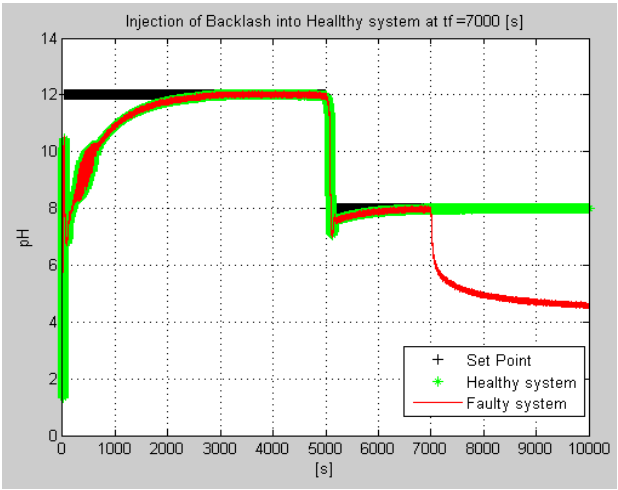


Figure 40: The MATLAB simulation results for healthy and faulty behavior caused by the backlash nonlinearity in the control valve actuator

Amongst the wavelets families functions the Morlet wavelet function shown in Figure 44 is suitable for continuous analysis using a continuous wavelet transform (CWT) [10], [17]. There is no scaling function associated with the Morlet wavelet. In many signal processing applications the Daubechies wavelet family functions are most used, especially for discrete wavelet transforms (DWT) [19]. To display detailed information about the Daubechies' least asymmetric orthogonal wavelets is used the MATLAB command *waveinfo*('sym'). To compute the wavelet and scaling function (if available), use *wavefun*( ):

```
[psi,xval] = wavefun('morl',10);
plot(xval,psi); title('Morlet Wavelet');
```

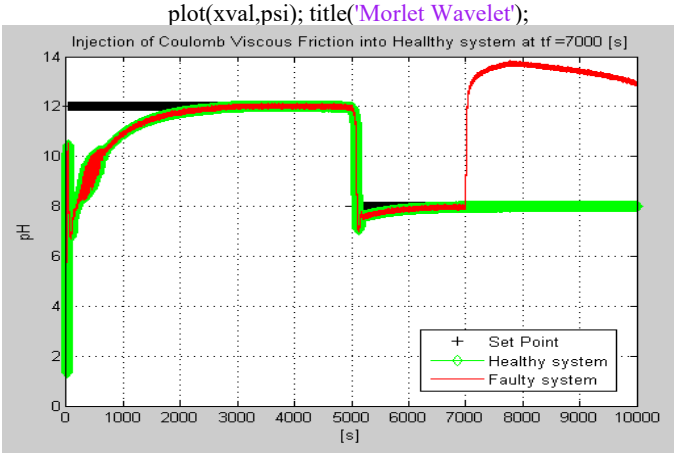


Figure 41: The MATLAB simulation results for healthy and faulty behavior caused by the Coulomb viscous friction in the control valve actuator

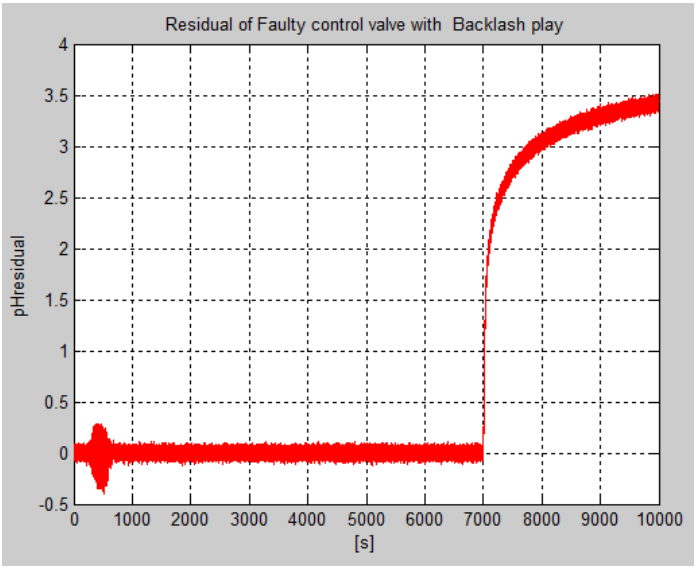


Figure 42: The MATLAB simulation results for output control system residual caused by the backlash nonlinearity in the control valve actuator

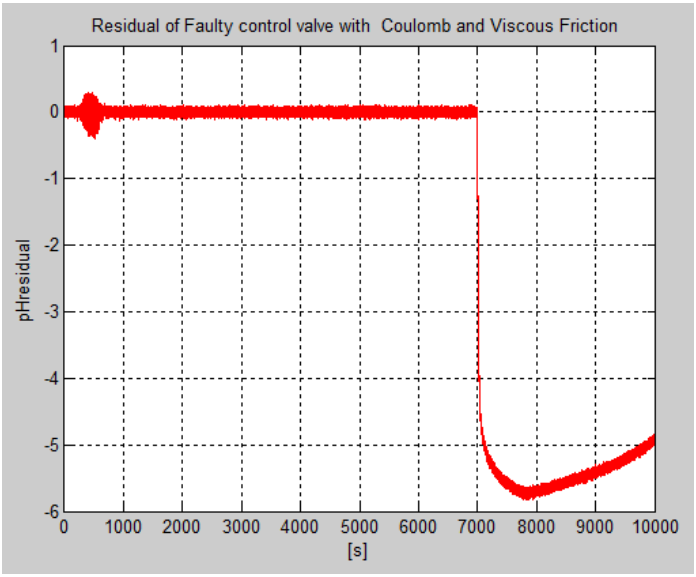


Figure 43: The MATLAB simulation results for output control system residual caused by the Coulomb viscous friction in the control valve actuator

For wavelets associated with a multiresolution analysis can be computed both the scaling function and wavelet. The following MATLAB code returns the scaling function and wavelet for the Daubechies' extremal phase wavelet with 4 vanishing moments [17], as is shown in Figure 45:

```
[phi,psi,xval] = wavefun('db4',10);
subplot(211);
plot(xval,phi);
title('db4 Scaling Function');
subplot(212);
plot(xval,psi);
title('db4 Wavelet');
```

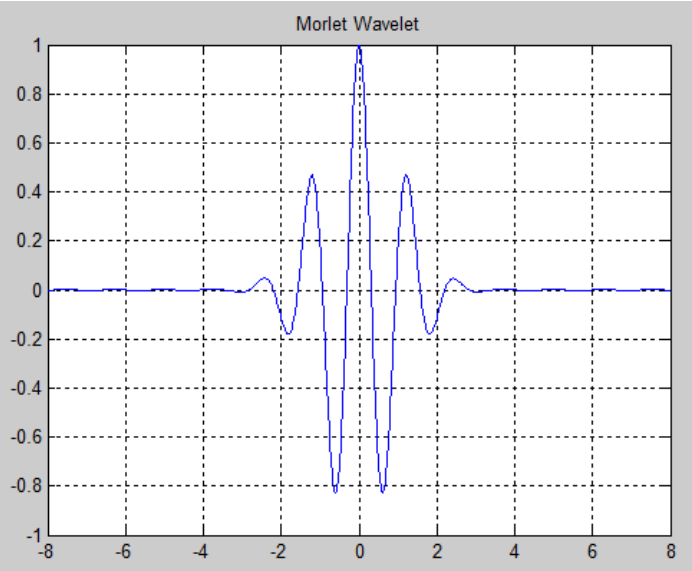


Figure 44: The Morlet wavelet function representation with 10 vanishing moments

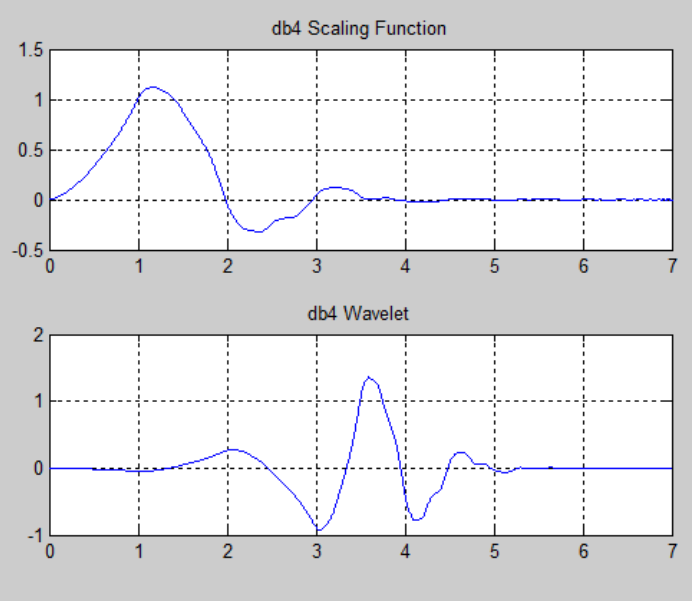


Figure 45: The Daubechies' extremal phase wavelet with 4 vanishing moments

In discrete wavelet analysis, the analysis and synthesis filters are of more interest than the associated scaling function and wavelet. For filter design and implementation you can use *wfilters*( ) to obtain the analysis and synthesis filters. For example, in order to obtain the decomposition (analysis) and reconstruction (synthesis) filters for the Bspline biorthogonal wavelet is required to specify 3 vanishing moments in the synthesis wavelet and 5 vanishing moments in the analysis wavelet with the following MATLAB cod lines to generate and to plot the filters' impulse responses, very useful in this section for fault detection based on the wavelets filters analysis and synthesis (Low Pass Filters and High Pass Filters, used for low frequencies and high frequencies signals filtration respectively) [17]:

```
[LoD, HiD, LoR, HiR] = wfilters('bior3.5');
subplot(221);
stem(LoD);
title('Lowpass Analysis Filter');
subplot(222);
stem(HiD);
title('Highpass Analysis Filter');
subplot(223);
stem(LoR);
title('Lowpass Synthesis Filter');
subplot(224);
stem(HiR);
 title('Highpass Synthesis Filter');
```

The simulation results are shown in Figure 46.

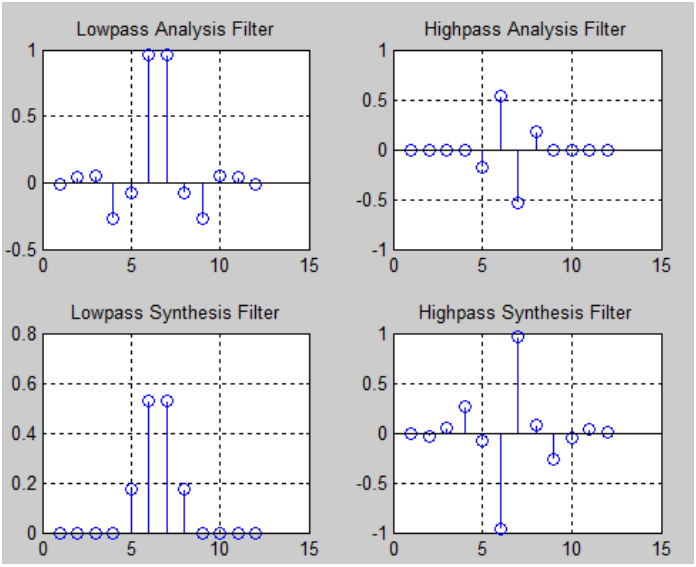


Figure 46: The Daubechies' extremal phase wavelet with 4 vanishing moments

The Daubechies wavelets (dbN) are the Daubechies' extremal phase wavelets, for which N refers to the number of vanishing moments. These filters are also referred to in the literature by the number of filter taps, which is 2N [17], [21].
The Symlet wavelets (symN) are also known as Daubechies' least-asymmetric wavelets. The symlets are more symmetric than the extremal phase wavelets. In symN, N is the number of vanishing moments. These filters are also referred to in the literature by the number of filter taps, which is 2N [17], [21].To obtain a survey of the main properties of this family, you must enter waveinfo('bior') at the MATLAB command line. More precisely, the orthogonal and biorthogonal filter banks are arrangements of lowpass, highpass, and bandpass filters that divide the signals data sets into subbands [17], [21]. If the subbands are not modified, these filters enable perfect reconstruction of the original data. In most applications, the data are processed differently in the different subbands and then reconstruct a modified version of the original data. Orthogonal filter banks do not have linear phase. Biorthogonal filter banks do have linear phase [17], [21]. The wavelet and scaling filters are specified by the number of the vanishing moments, which allows removing or retaining polynomial behavior in the signals data sets. Furthermore, lifting allows designing perfect reconstruction filter banks with specific properties. In order to obtain and use the most common orthogonal and biorthogonal wavelet filters can be used Wavelet Toolbox™ functions [17]. The design of custom perfect reconstruction filter bank can be performed through elementary lifting steps. In addition, can also be added your own custom wavelet filters. However, by using

the wavelet filter bank architecture depicted in Figure 47 it is possible to obtain residues that change in a noticeable manner in order to offer precious information about the time detection of the faults and its severity [19, 20]. The subband model is suggested in [19] of the form:

$$M(z) = (1 - z^{-1})^{-s}(a + bz^{-1}) \quad (5)$$

where s is an integer number and a, b are real numbers. In [19] is used the '**db8**' wavelet for wavelet filter bank design of level 3 decomposition [19] for a Single-Input Single-Output (SISO) plant extended in [20] for a Multivariable (MIMO) plant. A wavelet based-frequency subband analytical redundancy scheme to calculate the residuals for different faults is shown in Figure 48, and used for wavelet filter bank synthesis and analysis of level 3 decomposition in [19, 20] , and also in our case study. In this scheme G(z) and H(z) represent the z-transforms of the low pass filter (LPF) and high pass filter (HPF) respectively. A two-channel critically sampled filter bank filters the input signal using a lowpass and highpass filter [21]. The subband outputs of the filters are downsampled by two to preserve the overall number of samples. To reconstruct the input, upsample by two and then interpolate the results using the lowpass and highpass synthesis filters. If the filters satisfy certain properties, a perfect reconstruction of the input is achieved [21]. In Figure 49 is proved this perfect reconstruction by representing in the same graph the original faulty pH output signal (in the top) and the reconstruction faulty ph waveform response.

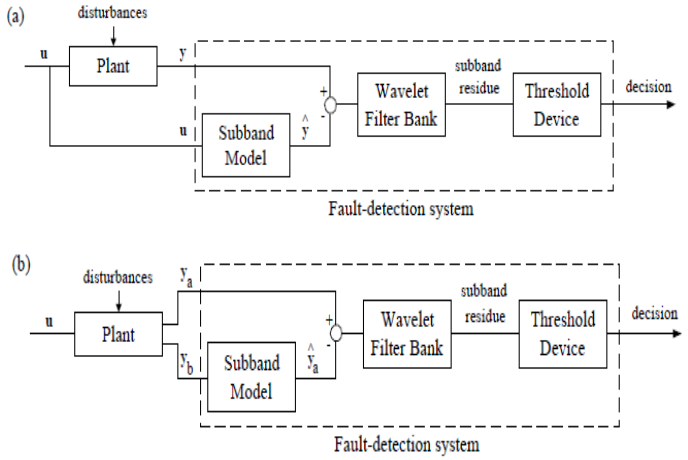


Figure 47: Original wavelet analytical redundancy architecture for (a) input-output and (b) output-output consistency (snapshot view from [19], [20])

At each successive level, the number of scaling and wavelet coefficients is downsampled by two so the total number of coefficients is preserved. In order to obtain the level three DWT of the pH faulty signal is using the 'sym4' orthogonal filter bank, using the MATLAB code line [21]:

[C,L] = wavedec(wecg,3,'sym4');

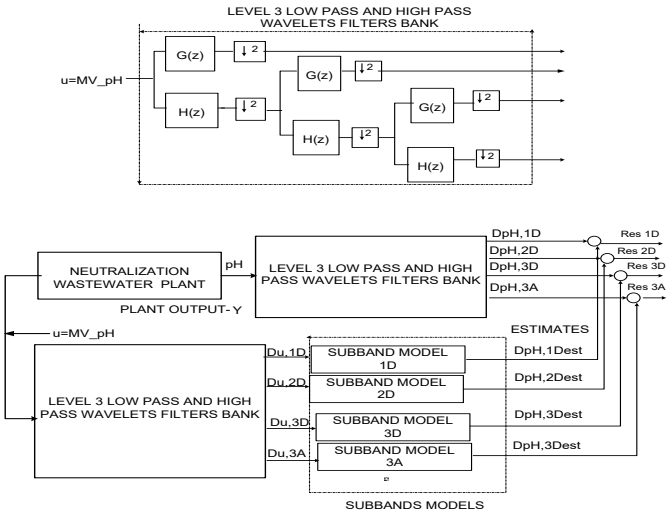


Figure 48: A wavelet based-frequency subband analytical redundancy scheme (suggested in [19, 20])

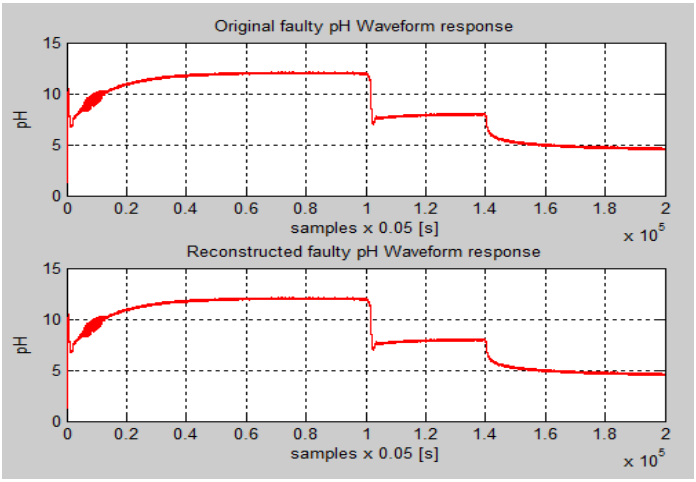


Figure 49: The original and reconstruct pH faulty signals

The number of coefficients by level is contained in the vector L. The first element of L is equal to 256, which represents the number of scaling coefficients at level 3 (the final level). The second element of L is the number of wavelet coefficients at level 3. Subsequent elements give the number of wavelet coefficients at higher levels until the final element of L is reached. The final element of L is equal to the number of samples in the original signal. The scaling and wavelet coefficients are stored in the vector C in the same order. In order to extract the scaling or wavelet coefficients, the MATLAB commands *appcoef* ( ) or *detcoef*.( ) can be used. All the wavelet coefficients are extracted in a cell array and final-level scaling coefficients [21]. In the Figures 50 and 51 are shown all these coefficients for faulty pH signal caused by backlash nonlinearity and for healthy pH signal (free faults) respectively.

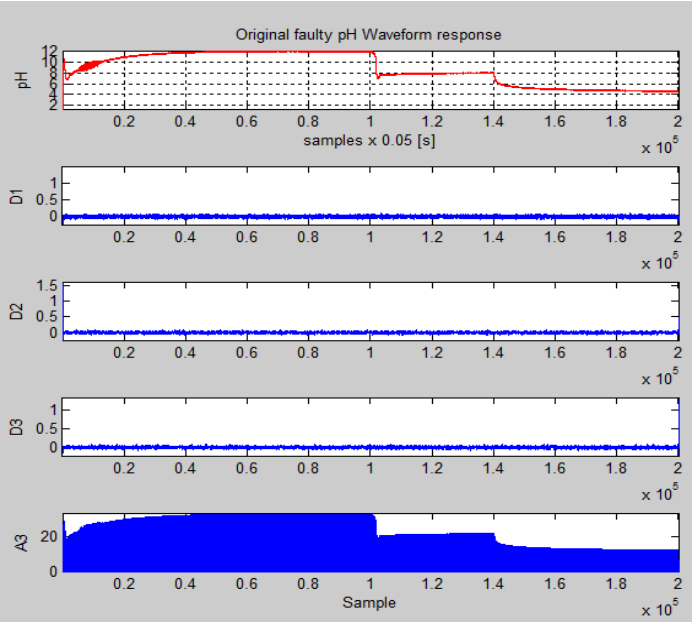


Figure 50: The detailed vectors coefficients D1, D2, D3 and approximation vector coefficients A3 for faulty pH signal caused by the backlash in control valve actuator.

The residuals of the detailed coefficients Res1D, Res2D, Res3D and also of the approximation coefficients Res3A are shown in the Figures 52 until 55. All these figures reveal a good detection of the fault injection instant, and also provide useful information about the fault severity, as is shown in Figure 55. The residuals for detailed vectors coefficients are almost zero if the noise will be filtrated. Similar, for the second nonlinearity in the control valve actuator the detection procedure of the faults is the same. The simulation results are shown in Figures 56 until 61.

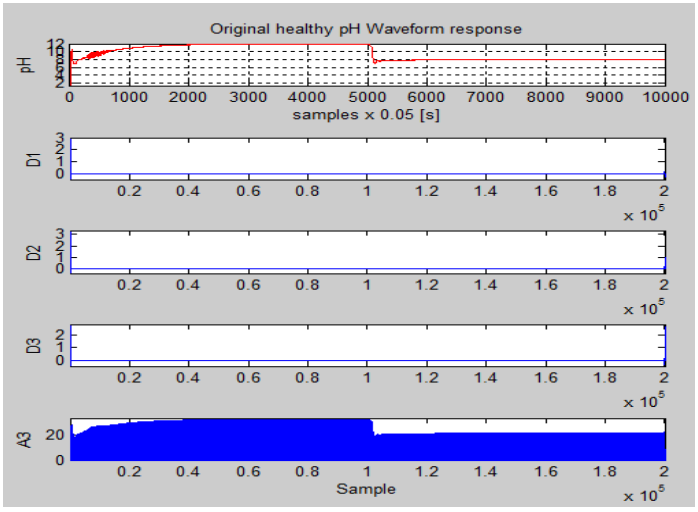


Figure 51: The detailed vectors coefficients D1, D2, D3 and approximation vector coefficients A3 for healthy pH signal

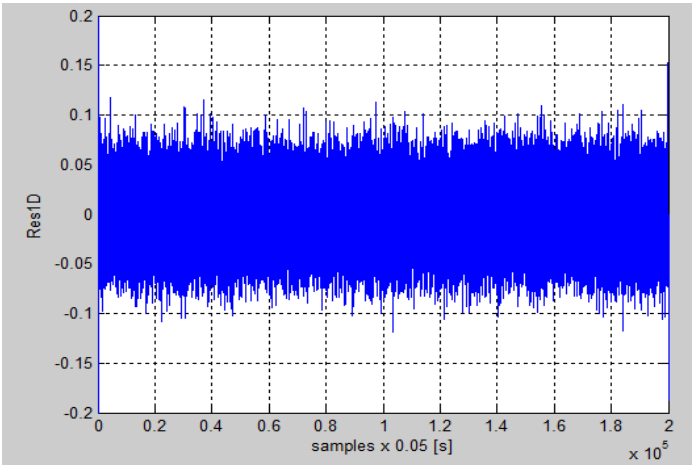


Figure 52: The residual of the detailed vector coefficients Res1D

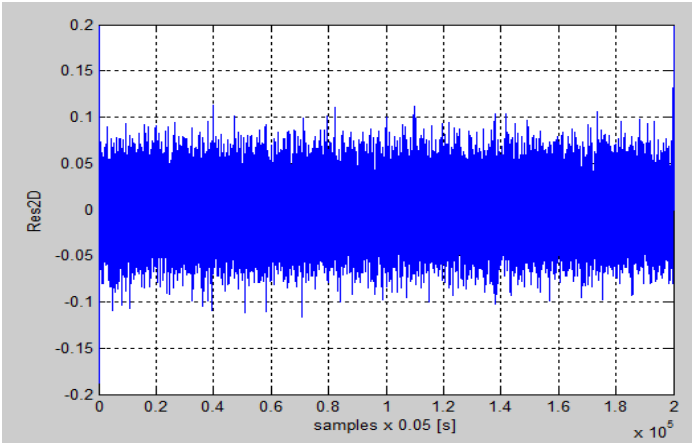


Figure 53: The residual of the detailed vector coefficients Res2D

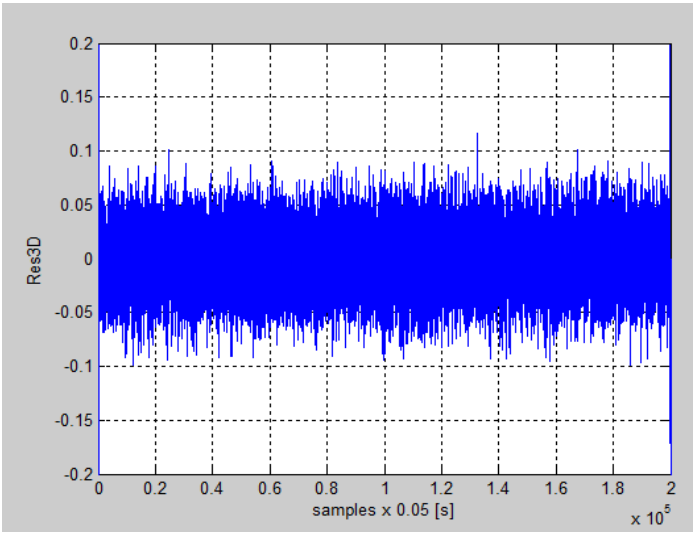


Figure 54: The residual of the detailed vector coefficients Res3D

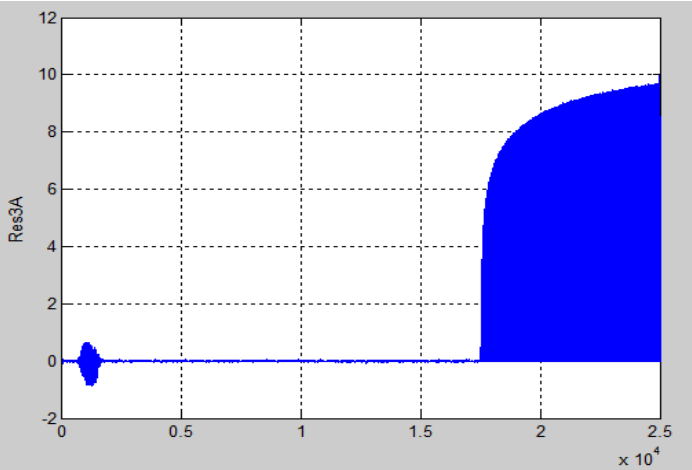


Figure 55: The residual of the approximation vector coefficients Res3A

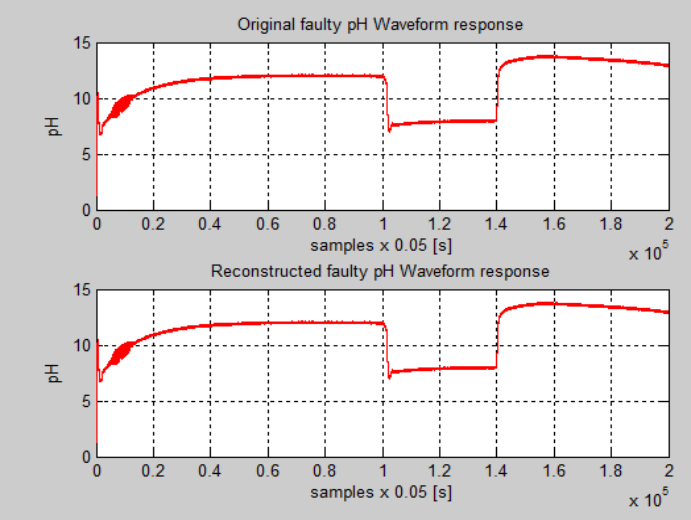


Figure 56: The original and reconstruct pH faulty signals caused by a Coulomb viscous friction in the control valve actuator.

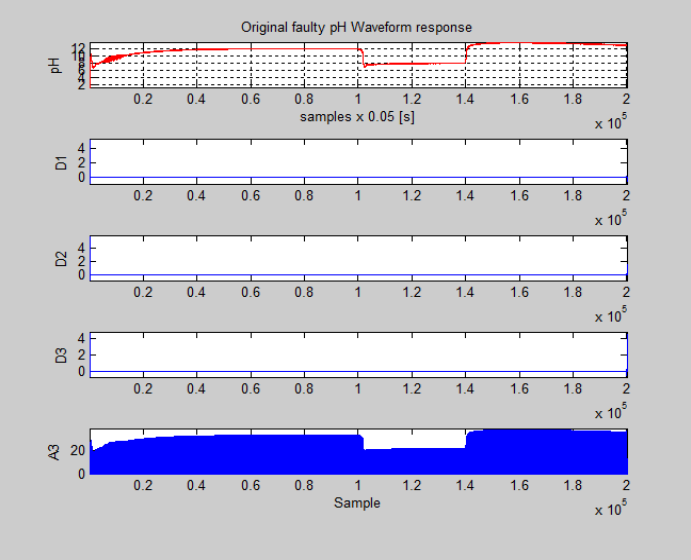


Figure 57: The detailed vectors coefficients D1, D2, D3 and approximation vector coefficients A3 for faulty pH signal caused by the Coulomb viscous friction in control valve actuator.

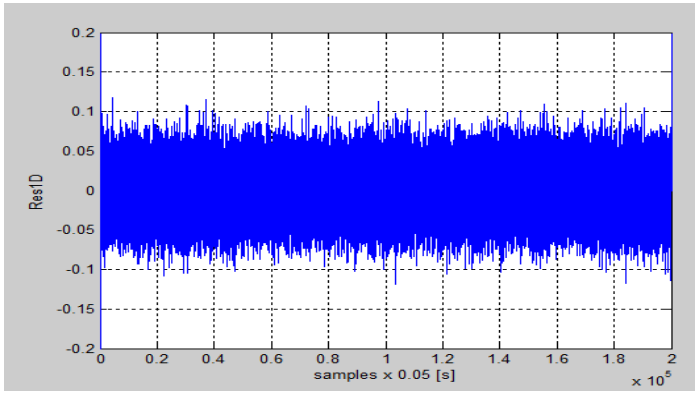


Figure 58: The residual of the detailed vector coefficients Res1D

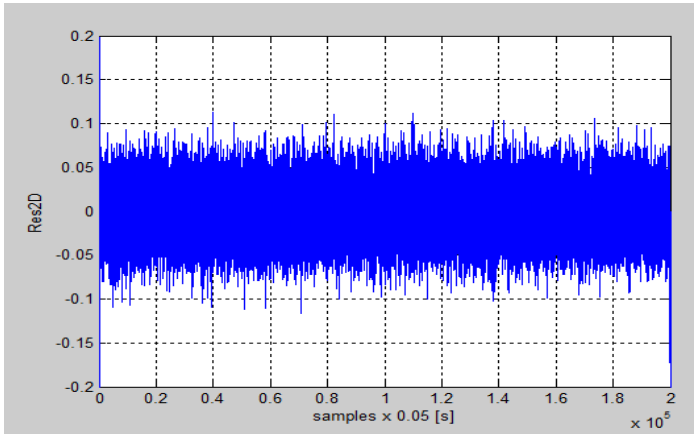


Figure 59: The residual of the detailed vector coefficients Res2D

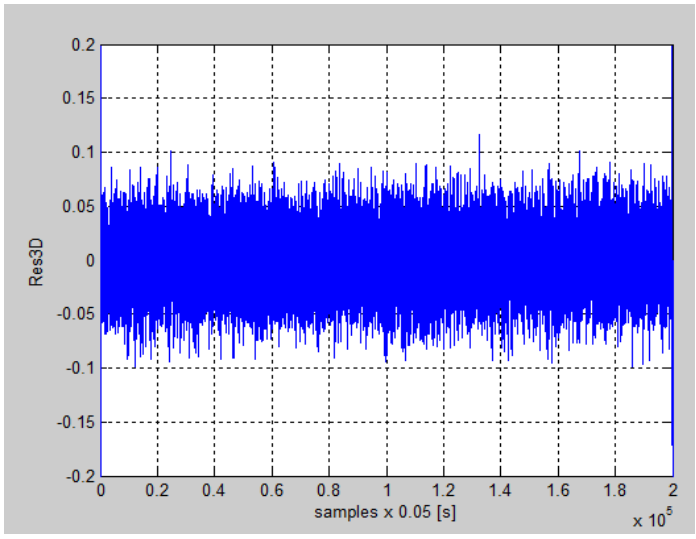


Figure 60: The residual of the detailed vector coefficients Res3D

8. Conclusions

In this research paper we open a new research direction in control systems applications field by performing a lot of investigations on the use of multisignal 1-D wavelet analysis to improve the accuracy, robustness, the design and the implementation in real-time of FDI techniques. These investigations are performed on the particular case study, mainly chosen to evaluate the impact of the uncertainties and the nonlinearities of the sensors and actuators on the overall performance of the control systems, namely a neutralization of

wastewater plant, in an attractive MATLAB SIMULINK simulation environment. The preliminary simulation results are encouraged and extensive investigations will be done in future work to extend the applications area. The 1-D wavelet analysis proved its effectiveness as a useful tool for signals processing, design and analysis based on wavelet transforms found in a wide range of control systems industrial applications. Based on the fact that in the real life there is a great similitude between the phenomena, we are motivated to extend the applicability of these techniques to solve similar applications from control systems field, such is done in our research work.

Conflict of Interest

The authors declare no conflict of interest.

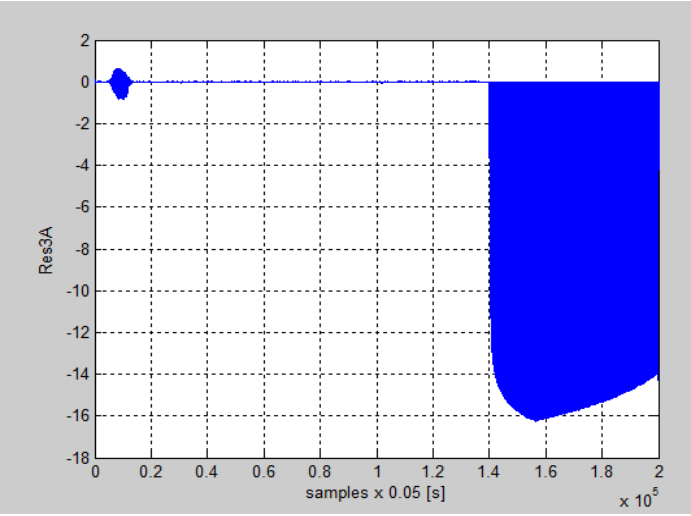


Figure 61: The residual of the approximation vector coefficients Res3A

References

[1] C.D. Bocaniala, J. Sa da Costa, and R. Louro, "A fuzzy classification solution for fault diagnosis of valve actuators", Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science book series, LNCS, Springer, Berlin, Heidelberg), vol.2773,Vasile Palade, Robert J. Howlett, and Lakhimi Jain (Eds.), pp. 741-747, 2003.

[2] N.Tudoroiu, M.Zaheeeruddin,"Fault detection and diagnosis of the valves actuators in HVAC systems, using frequency analysis", International Conference on Industrial Electronics and Control Applications, ICIECA 2005, 4 pages, Quito, 29 Nov.-2 Dec. 2005.

[3] D. Düstegör, E. Frisk, V. Cocquempot, M. Krysander, and M. Staroswiecki, "Structural analysis of fault isolability in the DAMADICS benchmark", Control Engineering Practice, vol. 14, no.6, pp.597- 608, June 2006.

[4] M. Farenzena and J. O. Trierweiler, "Valve backlash and stiction detection in integrating processes", presented at the 8th IFAC Symposium on Advanced Control of Chemical Processes, The International Federation of Automatic Control Furama Riverfront, Singapore, July 10-13, 2012, IFAC Symposium Preprints.

[5] T. Hägglund, " Automatic on-line estimation of backlash in control loops", Proc. Contr. J., vol. 17, no. 16, pp. 489 - 499, July 2007,

Published by Elsevier Ltd.

[6] S. Sivagamasundari, D. Sivakumar, " A practical modeling approach for stiction in control valves", presented at the International Conference on Modeling, Optimization and Computing (ICMOC 2012), Published by Elsevier Ltd., Procedia Engineering, Vol.38, 2012, pp.3308-3317.

[7] M.A.A. Shoukat Choudhury, N.F. Thornhill, and S. L. Shah, "Modeling valve stiction", Published by Elsevier, in Control Engineering Practice, vol.13, 2005, pp.641-658.

[8] http://www.springer.com/cda/content/document/cda_downloaddocument/9781848827745-c1.pdf?SGWID=0-0-45-812713-p173913817.

[9] Fang Liu, "Data-Based Fault Detection and Isolation Methods for a Nonlinear Ship Propulsion System", Masters of Applied Science Thesis, School of Engineering Science, July 2004, Simon Fraser University, Burnaby, BC, Canada.

[10] C. Kerstin Schremmer, "Multimedia Applications on the Wavelet Transform", PhD Thesis, University from Mannheim, 2001.

[11] I. Rosdiazli, "Practical Modelling and Control Implementation Studies on a pH Neutralization Process Pilot Plant, PhD Thesis, March 2008, University of Glasgow.

[12] Qian Xiao, Jianhui Wang, "The Multi-Frequency Stochastic Resonance Detection Based on Wavelet Transform in Weak Signal ", International Journal of Information and Systems Sciences, vol. 6, no. 4, 2010, pp. 377–384.

[13] https://www.mathworks.com/examples/wavelet

[14] https://www.mathworks.com/help/wavelet/examples/multisignal-1-d-wavelet-analysis.html ,accessed in May 25th 2017

[15] C. Garcia, R.J. Correa de Godoy, " Modelling and Simulation of pH Neutralization Plant Including the Process Instrumentation", in Applications of MATLAB in Science and Engineering, Tadeusz Michalowski-Editor, Ed. InTech, available from http://www.intechopen.com/books/applications-of-matlab-in-science-and-engineering/modelling-andsimulation-of-ph-neutralization-plant-including-the-process-instrumentation.

[16] A. Terry Bahill, "A Simple Adaptive Smith-Predictor for Controlling Time-Delay Systems", Tutorial, available on the web site: http://sysengr.engr.arizona.edu/publishedPapers/SmithPredictor.pdf, accessed in May 15th 2017.

[17] Michel Misiti, Yves Misiti, Georges Oppenheim, Jean-Michel Poggy, "Wavelet Toolbox's User Guide-MATLAB R2017a", The MathWorks Inc., 2017, available on the website: https://www.mathworks.com/help/pdf_doc/wavelet/wavelet_ug.pdf

[18] https://www.mathworks.com/help/wavelet/ug/wavelet-changepoint-detection.html

[19] H.M.Paiva, M.A.Q.Duarte, R.K.H. Galvano, S. Hadjiloucas, "Wavelet based detection of changes in the composition of RLC networks", Dielectrics, Journal of Physics: Conferences Series 472, 2013.

[20] H.M.Paiva, R.K.H. Galvano, L. Rodrigues, "A wavelet-based multivariable approach for fault detection in dynamic systems", Sba Controle & Automação, Natal , vol.20, no.4 , pp.455-464, 2009.

[21] MATLAB Wavelet Toolbox Documentation "Orthogonal and Biorthogonal Filter Banks",

https://www.mathworks.com/help/wavelet/ug/orthogonal-and-biorthogonal-filter-banks.html , MATLAB R2017a.

[22] M. Rafiqul Islam "Wavelets, its Application and Technique in signal and image processing", Global Journal of Computer Science and Technology, vol. 11. Issue 4, 2011, https://globaljournals.org/GJCST_Volume11/7-Wavelets-its-Application-and-Technique-in-signal-and-image-processing.pdf

[23] Hong, D., J. Wang, R. Gardner, "Real Analysis with an introduction to Wavelets and Applications", Elsevier Inc., ISBN: 0-1235-4861-6, pp. 256-307, 2005.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 2, No. 3, 1711-1726 (2017)
www.astesj.com
Special Issue on Recent Advances in Engineering Systems

ASTES Journal
ISSN: 2415-6698

Pluggable Controllers and Nano-Patterns in Java with Lola

Yossi Gil [1], Ori Marcovitch[1], Matteo Orrú[*,1], Ori Roth[1]
[1]*Computer Science Dept., CS Taub Building, The Technion—IIT, Haifa 3200003, Israel*

ARTICLEINFO
Article history:
Received: 11 June, 2017
Accepted: 21 July, 2017
Online: 09 September, 2017
Keywords:
Programming Languages
Preprocessors
Macro Languages
Pluggable Controllers
Language Augmentation
Nano-Patterns

ABSTRACT

Pluggable controllers are a different way to design control constructors such as `if`, `while`, `do`, *switch, and operators such as short circuit conjunction (*`&&`*) and the "*`?.`*" operator of the Swift programming language. Adoption of pluggable controllers enables the final user to modify and extend the control flow constructs (*`if`, `while`*, etc.) of an underlying programming language, the same way they can do if they implement functions such as* `printf` *and class* `String` *in a standard library.*

In modular, pluggable controller based language design, beside core control constructors, there are others, defined in standard libraries, with the purpose of augmenting and enriching the language. These pluggable controllers are extensible and replaceable. Being less intertwined in the main language, control constructor libraries can evolve independently from it, and their releases do not mandate new language releases.

We illustrate the implementation of pluggable controllers using Lola*, a powerful language-independent preprocessor and macro language. We demonstrate the introduction of new pluggable controllers with two case studies. The implementation of a Java stenography based on prevalent Java idioms, called "nano-patterns" or nanos, and the introduction in Java of new code constructs inspired by the Mathematica language's commands.*

1 Introduction

A recent publication [1] described a preliminary new perspective of viewing keywords such as `if` and `while` as library functions such as `printf`: standardized, but user extendable and replaceable. In this work, we expand on this vision, demonstrating the extension of Java with control keywords for two applications: a stenography for a concrete nano-patterns language whose prevalence in Java was previously demonstrated [2] and the implementation of Mathematica-like control structure in Java.

1.1 Control Constructors

Control Constructors are defined as those elements of a programming language that make it possible to assemble *commands*. In textbooks [3–5] as well as in classic works on programming languages [6–9], is reported that there are essentially three kinds of control constructors:

| ***Sequential*** | ***Iterative*** | ***Conditional*** |

A further distinction can be made between *atomic* and *compound* commands, where the latter are formed by primitive and other smaller compound commands. In Pascal, for example, atomic commands are the empty command, the assignment, and the procedure call. On the other hand, `Begin`...`end` is the sequential constructor, whereas `While`...`do`... is an iterative constructor, and, `If`...`then`... is a conditional constructor.

In real languages, there is no sharp separation between *expressions* and *commands*. Many *operators* that form expressions can act as control constructors. This is the case of standard short circuits operators such as "`&&`" and "`||`", and standard conditional operators such as "`· ? · : ·`", Other examples worth mentioning are represented by:

1. The "`,`" (comma) operator of C (the sequential comma operator of C).
2. The "`||`" operator, which "provides a default value", the Python variant of a short circuit.
3. The "`??`" operator (`null` coalescing) of C#.
4. The "`?.`" operator (*fluency on* `null`) of Swift.
5. The "`noexcept`" operator of C++, recently introduced, which guards against exceptions.

[*]Computer Science Dept. The Technion—IIT, Taub Building, Haifa 3200003, matteo.orru@cs.technion.ac.il

These operators can be expressed in terms of sequential, iterative, and conditional control constructors (henceforth, "*SIC*"). *Structured programming* [7] is defined by the notion of SIC, meaning that SIC are sufficient to impose structure on the unstructured. In fact, any program that has **goto** instructions in it can automatically be converted into an equivalent one that uses only SIC [10].

Concrete languages sometimes deviate from the traditional concept of SIC. A typical example is the **while** command of Python that has its peculiar **else**. Another example is represented by the different ways **switch** statements deal with fall-through cases, which is different in some languages. Variations and diversifications also appear in the semantics of

try...**catch**...**finally**

blocks. Language engineers, developers and practitioners are generally interested in using new language features that are likely to enhance their productivity, efficiency and reduce their errors. In general, the creativity of language engineers slows down after a languages' first release. Even small changes to a (successful) language definition might have unpredictable, potentially negative effects in the least expected places [11, p.497–508].

For the above reason, the long time taken before introducing the **switch**ing on strings feature in Java (see Fig. 1) is understandable. It took around twenty years from proposal[1] to implementation[2]—probably because of worries about its implications for engineers.

```
public static void main(String[] args) {
  for (String arg: args)
  if (arg.equals("-c") || arg.equals("--bytes"))
  . . .
  else if (arg.equals("-m") || arg.equals("--chars"))
  . . .
  else if (arg.equals("-w") || arg.equals("--words"))
  . . .
  else if (arg.equals("-l") || arg.equals("--lines"))
  . . .
  else if (arg.equals("--help"))
  . . .
  else if (arg.equals("--version"))
  . . .
  . . .
}
```

Figure 1: Java code with multiple comparisons of the same string variable with a sequence of string literals.

The drawback of the prudence characteristic of language architects, when it comes to introducing new features, is that it might negatively affect software systems. For example, consider that the late introduction of generics in Java forced developers to use the unsafe **Vector** as a substitute.

1.2 A Modular, Plugin-Oriented Approach

This paper raises the idea that the design of programming languages, specifically in respect to their control constructors' design, may follow a different, modular *plug-in* oriented approach. This proposal suggests that a programming language should have a number of core control constructors, whereas additional and more sophisticated control constructors, which we called *pluggable controllers*, can be defined in standard libraries of controllers, similar to what happens with functions such as **printf** in C or classes such as **String** in Java.

This approach promotes the decoupling between the controllers and the language architecture, leaving to both sides the freedom to evolve independently. This solution has advantages in different scenarios. The final user would be able to modify and extend the control flow constructs such as **if**, **while**, etc. of an underlying programming language. Language designers would be able to experiment new features before changing the language architecture. The proposed approach could also foster more sound discussions on the introduction of new features inside the users community.

The problem of the modularization of languages components is he main problem of *Modular Language Development*, a research branch that investigates how to componentize language design. Several solutions to the problem of componetization are available—see, for example, Cazzola and Vacchi [12] which adopt a solution based on *Traits* [13].

Another approach exploits the concept of *extensible language*, which presents some challenges such as that of adapting the parser according to language evolution [14] .

1.3 Lola

The imposition of a new code constructors on an underlying language can be done using Lola [15], the ***L**anguage **O**f **L**anguage **A**mendment*, which is a modern, language-independent, preprocessor and macro language, orientated to *language extension*. Lola makes it possible to augment and amend syntactical constructs of any host language.

Lola works like a filter [16] (see also [17, Sect. 3.1]) that converts an input stream comprising tokens of different kinds into an output stream. Lola allows designers to extend and enrich existing languages with new constructors without affecting languages architecture, since plugged controllers can be defined in Lola configuration files. In addition, with Lola, experiments on new language features can be conducted in controlled environments by scholars, designers and practitioners.

1.4 Nano-Patterns

A *nano-pattern* is a recurring solution adopted by developers, for a common task that involves control constructors. The fact that they are recurring solutions means that they are used often by developers. Another possible interpretation is that they are workarounds that developers found to overcome languages restrictions. In a previous study, it has been shown that there exist a number of prevalent nano-patterns. With the term "prevalent" we mean that they recur often in software systems. These solutions are potential candidates for inclusion as new language features—which can be done painlessly with a pluggable controllers approach.

[1] http://bugs.java.com/bugdatabase/view_bug.do?bug_id=1223179

[2] http://docs.oracle.com/javase/7/docs/technotes/guides/language/strings-switch.html

1.5 Contribution

The present paper makes several contributions:

- We propose a novel, modular approach to the design of control constructors, the use of *pluggable controllers*. This approach makes it easier to add new (experimental) features to an existing programming language.
- We introduce Lola, a new preprocessor and macro language, oriented to language augmentation. We illustrate its syntax, work flow and present some meaningful examples.
- We further discuss the concept of *nano-patterns* (from now on *nanos*), introduced in previous works [1] as recurring solutions adopted by developers, to common tasks that involve control constructors.
- We demonstrate how to impose pluggable controllers on top of Java using Lola. Specifically, we present two case studies that show a specific usage of control constructors: a new *stenograpy* for Java nano-patterns and the Java implementation of common Mathematica commands.

Outline: The paper is organized as follows. Sect. 2 illustrates some basic concepts to establish a common vocabulary. Sect. 3 presents Lola, the Language Of Language Amendment. Sect. 4 describes the two case studies above mentioned. Sect. 5 discusses some practical applications and outlines some promising avenues for researchers and practitioners. Sect. 6 reports on related work. Sect. 7 concludes and suggests future directions for this research.

2 Background

2.1 Pluggable Controllers

The following is an intuitive definition of pluggable controllers:

> **Pluggable Controllers**
>
> Controllers should be just like functions and classes found in a library: *standardized*, yet *extendable* and *replaceable*.

The term *controller* in this definition encompasses both classical control constructs such as `while` and `if` ⋯ `else`, *and* operators such as "`??`" that control the order of evaluation of their arguments. According to the pluggable controllers approach, a language has only essential, *built-in* SIC. A standard library of varieties of controllers complements the basic SIC: For example, built-in might be the following:

$$\{c_1; \cdots c_n\} \tag{2.1}$$

or

$$\texttt{while (}e\texttt{)}\ c \tag{2.2}$$

and

$$\texttt{if (}e\texttt{)}\ c_1\ \texttt{else}\ c_2, \tag{2.3}$$

while a standard library contains `for`, `do` and `switch`. In any case, the library is not sealed. On the contrary, developers may add their own control constructors, e.g., adding

$$\texttt{while (}e\texttt{)}\ c_1\ \texttt{else}\ c_2 \tag{2.4}$$

to Java. With pluggable controllers, extending `switch` to strings becomes possible using by library evolution rather than a new language version. Further, no disturbance to the language's core should occur by adding a "`?.`" operator to it. Pluggable controllers join the trend of parameterization of programming languages' elements.

Historically, standard procedures such as `Write` were hardwired in many programming languages, and the same happens today, even in contemporary languages such as AWK. Pascal was the first language to introduce pre-defined procedures such as `Writeln`, functions such as `Sin`, literals such as `true`, and types such as `Integer`. These predefined types, functions, literals or procedures can be overridden by developers when necessary. Likewise, in C, anyone can produce their own version of `printf` from the standard `<stdio.h>` library. In Java, the atomic types, e.g., `int`, `double` and `boolean`, are built-in, whereas their Go equivalent are *pre-defined*.

2.2 Nano-Patterns

The definition of nano-patterns (or simply nanos) was introduced at the end of the seminal work on Micro-Patterns [18] and lately used by authors such as Singer et al. [19] and Batarseh [20], who provided their definitions. In this paper we use the following definition:

> **Nano-Patterns** (*intuitive definition*)
>
> A *nano-pattern* (or *nano*) is a pattern of (typically less than a dozen) control constructs, which recurs frequently, serving common or similar purpose, yet cannot be abstracted easily by a function.

Our working hypothesis was that ordinary control constructors are designed with a prudent view of their usability. At the same time, nanos leverage a unique way of using control constructors that emerges from their function, heretofore not necessarily discerned by language designers. Moreover, the different ways in which control constructors may be used are not static but evolve over time. They are continually being developed and perfected in many specific domains.

A case in point regards the nanos of iterations that were not included in the design of traditional languages. Consider the design of a pluggable controller that captures the following nano[3] in Java code:

$$\texttt{while (}e\texttt{)}\ c_1\ \texttt{else}\ c_2 \tag{2.5}$$

[3] This is actually a control constructor in Python.

Introducing this pluggable controller in Java might be worth the effort depending on the frequency of the patterns using `if` and the auxiliary variables required to capture its behavior. The same applies to other constructs that capture recurring tasks such as "apply an and/or/sum" and other associative operations on `Iterable`s" or "iterate, zipper style, on two lists" A high frequency of reuse may suggest that developers would gain some benefits by the introduction of a pluggable controller. In order to avoid dangerous side effects on the language architecture, the proposed solution is to place controllers in a library rather than in the language core.

In a previous work, a catalog of 38 nanos was identified through a process that includes an initial *subjective* scrutiny and successive further evaluations using the *objective* prevalence threshold test against a *baseline* corpus [2]. All the found nanos are:

- *Traceable*, using an appropriate parser (our nano tracer).
- *Purposeful*, because they are aimed at performing a specific programming task.
- *Prevalent*, according to the definition that follows.

The *Prevalence* of a nano in a given project of a corpus is equivalent to the number of times that it recurs in that specific project. A nano is "prevalent" if its *prevalence* value is higher than a given *prevalence threshold*. The *prevalence threshold* ρ is defined as the fraction of projects of a corpus, for which nano prevalence is higher than a given *reuse threshold*. Here we are considering $\rho = 0.5$, as in reference [2]. The *reuse threshold* r_{th} of a project p is computed similarly to the way done for the popular h-index [21] of a researcher: A project p has a method reuse index r if it has r-methods that are reused r times or more. Thus a nano meets the prevalence threshold criterion if its prevalence is higher than the *reuse threshold* in, at least, 50% of the projects in a corpus C.

The prevalence threshold criterion is robust, objective and filters out irrelevant or domain specific candidate nanos. The reuse index is more robust than other statistics such as *mean* and *median* [22]. In fact, reuse is characterized by a long tail distribution as it happens in many software engineering contexts [23–26]. Being based on the h-index, also the r-index presents the same advantages, apart from the mentioned robustness, such as the fact that it captures the "core" of projects' methods, namely those that are more reused. The insensitiveness towards small values (low reuse of methods) allows to filter out those values that are related to project idiosyncratic features [27]. At the same time, the r-index "inherits" from the h-index some of its drawbacks. When it is used to compare different projects, it is needed to take into account that there are intrinsic differences between projects, such as their age. The *reuse threshold* tends to mitigate some of these drawbacks. Moreover, when comparing the r-indexes of two projects, the only thing that matters is that it refers to the "core" methods. Two projects with the same r-index are treated as equivalent, even if there are marked differences in terms of total number of reuse or maximum number of reuse. From a different perspective, this can be considered an advantage because it means that the r-index implicitly ignores outliers.

Table 1 lists 19 out of the 38 nanos in the original catalog [2]. These nanos are those that involve **SIC**s. This new catalog has two categories: *"Expressions and Command Elaborators"* and *"Operation and Operators on Multitudes"*. Nanos included in the first category are used to manage small errors and exceptional values, substituting missing values with default ones, guarding against null parameters or missing pre-conditions in method execution and handling unusual control flow. The second category covers simple operations on multitudes (i.e., arrays, lists, sets, etc.) including but not limited to: retrieving subsets of values, applying commands, computing cardinality of a multitude, etc. The second column provides the names of the nanos, whereas the third column presents their *Intent*, which is a description of their purpose, written in pseudo code. It is worth noting that often a nanos' name is obtained by juxtaposing the corresponding pseudo code keywords. For example, the following code describes the intent of the nano evaluateUnlessDefaultsTo:

$$[\ \textbf{Evaluate}\ ]\ e\ \textbf{unless}\ b\ [\ \textbf{default}\ e'\ ], \tag{2.6}$$

which captures the following Java snippet of code like:

```
pos > 0 ? pos - 1 : 0
```
(2.7)

The pseudo code adopts the following conventions: C–**c**ommand, e–**e**xpression, p–**p**redicate, M–**M**ultitude (i.e., an array, stream, list, set, collection, etc.), b–**B**oolean expression, i–**i**dentifier, T–**T**ype, and, X–e**X**ception type. The fourth column reports the prevalence. Its values are related to the first, *baseline*, corpus, called Gil-Lalouche Corpus (from now on GL Corpus). It is composed of 26 popular projects selected from the *GitHub's TrendingWikiBookJavaIdiom repositories* [4] augmented by the *GitHub Java Corpus* [5] list due to Allamanis and Sutton [28]. See Table 2 for a list of the projects. This corpus was previously used in other independent research [29].

[4] https://github.com/trending?l=javasince=monthly

[5] http://groups.inf.ed.ac.uk/cup/javaGithub/

Table 1: Language of Java Nano-Patterns

	Name	Intent[a]	Prevalence[b]
	Expression and Command Elaborators		
1	executeUnless	**[Execute]** C **unless** b **[Execute]** C **when** b	100%
2	whenHoldsOn	**When** $p(\cdot)$ **of** x,y,z: C_1 **of** w: C_2 **otherwise** C_3	100%
3	defaultsTo	e_1 **default** e_2 e_1 `??` e_2	81%
4	questionQuestion	e_1 **default** e_2 **else** e_3 e_1 `??` $e_2 : e_3$	96%
5	safeNavigation	e`?.`f e`?.`$m()$	50%
6	evaluateUnless-DefaultsTo	v = **[Evaluate]** e **unless** b **[default** e'**]**	92%
7	safeCast	T`?(`e`)`	96%
8	throwOnFalse	**Throw** X **when** b b `?!` X	100%
9	throwOnNull	**Throw** X **when** e **nulls** e `!!` X	81%
10	notNullRequired	**Return default when** a **nulls** T f`(safe` F_1`,...,safe` F_n`){...};`	88%
11	notNullAssumed	**Return default when** e **nulls** $e \neq$ `null || return default;`	92%
12	holdsOrReturn	**Return default when** b b `|| return default;`	100%
13	ignoringExceptions	{...} `ignoring` $X_1,\ldots,X_n$`;`	58%
14	letInNext	`let` $v \leftarrow e$ `in` C`;`	100%
	Operations and Operators on Multitudes		
15	forFromTo	**For** $i = 0,\ldots,n-1$ **[do]** ... `for` $(i$: e_1`..` $e_2)$ C`;` // $i = e_1,\ldots,e_2-1$ `for` $(i$: `..` $e)$ C`;` // $i = 0,1,\ldots,e-1$ `for` $(i$: e`..)` C`;` // $i = e,\ldots$ (∞-loop) `for (..)` C`;` // ∞-loop (no loop index)	73%
16	aggregate	**Aggregate** M **[st** $p(\cdot)$**] with** $A(\cdot,\cdot)$ M`.filter(...).reduce(...)`	58%
17	selectBy	**Select** $i \in M$ **s.t.** $p(i)$ $\{x \in M \mid p(x)\}$ $M/p(\cdot)$ M`.filter(...).collect(...)`	73%
18	forEach	**For** $i \in M$ **[s.t.** $p(i)$**]** **[do]** ... `for (..)` $p(\cdot)$`:`C M`.forEach(...)` M`.filter(...).forEach(...)`	96%
19	firstSuchThat	**First** $\cdot$ $M/p(\cdot)$ M`.filter(...).findFirst()`	50%

[a] e.g., pseudo-code, Java 8 streams, suggestion for extending plain Java syntax, etc.

[b] Prevalence score in the baseline corpus

Table 2: Baseline corpus of 26 OSS Java projects: *Identification*, *reproducibility* information and *work volume* indicators.

Project	First Version	Last Version	Last Hash	#Days	#Authors
Atmosphere[1]	10-04-30	14-04-28	557e1044	1,459	62
CraftBukkit	11-01-01	14-04-23	62ca8158	1,208	156
Cucumber-jvm	11-06-27	14-07-22	fd764318	1,120	93
Docx4j	12-05-12	14-07-04	8edaddfa	783	19
Elasticsearch	11-10-31	14-06-20	812972ab	963	129
Essentials	11-03-19	14-04-27	229ff9f0	1,134	67
Guava[1]	11-04-15	14-02-25	6fdaf506	1,047	12
Guice	07-12-19	14-07-01	76be88e8	2,386	17
Hadoop-common	11-08-25	14-08-21	0c648ba0	1,092	69
Hazelcast	09-07-21	14-07-05	3c4bc794	1,809	65
Hbase	12-05-26	14-01-30	c67682c8	613	25
Hector	10-12-05	14-05-28	f2fc542c	1,270	95
Hibernate-orm	09-07-07	14-07-02	b0a2ae9d	1,821	150
Jclouds	09-04-28	14-04-25	f1a0370b	1,823	100
Jna	11-06-22	14-07-07	a5942aaf	1,110	46
Junit[1]	07-12-07	14-05-03	56a03468	2,338	91
K-9	08-10-28	14-05-04	95f33c38	2,014	81
Lombok	09-10-14	14-07-01	3c4f6841	1,721	22
Mongo-java-driver	09-01-08	14-06-16	5565c46e	1,984	75
Netty	11-12-28	14-01-28	3061b154	762	72
Openmrs-core	10-08-16	14-06-18	05292d98	1,401	119
RxJava	13-01-23	14-04-25	12723ef0	456	47
Spring-framework	10-10-25	14-01-28	c5f908b1	1,190	53
Titan	13-01-04	14-04-17	55de01a3	468	17
Voldemort	01-01-01	14-04-28	fb3203f3	4,865	56
Wildfly	10-06-08	14-04-22	5a29c860	1,413	194
		Total:	*26 projects*	38,250	1,932

The complete study regarded a total of 78 Java projects, split into three parts: 26 belong GL Corpus, 26 are the most starred GitHub Java projects and the remaining 26 are the most starred Android GitHub projects at the time we collected the projects[6]. The GL corpus was used as a training corpus. At first the initial catalog of candidates was collected, through a process of *pattern harvesting*, by analyzing a set of six Java projects (partially overlapping with the GL corpus). Next, the prevalence of each nanos belonging the initial catalog of candidates, was computed on the GL corpus. Following this, the nanos whose prevalence is lower than the prevalence threshold, were discarded. Finally, the prevalence was computed for the projects of the remaining corpora (testing corpus). Information regarding the reproducibility appears in Table 2.

[6] February 2017

3 Lola

Lola combines computational expressiveness with minimal syntax. This is made possible through the adoption of Python as an underlying computation model. Lola also includes high level pattern matching, independence from the host language and a declarative nature (achieved with directives from C). Lola input is composed of *host tokens*–those of the host language, Python *snippets* of code, Lola *keywords* and *user defined* keywords. In order to distinguish the host-tokens from the Lola keyword (such as **`##Find`** and **`##replace`**), the former are defined in an XML configuration file.

The output stream is the result of the application of "directives" found in the input stream to subsequent input. The directives correspond to macros whose invocation is triggered by a pattern matching engine that relies on regular expressions over tokens. This makes it possible to augment the host language syntax without having to meddle with the language semantics. Macros are expanded to equivalent constructs written in the host language syntax. They can also be expanded in the augmented syntax, with the purpose of being further expanded by other Lola macros to code in the host language such as Java or C.

There are two kinds of directives: *Generator*s and *Lexie*s. Generators are constructs that use *Generating Expression*s (*GE*s) to return sequences of host-tokens inside the output stream. We can distinguish between two kinds of generators: *atomic* or *constructor*. Examples of atomic generators are **`##`**, **`##Include`** and **`##Import`**. Constructors are, for example, **`##If`**, **`##Unless`** and **`##Case`**. When the

preprocessor encounters generators, it pauses the process of copying host tokens from the input to the output stream and inserts into the stream the result of the generating expression.

Lexies are Lola's basic elements of computation and contain the instructions that determine the outcome of the Lola execution. In practice, they describe the augmentation to the host language. The augmentation is expressed in terms of *extended Regular Expressions* (*RE*s). When the preprocessor finds a lexi in the source code, it tries to match the *RE* reported in **##Find** directives against the subsequent tokens. Lexies are structured in sections, with a *Declarative Header*, an *Action Section*, and a *Declarative Footer*. Sections may occur in any order. Directives such as **##Find** and **##description** belong to the Declarative Header, which contains the *RE* that can be matched by the subsequent host language tokens. Section such as **##replace** and **##run** belong to the Action Section, which contains the elaborators for the code matching *RE*s in previous sections. The Declarative Footer usually contains directives such as **##example** and **##log**, used for documentation purposes.

*GE*s and *RE*s are both *Compound Object*s of Lola. GEs are used as conditional commands (i.e., **##If**, **##Unless**, etc.) or to iterate over the elements of a collection (usually a list) such as in the case of **##ForEach**. REs appears in a lexi after the **##Find** directive. When the input stream *RE* matches, Lola creates a Python object that contains information about the location in the input. This information is accessible by the other directive such as **##run** and **##replace** in other sections of the lexi. For example, it is possible to access the **self** and **__str__** attributes (and many others).

In sum, the Lola workflow is the following: Lola tries to match the pattern reported in the **##Find** directive. When it succeeds, this triggers, for example, the **##replace** directive reported in Fig. 4, which replaces the found pattern with code written in the host language. *Patterns* are extended regular expressions (REs). Whenever a snippet of code matches with a *RE* , Lola creates a Python *reifying* object, which can be manipulated. The computations performed by a lexi include code replacement, but a lexi can also invoke Python code.

Lola's syntax strives to adhere to English sentence structure. Furthermore, it uses capitalization conventions to make code easier to read: `@CamelCase` convention is used for constructors and `@camelCase` is for elaborators. The use of abbreviations or acronyms, including those of familiar terms such as EOF it is strongly discouraged. Lola can be used also for computing code metrics, enforcing code standards, and adding a C preprocessor functionality to any programming language. Nevertheless, its main purpose is to allow developers to introduce new keywords and operators, and in general, new syntax to the host language.

4 Case Studies

In the present Section we demonstrate the use of Lola to implement custom pluggable controllers with two specific goals:

1. To implement a Java Stenography.
2. To define new commands taken from a different programming language, Mathematica.

4.1 Stenography for Java Nano-Patterns

Stenography[7] is defined as *"an abbreviated symbolic writing method that increases speed and brevity of writing as compared to longhand, a more common method of writing a language"* [8]. The idea behind a stenography for Java nano-patterns, comes from the experience common to any developer. Everyone has written a lot of code is familiar with the feeling of repetitively writing small fragments of code. Although frequent fragments may take only few seconds to be written, if their frequency is high it may lead to a non-trivial effort for the developers, on the long run. With this mind, it is easy to understand that a stenography for nano-patterns is likely to increase programmers' productivity, by shortening the syntactic structure of several code constructs.

Nano-patterns are common Java idioms that have been proven to be recurring, namely prevalent, in a meaningful dataset of Java software systems. Since they refers to working code, their Lola implementation preserve the semantics in the host language. Table 1 provides the proposed stenography in the *Intent* column. Due to space constraints, we are going to illustrate two significant examples, one for each group. Specifically, for the first group we are illustrating notNullRequired nano whereas for the second group we are describing forEach.

Each snippet of code has a similar structure. The top part present a lexi, which reports the directives written in Lola language. These directives determine the substitutions of the tokens in the input stream. In the bottom part is reported the code governed by Lola. We can distinguish three kinds of keywords: *host specific* (Java keywords in this case), Lola's *built-in* (beginning with a double hash character **##**). These are *reserved* keywords. Finally we have *user defined* keywords, defined in Lola directives that, by convention, begin with a single hash character **#**. For example in Fig. 4 we have:

- **#safe** that is a new defined keyword.
- **##Find**, **##NoneOrMore**, **##Either**, **##or**, **##separator**, **##Any**, are Lola's keywords.
- **##ArgumentDeclaration**, **##Identifier**, **##SafeArgumentDeclarationList**, **##Type** are not primitive Lola's keywords, but are defined in a separate **##Find** directives contained in the **std.lola** library.
- **int** and **return** are Java keywords.

The first is an **##Import** directive:

##Import "std.lola" (4.1)

which is used to access to predefined directives. In fact, **##ArgumentDeclaration**, **##Identifier**,

[7] Also known as *shorthand writing*.

[8] https://en.wikipedia.org/wiki/Shorthand

##SafeArgumentDeclarationList, **##Type** are defined in the **std.lola** library.

The fundamental directive is **##Find**: It allows Lola to find sequences of tokens (patterns) in the input stream that match a specific extended regular expression (*RE*). For example in Fig. 5 the *RE* is the following:

$$for\ (\#\#Any(loop)\ |\ \#\#Expression(filter))\ \ \#\#Any(statements) \tag{4.2}$$

When a match is found Lola may trigger a number of actions, such as the substitution or deletion of elements, the recording of elements for later use, etc. The most common directive is the substitution according to the instruction found in the **##replace** directive (after the homonym reserved keyword). On occasion, it is possible to run Python script of code inside the **##run** directive. The mentioned Python code must be surrounded by curly braces.

Whenever a pattern matches, Lola creates a Python reifying object which can be manipulated in the **##run** section. The results of the matches are stored in variables whose identifiers correspond to the parameters of the directives. In case of multiple values, they are stored in lists.

For example, eq. (4.2) shows the use of an atomic *RE* , **##Any**. **##Any** is used twice one with **loop** parameter and the second with the **statements** parameter. **loop** and **statement** are the names of the variables in the reifying object. **##Expression** is a not a build keyword but is defined in the **std.lola** library.

4.1.1 The notNullRequired Nano

The notNullRequired is used to guard against **null** expressions. In case the expression is **null** the method returns. The lexi appears at the beginning of the code reported in Fig. 4. It introduces a user defined keyword called **safe**. The **#safe** modifier is applied to the parameters **x** and **s** of the method declaration at the bottom. It is used to check if the **##safe** parameters are **null**. In this case the method returns.

The first **##Find** directive defines the first pattern, which consists in a **#safe** modifier followed by the classic Java syntax for methods parameters. This is defined in the **##ArgumentDeclaration** directive in the imported **std.lola** library. Each match is stored in **SafeArgumentDeclaration**.

In the second **##Find**, *RE* matches if it finds:

1. one or more standard (unsafe) parameters declaration;
2. one or more **#safe** parameters declaration;
3. both the first and the second case interleaved (**##Either**);
4. none of them (**##NoneOrMore**).

In the third **##Find**, the *RE* matches the method declaration. Finally the **##replace** directive perform the substitution. It is worth to note the use of Python code, and specifically of the method **join** of the **String** class, inside a *list comprehension* expression, to generate the output. This script have access to **##SafeArgumentDeclaration** variable, to the element **l** of the **##SafeArgumentDeclarationList** **ls**. Each **l** has two fields (**name** and **type**, each of them has a **name** field).

```
int f(Integer x, String s) {
  Objects.requireNonNull(x);
  Objects.requireNonNull(s);
  return x + Integer.parse(s);
}
```

Figure 2: Java output of the notNullRequired nano

The Java implementation is reported in Fig. 2. It exploits the **java.util.Objects** class which includes 9 static methods to work on objects. For each method argument **x** with **#safe** modifier, adds a call to the static method **Objects.requireNonNull(x)** at the start of the method's body. This specific implementation throws **NullPointerException** when one of **#safe** arguments is **null**. However, we can think of other actions that can be done (i.e., printing a warning, returning null, etc.).

4.1.2 The forEach Nano

forEach implements a *filtered* **ForEach** loop. It is used to apply a command by iterating over a multitude or a subset of a multitude (as it happens in SQL). The lexi reported in Fig. 5 have a single **##Find** directive, that matches a pattern that includes an enhanced **for** loop and an expression separated by a *pipe* character **"|"**. The loop contains statements that are applied just if the expression after **"|"** is **true**.

```
for (final Integer i : nums()) {
  if (!(i%2==0))
  continue;
  f(i);
}
```

Figure 3: Java output of the forEach nano

Fig. 3 shows the Java code in output. The loop iterates over an Iterable returned by the nums() method, skipping the elements that do not satisfy the predicate *p* - in this case, if they are not even numbers. *p* can be a lambda expression.

4.2 Mathematica's Commands in Java

In the present section we present, as case study, the introduction of new commands in Java using Lola. We are particularly interested is on some commands of Mathematica, a popular language for technical computing. Mathematica presents several control structures that differs from those available in Java.

Table 3 reports the selection of the Mathematica's commands that we implemented. The second column reports the command in the Mathematica syntax, that is also the stenographic convention we adopted in the implementation. Descriptions are taken from the Mathematica documentation. Due to space constraints we are illustrating only one of the most meaningful commands, represented by the **Do** command.

```
##Import "std.lola"
##Find(SafeArgumentDeclaration) #safe ##ArgumentDeclaration
##Find(SafeArgumentDeclarationList) ##NoneOrMore ##Either ##ArgumentDeclaration(dec)
##or ##SafeArgumentDeclaration(dec) ##separator,
##Find ##Type(type) ##Identifier(name) (##SafeArgumentDeclarationList(ls)) {##Any(statements)}
##replace ##(type) ##(name) (##(",".join([l.type.name + "␣" + l.name.name for l in ls.decs]))) {
  ##("".join(["Objects.requireNonNull(" + l.name.name + ");" for l in ls.decs]))##(statements)}

int f(#safe Integer x, #safe String s) {
  return x + Integer.parse(s);
}
```

Figure 4: Java code governed by Lola for the notNullRequired nano

```
##Import "std.lola"
##Find for (##Any(loop) | ##Expression(filter)) { ##Any(statements) }
##replace for (##(loop)) {if (!(##(filter))) continue; ##(statements)}

for (final Integer i : nums() | i % 2 == 0) {
f(i);
}
```

Figure 5: Java code governed by Lola for the forEach nano

4.3 Loops: the Do Command

The **Do**[9] has multiple options and can be used in different ways. An example is reported in eq. (4.3): It iterates *n* times *expr*.

Do[expr, n] (4.3)

Fig. 6 reports the lexi and the Java code governed by Lola. It is possible to see the use of the **##run** directive: the **doName** variable is elaborated in the successive **##replace** directive as well as the **times** parameter of **##Literal** and the **ee** parameter of **##Any** in the **##Find** directive.

```
##Find do[##Any(e),{##Literal(times)}];
##run {
  if 'x' not in locals():
   x=0
  else:
   x=x+1
  doName='_i'+str(x)
}
##replace for(int ##(doName)=0;##(doName)<##(times);++##(doName)) {
  ##(e);
}

public void foo(){
  do[f(),{5}];
  do[g(),{15}];
}
```

Figure 6: Lexi for Do

Fig. 7 reports the output of the Lola engine. **f()** and **g()** are iterated *n* times (having *n* different values for each **for** loop).

```
public void foo(){
  for(int _i0=0; _i0<5; ++_i0) {f();}
  for(int _i1=0; _i1<15;++_i1) {g();}
}
```

Figure 7: Java output for Do

5 Discussion

Several opportunities may derive from the adoption of a pluggable, modular approach for programming language design. With every likelihood, some innovations in languages design might have been introduced faster and earlier. Also discussions around the new features would have been more sound, because supported by more realistic evidence of their feasibility and impact on languages architecture. In this work we are specifically interested in the design of control constructors. Lola is certainly helpful in this regards. Its preprocessing and macro language capabilities allow the final user to augment a language with new controllers (in a pluggable way) without affecting the language architecture.

With every likelihood using Lola we could have had **switch** on strings by augmenting the language with a library without changing the language version. It would have been the same for the C#'s **?.** operator[10]. Other improvements made possible by Lola is it multi-way, non-fall through branching conditional operator. And this would have been made using standard libraries of varieties, without changing language's core.

Some language extensions, such as the **?.** operator, are expected to be widely adopted by any kind of user, meaning that everyone is expected to eventually use them. Other application may be related to specific tasks. Programmers writing tests for an experts system may wish to define their own control constructors to support declarative tests such as

```
#Tweaking "int␣i=3;i+=2;" #gives "int␣i=5;";
```

Developers trained in the functional programming school, may find useful to use also in Java list expressions such as:

```
#sum #apply (_) -> 1./(_*_) #to primes();
```

Another possible application is represented by the development of libraries to deal with common tasks such as logging or interacting with SQL . As an example, consider Mockito, a popular "mocking framework for unit tests

[9] https://reference.wolfram.com/language/tutorial/LoopsAndControlStructures.html
[10] https://msdn.microsoft.com/en-us/library/dn986595.aspx

Table 3: Mathematica commands in Java

	Name	Command[a]	Description[b]
1	If	**If**[*condition*,*t*,*f*]	Gives *t* if *condition* evaluates to True, and *f* if it evaluates to False
		If[*condition*,*t*,*f*,*u*]	Gives *u* if *condition* evaluates to neither True nor False
2	Do	**Do**[*expr*,*n*]	Evaluates *expr* *n* times
		Do[*expr*,{*i*,i_{max}}]	Evaluates *expr* with the variable *i* successively taking on the values 1 Through i_{max} (in steps of 1)
		Do[*expr*,{*i*,i_{min},i_{max}}]	Starts with $i = i_{min}$
		Do[*expr*,{*i*,i_{min},i_{max},*di*}]	Uses steps di
		Do[*expr*,{*i*,{i_1,i_2,...}}]	Uses the successive values i_1, i_2, ...
		Do[*expr*,{*i*,i_{min},i_{max}},{*j*,j_{min},j_{max}},..]	Evaluates *expr* looping over different values of *j* etc., for each *i*
3	Nest	**Nest**[*f*,*expr*,*n*]	Gives an expression with *f* applied *n* times to *expr*
4	NestWhile	**NestWhile**[*f*,*expr*,*test*]	Starts with *expr*, then repeatedly applies *f* until applying test to the result no longer yields True
5	Which	**Which**[$test_1$,$value_1$,$test_2$,$value_2$,...]	Evaluates each of the $test_i$ in turn, returning the value of the $value_i$ corresponding to the first one that yields True

[a] Mathematica instruction.
[b] As reported in Mathematica's documentation.

in Java". Using a future (not the current) release of Lola, Mockito's developers would be able to rewrite the following snippet of code[11]:

```
Iterator i=mock(Iterator.class);
when(i.next()).thenReturn("Hello").thenReturn("World");
String result=i.next()+"␣"+i.next();
assertEquals("Hello␣World", result);
```

in the following way:

```
Iterator i=mock(Iterator.class);
#mock Iterator
#upon next() #return "Hello," #then "World!"
#affirm next() + "␣" + next() #is "Hello,␣World\n";
```

which is arguably more explanatory. With the present version of Lola it is not possible to implement the latest code, but it will be possible in the next release, currently under development.

In the Lola version, the following

```
#inlining...#to...
```

is a user defined control constructor, which can be seen as a syntactic sugar of the following instruction:

```
inliningInto("int␣i=3;i+=2;", "int␣i=5;");
```

In another form, using an appropriate fluent API library we have:

```
inlining("int␣i=3;i+=2;").to("int␣i=5;");
```

Lola can be seen as a special case of syntactic sugaring, task that is well performed by tools such as SugarJ [30], Racket [31] or Occam through Camlp4 [32]. Since Lola's focus is specifically on language extension it is possible to envision a rather coherent ensemble of applications of the idea of language extension, such as a definition of a DSL like fluent API. Lola can certainly be used during the DSLs development. At the same time, Lola itself can be seen as a DSL which uses *Macro Processing* as implementation approach [33].

Many other applications of Lola are possible in the field of testing, logging, design-by-contract, etc. (see, as an example, the recent work on *Seamless Requirement* from Naumchev and Meyer [34]) . Our current empirical study of nano-patterns in Java [2] indicates that nano-patterns occur in two thirds of methods, about half of the statements, third of conditional statements and 90% of all iterative statements, in the Gil-Lalouche corpus [29].

The basic scenario for Lola is that of language extension, amendment or augmentation. Users involved are developers, language engineers or advanced users interested in extending a General Purpose (GPL) or Domain Specific (DSL) host language. Apart from learning Lola's syntax (which, in the authors' opinion, should not have a steep learning curve for an average developer) the users need to know Python. This can be seen as a drawback of Lola. However, this issue is common to other similar solutions such as SugarJ [30], which presumes the knowledge of SDF [35] and Stratego [36]. On the other end, the use of Python might represent an opportunity, since its popularity [37,38], exposes Lola to a wider developers' community.

If language specification changes, this might in principle alter the behavior of the stenographic form of a nano-pattern imposed by Lola. In general, it is developers' responsibility to adjust Lola libraries according to the specification changes, so to guarantee *upward compatibility*. Nano-patterns have been proven to be popular solutions to small programming tasks, as results from their prevalence values. If we look at the baseline corpus described in Table 2, which represents a subset of the entire analyzed corpus, we find that the average *work volume* is, on average, 327.46 days (median 199 days), with a time range that spans from 2011 to 2014. During this time the Java Language Specification changed from version 3 (JLS3, 2004) [39] to 4 (JLS4, 2011) [40]. The JSL4 specification introduced important features, such as binary literals, diamond operators for generic type inference, etc.

Dyer et al., in their large scale investigation of Java projects hosted on SourceForge [12], found that new features are used by developers *before* the *official* release of the specification, taking advantage of the beta/pre-releases [41]. More likely than not, also the developers who worked on the projects included in the baseline corpus, adopted new language features during the development process. This means that the analysis is already, *implicitly*, taking into account the upward compatibility. In fact, the most prevalent nano-patterns emerged from source code that is likely to include both old and new features. This might suggest that either

[11] https://gojko.net/2009/10/23/mockito-in-six-easy-examples/
[12] https://sourceforge.net/

nano-patterns capture language constructs that were not interested by changes in the specification, or those that were affected by the mentioned changes, if present, did not pass our prevalence test.

Moreover, Dyer et al. found also that the majority of newly introduced language features are rarely used by developers [41], with few meaningful exception. According to Qiu et al. [42], who conducted a large-scale study on the use of Java constructs, the distribution of syntactic rules usage is *Zipfian*, with 20% of the most-used rules accounting for 85% of all rule usage, whereas the 65% of the least-used rules is used less than 5%. The same authors show that the adoption of new rules varies over time and it is contextual [42]. These findings might suggest that the problem of upward compatibility is not so significant in practice, being in fact mitigated by the tendency of developers to be, to a certain extent, recalcitrant to employ newly introduced features. Moreover, the decision to adopt an *h*-index variation as a reuse measure, is aimed at balancing the effect of such kind of statistical distribution, as discussed in Sect. 2.2.

6 Related Work

6.1 Lola

6.1.1 Embedded languages

Lola can be seen as an *embedded language*. The embedded text is distinguished by the host text because it is preceded by a hash character and ends with an un-escaped end-of-line character. The most trivial example involves comments. To a certain extent, also comments and string literals can be seen as embedded language. In Java, for example, comments begin and end with specific character sequences, `/*` and `*/`, respectively for the beginning and end of the comment. The same applies to JavaDoc comments. Inside a comment, text is treated differently than in regular code.

The most relevant example is represented by PHP, where commands are embedded in HTML [43] pages. Other meaningful examples are ASP [44] and JSP [45], which adopted the same idea. More recently we have a number examples of DSL languages which extend or enhance a host GPL language. They usually target specific problem, like interaction with databases as in the case of J% [46].

Lola is not an embedded language on its own. It actually embeds both the host language and Python. The host language is seen as a stream that can be manipulated, whereas Python snippets are used to compute the code processing.

6.1.2 Preprocessors

There are several preprocessors available. The C preprocessor [47], introduced in the early stages of its host language, C, is arguably the most famous text preprocessor, often identified by its acronym, CPP [48, 49]. It is almost entirely independent from the host programming language [50], to the point that it can be viewed as an independent programming language. As a programming language, the C preprocessor is limited. For example, it cannot perform iterations (loops) and it does not have conditionals on macro parameters and it does not allow recursive structure. It also has a minimal "library".

Other popular preprocessors are PL/1 and M4 [51]. The PL/1 preprocessor is similar to the C preprocessor in some respects (i.e., built-in types, directives, etc.) whereas M4 is a general language-independent preprocessor, which facilitates more complete processing (it allows recursive calls and has conditional constructs) if compared to the other above mentioned [52] preprocessors. M5 [53] is an improved version of M4.

Many preprocessors use a general purpose programming language, such as Python (e.g., Cog[13], jinja[14], django[15], PYM [54], pyexpander[16], pypreprocessor[17], Preprocessor in Python[18] and Cheetah[19]). Others use Perl (e.g., PerlPP[20]).

6.1.3 Macro Languages

There are several problems with macro systems. They might capture names that are already used (Hygenic Macros solve this problem [55]). Macro systems do not usually differentiate between lexical elements of the hosting language such as expressions, identifiers, constants, etc. One preprocessor that differs in this aspect is the Marco preprocessor [56], which has a way to reduce the coupling between the host language to the macro system.

6.2 Nano-Patterns

The first appearance of the term nano-patterns was in the conclusion of a work by Gil and Maman micro-patterns [57]. Singer et al. [19], starting from a work by Høst and Østvold [58], collected a language of 17 nano-patterns, which the authors demonstrated to be *prevalent* (at an 80% level), *traceable*, and, *purposeful*. The names of their found nano-patterns underscore their purposefulness. Singer et al.'s catalog is at the base of studies on the relationship between nano-patterns and defectiveness [59] or vulnerabilities [60].

Differently from our catalog, Singer et al.'s catalog regards properties of methods, whereas the nanos in our catalog are found at the method, command, expression and field

[13] `http://nedbatchelder.com/code/cog/`, Cog by Ned Batchelder.
[14] `http://jinja.pocoo.org/`, jinja by Armin Ronacher.
[15] `https://docs.djangoproject.com/en/1.7/topics/templates/`, Django Software Foundation.
[16] `http://pyexpander.sourceforge.net/`, pyexpander by Goetz Pfeiffer.
[17] `https://code.google.com/p/pypreprocessor/`, pypreprocessor by Evan Plaice.
[18] `http://orbeckst.github.io/GromacsWrapper/gromacs/core/fileformats/preprocessor.html`, Preprocessor in Python by Evan Plaice and Oliver Beckstein.
[19] `http://www.cheetahtemplate.org/`, Cheetah by Tavis Rudd.
[20] `https://github.com/d-ash/perlpp`, PerlPP by Andrew Shubin.

levels [2]. Singer et al.'s nanos are orthogonal. Each nano-pattern represents a binary property, meaning that it can be present or not. This property lead to the definition of composite nano-patterns, which are those that are a combination of *fundamental* nano-patterns.

Other works on nano-patterns are those of Batarseh [20], who presents a language formed by 16 method properties and Lee et al. [61], who report 67 attributes of different kind of nanos: method signature, body and ties of body and behavior.

6.2.1 Other Kinds of Patterns

In the software engineering literature (scientific and not), there has been a surge of interest regarding the theory and application of "design patterns". Such patterns are a specific arrangement of object oriented components (classes, methods, inheritance) and represent recurring solutions to common design problems [62, 63]. Research interest revolves, for example, around design patterns automatic detection.

At a lower level of granularity stands the "micro patterns" (or μ-patterns), which are predicated on OO types (classes). Micro patterns reflect a specific use of OO features, such as the absence of methods, inheritance, etc. [57]. Differently from design patterns, the 27 micro patterns in Gil and Maman's catalog are *prevalent*, which means that they are present in around 75% of all classes, as empirical studies have shown [57]. Micro patterns are *traceable*, in the sense that they can be automatically recognized. In opposition, design patterns are not traceable [64] and many attempts have been made to formalize [65–69] and to automatically detect them [70–75].

Design patterns are also purposeful, which means that they deal with a specific problem. This is not the case of μ-patterns (they just track the presence of a coding technique), albeit the prevalence might suggest that a μ-pattern is used on purpose. The lack of purpose is also a characteristic of nano-patterns, although for a different reason. Their purpose is related to the small programming task that they carry out, and can often be learned from their name, but it is usually unrelated to the system where they occur.

6.2.2 Other Pattern-Like Constructs

A concept similar to nano-patterns is that of *idiom*s, which are, quoting Allamanis and Sutton, "a syntactic fragment that recurs frequently across software projects and has a single semantic purpose" [76]. The main difference between idioms and nano-patterns is that the latter are not single fragments but predicates.

Notably we can find idiom catalogs for different languages, e.g., [77–80]. Also, IDEs provide support for the definition and customization of idioms [81, 82]. Some instructional/educational material on idioms is found in work whose aim is to teach idioms as programming tips [83–85].

Sutton et al. investigated the presence of idioms in generic C++ libraries, finding a high coverage (circa 85% of classes showing idioms) [86]. In a recent work, Allamanis and Sutton applied machine learning techniques to *automatically* detect idioms in source code. The discovered idioms included "cross-projects idioms that represent important program concepts like object creation, exception handling, etc." with a prevalence ranging between 3% to 31%, depending on some factors (i.e., the training and testing dataset, the used parameters, etc.) From their work, we can see that the prevalence of idioms tends be lower than that of nanos [76].

Another difference between the two constructs is that nanos are *sought*, whereas idioms are discovered. Some of the found idioms show semantic purposes, in the sense that they are used for object creation, exception handling, and resource management. In the literature on idioms, other relevant studies are those of Koening [87], Langer [88] and Willis [89].

6.2.3 Vocabulary *vs.* Structure

Vocabulary: Some related work deals with code nomenclature or the study of the vocabulary that developers use to describe program elements. Linstead et al. studied source code vocabulary and discovered the existence of naming trends related to specific syntactic elements such as classes and interfaces [90]. Enslen et al. worked on an optimized algorithm to split identifiers into words' sequences [91]. Abebe et al. studied the evolution of vocabulary used by developers along a time line [92]. Newman et al. studied how to determine and assign lexical categories starting from source code [93].

Høst and Østvold investigated the implementation of methods from a corpus of Java applications, to determine which word is the best for a method naming [94]. In a subsequent work, they dealt with generation of *"a semantics which captures our common interpretation of method names"* [58]. They worked on *traceable pattern*s and proposed a set of *traceable attributes* that they claim can be useful building blocks of nano-patterns.

It is worth noting that the authors *explicitly* claim they were not proposing nano-patterns. The way vocabulary matches the structure or not can also lead to "naming bugs", a problem that can be mitigated by an automatic procedure. Høst and Østvold proposed automatic tool [95] to suggest proper names. Kashiwabara et al., in different works, presented techniques to identify candidate verbs for methods [96, 97]. There were some attempts to use recurring *structure* for software engineering purposes. Examples are works on *beacons* [98, 99]—stereotypical, recurring segments of code that are quickly recognized by experienced developers.

Structure: Some researchers in the area focus their efforts on investigating the recurring elements of the syntactical *structure* of programs. A concept similar to nano-patterns is that of *stereotype*. A stereotype is a syntactical structure that capture the intent of methods and classes and can be considered an extension of the micro-nano-pattern concept [100].

Andras et al. and Moreno et al. compared the outcome of the run-time with their stereotype [101], to investigate the consistency between method design and implementation [101, 102]. Qiu et al. showed that the use of syntacti-

cal rules in actual programs follows a Zipf-like distribution (just as happens with words in natural language). In other words, small sets of syntactic structures tend to govern entire projects [103]. Abd-El-Hafiz et al. classified loops by complexity levels [104]. Wang et al. introduced an automatic approach to determine high level action through the analysis of a given loop [105]. Mens et al. proposed to describe patterns with a declarative meta-programming language similar to Prolog.

7 Conclusion

We introduced the concept of pluggable controllers as a way to facilitate the introduction of new constructors, and explained how they can be implemented using Lola, the Language of Language Amendment, a powerful preprocessor and macro-language. Lola lets developers augment or even amend language constructors without affecting the language architecture.

This argument was illustrated by two cases study: A stenography for Java nanos and the introduction of Mathematica's commands in Java. Nanos are recurring idioms devised by developers to perform simple tasks. In a recent study they revealed to be prevalent in Java, under a given definition of prevalence. We presented 19 nanos that belong to a wider catalog [2] and illustrated the implementation of Java stenography based on them.

Their prevalence make them good candidates to become pluggable controllers. A stenography for nanos, such that illustrated in this work for Java—and its subsequent definition in standard libraries of pluggable controllers, is likely to improve developers efficiency by reducing coding effort.

In the present paper we reported a peculiar application of Lola. However, Lola's applications are not limited to the extension of SIC as pluggable controllers. For example, with Lola it is possible to compute the Halstead metrics [106], simulate Aspect Oriented Programming, add syntactic sugar, and many other kinds of language augmentation [15].

Relying on a powerful preprocessing engine, Lola makes it easy to introduce augmentation. Formal and precise proofs of semantics are however difficult: Since the semantics of the underlying language is not available to preprocessors, even the formulation of precise statements on semantics seems impossible. More so, with Lola which is language independent, and draws much of its power from the loose coupling with the underlying semantics and changes made to it.

7.1 Future Work on Lola

Future works will involve both further improvement of Lola and research on the application of Lola to help developers' work.

- Currently, Lola supports *trivia* (elements such as space, tabs, etc.), though it is limited to just a few of them such as `##NewLine` and `##EndOfFile`, and must be improved. This type of support may be needed to process comments, JavaDoc (for Java) and the like.
- Pattern matching features should be improved in order to treat string literals. The introduction of features that allow the host language to be changed on the fly would enable application of Lola's preprocessing to mixed code (i.e., Java code with an SQL queries as happens using JDBC, etc.)
- Another promising research avenue emerges if we ponder the possibility of applying Lola to itself.
- We would like to allow users to define new Lola keywords from within Lola itself (now only possible with simple Python classes).
- Some enhancement are required to improve usability. For example, using single and double # to distinguish Lola keywords from host language identifiers might lead to some confusion. We are working on a different solution, such as using another character (e.g., @) instead of #.

7.2 Future Work on Nanos

Further research it is needed to improve the prevalence values of nano-patterns by specifically looking for new candidates in other corpora. In this work, we used a lightweight nano tracer to track the nanos in source code. Additional work it is needed to improve search possibilities in several aspects (i.e., improving the namespace analysis, etc.). The subjective and sample bias in the harvest may have led us to miss some patterns. Further efforts are needed to verify this suspicion and to extend the basic catalog.

Conflict of Interest The authors declare no conflict of interest.

Acknowledgment The authors thank the anonymous reviewers for their valuable suggestions. Inspiring discussions with Tomer Levy are gratefully and intentionally acknowledged. This research was supported by THE ISRAEL SCIENCE FOUNDATION (grant No. 1803/13 *).

References

[1] Yossi Gil, Ori Marcovitch, and Matteo Orrù. Pluggable controllers and nano-patterns. In Martin Pinzger, Gabriele Bavota, and Andrian Marcus, editors, *IEEE 24th International Conference on Software Analysis, Evolution and Reengineering, SANER 2017, Klagenfurt, Austria, February 20-24, 2017*, pages 447–451. IEEE Computer Society, 2017.

[2] Yossi Gil, Ori Marcovitch, and Matteo Orrù. A Nano-Patterns Language for Java. Draft available at https://-goo.gl/tSte12, 2017.

[3] David A. Watt. *Programming Language Design Concepts*. John Wiley & Sons, 2004.

[4] Raphael A. Finkel. *Advanced Programming Language Design*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1995.

[5] Ravi Sethi. *Programming Languages: Concepts and Constructs*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1989.

[6] Donald E. Knuth. Structured programming with Go to statements. *ACM Comput. Surv.*, 6(4):261–301, December 1974. ACM, New York, NY, USA.

[7] H.D. Mills. Mathematical foundations for structured programming. Technical report, IBM rep. FSC 72-6012, IBM Fed. Syst. Div., Gaithersburg, Md., 1972.

[8] Niklaus Wirth. On the composition of well-structured programs. *ACM Comput. Surv.*, 6(4):247–259, December 1974. ACM,New York City, NY, USA.

[9] J. R. Donaldson. Structured programming. In Edward Nash Yourdon, editor, *Classics in Software Engineering*, pages 179–185. Yourdon Press, Upper Saddle River, NJ, USA, 1979.

[10] Corrado Böhm and Giuseppe Jacopini. Flow diagrams, turing machines and languages with only two formation rules. *Commun. ACM*, 9(5):366–371, may 1966. ACM,New York City, NY, USA.

[11] Bertrand Meyer. *EIFFEL the Language*. Object-Oriented Series. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1992.

[12] Walter Cazzola and Edoardo Vacchi. Language components for modular DSLs using traits. *Computer Languages, Systems & Structures*, 45:16–34, apr 2016. Springer Publishing, New York City, NY, USA.

[13] Nathanael Schrli, Stéphane Ducasse, and Oscar Nierstrasz. Traits: Composable units of behavior. In *(ECOOP'03)*, volume 2743, pages 248–274, 2003-07.

[14] Leonardo V.S. Reis, Vladimir O. Di Iorio, and Roberto S. Bigonha. An on-the-fly grammar modification mechanism for composing and defining extensible languages. *Computer Languages, Systems & Structures*, 42:46–59, jul 2015. Springer Publishing, New York City, NY, USA.

[15] Iddo E. Zmiry. Lola: a Programming Language for Augmenting Programming Languages. Master's thesis, Technion—Israel Institute of Technology, 2016. url: https://drive.google.com/file/d/0B3645jTHku6WZzVkMl9uVGlTQ2M/view.

[16] Dennis M. Ritchie. The evolution of the UNIX time-sharing system. In *Language Design and Programming Methodology*, pages 25–35. Springer, 1980.

[17] David Garlan and Mary Shaw. An introduction to software architecture. Technical report, Carnegie Mellon University, Pittsburgh, PA, USA, 1994.

[18] Joseph (Yossi) Gil and Itay Maman. Micro patterns in java code. In *Proceedings of the 20th Annual ACM SIGPLAN Conference on Object-oriented Programming, Systems, Languages, and Applications*, OOPSLA '05, pages 97–116, New York, NY, USA, 2005. ACM.

[19] Jeremy Singer, Gavin Brown, Mikel Luján, Adam Pocock, and Paraskevas Yiapanis. Fundamental nano-patterns to characterize and classify java methods. *Elect. Notes Theor. Comp. Sci.*, 253(7):191–204, September 2010. Springer Publishing, New York City, NY, USA.

[20] Feras Batarseh. Java nano patterns: A set of reusable objects. In *Proceedings of the 48th Annual Southeast Regional Conference*, ACM SE '10, pages 60:1–60:4, New York, NY, USA, 2010. ACM.

[21] J. E. Hirsch. An index to quantify an individual's scientific research output. *Proc. National Academy of Sciences of the USA of America*, 102(46):16569–16572, 2005. National Academy of Sciences.

[22] Leo Egghe and Ronald Rousseau. An infometric model for the Hirsch-index. *Scientometrics*, 69(1):121–129, 2006. Springer Publishing, New York City, NY, USA.

[23] Tal Cohen and Joseph Gil. Self-calibration of metrics of Java methods. In *TOOLS Pacific'00: 37th Int. Conf. Tech. OO Lang. & Syst.*, pages 94–107, Washington, DC, 2000. IEEE. IEEE,Washington, DC, USA.

[24] G. Concas, M. Marchesi, S. Pinna, and N. Serra. Power-laws in a large oo soft. syst. *IEEE*, 33(10):687–708, October 2007.

[25] G. Concas, M. Marchesi, A. Murgia, R. Tonelli, and I. Turnu. On the distribution of bugs in the eclipse syst. *IEEE*, 37(6):872–877, November 2011.

[26] I. Turnu, G. Concas, M. Marchesi, S. Pinna, and R. Tonelli. A modified yule process to model the evolution of some OO system properties. *Inf. Sc.*, 181(4):883–902, 2011. Springer Publishing, New York City, NY, USA.

[27] S. Alonso, F.J. Cabrerizo, E. Herrera-Viedma, and F. Herrera. h-index: A review focused in its variants, computation and standardization for different scientific fields. *Journal of Informetrics*, 3(4):273 – 289, 2009.

[28] Miltiadis Allamanis and Charles A. Sutton. Mining source code repositories at massive scale using language modeling. In Thomas Zimmermann, Massimiliano Di Penta, and Sunghun Kim, editors, *Proc. the 10th Working Conf. Mining Soft. Repositories, MSR '13, San Francisco, CA, USA, May 18-19, 2013*, pages 207–216, NY/USA, 2013. IEEE.

[29] Joseph (Yossi) Gil and Gal Lalouche. When do soft. complexity metrics mean nothing?—when examined out of context. *J. Object Tech.*, 15(1):2:1–25, 2016.

[30] Sebastian Erdweg, Tillmann Rendel, Christian Kästner, and Klaus Ostermann. Sugarj: Library-based syntactic language extensibility. *SIGPLAN Not.*, 46(10):391–406, oct 2011. ACM, New York, NY, USA.

[31] Matthew Flatt. Creating languages in Racket. *Communications of the ACM*, 55(1):48–56, January 2012. ACM,New York City, NY, USA.

[32] Daniel de Rauglaudre. Camlp4–reference manual. `http://caml.inria.fr/pub/docs/manual-camlp4/index.html`, 2003. Accessed: 2016-08-26.

[33] Tomaž Kosar, Pablo E. Martínez López, Pablo A. Barrientos, and Marjan Mernik. A preliminary study on various implementation approaches of domain-specific language. *Information and Software Technology*, 50(5):390–405, apr 2008.

[34] A. Naumchev and B. Meyer. Seamless requirements. *Computer Languages, Systems & Structures*, 49:119 – 132, 2017. Springer Publishing, New York City, NY, USA.

[35] J. Heering, P. R. H. Hendriks, P. Klint, and J. Rekers. The syntax definition formalism sdf—reference manual—. *SIGPLAN Not.*, 24(11):43–75, November 1989.

[36] Martin Bravenboer, Karl Trygve Kalleberg, Rob Vermaas, and Eelco Visser. Stratego/xt 0.17. a language and toolset for program transformation. *Science of Computer Programming*, 72(1):52 – 70, 2008. Special Issue on Second issue of experimental software and toolkits (EST).

[37] PYPL PopularitY of Programming Language. `http://pypl.github.io/PYPL.html`. Accessed: 2017-07-31.

[38] TIOBE Index for July 2017. `https://www.tiobe.com/tiobe-index/`. Accessed: 2017-07-31.

[39] James Gosling, Bill Joy, Guy Steele, and Gilad Bracha. *Java(TM) Language Specification, The (3rd Edition) (Java (Addison-Wesley))*. Addison-Wesley Professional, 2005.

[40] James Gosling, Bill Joy, Guy L. Steele, Jr., Gilad Bracha, and Alex Buckley. *The Java Language Specification, Java SE 7 Edition*. Addison-Wesley Professional, 1st edition, 2013.

[41] Robert Dyer, Hridesh Rajan, Hoan Anh Nguyen, and Tien N. Nguyen. Mining billions of ast nodes to study actual and potential usage of java language features. In *Proceedings of the 36th International Conference on Software Engineering*, ICSE 2014, pages 779–790, New York, NY, USA, 2014. ACM.

[42] Dong Qiu, Bixin Li, Earl T. Barr, and Zhendong Su. Understanding the syntactic rule usage in java. *Journal of Systems and Software*, 123:160 – 172, 2017.

[43] Dave Raggett, Arnaud Le Hors, and Ian Jacobs. HTML 4.01 specification. W3C Recommendation, W3C - World Wide Web Consortium, December 1999.

[44] George Reilly Brian Francis and Dino Esposito. *Professional Active Server Pages 3.0*. Wrox Press, 1999-09.

[45] Hans Bergsten. *JavaServer Pages*. O'Reilly Media, 3rd edition, 2003-12.

[46] Vassilios Karakoidas, Dimitris Mitropoulos, Panagiotis Louridas, and Diomidis Spinellis. A type-safe embedding of SQL into Java using the extensible compiler framework J%. *Computer Languages, Systems & Structures*, 41:1–20, apr 2015. Elsevier Science Publishers, Amsterdam, The Netherlands.

[47] GNU. C preprocessor. `https://gcc.gnu.org/onlinedocs/cpp/`.

[48] Dennis Ritchie. he Development of the C Language. In John A. N. Lee and Jean E. Sammet, editors, *HOPL Preprints*, pages 201–208. ACM, 1993.

[49] A. Snyder. A portable compiler for the language C. Technical report, Massachusetts Institute of Technology, Cambridge, MA, USA, 1975.

[50] GNU. The C preprocessor – overview. `https://gcc.gnu.org/onlinedocs/cpp/Overview.html#Overview`.

[51] GNU. M4 manual. `http://www.gnu.org/software/m4/manual/m4.html`.

[52] GNU. M4 manual – history section. `https://www.gnu.org/software/m4/manual/m4.html#History`.

[53] A. Dain Samples. *User's Guide to the M5 Macro Language*. Number UCB/CSD-91-621 in Report // Computer Science Division (EECS), UCB/CSD. University of California, Berkeley, Computer Science Division, 2nd edition, 1991.

[54] Robert F Tobler. Pym-a macro preprocessor based on python. In *Proceedings of the 9th International Python Conference, Long Beach, California*, 2001.

[55] Eugene Kohlbecker, Daniel P Friedman, Matthias Felleisen, and Bruce Duba. Hygienic macro expansion. In *Proceedings of the 1986 ACM conference on Lisp and functional programming*, pages 151–161, 1986. ACM, New York, NY, USA.

[56] Byeongcheol Lee, Robert Grimm, Martin Hirzel, and Kathryn S McKinley. Marco: Safe, expressive macros for any language. In *ECOOP 2012–Object-Oriented Programming*, pages 589–613. Springer, 2012.

[57] Joseph (Yossi) Gil and Itay Maman. Micro patterns in Java code. In Ralph Johnson and Richard P. Gabriel, editors, *Proc. the 20th Annual OO Prog., Syst., Lang., and App., OOPSLA 2005*, volume 40, pages 97–116, NY/USA, October 2005. ACM.

[58] Einar W. Høst and Bjarte M. Østvold. The java prog.'s phrase book. In Dragan Gasevic, Ralf Lämmel, and Eric Van Wyk, editors, *Soft. Language Eng., First Int. Conf., SLE 2008, Toulouse, France, Sep 29-30, 2008. Revised Selected Papers*, volume 5452 of *Lecture Notes in Comp. Science*, pages 322–341, NY/USA, 2008. Springer.

[59] A.K. Deo and B.J. Williams. Preliminary study on assessing software defects using nano-pattern detection. In *24th International Conference on Software Engineering and Data Engineering, SEDE 2015*, 2015.

[60] K. Z. Sultana, A. Deo, and B. J. Williams. A preliminary study examining relationships between nano-patterns and software security vulnerabilities. In *2016 IEEE 40th Annual Computer Software and Applications Conference (COMPSAC)*, volume 1, pages 257–262, June 2016. IEEE,Washington, DC, USA.

[61] Illo Lee, Suntae Kim, Sooyong Park, and Younghwa Cho. *Attributes for Characterizing Java Methods*, pages 185–191. Springer, Berlin, Heidelberg, 2016.

[62] Erich Gamma, Richard Helm, Ralph Johnson, and John Vlissides. *Design Patterns: Elements of Reusable Object-Oriented Software*. Professional Computing series. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1995.

[63] Erich Gamma. Going beyond objects with design patterns. In *ECOOP'97—OO Prog.: 11th European Conf. Jyväskylä, Finland, Jun 9–13, 1997 Proc.*, pages 530–530, Berlin, Heidelberg, 1997. Springer.

[64] Peter Van Emde Boas. Resistance is Futile; formal linguistic observations on design patterns. techreport ILLC-CT-1997-03, The Institute For Logic, Language, and Computation (ILLC), University of Amsterdam, 1997-02.

[65] Amnon H. Eden, Joseph (Yossi) Gil, Yoram Hirshfeld, and Amiram Yehudai. Motifs in object oriented architecture, 1999.

[66] Amnon H. Eden, Joseph (Yossi) Gil, and Amiram Yehudai. A formal language for design patterns. In *(PLoP'96)*, 1996-09.

[67] Amnon H. Eden. Formal specification of object-oriented design. In *Int. Conf. on Multidisciplinary Design in Engineering (CSME-MDE'01)*, pages 21–22, 2001-11.

[68] Amnon H. Eden. A visual formalism for object-oriented architecture. In *(IDPT'02)*, 2002-06.

[69] Amnon H. Eden. Giving "the quality" a name - precise specification of design patterns - a second look at the manuscripts. *Journal of Object-Oriented Programming*, 11(3), 1998-06.

[70] Rudolf Ferenc, Árpád Beszédes, Lajos Jeno Fülöp, and Janos Lele. Design pattern mining enhanced by machine learning. In *21st "IEEE" 2005), 25-30 Sep 2005, Budapest, Hungary*, pages 295–304, NY/USA, 2005. IEEE.

[71] Yann-Gaël Guéhéneuc, Jean-Yves Guyomarc'h, and Houari Sahraoui. Improving design-pattern identification: a new approach & an exploratory study. *Soft. Quality J.*, 18(1):145–174, 2010. Kluwer Academic Publishers, Hingham, MA, USA.

[72] Marco Zanoni, Francesca Arcelli Fontana, and Fabio Stella. On applying machine learning techniques for design pattern detection. *J. Syst. and Soft.*, 103:102–117, 2015. Springer Publishing, New York City, NY, USA.

[73] Yann-Gaël Guéhéneuc, Houari A. Sahraoui, and Farouk Zaidi. Fingerprinting design patterns. In *11th Working Conf. Rev. Eng., WCRE 2004, Delft, The Netherlands, Nov 8-12, 2004*, pages 172–181, NY/USA, 2004. IEEE.

[74] Marek Vokác. An efficient tool for recovering design patterns from c++ code. *Journal of Object Technology*, 5(1):139–157, 2006.

[75] Linzhang Wang, Zhixiong Han, Jiantao He, Hanfei Wang, and Xuandong Li. Recovering design patterns to support program comprehension. In *Proceedings of the 2Nd International Workshop on Evidential Assessment of Software Technologies*, EAST '12, pages 49–54, New York, NY, USA, 2012. ACM.

[76] Miltiadis Allamanis and Charles A. Sutton. Mining idioms from source code. In Shing-Chi Cheung, Alessandro Orso, and Margaret-Anne D. Storey, editors, *Proc. the 22nd "ACM" Found. Soft. Eng., (FSE-22)*, pages 472–483, Hong Kong, November 2014. ACM.

[77] Wikibooks. More C++ Idioms, 2017. https://en.wikibooks.org/wiki/More_C%2B%2B_Idioms.

[78] Java Idioms Editor. Java Idioms, 2017. http://c2.com/ppr/wiki/JavaIdioms/JavaIdioms.html.

[79] S. Chuan. JavaScript Patterns Collection, 2014. http://shichuan.github.io/javascript-patterns/.

[80] R. Waldron. Principles of writing consistent, idiomatic JavaScript, 2014. https://github.com/rwaldron/idiomatic.js/.

[81] E. Recommenders-Contributors. Eclipse SnipMatch, 2014.

[82] JetBrains. High-speed coding with Custom Live Templates., 2014. http://bit.ly/1o8R8Do.

[83] Nathan Gurewich and Ori Gurewich. *Java Manual of Style*. Ziff-Davis Publishing Co., April 1996.

[84] Chris Laffra. *Advanced Java: Idioms, Pitfalls, Styles & Prog. Tips*. Prentice Hall Ptr, September 1996.

[85] Craig Larman and Rhett Guthrie. *Java 2 Performance & Idiom Guide*. Prentice Hall Ptr, NJ/USA, 1999.

[86] A. Sutton, R. Holeman, and J. I. Maletic. Identification of idiom usage in c++ generic libraries. In *2010 IEEE 18th Int. Conf. Program Comprehension*, pages 160–169, New York,NY, June 2010. IEEE.

[87] Andrew Koenig. Idiomatic design. *Comm. ACM*, pages 14–19, 1995. ACM, New York, NY, USA.

[88] A. Langer. Java programming idioms. In *Proc. OO Lang. & Syst.. TOOLS 38*, pages 197–198, 2001. IEEE,Washington, DC, USA.

[89] L. M. Wills. Automated program recognition: A feasibility demonstration. *Artif. Intell.*, 45(1-2):113–171, September 1990. Elsevier Science Publishers Ltd. Essex, UK.

[90] Erik Linstead, Lindsey Hughes, Cristina Lopes, and Pierre Baldi. Exploring Java software vocabulary: A search & mining perspective. In *Proc. Int. Conf. Soft. Eng.*, pages 29–32, 2009. IEEE,Washington, DC, USA.

[91] E. Enslen, E. Hill, L. Pollock, and K. Vijay-Shanker. Mining source code to automatically split identifiers for software analysis. In *2009 6th IEEE International Working Conference on Mining Software Repositories*, pages 71–80, Washington, DC, USA, May 2009. IEEE.

[92] S. L. Abebe, S. Haiduc, A. Marcus, P. Tonella, and G. Antoniol. Analyzing the evolution of the source code vocabulary. In *2009 13th European Conference on Software Maintenance and Reengineering*, pages 189–198, Washington, DC, USA, March 2009. IEEE.

[93] C. D. Newman, R. S. AlSuhaibani, M. L. Collard, and J. I. Maletic. Lexical categories for source code identifiers. In *2017 IEEE 24th International Conference on Software Analysis, Evolution and Reengineering (SANER)*, pages 228–239, Feb 2017. IEEE,Washington, DC, USA.

[94] E. W. Host and B. M. Ostvold. The programmer's lexicon, volume i: The verbs. In *Seventh IEEE International Working Conference on Source Code Analysis and Manipulation (SCAM 2007)*, pages 193–202, Sept 2007. IEEE, Washington, DC, USA.

[95] Einar W. Høst and Bjarte M. Østvold. Debugging method names. In *Proceedings of the 23rd European Conference on ECOOP 2009 — Object-Oriented Programming*, Genoa, pages 294–317, Berlin, Heidelberg, 2009. Springer-Verlag.

[96] Y. Kashiwabara, Y. Onizuka, T. Ishio, Y. Hayase, T. Yamamoto, and K. Inoue. Recommending verbs for rename method using association rule mining. In *2014 Software Evolution Week - IEEE Conference on Software Maintenance, Reengineering, and Reverse Engineering (CSMR-WCRE)*, pages 323–327, Feb 2014. IEEE,Washington, DC, USA.

[97] Yuki Kashiwabara, Takashi Ishio, Hideaki Hata, and Katsuro Inoue. Method verb recommendation using association rule mining in a set of existing projects. *IEICE Transactions on Information and Systems*, E98.D(3):627–636, 2015.

[98] Ruven Brooks. Towards a theory of the comprehension of computer programs. *Int. J. Man-Machine Studies*, 18(6):543–554, 1983. Elsevier Science Publishers, Amsterdam, The Netherlands.

[99] Martha E. Crosby, Jean Scholtz, and Susan Wiedenbeck. The roles beacons play in comprehension for novice & expert programmers. In J. Kuljis, L. Baldwin, and R. Scoble, editors, *Proc. 14th Ann. Workshop of the Psychology of Programmers Interest Group (PPIG)*, pages 58–73, 2002.

[100] N. Dragan, M. L. Collard, and J. I. Maletic. Reverse engineering method stereotypes. In *2006 22nd IEEE International Conference on Software Maintenance*, pages 24–34, Washington, DC, USA, Sept 2006. IEEE.

[101] P. Andras, A. Pakhira, L. Moreno, and A. Marcus. A measure to assess the behavior of method stereotypes in OO software. In *2013 4th International Workshop on Emerging Trends in Software Metrics (WETSoM)*, pages 7–13, Washington, DC, USA, May 2013. IEEE.

[102] L. Moreno and A. Marcus. Jstereocode: automatically identifying method and class stereotypes in java code. In *2012 Proceedings of the 27th IEEE/ACM International Conference on Automated Software Engineering*, pages 358–361, New York, NY, Sept 2012. IEEE.

[103] Dong Qiu, Bixin Li, Earl T. Barr, and Zhendong Su. Understanding the syntactic rule usage in Java. *J. Syst. and Soft.*, 123:160–172, 2017. Springer Publishing, New York City, NY, USA.

[104] S. K. Abd-El-Hafiz and V. R. Basili. A knowledge-based approach to the analysis of loops. *IEEE*, 22(5):339–360, May 1996.

[105] X. Wang, L. Pollock, and K. Vijay-Shanker. Developing a model of loop actions by mining loop characteristics from a large code corpus. In *2015 IEEE 31st Int. Conf. Soft. Maint. & Evolution, ICSME 2015 - Proc.*, 2015. IEEE,Washington, DC, USA.

[106] Maurice H. Halstead. *Elements of Software Science*. Operating and Programming Systems. Elsevier Science Inc., New York, NY, USA, 1977.

ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 2, No. 3, 1727-1740 (2017)
www.astesj.com
Special Issue on Recent Advances in Engineering Systems

ASTES Journal
ISSN: 2415-6698

On Modeling Affect in Audio with Non-Linear Symbolic Dynamics

Pauline Mouawad[*,1], Shlomo Dubnov[2]

[1]*Laboratoire Bordelais de Recherche en Informatique, University of Bordeaux, 33400, France*

[2]*Center for Research in Entertainment and Learning, University of California San Diego, CA 92093, USA*

ARTICLEINFO

Article history:
Received: 31 May, 2017
Accepted: 01 August, 2017
Online: 09 September, 2017

Keywords :
Nonlinear dynamics
Recurrence plots
Symbolization
Dynamical invariants
Variable Markov Oracle
Information rate
Embedding

ABSTRACT

The discovery of semantic information from complex signals is a task concerned with connecting humans' perceptions and/or intentions with the signals content. In the case of audio signals, complex perceptions are appraised in a listener's mind, that trigger affective responses that may be relevant for well-being and survival. In this paper we are interested in the broader question of relations between uncertainty in data as measured using various information criteria and emotions, and we propose a novel method that combines nonlinear dynamics analysis with a method of adaptive time series symbolization that finds the meaningful audio structure in terms of symbolized recurrence properties. In a first phase we obtain symbolic recurrence quantification measures from symbolic recurrence plots, without the need to reconstruct the phase space with embedding. Then we estimate symbolic dynamical invariants from symbolized time series, after embedding. The invariants are: correlation dimension, correlation entropy and Lyapunov exponent. Through their application for the logistic map, we show that our measures are in agreement with known methods from literature. We further show that one symbolic recurrence measure, namely the symbolic Shannon entropy, correlates positively with the positive Lyapunov exponents. Finally we evaluate the performance of our measures in emotion recognition through the implementation of classification tasks for different types of audio signals, and show that in some cases, they perform better than state-of-the-art methods that rely on low-level acoustic features.

1 Introduction

This paper is an extension of work originally presented in IEEE 11th ICSC [1].

The task of capturing emotional meaning from audio structure while disregarding trivial or irrelevant information is a complex process that cannot be inferred using low-level acoustics. Recent advances in research on sound dynamics have shown that nonlinear phenomena exist in complex audio signals [2, 3, 4, 5, 6]. Such complex information is shaped in the nonlinear dynamical structure of audio content that is brought together by repeating patterns evolving in a temporal order. Nonlinear dynamics analysis consists of a set of methods that unravel these fine-grained patterns, and study their role in conveying meaningful information[7]. Particularly, it was shown that the strange attractors of the sounds' dynamical systems carry perceptual meaning [8]. As a consequence recent studies were successful at applying methods of nonlinear dynamics to capture voice pathologies in speech signals [9, 3, 4, 10], recognize environmental sounds [11, 12] and discriminate between different singing styles [2]. Despite these advances, very few researches have applied nonlinear dynamics for modeling emotion in audio signals. In [6], measures of the geometrical properties of the phase space reconstruction are employed to recognize affect in speech; in [13], recurrence properties of the vowel accurately describe the dynamic behaviour of six basic emotions.

In our former work [1], we proposed a novel learning framework of affective auditory scene analysis us-

[*]Pauline Mouawad, 36 Rue Marc Sangnier, 33400 Talence, France, pauline.mouawad@u-bordeaux.fr

https://dx.doi.org/10.25046/aj0203212

ing a recently developed method of non-linear dynamic signal analysis, the Variable Markov Oracle (VMO), that finds the best audio structure representation in terms of the symbolized recurring patterns, while preserving their temporal order. We obtained symbolic recurrence quantification analysis (RQA) measures without reconstructing the phase space with embedding. Our contribution over previous recurrence analysis methods is that our model explored the dynamics of the most informative recurrent patterns in the signal, by means of symbolic RQA features. We showed that measures of periodicity and complexity derived from our model are relevant for the characterization of affect in auditory scenes, and that they perform better than state-of-the-art methods relying on low-level acoustic features.

In this paper we contribute to the ongoing research on affective semantics from sound by proposing a new set of symbolic dynamical invariants computed from the VMO. After the symbolization step, we propose a novel method of computing symbolic dynamical invariants from the symbolized sequences of the time series, after performing an embedding step. The invariants are: the correlation dimension (D2), correlation entropy (K2) and the Lyapunov Exponent (LE). In literature, the correlation dimension and Lyapunov Exponent have been successful in discriminating voice quality as well as in characterizing pathologies in voice [14, 9]. Furthermore, D2 and K2 are effective in the detection of emotion in speech [3, 4]. In the music domain, LE and D2 were used to characterize the clarinet tone [15], however rare are the literatures that investigate the potential of dynamical invariants in characterizing emotion in music.

The contribution of this paper is a novel method of nonlinear dynamics that derives symbolic dynamical measures from an adaptive time series symbolization method (VMO), with and without phase space reconstruction. First, it derives symbolic RQA measures from a symbolization of the signal's feature frames without embedding [1]. Second, it estimates symbolic dynamical invariants from the symbolized time series after embedding. Then we estimate our symbolic dynamical measures for the logistic map and show that they are in agreement with known methods from literature. Finally we test the performance of our symbolic complexity measures in predicting emotion by performing classification tasks on four types of sound stimuli. The advantage of our symbolic complexity measures is that first they quantify the dynamics of the most meaningful recurring patterns in the signal; second, their number is determined and hence we don't have to address the problem of the dimensionality of the dataset, therefore no feature selection methods are employed; third we show that they are efficient in recognizing emotions across three types of stimuli, which suggests that their performance does not depend on the type of sound under study.

2 Theoretical Background

2.1 Nonlinear Time Series Analysis

Nonlinear time series analysis (NLTSA) consists of a set of methods that characterize dynamical information from time-ordered values in a dataset. It is based on the fact that the real underlying dynamical state of a complex system is often unknown, and that all the information needed to determine the future behaviour of the system's state is independent of its past, and can be predicted based on knowledge of the present state, which is the observable measured by the time series.

In order to learn about the underlying dynamics of time-ordered data such as audio signals, it is necessary to reconstruct the phase space.

Phase Space Reconstruction The states of dynamical systems change in time, and their time evolution is defined geometrically in the shape of trajectories that belong to a *phase space* known as *strange attractor*.

In practice, we do not have a full knowledge of the dynamical system in order to reconstruct its phase space. But we do have a time-discrete measurement of one observable, which results in a scalar and discrete time series, that is used to reconstruct the original system's dynamics, through the reconstruction of its phase space via embedding. The embedding theorems guarantee that for noise-free data, there is a dimension m such that the embedded vectors are equivalent to the original phase space vectors [16].

To reconstruct the phase space of a system from a time series, the Takens' embedding theorem is used [17] and the framework is the following [18]:

Let $x(t)$ be a trajectory of a dynamical system and $s(t) = s(x(t))$ the result of a scalar measurement on it. Then a delay reconstruction with time delay τ and embedding dimension m is given by:

$$s(t) = (s(t-(m-1)\tau), s(t-(m-2)\tau), ..., s(t)) \quad (1)$$

Embedding parameters One of the main challenges of the delay-coordinate embedding theorem, is choosing appropriate values of dimension m and time lag τ [16].

Several methods exist that derive m and τ but we are naming the most widely used ones in literature. First τ is estimated: if τ is very small, consecutive elements of the delay vectors will highly correspond, and all the vectors will be clustered around the main diagonal, unless m is very large. If τ is very large, consecutive elements are independent, and the points will fill a large space in the phase space [16]. Two functions can be used to determine τ: the first zero of the autocorrelation function of the time series and the first minimum of the mutual information function (FMMI). In this work we use the FMMI.

Once τ is chosen, the next step is to estimate the embedding dimension m. If m is too large, the embedded data will be redundant, which will confuse the performance of prediction algorithms. Two widely known methods can be used: the false-nearest neighbour algorithm (FNN) [19] and the 'asymptotic invariant approach'. The FNN method is used in this work since it is the most widely used one, and m is chosen where the number of false neighbours drops to zero.

The resulting is equivalent to the attractor in the original phase space if $m \leq 2d+1$, were d is the dimension of the original phase space. In general we don't know what the value of d is, but using the FNN method, m is guaranteed to fulfil that requirement.

2.2 Recurrence Plots

Recurrence is a fundamental property of most dynamical systems. In fact it is due to the systems' recurrence to former states that we know how to predict the future state of the system. Recurrence takes place in the system's phase space, and the tool that measures it is called a recurrence plot (RP) [20].

Given a trajectory $\vec{x_i} \in \mathbb{R}^d$ in a d-dimensional phase space of a dynamical system, the RP is a two-dimensional visualization of the square recurrence matrix of the embedded time series defined by:

$$R_{i,j}^{m,\varepsilon} = \Theta(\varepsilon - \left\|\vec{x_i} - \vec{x_j}\right\|), \quad i,j = 1,...,N \tag{2}$$

where $\vec{x_i}$ and $\vec{x_j}$ are phase space trajectories in an m-dimensional phase space, N is the number of measured points in a trajectory, ε is a threshold distance, $\Theta(.)$ the Heaviside function such that: $\Theta(x) = 0$ if $x < 0$ and $\Theta(x) = 1$ otherwise, and $\|.\|$ is some appropriate choice of a norm, such as the L_2-norm, otherwise known as the Euclidean distance. Both axes of the RP are time axes. The dots or pixels located at (i,j) and (j,i) on the RP are black if the distance between points x_i and x_j in the phase space fall inside a *ball* or *threshold corridor* of radius ε, the threshold distance [21, 22]. In this case, the black points refer to *recurring* states also termed ε-recurrent states since they occur in an ε-neighbourhood. The ε-recurrent states are represented by the relation[20]:

$$\vec{x_i} \approx \vec{x_j} \iff R_{i,j} \equiv 1. \tag{3}$$

The dots are white if $R_{i,j} \equiv 0$. The RP always displays a main black diagonal line called the line of identity (LOI), since $R_{i,i} \equiv 1$ by definition. For more in-depth description of the RP properties, the reader is referred to [20].

2.3 Recurrence Quantification Analysis

In order to derive meaning from the structures of the RP, various complexity measures are computed that quantify those structures. Such quantification is important since it will be employed to characterize the dynamical information and to perform predictions. These statistical measures are known as Recurrence Quantification Analysis (RQA) and are based on the density of recurrence points, the diagonal and vertical line structures in the RP [21, 23, 24]. RQA can be applied to non-stationary processes in continuous or discrete time series. For example, the metric *determinism* can discriminate signals from noise, and is valuable in pattern mining and classification tasks.

2.3.1 Measures based on the density of recurrence points

Given an RP thresholded at ε (Eq. 2), the Recurrence Rate (RR) measures the density of recurrence points in the RP:

$$RR = \frac{1}{N^2} \sum_{i,j=1}^{N} R_{i,j} \tag{4}$$

The RR measure corresponds to the correlation sum (D2) measure, but D2 excludes the main diagonal line (LOI):

$$D2 = \frac{1}{N(N-1)} \sum_{\substack{i,j=1 \\ j \neq i}}^{N} R_{i,j} \tag{5}$$

2.3.2 Diagonal lines based measures

Given the histogram $P(l)$ of diagonal lines of length l, the following measures are computed:

Determinism (DET) is the percentage of points in diagonal line of at least length $l = l_{min}$, i.e. the ratio of recurrence points in the diagonals to all recurrence points, and is a measure of the predictability of the system. Processes with chaotic behaviour cause none or very short diagonals. Deterministic processes cause longer diagonals and less isolated recurrence points.

$$DET = \frac{\sum_{l=l_{min}}^{N} lP(l)}{\sum_{l=1}^{N} lP(l)} \tag{6}$$

The average length of diagonal line length L is the average time during which two segments of a trajectory are close to each other, and it refers to the mean prediction time. The length l of diagonal lines refer to the number l of time steps during which a segment of the trajectory is close to another segment of the trajectory at a different time. Therefore the diagonal lines are related to the divergence of the trajectory segments.

$$L = \frac{\sum_{l=l_{min}}^{N} lP(l)}{\sum_{l=l_{min}}^{N} P(l)} \tag{7}$$

Then the length L_{max} of the longest diagonal line in the RP excluding LOI is derived:

$$L_{max} = max(\{l_i\}_{i=1}^{N_l}) \tag{8}$$

And the inverse of L_{max} indicates the divergence (DIV) of the phase space trajectory. The faster the

trajectory segments diverge, the diagonal lines will be shorter, and the value of DIV will be higher:

$$DIV = \frac{1}{L_{max}} \tag{9}$$

The next measure is the Shannon entropy of diagonal line length distribution in the RP (S_{RP}), which is the probability $p(l) = P(l)/N_l$ to find a diagonal line of exactly length l in the RP. It is a measure of complexity in the RP in terms of the diagonal lines, such that, for uncorrelated noise the value of S_{RP} will be small, which indicates a low complexity.

It is defined as:

$$S_{RP} = -\sum_{l=l_{min}}^{N} p(l)\ ln\ p(l) \tag{10}$$

The $RATIO$ is a measure that uncovers transitions in the system's dynamics:

$$RATIO = \frac{DET}{RR} \tag{11}$$

2.3.3 Vertical lines based measures

Measures based on vertical structures in the RP uncover chaos-chaos transitions [20] in a dynamical system that are not found using diagonal line based measures. These are *laminarity* and *trapping time*.

The laminarity (LAM) refers to the occurrence of laminar states in the system independently of their lengths. If the RP contains less vertical lines and more single recurrence points, then the value of LAM will be low. Its definition is analogous to the definition of DET for vertical lines of minimal length $v = v_{min}$.

$$LAM = \frac{\sum_{v=v_{min}}^{N} vP(v)}{\sum_{v=1}^{N} vP(v)} \tag{12}$$

The trapping time measure (TT) is the average length of vertical lines, and estimates the mean time that the system's state will be trapeed:

$$TT = \frac{\sum_{v=v_{min}}^{N} vP(v)}{\sum_{v=v_{min}}^{N} P(v)} \tag{13}$$

3 Our Approach

One key concern when using RPs is finding the threshold to make sure that the RP exhibits enough recurrence points. Another difficulty to address is the length of the sequence used to generate the RP. This is considered as a second embedding step that is different from the phase space embedding, however in traditional RP construction methods these two steps are indistinguishable, as the RP is constructed first and then the recurring patterns are found by looking for diagonal lines.

In this work we propose a novel method that does not require a phase space reconstruction with embedding. This is done using the Variable Markov Oracle (VMO) [25], a suffix automaton that reduces a multivariate time series down to a symbolic sequence while retaining the recurring sub-sequences. Accordingly, we consider recurrences of symbolic sequences without a need to estimate a threshold, since this step is implicitly done during the symbolization, based on a mutual information criterion that estimates the optimal threshold in terms of maximizing Information Rate (IR) [26]. IR considers the mutual information between past and present in a signal. In the next section we describe this approach.

In a first phase, we estimate symbolic RQA from the symbolic RPs generated from the VMO without embedding. In a second phase, we estimate symbolic dynamical invariants from the VMO generated symbolic recurrences, after applying an embedding.

3.1 The Variable Markov Model

The Variable Markov Oracle (VMO) [25] is a suffix tree data structure that is derived from Factor Oracle (FO) [27, 28] as well as Audio Oracle (AO) [29].

FO is a suffix automaton that finds factors (*repeated substrings*) in a word (or sequence of symbols), as well as patterns (*repeated suffixes*) [27]. It has been employed mainly for optimal string matching algorithms, such as biosequence pattern matching. Assayag et al. 2004 showed how the FO can be adapted to learn symbolic musical sequences and generate symbolic musical improvisations in real-time [28].

AO is an extension of FO for audio signals, that is independent of the audio feature representation. AO extends the applications of FO to multivariate time series such as an audio signal sampled at discrete times. Based on a distance measure, the AO structure finds and links all the possible combinations of audio subclips that are similar. AO has been successfully applied to audio generation.

3.1.1 VMO Construction

VMO inherits the strengths of both FO and AO. The important improvement over its predecessors, is that VMO assigns symbols to the signal frames connected by suffix links during AO construction: it accepts a signal O as input, outputs the oracle structure, and keeps track of the sequence of assigned labels $Q = q_1,...,q_N$ as well as a list of pointers to their corresponding observations $O = O[1],...,O[N]$. As such VMO performs a symbolization of a signal's time series by storing the information regarding the repeated substrings via the suffix links created during AO construction and upgrades AO by assigning labels to the frames connected by suffix links.

The notations of the forward and suffix links remain the same as in FO construction. The detailed algorithm is found in [30, 25].

As mentioned earlier, a similarity threshold θ is introduced to determine if a signal sample $O[i]$ is similar to another sample $O[j]$.

In order to find the best symbolization of the signal, different VMO models can be created with different θ values. There is a tradeoff to consider when choosing θ values. If θ is very low, every frame will be different than every other frame, and VMO assigns a different symbol to each frame in O. If θ is very high, frames that are different are considered similar, and the same symbol is assigned to every frame in O. In both cases no structure in the time series can be captured by VMO.

Hence θ should be determined before VMO construction. Dubnov et al. have shown that the value of θ can be resolved by computing the Information Rate (IR) over candidate θ values [31]. The optimal θ value is the one that yields a highest IR value.

Information Rate IR is an information theoretic metric that measures the information content of a time series.

Let $x_1^N = x_1, x_2, ..., x_N$ be a time series x with N observations, where $H(x) = -\sum P(x) log_2 P(x)$ is the entropy of x, then the definition of IR is [32]:

$$IR(x_1^{n-1}, x_n) = H(x_n) - H(x_n|x_1^{n-1}) \tag{14}$$

And it is approximated by replacing the entropy terms in equation 14 by a complexity measure C associated with a compression algorithm [32]. The complexity measure is the number of bits used to compress x_n independently using the past observations x_1^{n-1}:

$$IR(x_1^{n-1}, x_n) \approx C(x_n) - C(x_n|x_1^{n-1}) \tag{15}$$

Compror is a lossless compression algorithm based on FO and the length of the longest repeated suffix link (**lrs**). Details on *Compror* as well as on the method of combining Compror with AO and IR are found in [33]

IR is the mutual information between past and present observation in a signal $O[t]$ and is maximized when there is balance between variation and repetition in the symbolized signal. This means that a VMO with a higher IR value captures more of the repeating patterns than a VMO with a lower IR value [32]. Figure 1 shows two oracle structures obtained with two extreme θ values. Figure 2 visualizes the sum of IR values versus different values of θ.

3.2 Symbolic Recurrence Plots from VMO

From the generated VMO-symbolized time series, we obtain the symbolic RP (RP_S hereafter), plotted from the binary self-similarity matrix. The index of a suffix link is a point on the RP_S and a repeated sequence is detected as a line since it includes repetitions of length 1, 2, up to the longest repeated length. This makes VMO effectively find a repetition for variable length non-uniform embedding.

We redefine the symbolic RP_S obtained from the optimal VMO model of the signal's time series:

$$R_{i,j}^{\sigma_M,\theta} = \Theta(\theta - d(\sigma_{q_i}, \sigma_{q_j})) \quad i,j = 1,...,N \tag{16}$$

Such that:

$$R_{i,j}^{\sigma_M,\theta} = \begin{cases} 1 & if\ d(\sigma_{q_i}, \sigma_{q_j}) is \leq \theta \\ 0 & otherwise \end{cases} \tag{17}$$

Where N is the number of states considered, σ_M refers to the Mth symbolized substring, Θ is the Heaviside step function (i.e. $\Theta(x) = 0$ if $x < 0$, and $\Theta(x) = 1$ otherwise). θ is a threshold distance, and $d(\sigma_{q_i}, \sigma_{q_j})$ is a distance metric between pairs of symbolized substrings q_i at $t = i$ and q_j at $t = j$.

3.3 Feature Extraction

In our experiments, we derive two sets of complexity features.

Symbolic RQA measures In the first phase we estimate symbolic RQA measures (RQA_S hereafter). Standard ways to consider similarity in audio signals is through time-frequency representation. In a preprocessing stage, the time series is transformed into a constant-Q transform (CQT) feature vector. CQT is a logarithmic spacing of filter center frequencies versus bandwidths, that represents the audio signal in a form that approximates human auditory analysis. Then the CQT feature vector is passed as input to the VMO constuction algorithm, that generates several symbolizations of the features in terms of their recurrence properties. by means of IR, the optimal threshold θ is evaluated to obtain the optimal VMO symbolization model $\mathcal{M}_S$. Then the symbolic RP_S is generated from the self-similarity matrix obtained from the longest repeated substrings (LRS) of $\mathcal{M}_S$. Then RQA_S estimates are obtained. Although standard RQA metrics are not invariants, in our case, since their estimation is independent of embedding, our RQA_S can be considered as invariants.

Symbolic Dynamical Invariants In the second phase we estimate symbolic dynamical invariants: the correlation dimension (D2), correlation entropy (K2) and the Lyapunov Exponent (LE).

Given a one-dimensional time series obtained from an audio signal, first we embed it using Takens' time-delay embedding method. The dimension m is determined by the false-nearest neighbour algorithm, and the value is chosen where the false nearest neighbours are zero. The value of the time delay τ is defined by the first minimum of the mutual information function. Next, we symbolize it with VMO, and select $\mathcal{M}_S$ by means of IR. Then from the selected VMO model we obtain a representation of the LRS found in the series. From this representation we proceed to extract D2, K2 and LE. The framework is depicted in figure 3.

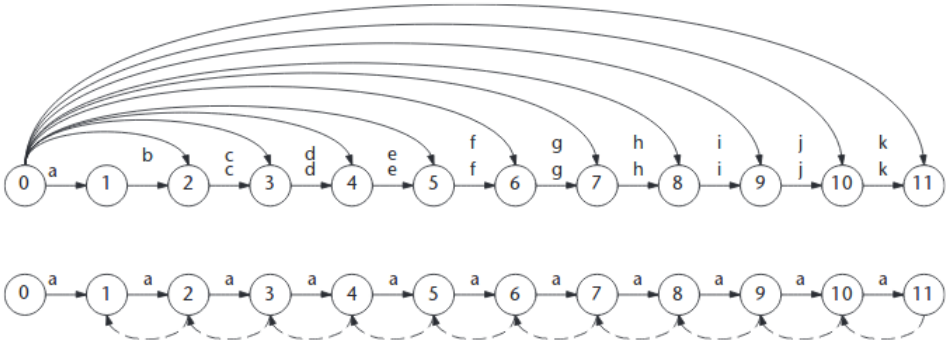


Figure 1: Two oracle structures. The top oracle has a very low θ value. The bottom oracle has a very high θ value [30]

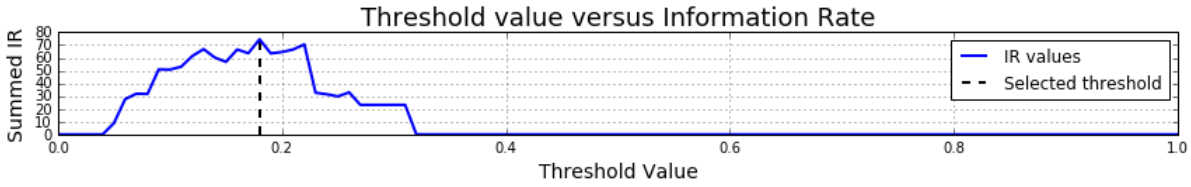


Figure 2: IR values on vertical axis and on horizontal axis. The solid curve in blue shows the relations between the two measures and the dashed black line indicates the selected θ by locating the maximal IR value. [32]

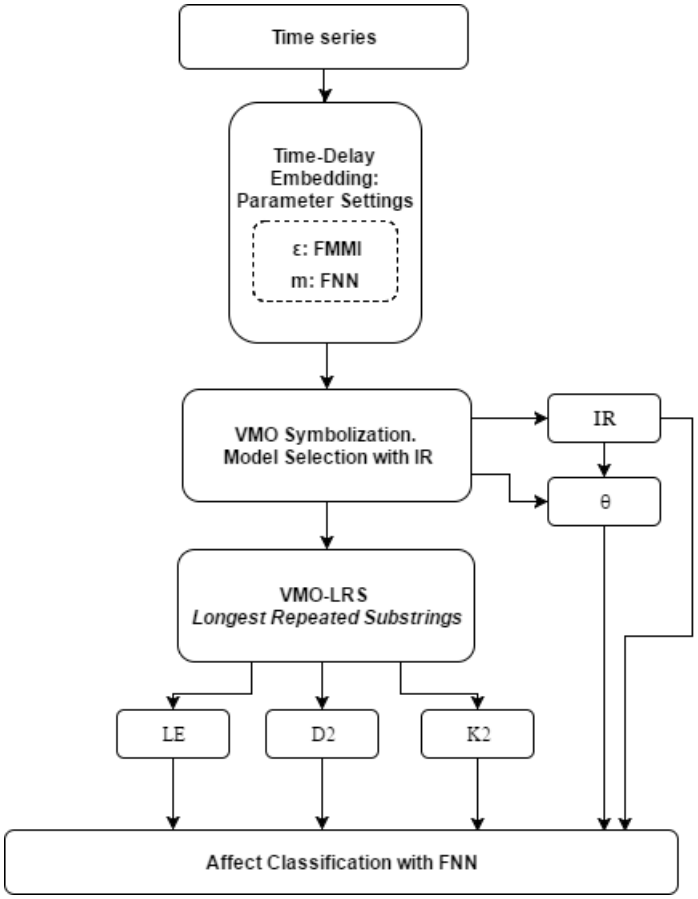


Figure 3: Dynamical Invariants extraction framework

Normally, in order to derive the LE using Rosenstein's algorithm or Eckmann's algorithm, either algorithm operates directly on the time series after embedding, and then computes the LE. In our approach we obtain our symbolic invariants from the LRS of the optimal VMO model $\mathcal{M}_S$. This is a novel aspect where we obtain symbolic dynamical invariants that describe the dynamical behaviour of only the most meaningful recurrences found in the series.

Before employing the dynamical invariants in emotion prediction tasks, we first probe to what extent our estimates are in agreement with known methods by illustrating their application to the logistic map. Then we question their role in discriminating emotion in voice, auditory scenes, instrumental music as well as in film music.

Correlation Dimension The correlation dimension (D2) is a geometric measure that tells how complex are the dynamics of the system: a more complex system has a higher dimension, which in is estimated from the symbolic recurrence plot. D2 estimates the complexity of the dynamics: a higher D2 indicates a more complex system's dynamics. $D2_S$ is computed from the symbolic RP_S by the correlation sum [20]:

$$D2_S = \frac{1}{N(N-1)} \sum_{\substack{i,j=1 \\ j \neq i}}^{N} \Theta(\theta - d(\sigma_{q_i}, \sigma_{q_j})) \qquad (18)$$

Correlation Entropy The correlation entropy (K2) also known as 2nd order Rényi entropy quantifies the loss of information in time, in a dynamic system. It is estimated from the symbolic RP_S as:

$$K2_S = -ln(D2_S) \qquad (19)$$

Lyapunov Exponent Lyapunov exponents (LE) estimate the amount of chaos in a dynamical system by quantifying the exponential divergence of initially close phase-space trajectories. A system with one or more positive LEs is defined to be chaotic.

4 Example Application

In this section we illustrate the application of the RQA_S as well as the symbolic dynamical invariants for synthetic examples of the logistic map. Logistic map is a well known model that exhibits different dynamic properties as a function of a single parameter, and allows comparing standard non-linear dynamic information measures with the data analysis methods we develop here. The logistic map model allows us to investigate some intriguing computational properties of sequences that happen at the onset of chaos (add references).

4.1 Symbolic RQA_S for the logistic map

Mathematically, the equation of the logistic map is defined as:

$$x_{i+1} = ax_i(1 - x_i) \qquad (20)$$

where x_i is a real number between zero and one and a is a positive constant. We generate multiple time series from the logistic map and define the control parameter $r \in [3.5, 4]$, with $\Delta r = 0.0005$, so that for each r we have a separate time series T of length 1000. The values of the parameters are set in order to compare the results with [20] and accordingly, we embed the time series with dimension $m = 3$ and time delay $\tau = 1$.

Figure 4 shows plots of our VMO-derived RQA_S measures. A comparison of the plots below is made with similar plots in [20] derived directly from the time series after embedding.

Plots (A), (C), (E) of figure 4 show the measures DET_S, $Lmax_S$ and L_S respectively that are based on the diagonal lines. They have similar maximas that indicate the periodic-chaos/chaos-periodic transitions. $Lmax_S$ detects all such transitions, but DET_S and L_S do not find them all.

Similarly, the chaos-chaos transitions to the laminar states are depicted by the measures based on the vertical structures, shown in plots (B), (D), (F): LAM_S, TT_S and $Vmax_S$. The difference between LAM_S and $Vmax_S$ is that LAM_S only measures the amount of laminar states, while $Vmax_S$ estimates the maximum duration of the laminar states. The lines in $Vmax_S$ plot show significant drops within periodic windows, indicating that the chaos-order transitions are also identified. This is in agreement with [20] who states that RQA measures are able to identify bifurcation points. However the LAM_S plot shows a different structure, it displays minimas or drops that correspond to the chaos-chaos transitions, while in the referenced work the LAM plot displays maximas or peaks at the same locations. This may be due to the fact that our LAM_S is derived from a symbolic representation of the series rather than the data itself. However as in the referenced paper, LAM_S is different from the other two vertical-based measures $Vmax_S$ and TT_S, in that it does not peak at inner crises, possibly because it is more robust against noise in the data. Finally similarly to the method in [20], with our symbolization method a 1000 data points are enough to derive the RP-based measures.

4.2 Symbolic dynamical invariants for the logistic map

To test the suitability of our symbolic complexity measures, we compute them for the logistic map. For each of LE, D2, K2, IR as well as θ, there is one value per time series.

Figure 5 portrays plots of the LE computed directly from the time series after embedding, the Shannon entropy (S_{RP_S}) estimated from the RP_S, the sym-

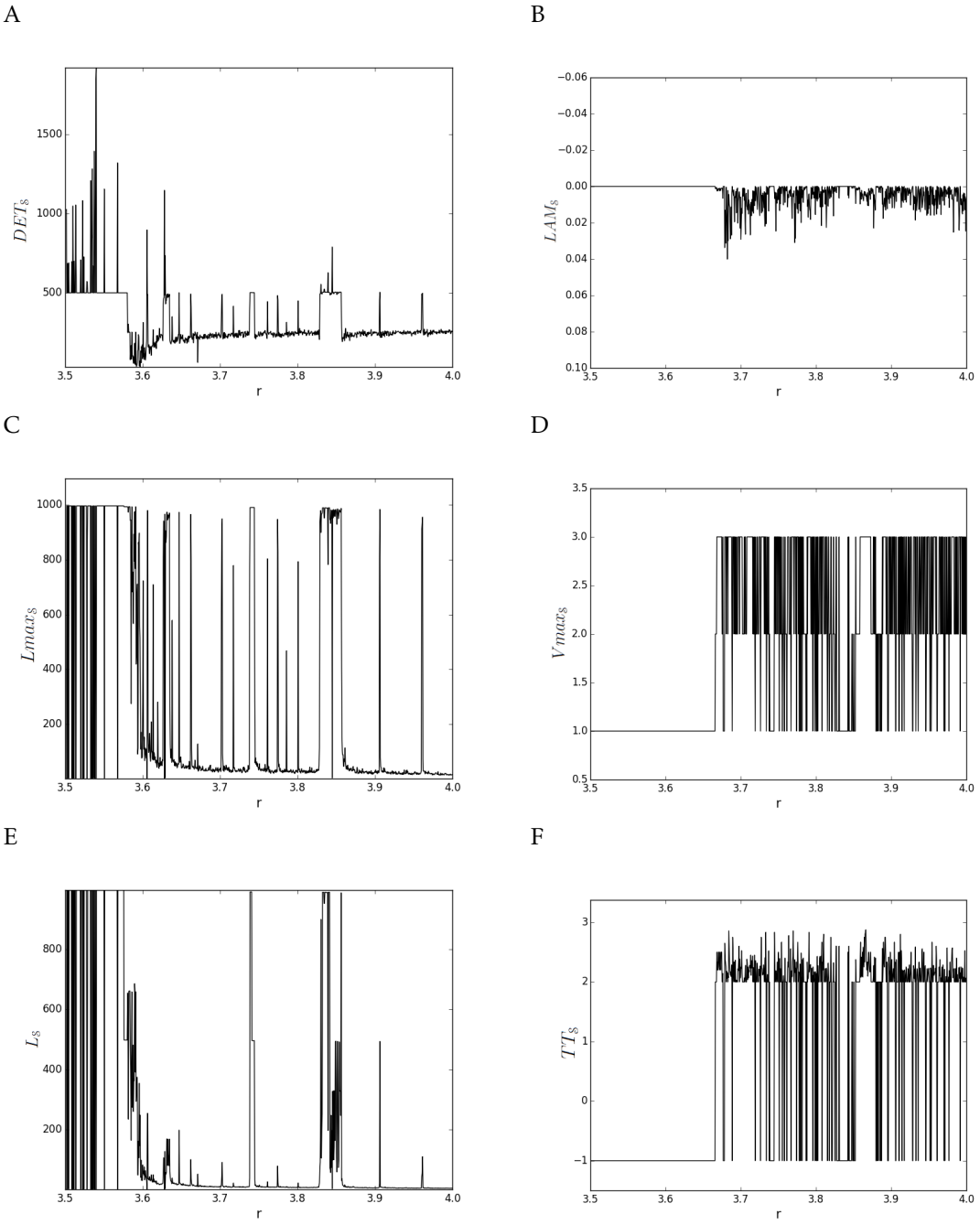


Figure 4: RQA_S from the RP_S of the logistic map: $r \in [3.5, 4.0]$, $\Delta r = 0.0005$ and $T = 1000$. (A) Determinism. (B) Maximal diagonal line length. (C) Average diagonal line length. (D) Laminarity. (E) Maximal vertical line length. (F) Trapping time.

bolic LE_S and $K2_S$. The plots are commented in subsection 4.3.

The formal relationship between the correlation entropy K2 and the Lyapunov Exponents LE is [20]:

$$K_2 \leq \sum_{\lambda_i > 0} \lambda_i \tag{21}$$

where λ_i denote the Lyapunov exponents. From Eq.(21) one sees that K2 is a lower bound for the sum of the positive Lyapunov exponents.

4.3 Qualitative Comparison of Dynamical Invariants

There are different dynamical regimes and transitions that occur between the values in the range of r of the logistic map. They appear in the form of accumulation points, periodic and chaotic states, band merging points, period doublings and various order-chaos, chaos-order as well as chaos-chaos transitions [34, 20].

In [35], authors applied various measures to derive statistical meaning from the graphical structures of symbolic recurrence plots. One such measure is the Shannon entropy of diagonal line length distribution in the RP (S_{RP_S}).

The Shannon entropy is a measure of the complexity of the RP, such that for uncorrelated noise it takes a small value, which indicates a low complexity [20]. However it is not a dynamical invariant [36].

According to Eckmann et al. 1987, the diagonal line lengths on RPs are related to the inverse of the largest LE [37]. This is true for some cases despite the fact that empirical studies have shown that S_{RP_S} is capable of identifying dynamical transitions, and therefore should grow as the system's complexity grows. [38].

Since the Shannon entropy quantifies the complexity of the dynamical system being studied, it is expected that its values increase when the system develops, that is when it varies from non-chaotic to chaotic regime [38]. Hence it is expected to be positively correlated with the LE, rather than negatively correlated. And within periodic windows, the entropy should considerably decrease.

Therefore in a recent work, [39] has proposed another estimation of the Shannon entropy from RPs from the relative frequency of the occurrence of the diagonal segments of nonrecurrent points formed by white dots, that are a signature of complexity within the data. In this case a one-to-one correspondence was seen between the new Shannon entropy estimate and the positive LE. That is, the Shannon entropy increased as the bifurcation parameter of the logistic equation increased, as illustrated in the plots found in [39]. However, [38] claim that the definition of the entropy from the white nonrecurrent dots does not solve the problem of the negative correlation between the entropy and the positive LE.

Here we compute the positive LE from the logistic map time series, after embedding and symbolization with VMO. The LE is computed from the VMO-derived LRS, and the Shannon entropy is estimated from the VMO-derived symbolic RP_S.

Figure 5 (A) shows the plot of the LE obtained from the time series of the logistic map. Plot (C) displays the Shannon entropy S_{RP_S} estimated from the black diagonal lines of the symbolic recurrence plot. The S_{RP_S} plot detects the chaos-chaos transitions as well as the periodic-chaos and chaos-periodic transitions. The plot shows that the S_{RP_S} correlates positively with the LE rather than with its inverse. We note though that for $r \in [3.9, 4.0]$ the values of the entropy seems to slightly decrease with the chaotic behaviour of LE. Further investigation is needed to understand this behaviour, that seems to correlate with the inverse of LE for that particular region of r.

It is clear that with our method, while maintaining the computation of the entropy from the black recurrence dots of the symbolic RP_S generated from the LRS of the symbolized time series, we obtain an estimate of the diagonal lines entropy that correlates positively with the Lyapunov exponents plot, with the exception of the region of $r \in [3.9, 4.0]$.

Plot (B) illustrates the LE_S obtained from the VMO-derived LRS. Plot (D) is the K2 plot derived from the RP_S and shows that the K2 plot is a lower bound of the LE_S plot. This verifies the relation expressed in Eq. 21, where K2 is defined as being a lower bound on the sum of the positive LEs.

Plot (E) displays the IR values that correspond to the optimal VMO model for each time series generated from the logistic map. By contrasting it with plot (A) of figure 4, we notice that the IR plot correlates positively to the determinism measure. Both show maximas or peaks at the chaos-order transitions. Additionally, IR captures the chaos-chaos transitions as well.

5 Model of Affect

Model of Affect. Affect can be described with basic emotional categories or emotional dimensions. We use categorical representations of Russel's two-dimensional model of valence and arousal (VA) described by pleasant-unpleasant for V, and awake-tired for A [40].

Stimuli. Four databases are used for experimental validation of the proposed symbolic dynamical features. The International Affective Digitized Sounds (IADS-2) [41] consists of a set of standardized, affective environmental sounds that span a broad range of semantic categories, and are rated on the two-dimensional model of VA.

The Montreal Affective Voices (MAV) [42] consist of 90 nonverbal affect bursts, enacted by 10 different actors in the following 8 categorical emotions: anger, disgust, feat, pain, sadness, surprise, happiness and pleasure, in addition to the neutral expression. The durations of the vocalizations vary between 0.385 to 2.229 seconds sampled at 44100 Hz. In order to per-

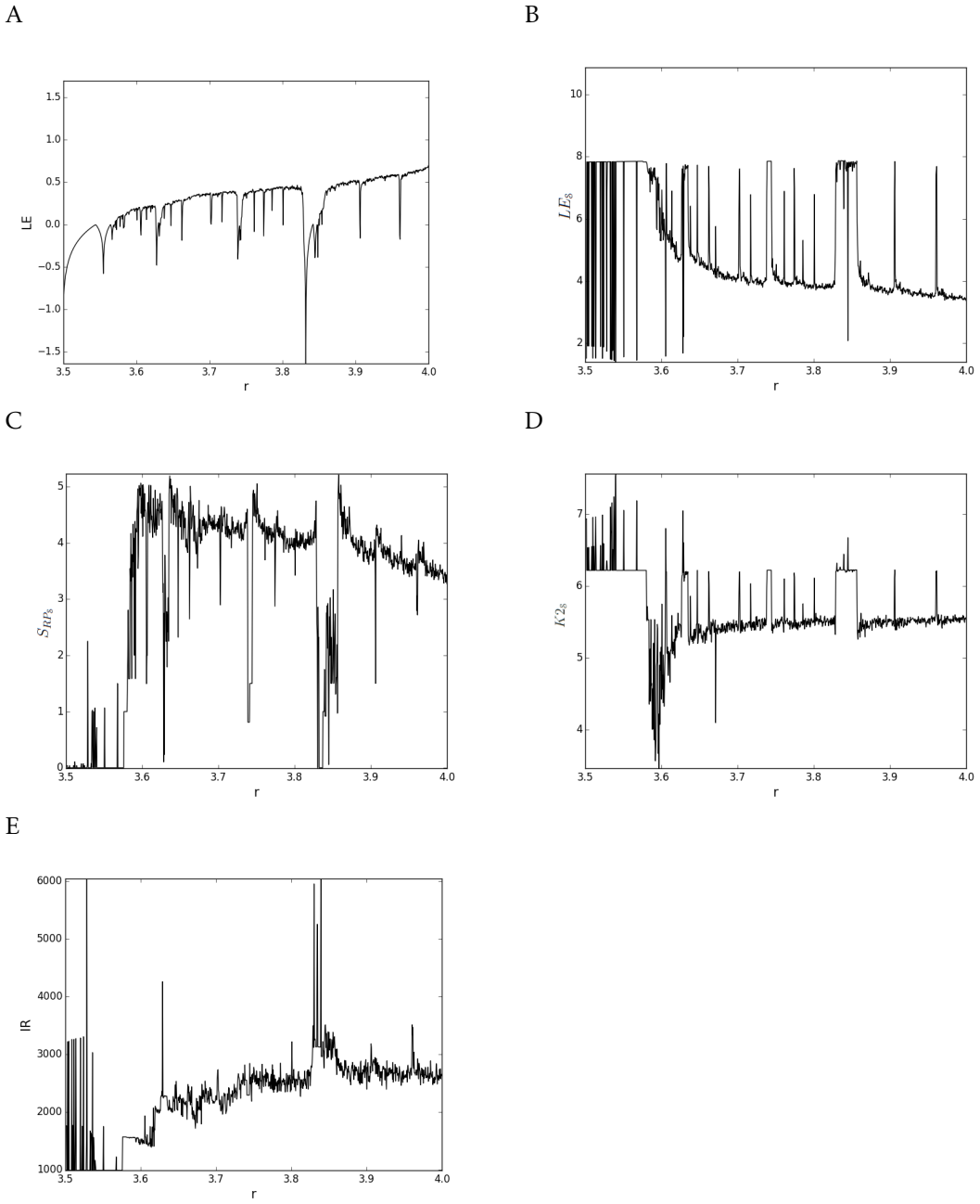


Figure 5: Dynamical Invariants of logistic map: $r \in [3.5, 4.0]$, $\Delta r = 0.0005$. (A) LE from the time series. (B) LE_S from LRS. (C) Shannon entropy from RP_S. (D) $K2_S$ from RP_S. (E) Information rate from RP_S.

form a comparative analysis with different databases and to ensure consistency in the affective annotations, the emotional categories of MAV sounds were mapped on the VA model described by: pleasant-unpleasant (V) and awake-tired (A).

The Musical Emotional Bursts (MEB) [43] consist of 80 short instrumental musical clips, played using the clarinet and the violin, in the following 3 categorical emotions: happiness, sadness and fear, plus the neutral expression. The mean duration of the MEB clips is of 1.6 seconds. The discrete emotions were mapped on the VA model.

The Film Music Excerpts (FME) [44] consist of 360 musical excerpts at 44100 Hz sampling rate, annotated on the VA dimensional model of emotions described with: high valence, low valence, high energy, low energy. The FME is designed to provide musical stimuli that allow a systematic comparison of the perceived emotions in music.

Dataset. Two datasets are evaluated: the symbolic RQA_S measures for the IADS, MAV, MEB and FME stimuli, and symbolic dynamical invariants for the IADS, MAV and FME stimuli.

Classification Model. In this work we use a feedforward artificial neural net (ANN) [45]. The feedforward ANN is chosen for its simplicity and suitability for the problem studied. The network has one hidden layer and one output layer. It was trained with the Levenberg- Marquardt as well as the scale conjugate gradient backpropagation learning algorithms, and the validation was performed using the mean squared error function (MSE).

6 Experiments

The experimental work as well as the results obtained are described in this section.

6.1 Experimental work

For the computation of the RQA_S dataset, the CQT is obtained at 44100 Hz sampling rate, hop length of 512 and 84 bins. Then we apply the symbolization process and estimate the RQA_S. For the computation of the dynamical invariants, the time series is embedded first and then we apply the symbolization. The datasets are normalized before training so that column features are scaled to have standard deviation 1, and centered to have mean 0.

Classification. The dataset is divided into 70% training, 15% validation and 15% testing. The classification tasks are conducted in a multi-class one-versus-all fashion whereby each of the six affective sub-dimensions is in turn considered as positive and negative class. The final results are then averaged to get the classifier's performance on VA. The Neural Network Toolbox in MATLAB is used. No feature selection was made prior to classification.

Performance evaluation. If the dataset is too small or biased, over-fitting can occur. Therefore in addition to the MSE function, we applied the Adaptive Synthetic Sampling (ADASYN) algorithm to rebalance the datasets. The classifier's performance is evaluated before and after dataset rebalancing, using a combination of performance metrics taken from the confusion matrix. These are: accuracy (ACC), precision, recall, F1-measure, F2-measure, area under the receiver operating characteristic curve (AUC), as well as Cohen's Kappa. Accuracy and precision are highly sensitive to data imbalance therefore three additional measures are computed: Cohen's Kappa (κ), F1-measure and F2-measure.

6.2 Results

Classification performance

In order to evaluate the generalizability of our features, we tested them on four different types of stimuli.

The classification performance rates of the RQA_S measures are reported in tables 1 to 5, and performance rates of the dynamical invariants are in tables 6 and 8.

In table 1, the prediction accuracies on VA are (74%, 90%) respectively for the auditory scenes (IADS). Comparing to other existing work, in [46] a classification task was evaluated for the IADS database using a set of 101 acoustic features, and achieved a performance of less than 50% accuracy. We note some poor metrics values such as κ is very low for the arousal dimension indicating an agreement close to chance level; AUC is 0.66 for valence. However the values of F_1, F_2, precision and recall are fairly high. These results could be consolidated or checked if the same tests are repeated with a much larger dataset.

In tables 2 and 3 high prediction accuracies of (97%, 90%) and (81%, 90%) are attained respectively on VA for the musical clips. The values of the remaining metrics are higher for the violin music than for the clarinet music. Although in both cases $\kappa > 0.40$ for both VA which shows an agreement above chance level between observed and predicted values.

Table 4 shows a success rate of (79%, 93%) on VA for the MAV dataset using RQA_S. This rate is verified with the values of the remaining metrics which are fairly high: both F_1 and F_2 are > 0.70, $\kappa > 0.40$, and the AUC is close to 1 for arousal and 0.79 on valence.

In table 5, the classification results on the FME dataset using the RQA_S metrics are of (65%, 77%) on VA. This shows that the RQA_S do not capture well valence in the film music excerpts. Further investigations are needed in this respect, to determine what is impacting the differences in the recognition rates between the short music clips and the short film music excerpts.

The classification success rates using the dynamical invariants achieve (72%, 82%) for IADS (table 6), (69%, 73%) for FME (table 7) and (84%, 91%) for MAV (table 4). The invariants perform well for IADS and MAV, however they have a rather low rate for FME,

Table 1: RQA_S performance measures for IADS on VA

Affect	ACC	PPV	TPR	F_1	F_2	κ	AUC
Valence	0.74	0.93	0.73	0.82	0.76	0.17	0.66
Arousal	0.90	1.00	0.90	0.94	0.92	0.06	0.72

Table 2: RQA_S performance measures for Violin music on VA

Affect	ACC	PPV	TPR	F_1	F_2	κ	AUC
Valence	0.97	0.95	0.98	0.96	0.97	0.93	0.98
Arousal	0.90	0.73	1.00	0.83	0.92	0.76	0.92

Table 3: RQA_S performance measures for Clarinet music on VA

Affect	ACC	PPV	TPR	F_1	F_2	κ	AUC
Valence	0.81	0.75	0.71	0.73	0.72	0.56	0.85
Arousal	0.90	0.83	0.89	0.86	0.88	0.78	0.95

Table 4: RQA_S performance measures for MAV on VA

Affect	ACC	PPV	TPR	F_1	F_2	κ	AUC
Valence	0.79	0.78	0.77	0.74	0.75	0.47	0.79
Arousal	0.93	0.78	0.81	0.79	0.80	0.75	0.96

Table 5: RQA_S performance measures for FME on VA

Affect	ACC	PPV	TPR	F_1	F_2	κ	AUC
Valence	0.65	0.60	0.87	0.71	0.80	0.30	0.71
Arousal	0.77	0.87	0.46	0.60	0.51	0.47	0.83

Table 6: Dynamical invariants performance measures on IADS

Affect	ACC	PPV	TPR	F_1	F_2	κ	AUC
Valence	0.72	0.91	0.71	0.79	0.74	0.14	0.64
Arousal	0.82	0.78	0.76	0.75	0.75	0.09	0.61

Table 7: Dynamical invariants performance measures for FME on VA

Affect	ACC	PPV	TPR	F_1	F_2	κ	AUC
Valence	0.69	0.69	0.70	0.69	0.70	0.39	0.75
Arousal	0.73	0.72	0.46	0.56	0.50	0.38	0.79

Table 8: Dynamical invariants performance measures for MAV on VA

Affect	ACC	PPV	TPR	F_1	F_2	κ	AUC
Valence	0.84	0.72	0.80	0.75	0.78	0.59	0.91
Arousal	0.91	0.61	0.90	0.72	0.82	0.45	0.91

although they perform better than RQA_S on valence.

In [47], an emotion classification task is made on the same FME stimuli, using 200 acoustic features and five emotional categories: anger, fear, happiness, sadness and tenderness. The recognition rates using support vector machines (SVM) are: 65%, 67%, 59%, 69% and 67% for the five emotions respectively. Although in our case we did not use discrete emotions but the VA model, however with only five nonlinear dynamical features our recognition rates are (69%, 73%) and using our RQA_S the rates are (65%, 77%).

7 Conclusive Remarks and Discussion

In this work we proposed a novel method to estimate complexity measures from the symbolic RP_S of the VMO symbolized time series of audio signals. The symbolic RQA_S measures were estimated without phase space reconstruction. The symbolic dynamical invariants were estimated after embedding. We estimated our dynamical measures for the symbolized time series of the logistic map and through a qualitative analysis of the respective plots, we showed that our symbolic measures are in agreement with the same measures obtained using different methods in literature. Furthermore, we estimated the symbolic Shannon entropy S_{RP_S} of the RP diagonal lines from the LRS of the optimal VMO model, and showed that it correlates positively with the Lyapunov exponents except for the region $r \in [3.9, 4.0]$.

In order to evaluate the performance of our symbolic dynamical measures in characterizing affect in various sounds, we conducted classification tasks on four types of stimuli. High emotion recognition rates were achieved for the IADS, MAV and MEB datasets for both sets of symbolic measures. This highlights the powerful performance of the measures in emotion recognition independently of the type of stimuli and shows that they generalize well across different types of stimuli. Furthermore it encourages future work to test the features in large scale tasks to determine if they can gain consensus as a general-purpose feature set.

However we obtained rather low recognition rates for film music excerpts on valence, and an average rate for arousal of 77% and 73%. Further work is needed to explore why both sets of complexity measures, the RQA_S as well as the dynamical invariants obtained in this work, achieve rather low recognition rates on film music excerpts, compared to the results obtained on the IADS, MAV as well as the music clips datasets. Obviously the dynamics differ across different sounds, but it would be interesting for future work to further investigate what particular aspects of the dynamics carry affective information. It would also be interesting to determine how such knowledge differs across different types of audio signals.

Conflict of Interest The authors declare no conflict of interest.

References

[1] Pauline Mouawad and Shlomo Dubnov. Novel method of nonlinear symbolic dynamics for semantic analysis of auditory scenes. In *Semantic Computing (ICSC), 2017 IEEE 11th International Conference on*, pages 433–438. IEEE, 2017.

[2] Caitlin J Butte, Yu Zhang, Huangqiang Song, and Jack J Jiang. Perturbation and nonlinear dynamic analysis of different singing styles. *J Voice*, 23(6):647–652, 2009.

[3] Patricia Henríquez, Jesús B Alonso, Miguel A Ferrer, Carlos M Travieso, and Juan R Orozco-Arroyave. Application of nonlinear dynamics characterization to emotional speech. In *International Conference on Nonlinear Speech Processing*, pages 127–136. Springer, 2011.

[4] Patricia Henríquez, Jesús B Alonso, Miguel A Ferrer, Carlos M Travieso, and Juan R Orozco-Arroyave. Nonlinear dynamics characterization of emotional speech. *Neurocomputing*, 132:126–135, 2014.

[5] Hanspeter Herzel. Bifurcations and chaos in voice signals. *Appl. Mech. Rev*, 46(7):399–413, 1993.

[6] Ali Shahzadi, Alireza Ahmadyfard, Ali Harimi, and Khashayar Yaghmaie. Speech emotion recognition using nonlinear dynamics features. *Turk J Elec Eng & Comp Sci.*, 23(Sup. 1):2056–2073, 2015.

[7] Peter beim Graben and Axel Hutt. Detecting recurrence domains of dynamical systems by symbolic dynamics. *Phys. Rev. Lett.*, 110(15):154101, 2013.

[8] Julyan HE Cartwright, Diego L González, and Oreste Piro. Pitch perception: A dynamical-systems perspective. *Proceedings of the National Academy of Sciences*, 98(9):4855–4859, 2001.

[9] Patricia Henríquez, Jesús B Alonso, Miguel A Ferrer, Carlos M Travieso, Juan I Godino-Llorente, and Fernando Díaz-de María. Characterization of healthy and pathological voice through measures based on nonlinear dynamics. *IEEE Trans. Audio, Speech, Language Process.*, 17(6):1186–1195, 2009.

[10] C de A Washington, FM de Assis, BG Aguiar Neto, Silvana C Costa, and Vinıcius JD Vieira. Pathological voice classification based on recurrence quantification measures. 2012.

[11] Gerard Roma, Waldo Nogueira, and Perfecto Herrera. Recurrence quantification analysis features for environmental sound recognition. In *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 1–4. IEEE, 2013.

[12] Gerard Roma, Waldo Nogueira, Perfecto Herrera, and Roc de Boronat. Recurrence quantification analysis features for auditory scene classification. *IEEE AASP Challenge: Detection and Classification of Acoustic Scenes and Events, Tech. Rep*, 2013.

[13] Angela Lombardi, Pietro Guccione, and Cataldo Guaragnella. Exploring recurrence properties of vowels for analysis of emotions in speech. *Sensors & Transducers*, 204(9):45, 2016.

[14] Jesús B Alonso, Fernando Díaz-de María, Carlos M Travieso, and Miguel Angel Ferrer. Using nonlinear features for voice disorder detection. In *ISCA tutorial and research workshop (ITRW) on non-linear speech processing*, 2005.

[15] Teresa D Wilson and Douglas H Keefe. Characterizing the clarinet tone: Measurements of lyapunov exponents, correlation dimension, and unsteadiness. *The Journal of the Acoustical Society of America*, 104(1):550–561, 1998.

[16] Holger Kantz and Thomas Schreiber. *Nonlinear time series analysis*, volume 7. Cambridge university press, 2004.

[17] Floris Takens. Detecting strange attractors in turbulence. In *Dynamical systems and turbulence, Warwick 1980*, pages 366–381. Springer, 1981.

[18] Thomas Schreiber. Interdisciplinary application of nonlinear time series methods. *Phys. Rep.*, 308(1):1–64, 1999.

[19] Matthew B Kennel, Reggie Brown, and Henry DI Abarbanel. Determining embedding dimension for phase-space reconstruction using a geometrical construction. *Physical review A*, 45(6):3403, 1992.

[20] Norbert Marwan, M Carmen Romano, Marco Thiel, and Jürgen Kurths. Recurrence plots for the analysis of complex systems. *Physics reports*, 438(5):237–329, 2007.

[21] Elizabeth Bradley and Folger Kantz. Nonlinear time-series analysis revisited. *Chaos*, 25(9):097610, 2015.

[22] Jerome Rolink, Martin Kutz, Pedro Fonseca, Xi Long, Berno Misgeld, and Steffen Leonhardt. Recurrence quantification analysis across sleep stages. *BIOMED SIGNAL PROCES*, 20:107–116, 2015.

[23] David Schultz, Stephan Spiegel, Norbert Marwan, and Sahin Albayrak. Approximation of diagonal line based measures in recurrence quantification analysis. *Phys. Lett. A*, 379(14):997–1011, 2015.

[24] Charles L Webber, Norbert Marwan, Angelo Facchini, and Alessandro Giuliani. Simpler methods do it better: success of recurrence quantification analysis as a general purpose data analysis tool. *Phys. Lett. A*, 373(41):3753–3756, 2009.

[25] Cheng-i Wang and Shlomo Dubnov. The variable markov oracle: Algorithms for human gesture applications. *IEEE MultiMedia*, 22(4):52–67, 2015.

[26] Shlomo Dubnov. Spectral anticipations. *Comput. Music J.*, 30(2):63–83, 2006.

[27] Cyril Allauzen, Maxime Crochemore, and Mathieu Raffinot. Factor oracle: A new structure for pattern matching. In *International Conference on Current Trends in Theory and Practice of Computer Science*, pages 295–310. Springer, 1999.

[28] Gérard Assayag and Shlomo Dubnov. Using factor oracles for machine improvisation. *Soft Comp.*, 8(9):604–610, 2004.

[29] Shlomo Dubnov, Gerard Assayag, and Arshia Cont. Audio oracle: A new algorithm for fast learning of audio structures. In *Proceedings of International Computer Music Conference (ICMC)*. ICMA, 2007.

[30] Cheng-i Wang and Shlomo Dubnov. Guided music synthesis with variable markov oracle. In *The 3rd International Workshop on Musical Metacreation, 10th Artificial Intelligence and Interactive Digital Entertainment Conference*, 2014.

[31] Shlomo Dubnov, Gérard Assayag, and Arshia Cont. Audio oracle analysis of musical information rate. In *Semantic Computing (ICSC), 2011 Fifth IEEE International Conference on*, pages 567–571. IEEE, 2011.

[32] Cheng-i Wang and Shlomo Dubnov. Pattern discovery from audio recordings by variable markov oracle: A music information dynamics approach. In *Proc IEEE Int Conf Acoust Speech Signal Process.*, pages 683–687. IEEE, 2015.

[33] Arnaud Lefebvre and Thierry Lecroq. Compror: on-line lossless data compression with a factor oracle. *Inform. Process. Lett.*, 83(1):1–6, 2002.

[34] Pierre Collet and J-P Eckmann. *Iterated maps on the interval as dynamical systems*. Springer Science & Business Media, 2009.

[35] LL Trulla, A Giuliani, JP Zbilut, and CL Webber. Recurrence quantification analysis of the logistic equation with transients. *Phys. Lett. A*, 223(4):255–260, 1996.

[36] H Rabarimanantsoa, L Achour, C Letellier, A Cuvelier, and J-F Muir. Recurrence plots and shannon entropy for a dynamical analysis of asynchronisms in noninvasive mechanical ventilation. *Chaos*, 17(1):013115, 2007.

[37] J-P Eckmann, S Oliffson Kamphorst, and David Ruelle. Recurrence plots of dynamical systems. *EPL (Europhysics Letters)*, 4(9):973, 1987.

[38] Deniz Eroglu, Thomas K DM Peron, Nobert Marwan, Francisco A Rodrigues, Luciano da F Costa, Michael Sebek, István Z Kiss, and Jürgen Kurths. Entropy of weighted recurrence plots. *Physical Review E*, 90(4):042919, 2014.

[39] Christophe Letellier. Estimating the shannon entropy: recurrence plots versus symbolic dynamics. *Physical review letters*, 96(25):254102, 2006.

[40] Ulrich Schimmack and Alexander Grob. Dimensional models of core affect: A quantitative comparison by means of structural equation modeling. *Eur. J. Pers.*, 14(4):325–345, 2000.

[41] Margaret M Bradley and Peter J Lang. The international affective digitized sounds (; iads-2): Affective ratings of sounds and instruction manual. *University of Florida, Gainesville, FL, Tech. Rep. B-3*, 2007.

[42] Pascal Belin, Sarah Fillion-Bilodeau, and Frédéric Gosselin. The montreal affective voices: a validated set of nonverbal affect bursts for research on auditory affective processing. *Behav Res Methods.*, 40(2):531–539, 2008.

[43] S Paquette, I Peretz, and P Belin. The musical emotional bursts: a validated set of musical affect bursts to investigate auditory affective processing. *Frontiers in Psychology*, 4(509):1–7, 2013.

[44] Tuomas Eerola and Jonna K Vuoskoski. A comparison of the discrete and dimensional models of emotion in music. *Psychol. Music*, 39(1):18–49, 2011.

[45] Philip J Drew and John RT Monson. Artificial neural networks. *Surgery*, 127(1):3–11, 2000.

[46] Konstantinos Drossos, Andreas Floros, and Nikolaos-Grigorios Kanellopoulos. Affective acoustic ecology: towards emotionally enhanced sound events. In *Proceedings of the 7th Audio Mostly Conference: A Conference on Interaction with Sound*, pages 109–116. ACM, 2012.

[47] Cyril Laurier, Olivier Lartillot, Tuomas Eerola, and Petri Toiviainen. Exploring relationships between audio features and emotion in music. In *ESCOM 2009: 7th Triennial Conference of European Society for the Cognitive Sciences of Music*, 2009.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1741-1747 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Retrieving Dialogue History in Deep Neural Networks for Spoken Language Understanding

Myoung-Wan Koo*, [1], Guanghao Xu[1], Hyunjung Lee[2], Jungyun Seo[1]

[1]*Department of Computer Science and Engineering, Sogang University, 04107, Republic of Korea*

[2]*Institut für Linguistik, Universität Leipzig, 04107, Germany*

ARTICLE INFO

Article history:
Received: 30 May, 2017
Accepted: 13 August, 2017
Online: 15 September, 2017

Keywords:
Spoken Language Understanding
Convolutional Neural Network
Recurrent Neural Network

ABSTRACT

In this paper, we propose a revised version of the semantic decoder for multi-label classification task in the spoken language understanding (SLU) pilot task of the Dialog State Tracking Challenge 5 (DSTC5). Our model concatenates two deep neural networks - a Convolutional Neural Network (CNN) and a Recurrent Neural Networks (RNN) - for detecting semantic meaning of incoming utterance with the assistance of algorithm adaptation method. In order to evaluate the robustness of our proposed models, comparative experiments on the DSTC5 dialogue datasets are conducted. Experimental results show that the proposed models outperform most of the submitted models in the DSTC5 in terms of F1-score. Without any manually designed features or delexicalization, our model has proven its efficiency of tackling the multi-label SLU task, using only publicly available pre-trained word vectors. Our model is capable of retrieving the dialogue history, and thereby it could build the concise concept structure by employing the pragmatic intention as well as semantic meaning of utterances. The architecture of our semantic decoder has a potential to be applicable to other variety of human-to-human dialogues to achieve SLU.

1. Introduction

The spoken language understanding (SLU) has been one of the fundamental components of an end-to-end dialogue system. The Dialog State Tracking Challenge 5 (DSTC5) released a pilot SLU task, which requires to extract semantic meaning of users' utterances in task-oriented dialogues and to fill the slots with speech acts. Unlike the previous challenges (DSTC 2&3) where the human-to-system dialogues were targeted, the corpus of DSTC 5 has been no more than challengeable due to the following points: human-to-human dialogues, cross-linguistic data and multi-label classification task. As its corpus is built by collecting human-to-human conversation in a natural setting, more than one speech act can be annotated to a single utterance.

Xu et al. propose a Convolutional Neural Network (CNN) model with a threshold predictor to tackle a multi-label speech act classification task on DSTC5 corpus [1]. With the assistance of algorithm adaptation method, the model they propose is adapted for the multi-label classification task without any manually designed features. Although the model, however, has advantage of handling the multi-label classification task, there still remain rooms for improvement to achieve state-of-art SLU.

The aim of this study is to improve the model proposed by Xu et al. by building a more robust and concise semantic classifier. In this revised model two deep neural networks, Convolutional Neural Network and Recurrent Neural Network, are conjoined to conduct a multi-label speech act classification task in an improved fashion. Our newly revised model shows a synergy effect of retrieving dialog context as well as exploiting a current utterance. In addition, a threshold learning mechanism is engaged to enable our proposed model to produce an output of a set of multiple labels called speech acts.

The rest of this paper is organized as follows. Section 2 gives a brief review of DSTC5 dataset and some related works in the SLU pilot task. In Section 3, we introduce the architecture of our proposed model and a threshold predictor. The Section 4 gives a detailed description of the DSTC5 dataset and describes how we set up the experiments for training data and evaluation process. In Section 5, we provide our experimental results to optimize the

*Corresponding Author: Myoung-Wan Koo, Department of Computer Science and Engineering, Sogang University, Republic of Korea, mwkoo@sogang.ac.kr

Table 1. An example of test utterances annotated with speech act information.

Speaker	Chinese Utterances and their Translated English Sentences.	Speech Act Category (Attribute)
Guide	um, sentosa the universal studios in the matter. You see it, the whole family. 嗯，圣淘沙里面的环球影城啦，你看啦，一家大小。	FOL (ACK) FOL (INFO)
Tourist	there are still in the place where I can recommend? 还有地方可以介绍的吗？	QST (RECOMMEND) QST (WHERE)
Guide	yes, we have, um, the zoo. the daytime the zoo. 嗯，我们有一个动物园，那个日间动物园。	FOL (RECOMMEND) FOL (WHERE)
Tourist	how big is the singapore? 新加坡有多大？	QST (INFO)

performance of our CNN-RNN classifier on the DSTC5 SLU task. The Section 6 discusses the differences between the previous and current models and the strength of concatenating two deep neural networks in SLU task. The Section 7 concludes this paper[1].

2. Background

2.1. The DSTC5 Dataset

The DSTC5 provides the TourSG corpus, which consists of dialogue sessions collected from Skype calls between tour guides and tourists focusing on offering touristic information of Singapore[2]. For the SLU task, the system is given the utterances from both the tourist and the guide as its input, and the system subsequently tags the utterances spoken by both the speakers with appropriate *speech acts categories* and *attributes*.

Each sub-utterance belongs to one of the four basic *speech act categories* that denote general information of current dialogue flow. More specific speech act information can be annotated by the combination with the *speech act attributes*. Therefore, a classifier is demanded for classifying a set of labels consisting of speech act categories and attributes tagged to a single utterance. Reference [3] gives complete list of speech act categories and attributes.

Table 1 shows Chinese test utterances and ones translated in English that annotated with their corresponding speech act categories and attributes.

2.2. Other models in pilot SLU task of the DSTC5

A simple baseline model for the SLU task is provided by the committee of the DSTC 5, which uses a binary relevance (BR) approach and trains a set of linear support vector machines (SVM) for multi-label speech act classification. The baseline model is built within traditional TF-IDF approach which mainly depends on keywords that appeared in the utterances per speaker. This baseline model, however, has a crucial deficiency in detecting semantic meanings of utterances appropriately, as it only superficially decodes meaning relying on words on the surface level.

Ushio et al., which takes the first place in this challenge, proposes a local co-activate multi-task learning model (LC-MTL) for capturing structured speech acts by using recurrent convolutional neural networks[4]. This model consists of a CNN, which represents incoming utterances as sentence vectors, and two LSTMs, which locally co-activates neurons in hidden layer between speech act categories and attributes of the corresponding utterances.

Xu et al also proposes a CNN model with a threshold predictor, which enables to predict more than one speech act annotated to an utterance. It has been proven that CNN is capable to detect the core semantic meaning without any knowledge on the syntactic structures of a language or any manually designed features. This model, however, has faced the ambiguity issue, as the output node of CNN only depend on each current utterance to produce proper scores for each set of labels. In Section 3 we will introduce the revised version of the semantic decoder that has a power to disambiguate the meaning of utterances.

2.3. Related Works

So far there have been many researches grounded on CNN architecture with regard to the language processing tasks. The CNN proposed by Kim achieves good performance across several datasets despite of its simple architecture which consists of a convolutional layer, a max-pooling layer and a softmax classifier. Wang et al. proposed another architecture of CNN with an additional semantic layer, which exploits the contextual information from short texts[6]. Another CNN model with an unsupervised “region embedding”, proposed by Johnson, works well for long text classification task like movie reviews[7]. Zhang et al. explored the effects of hyperparameters in one-layer CNN architectures, and reported their impact on performance over multiple runs[8].

In addition, recent advances in deep learning with RNN including LSTM have also achieved impressive improvements on various natural language processing tasks such as language modeling, sequence tagging and machine translation. Language model based on RNN(RNN-LM) was proposed by Mikolov et al., which significantly outperformed the previous back-off models, even in case when RNN-LM was trained on much less data than back-off models were [9]. Another variant of RNN-LM based on LSTM proposed by Sundermeyer, showed additional improvements of 8% relative in perplexity over the RNN-LM[10]. Huang et al. proposed a sequence tagging model using a bidirectional LSTM, which was capable of using effectively both previous and oncoming input features. Their model have achieved state of the art accuracy on various sequence tagging tasks such as POS tagging, chunking and named entity recognition(NER)[11]. Sutskever et al. proposed a sequence to sequence learning model

[1] The source code of the proposed system is available on https://github.com/hkhpub/dstc5-slu

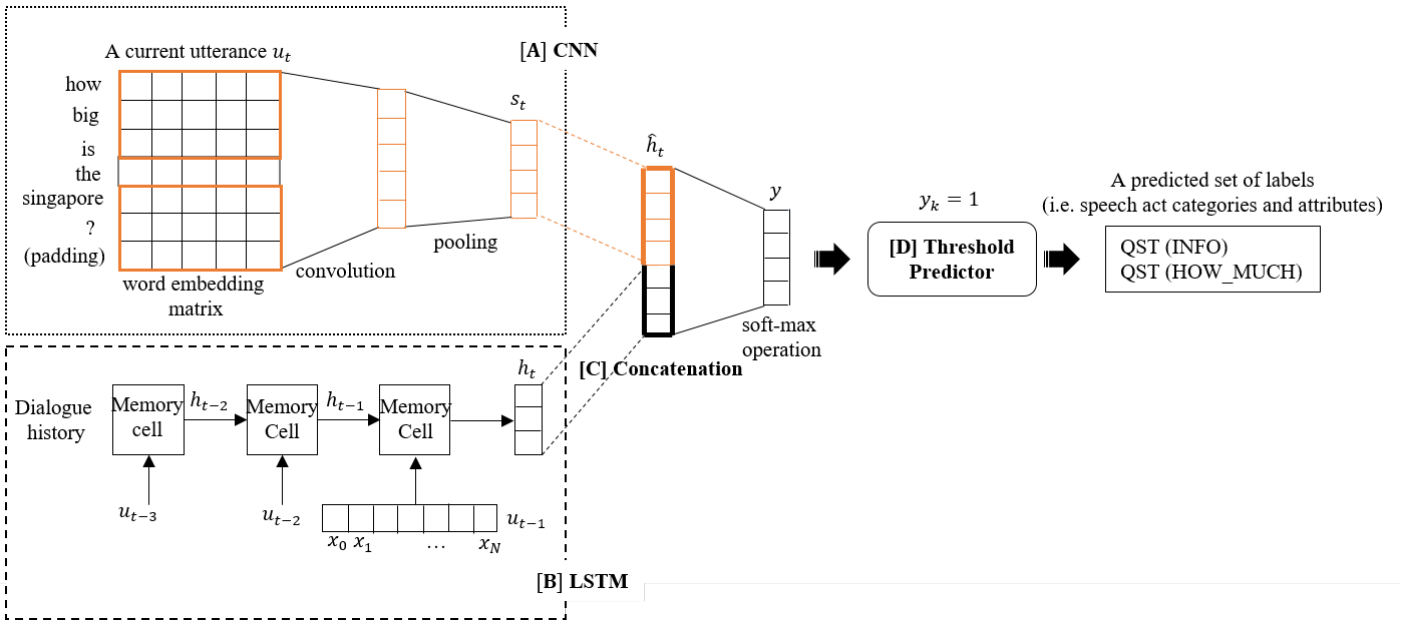


Figure 1. The architecture of our proposed model.

grounded on LSTM. They achieved a higher BLEU score on a translation task from English to French on the WMT-14 dataset, compared to a phrase-based machine translation system[12].

Several researches have been recently conducted by combining CNN and RNN models. Vinyals et al. proposes an conjoined model for image captioning, which encodes image features using deep convolutional networks and automatically generates captions with recurrent networks[13]. Kim proposes a recurrent convolutional network model, in which the penultimate layer of CNN is connected to the recurrent layers of the RNN model in order to track a topic of a dialogue segment in human-human conversations[14]. Unlike the previous CNN-RNN models, another jointed CNN and RNN model is proposed by Barahona et al., where the model is optimized with two distinctive inputs: a current user' utterance and dialog act-slot pairs of previous system's utterances [15]. In the task of decoding semantic meaning of spoken languages each input is utilized in sentence representation and context representation, respectively.

3. Model Architecture

In this section, we propose a jointed model which combines Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) . The overall architecture of our semantic decoder is depicted in Figure 1. Our model consists of three modules: (i) a CNN with multiple filters that encodes semantic meaning of current utterance, output nodes that produces scores for each label, (ii) a RNN with recurrent connections between neurons that store contextual information of dialogues and (iii) a multi-label threshold predictor that generates a reference point using the scores of the labels. The threshold is then used to for the system to decide whether each label is as relevant or irrelevant. Specifically, we newly adopt the RNN to improve the previous semantic decoder of Xu et al. with assistance of contextual information drawn from dialogues so that the disambiguity problem is raveled out.

3.1. CNN Architecture

Coming up with the architecture of CNN [5], which is specifically illustrated in the part [A] of Figure 1, given our CNN classifier is capable of optimizing its parameters with respect to multi-label cross entropy loss function. Formally, let $x_i \in \mathbb{R}^k$ be the k-dimensional word embedding vector corresponding to i-th word in a given utterance. An utterance u_t of length n are represented as a $n \times k$ matrix:

$$u_t = x_1 \oplus x_2 \oplus \ldots \oplus x_n \quad (1)$$

where t is the index of dialogue turn and $\oplus$ is the concatenation operator. A convolutional operation involves a filter $w \in \mathbb{R}^{h \times k}$, which is applied to a window of h adjacent words to produce a new feature. A feature c_i is generated by applying a hyperbolic tangent function f:

$$c_i = f(w \cdot x_{i:i+h-1} + b) \quad (2)$$

where $b \in \mathbb{R}$ is a bias term. The filter is applied to every possible window of words in the utterance to produce a feature map:

$$c = [c_1, c_2, \ldots, c_{n-h+1}] \quad (3)$$

A max-over-time pooling is then operated to take the maximum value $\hat{c} = \max\{c\}$ as a representative feature for the filter.

Following the same procedure as described above, multiple filters with varying window size h are integrally engaged into multiple adjacent features. These features are then concatenated into a fixed-length and 'top-level' feature vector s_t which automatically encodes most of representative features from a given utterance at dialogue turn t.

3.2. RNN Architecture

Since each utterance is dependent on the previous utterances in a conversation, referring to dialogue context from previous dialogue utterances is essential to wholly understand the meaning

Table 2. Statistics of dstc5 datasets.

Datasets (Speaker)	Language	*M*	*L*	*C*
Train (Tourist)	English	14,226	74/88	1.19
Test (Tourist)	Chinese	4,085	61/88	1.16
Train (Guide)	English	19,916	69/88	1.24
Test (Guide)	Chinese	8,555	71/88	1.21

M: Number of utterances.
L: Size of label set (size/total).
C: Average number of labels per utterance.

of a current utterance. To access to dialogue context, we employ a long short term memory (LSTM). It is much better for preserving information over long periods of time than vanilla RNN due to its ability to deal with vanishing and exploding gradients[16].

To track the dialogue context, two questions must be taken into consideration: (i) to what extent we should track previous utterances as dialogue history and (ii) in which form those utterances are fed into the LSTM as inputs. For length L of the dialogue history, we treat it as a free parameter and determine its value empirically. Once the length L is determined, the dialogue history is represented as $\{u_{t-L}, u_{t-L+1}, \dots, u_{t-1}\}$ where t is the index of dialogue turn.

As depicted in the part [B] of Figure 2, the structure of LSTM is divided into a memory cell c_l and three gates: a forget gate f_l, an input gate i_l and an output gate o_l. Three kinds of gates functions to decide which amount of information the memory cell should keep or forget at a time step l, where l denotes for each word in the utterances given a length L. The input x_l and the output h_l of LSTM are updated as follows:

$$i_l = \sigma(W^i \cdot x_l + U^i \cdot h_{l-1} + b^i) \quad (4)$$

$$f_l = \sigma(W^f \cdot x_l + U^f \cdot h_{l-1} + b^f) \quad (5)$$

$$o_l = \sigma(W^o \cdot x_l + U^o \cdot h_{l-1} + b^o) \quad (6)$$

$$g_l = tanh(W^g \cdot x_l + U^g \cdot h_{l-1} + b^g) \quad (7)$$

$$c_l = f_l \odot c_{l-1} + i_l \odot g_l \quad (8)$$

$$h_l = o_l \odot tanh\,(c_t) \quad (9)$$

where x_l is the input at the current time step, h_l is the hidden unit at time step l, b is a bias term, $\sigma(\cdot)$ is a logistic sigmoid function and $\odot$ denotes a point-wise multiplication operation. Each word of the utterances in the dialogue history is represented as word embedding vectors and fed into the LSTM sequentially as an input x_l. The last hidden unit h_l of LSTM is obtained which encodes the dialogue context information for current utterance u_t.

3.3. Combining CNN and LSTM

To utilize both the representative feature of current utterance and context information of previous utterances, we concatenate the hidden unit h_t of LSTM to the 'top-level' feature vector s_t modeled by the CNN, as illustrated in [C] of Figure 1. Then, the penultimate layer consists of the concatenated vector $\hat{h}_t = s_t \oplus h_t$, which is passed to a fully connected output layer.

$$\hat{y} = W \cdot \hat{h} + b \quad (10)$$

Then the softmax is operated to normalize the output vector $\hat{y}$ to the probability distribution as follows:

$$P(y_k = 1) = \frac{\exp(\hat{y}_k)}{\sum_j \exp(\hat{y}_j)} \quad (11)$$

where k denotes the index of the multi-hot vector y, which represents the pairs of speech act attribute and category information of utterance.

3.4. Multi-label Threshold Predictor

The output probability distribution $p(y|u)$ from the softmax layer is used for multi-label prediction, while the proposed model is trained and used in prediction for a given utterance u. A relevant label set Y for an utterance u is determined by a threshold t as follows:

$$Y = \{k | p_k > t,\ k \in L\}. \quad (12).$$

The threshold learning mechanism used in the literature [17, 18] is adopted in [D] of Figure 1, which models t with a linear regression model. The learning procedure is described as follows: For each training example $(u_m,\ Y_m)$, we set the target threshold value t_m which minimizes the count of misclassified labels as follows:

$$t_m = \underset{t}{\operatorname{argmin}}(|\{k | k \in Y_m, p_k^m \le t\}| + |\{l | l \in \overline{Y_m}, p_l^m \ge t\}|) \quad (13)$$

where p_k^m, p_l^m is the output probability of relevant label k and irrelevant label l associated with utterance u_m respectively. The target threshold values t_m is used in learning the θ of the threshold predictor $T(\boldsymbol{u}; \theta)$:

$$E(\theta) = \frac{1}{2}\sum_{m=1}^{M}(T(\boldsymbol{x}_m; \theta) - t_m)^2 + \frac{\lambda}{2}\|\theta\|^2, \quad (14)$$

where λ is the regularization parameter and M is the number of utterances in the train set. At the test time, the learned threshold is used to choose the relevant labels Y of a test utterance, as illustrated in (14).

4. Experimental Setup

In this section we introduce the DSTC5 dataset and describe the different models with their corresponding performances.

4.1. Statistics of DSTC5 Datasets

The summary statistics of the SLU datasets for the both speakers of the DSTC5 after tokenization are given in Table 2. For the case of *Guide*, one interesting point to note is that the size of sets of labels in the train set is smaller than that in the test set , which means that there is no way for the classifier to learn cases of certain labels assigned to utterances during the training and predict correct speech acts on the test dataset of *Guide*.

4.2. Hyper-parameters

In our experiments, we use: filter windows of 2, 3, 4 with 200 feature maps for the CNN, dimension of 100 for the hidden unit of LSTM. Those values are chosen by adopting a rough grid search[8]. The model undergoes training through stochastic gradient descent over shuffled mini-batches with Adam optimization algorithm. The model stops the iterant processes of learning by an early stopping mechanism.

As we mentioned in Section 3.2 that the length of dialogue context *L* is treated as a free parameter, the optimal value of *L* is

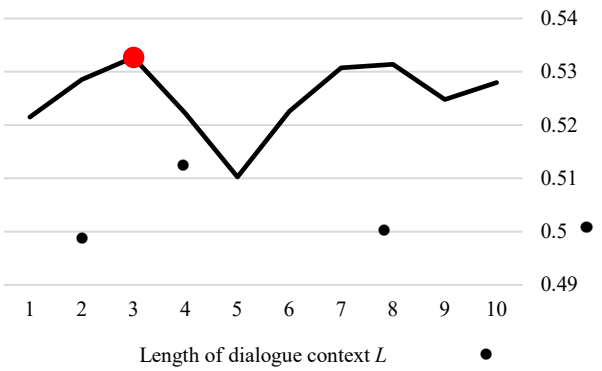


Figure 2. The performance of CNN-LSTM model based on various lengths of dialogue context on Tourist dataset.

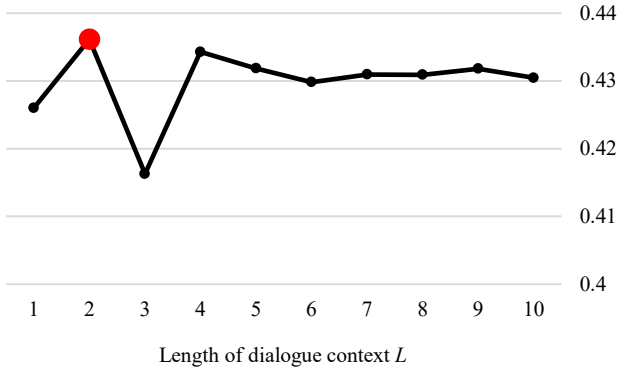


Figure 3. The performance of CNN-LSTM model based on various lengths of dialogue context on Guide dataset.

derived by conducting a grid search, where L ranges from 1 to 10 on both *Guide* and *Tourist* datasets. Figure 2 and Figure 3 illustrate the performance changes with respect to the length of dialogue context in the *Guide* and *Tourist* dataset respectively. It is shown that the optimal length of dialogue context would be L=3 and L=2 for the *Tourist* and *Guide* dataset, respectively, for making a dialogue context that improves the classification performance.

4.3. Word Embedding Vectors

GloVe[19] and *Word2Vec*[20] are the two most popular word embedding algorithms aiming at mapping semantic meaning of words in a geometric space. We initialize our proposed models with two publicly available pre-trained word vectors and both word embedding vectors have dimensionality of 300; *GloVe* that are trained on 6 billion words from Wikipedia 2014 and Gigaword5[2] and *Word2Vec* that are trained on 100 billion words from Google News[3]. In the preliminary experiments, it is observed that all the proposed models trained on top of the pre-trained *Word2Vec* show slightly better performances over those on top of pre-trained *GloVe*. In this sense only the performances of *Word2Vec* based models will be presented in this paper.

4.4. Model Variations

We evaluate three models with different architecture:

- CNN (multiclass): the model that predicts only one speech act category and attribute for given a *speaker's* utterance.
- CNN-LSTM (multiclass): the combined model which exploits dialogue context information from previous *speaker's* utterances
- CNN-LSTM (thresholding): the combined model with a threshold predictor that classifies multiple labels of speech act categories and attributes.

4.5. Evaluation Metrics

In the SLU task, a system is required to match relevant speech acts for a given unlabeled utterance spoken by the target role speaker. The following evaluation metrics are used in DSTC5.

- Precision: Fraction of speech act labels that are correctly predicted.
- Recall: Fraction of speech act labels in the gold standard that are correctly predicted.
- F-measure: The harmonic mean between precision and recall.

5. Results

Table 3 and Table 4 summarize the comparative experimental results of our models for classifying speech acts categories and attributes on *Guide* and *Tourist* datasets, respectively. On the *Tourist* dataset, the CNN-LSTM (multi-class) model shows promising results. Note that CNN-LSTM (multi-class) models outperform most of the models submitted in the DSTC5 for both speakers even without utilizing threshold learning mechanism. In terms of the F1-score, the model CNN-LSTM (thresholding) significantly outperforms all the other models on *Tourist* dataset. For the case of *Guide* dataset, our model is slightly behind of Team 2's model, but still highly comparable.

These results suggest that for conducting the SLU task of DSTC5 the CNN is a suitable model to predict and fill the slots of speech act categories and attributes. A threshold predictor enables the CNN models to classify a set of multiple labels on each utterances. It is worthy nothing that our model uses only publicly available word-embedding vectors without having any manually designed features or using delexicalization. Last but the most important thing is the dialogue context information encoded with the LSTM helps to improve the performance of conducting SLU task.

6. Discussion

At this point we circle back to the very beginning in order to understand what limitation had been posed on Xu et al.'s model and how our CNN-RNN model could tackle those issues[4]. The basic ideas of our study start from the properties of utterances. Each utterance is a sub-part of a dialogue and inevitably depends on the previous utterances in the intertwined way. In order to capture the semantic meaning from utterances the semantic

[2] http://nlp.stanford.edu/projects/glove

[3] https://code.google.com/archive/p/word2vec

[4] A reviewer suggested that the problem statement of previous model should be addressed. We thank a reviewer for bringing the comparison between previous and current models to our attention.

Table 3. Comparative results for *Tourist* dataset

Model	Precision	Recall	F1-measure
Baseline (SVM)	0.3694	0.1828	0.2446
Team 2 (LC-MTL)	0.5331	0.5263	0.5297
Team 3	0.4591	0.4241	0.4409
Team 5	0.5026	0.4484	0.4739
Team 7	0.5079	0.4156	0.4571
Xu (2017)	0.5010	0.5624	0.5299
CNN (multi-class)	**0.5462**	**0.4873**	**0.5151**
CNN-LSTM (multi-class)	**0.5603**	**0.4999**	**0.5284**
CNN-LSTM (thresholding)	**0.5455**	**0.5276**	**0.5364**

Table 4. Comparative results for *Guide* dataset

Model	Precision	Recall	F1-measure
Baseline (SVM)	0.4588	0.2480	0.3219
Team 2 (LC-MTL)	0.5127	0.4251	0.4648
Team 3	0.4340	0.3635	0.3956
Team 5	0.4639	0.3820	0.4190
Team 7	0.5007	0.2976	0.3733
Xu (2017)	0.4239	0.4295	0.4266
CNN (multi-class)	**0.4768**	**0.3927**	**0.4307**
CNN-LSTM (multi-class)	**0.4837**	**0.3983**	**0.4369**
CNN-LSTM (thresholding)	**0.4630**	**0.4256**	**0.4424**

Table 5. An example of test utterances annotated by different models.

No.	Utterances	Speech Act Category (Attribute)	
		CNN	**CNN-LSTM**
1	嗯。 uh-huh .	FOL (INFO)	FOL (ACK)
2	嗯。 uh-huh .	FOL (ACK)	RES (CONFIRM)
3	我说她们在新加坡的主要交通工具是什么? i said , what 's the main means of transport them in Singapore?	QST (WHAT)	QST (INFO)
4	政府 国家给你们出的吗? the government will give you out of the country?	QST (INFO)	QST (CONFIRM)
5	我们这样谈话的。 we talk like this.	FOL (PREFERENCE)	FOL (INFO)
6	啊，不贵 ah , it 's not expensive.	FOL (INFO)	FOL (POSITIVE)

decoder must be robust enough (i) to detect the semantic meaning from utterances regardless of the specific word order – *propositional meaning,* and (ii) to understand the intention when a speaker utters them in the middle of dialogues – *pragmatic information*.

With these preliminaries in place Xu et al.'s model, which is grounded on CNN, overlooked the importance of pragmatic information, although it has specialized capability of extracting out of the necessary information from the current utterance. Our model has concatenated RNN to CNN before the soft-max operation so that it can retrieve the contextual meaning from the previous utterances in the dialogues and build more complete concept structure.

As illustrated in Table 5, our proposed CNN-LSTM model could correctly predict the labels which are supposed to be annotated. Consider the example of an utterance 嗯 '*uh-huh'*. If we only have a look this utterance itself, there is no way to disambiguate the meaning between acknowledge and confirmation. In the case of the example 4, since it has already uttered in the previous turn and both two participants of this dialogue know this fact, the confirmational phrase such as 'you meant … is it right?' is omitted to avoid the redundancy. This pragmatic information is stored in the long-term memory and utilized to understand the genuine meaning of corresponding utterance.

7. Conclusion

This paper has extended the work of Xu et al. on the SLU task of DSTC5, and has presented a semantic decoder using deep neural networks. We have compared different models for combining a threshold predictor and long short-term memory. Our concise model proved its efficiency by outperforming the models including the baseline on the DSTC5 and those submitted in the challenge. We observed that understanding the semantic meaning and building a concept structure of a certain utterance is effectively obtained by utilizing both current and previous information. It addresses that the architecture of our concatenated CNN-LSTM model may be applicable to other types of dialogues corpora for achieving SLU.

Acknowledgement

This work was supported by the National Research Council of Science & Technology (NST) grant by the Korea government (MSIP) (No. CRC-15-04-KIST). We sincerely thank two anonymous reviewers and external readers, and the audiences of 2017 Bigcomp conference for their detailed comments. All errors and misrepresentations, if any, are solely ours.

References

[1] G. Xu, H. Lee, M. Koo, J. Seo, "Convolutional Neural Network using a Threshold Predictor for Multi-label Speech Act Classification" in Big Data and Smart Computing (BigComp), 2017 IEEE International Conference, pp. 126-130, 2017.

[2] S. Kim, L. F. D'Haro, R. E. Banchs, J. D. Williams, M. Henderson, K. Yoshino, "The Fifth Dialog State Tracking Challenge" in Proceedings of the 2016 IEEE Workshop on Spoken Language Technology (SLT), pp. 511-517, 2016.

[3] S. Kim, L. F. D'Haro, R. E. Banchs, J. D. Williams M. Henderson, K. Yoshino, Dialog State Tracking Challenge 5 handbook v3.0, 2016.

[4] T. Ushio, H. Shi, M. Endo, K. Yamagami, N. Horii, "Recurrent Convolutional Neural Networks for Structured Speech Act Tagging" in Proceedings of the 2016 IEEE Workshop on Spoken Language Technology (SLT), pp. 518-524, 2016.

[5] Y. Kim, "Convolutional Neural Networks for Sentence Classification" in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1746-1751, 2014.

[6] P. Wang, J. Xu, B. Xu, C.L. Liu, H. Zhang, F. Wang and H. Hao, "Semantic Clustering and Convolutional Neural Network for Short Text Categorization" in Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics(ACL), pp. 352-357, 2015.

[7] R. Johnson, and T. Zhang, "Semi-supervised convolutional neural networks for text categorization via region embedding" In Advances in neural information processing systems, pp. 919-927, 2015.

[8] Y. Zhang, B. C. Wallace, "A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification" in arXiv preprint arXiv:1510.03820, 2016.

[9] T. Mikolov, M. Karafiát, L. Burget, J. Cernocký and S. Khudanpur, "Recurrent Neural Network based Language Model", In Proceedings of Interspeech, 2010.

[10] M. Sundermeyer, R. Schluter, and H. Ney, "LSTM neural networks for language modeling", In Proceedings of Interspeech, 2010.

[11] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF models for sequence tagging", In arXiv preprint arXiv:1508.01991, 2015.

[12] I. Sutskever, O. Vinyals and Q. V. Le, "Sequence to sequence learning with neural networks", NIPS Technical report, 2014.

[13] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, "Show and tell: Lessions learned from the 2015 mscoco image captioning challenge" IEEE transactions on pattern analysis and machine intelligence, **39**(4), 652-663, 2017.

[14] S. Kim, R. E. Banchs, H. Li, "Exploring Convolutional and Recurrent Neural Networks in Sequential Labelling for Dialogue Topic Tracking" in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 963-973, 2016.

[15] L. M. R. Barahona, M. Gasic, N. Mrkšić, P. H. Su, S. Ultes, T. H. Wen, S. Young, "Exploring Sentence and Context Representations in Deep Neural Models for Spoken Language Understanding" in Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pp. 258-267, 2016.

[16] S. Hochreiter, J. Schmidhuber, "Long short-term memory" Neural computation, **9**(8), 1735-1780, 1997.

[17] A. Elisseeff, J. Weston, "A kernel method for multi-labelled classification" Advances in Neural Information Processing Systems, **14**, 681-687, 2001.

[18] J. Nam, J. Kim, E. L. Mencía, I. Gurevych, and J. Furnkranz, "Large-scale multi-label text classification – revisiting neural networks" Machine Learning and Knowledge Discovery in Databases, Springer, 437-452, 2014.

[19] P. Jeffery, R. Socher, C. D. Manning, "Glove: Global vectors for word representation" in Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP), vol. 14, pp. 1532-1543, 2014.

[20] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, "Distributed representations of words and phrases and their compositionality" Advances in Neural Information Processing Systems, 3111-3119, 2013.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1748-1757 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

A Novel Approach to Sensor and Actuator Integrity Monitoring and Fault Magnitude Reconstruction

Weihua Li[1], Chiman Kwan[2,*]

[1]*Envision Energy Inc., Houston, TX 77002, USA*

[2]*Signal Processing, Inc., Rockville, MD 20850, USA*

ARTICLE INFO

Article history:
Received: 05 June, 2017
Accepted: 29 August, 2017
Online: 15 September, 2017

Keywords:
Integrity monitoring, detection isolation and reconstruction of faulty sensor
robustness to disturbance
dynamical systems

ABSTRACT

This paper proposes a novel approach to sensor and actuator integrity monitoring in a dynamic system. Multiple sensor and actuator faults can be detected. Furthermore, faulty sensors and actuators are isolated by contribution analysis. Most importantly, fault magnitudes can be correctly estimated and failed sensors or actuators outputs can be reconstructed. The proposed approach is robust to disturbances, minimizes false alarms, while achieving maximized sensitivity to any faults. Numerical examples justify correctness and validity of the developed methodology.

1. Introduction

Sensors and actuators are critical components in complex systems. For instance, in an airplane, effective flight control is impossible if sensors and/or actuators are malfunctioning [1-9]. Sensors and actuators can fail, and their failures have a significant impact on the performance of a system. In some applications such as arcing detection in power network [10] and motor corona monitoring [11], sensor and actuator condition plays critical role. In the worst case, the failure even can affect safe operation of the system, leading to a catastrophic event. Sensor failures may include precision degradation, drift, frozen reading, and complete failure [12]. Similarly, actuator failures may include limited range of motion, e.g., valve stiction, and complete failure [13]. It is a challenging task to detect sensor and actuator failures [1-5, 14-15] because sensor outputs contain information from a multitude of sources: normal system outputs, faulty sensor signals, and signals due to noise and external disturbances. Isolating different signals requires utilization of the input-output relationship of the system. Conventional approaches to increasing reliability of aircraft systems include installing redundant sensors, which will add more weights, costs, complexity, and most importantly, additional reliability problems.

Sensor fault detection and diagnosis has been a research topic for decades, and many articles have been published. Interested readers are referred to the survey paper by Frank [16]. While fault detection is relatively easy, isolation of multiple faults is still a challenge to many existing schemes. This paper proposes a novel and systematic approach to sensor and actuator integrity monitoring in systems such as jet engine of an aircraft. The first part is a fault detection scheme that can detect the presence of both sensor and actuator failures with maximized sensitivity. The newly proposed approach is based on analytic redundancy, which means no additional redundant sensors are needed. System input-output relationships (state-space dynamical model of the system) are used to help differentiate different sensor and actuator failures. The second step in the proposed scheme is the fault isolation. A proven technique known as contribution analysis [17-18] that has been widely used in process control industry is utilized for root cause search of sensor faults. Contribution analysis is a practical technique for fault isolation, particularly useful for multiple faults. Moreover, the concept of the structured residuals [19] is applied for actuator fault isolation. Finally, in line with the fault isolation results, a systematic procedure is conducted to reconstruct the faulty sensors/actuators such that the considered system, e.g., an aircraft, can keep functioning with acceptable performance in presence of sensor and/or actuator failures. Extensive simulations using a jet engine model clearly demonstrated the correctness and effectiveness of the proposed algorithms.

*Corresponding Author: Chiman Kwan, Signal Processing, Inc., Rockville, MD 20850, USA | Email: chiman.kwan@signalpro.net

We would like to mention that this paper is an expanded version of our earlier work [20]. More details and simulations have been added to enrich the technical contents. The paper is organized as follows. Section II describes the problem formulation by introducing notations and terminologies. Section III, IV, and V summarize the fault detection, fault isolation, and fault reconstruction algorithms. Section VI includes simulation results of applying the proposed algorithms towards a jet engine. Finally, conclusions are drawn in Section VII and future work is discussed.

2. Problem formulation

Consider the following dynamical system with multiple inputs (MI) and multiple outputs (MO),

$$\begin{aligned} \mathbf{x}_{k+1} &= \mathbf{A}\mathbf{x}_k + \mathbf{B}\mathbf{u}_k + \boldsymbol{\omega}_k \\ \mathbf{y}_k &= \mathbf{C}\mathbf{x}_k + \mathbf{D}\mathbf{u}_k + \boldsymbol{\varepsilon}_k \end{aligned} \tag{1}$$

where $\mathbf{x}_k \in \Re^n$ is the state variable vector, $\mathbf{u}_k \in \Re^l$ the input vector, and $\mathbf{y}_k \in \Re^m$ the output vector, at time k; $\{\mathbf{A},\ \mathbf{B},\ \mathbf{C}, \mathbf{D}\}$ are system matrices with compatible dimensions; $\boldsymbol{\omega}_k$ and $\boldsymbol{\varepsilon}_k$ are process disturbance and measurement noise, respectively. It is assumed that $\boldsymbol{\omega}_k$ and $\boldsymbol{\varepsilon}_k$ are independent random variables with zero mean values and covariance matrices

$$E\left\{\boldsymbol{\omega}_k \boldsymbol{\omega}_k^T\right\} = \mathbf{R}_\omega \in \Re^{n\times n}, \quad E\left\{\boldsymbol{\varepsilon}_k \boldsymbol{\varepsilon}_k^T\right\} = \mathbf{R}_\varepsilon \in \Re^{m\times m}.$$

It is further assumed that $\{\mathbf{A},\ \mathbf{B},\ \mathbf{C}, \mathbf{D}\}$, $\mathbf{R}_\omega$, and $\mathbf{R}_\varepsilon$ are known because they can be either derived or identified from historical data by using the subspace method of identification [21].

Note that (1) represents a failure-free MIMO system. In the presence of sensor failures, $\Delta\mathbf{y}_k \in \Re^{n_y}$ and actuator failures $\Delta\mathbf{u}_k \in \Re^{n_u}$, one has

$$\begin{aligned} \mathbf{x}_{k+1} &= \mathbf{A}\mathbf{x}_k + \mathbf{B}\mathbf{u}_k + \mathbf{B}\mathbf{M}_u\Delta\mathbf{u}_k + \boldsymbol{\omega}_k \\ \mathbf{y}_k &= \mathbf{C}\mathbf{x}_k + \mathbf{D}\mathbf{u}_k + \mathbf{D}\mathbf{M}_u\Delta\mathbf{u}_k + \mathbf{M}_y\Delta\mathbf{y}_k + \boldsymbol{\varepsilon}_k \end{aligned} \tag{2}$$

For a system represented by (2), given available data $\{\mathbf{u}_k, \mathbf{y}_k\}$, it is expected to detect if the magnitude of $\Delta\mathbf{u}_k$ or $\Delta\mathbf{y}_k$ is non-zero. If the answer is yes, the magnitude is estimated and the failure direction matrix $\mathbf{M}_u$ or $\mathbf{M}_y$ is identified. Finally, the failed sensors or actuators will be reconstructed on-line and in real time such that all the sensors are always functional with accepted performance. Herein, it must be addressed that $\{\mathbf{u}_k\}$ comes from controllers. Therefore, it is always available and does not include any actuator failures. On the other hand, $\{\mathbf{y}_k\}$ is measured, also always available, and can contain sensor failures. The variables and terms in this paper are summarized in Appendix.

3. Detection of Faulty Sensors/Actuators

Given an integer s (its determination will be discussed later), performing an algebraic manipulation on (2) leads to

$$\begin{aligned} \mathbf{y}_{k-s+i} = \mathbf{C}\mathbf{A}^i\mathbf{x}_{k-s} + \mathbf{C}\sum_{t=k-s}^{k-s+i-1}\mathbf{A}^{k-s+i-1-t}\left(\mathbf{B}\mathbf{u}_t + \mathbf{B}\mathbf{M}_u\Delta\mathbf{u}_t + \boldsymbol{\omega}_t\right) + \\ \mathbf{D}\mathbf{u}_{k-s+i} + \mathbf{D}\mathbf{M}_u\Delta\mathbf{u}_{k-s+i} + \mathbf{M}_y\Delta\mathbf{y}_{k-s+i} + \boldsymbol{\varepsilon}_{k-s+i} \end{aligned} \tag{3}$$

where $i \in [0, s]$. Note that at $i = 0$,

$$\begin{aligned} \mathbf{y}_{k-s} = \mathbf{C}\mathbf{x}_{k-s} + \mathbf{D}\mathbf{u}_{k-s} + \\ \mathbf{D}\mathbf{M}_u\Delta\mathbf{u}_{k-s} + \mathbf{M}_y\Delta\mathbf{y}_{k-s} + \boldsymbol{\varepsilon}_{k-s} \end{aligned} \tag{4}$$

Define a stacked vector

$$\mathbf{y}_{k-s:k} = \begin{bmatrix} \mathbf{y}_{k-s}^T & \mathbf{y}_{k-s+1}^T & \cdots & \mathbf{y}_k^T \end{bmatrix} \in \Re^{m(s+1)} \tag{5}$$

Similarly, define five more stacked vectors

$$\begin{aligned} &\mathbf{u}_{k-s:k} \in \Re^{l(s+1)},\ \Delta\mathbf{u}_{k-s:k} \in \Re^{n_u(s+1)},\ \Delta\mathbf{y}_{k-s:k} \in \Re^{n_y(s+1)}, \\ &\boldsymbol{\omega}_{k-s:k} \in \Re^{n(s+1)},\ \boldsymbol{\varepsilon}_{k-s:k} \in \Re^{m(s+1)}. \end{aligned} \tag{6}$$

It turns out from (3)-(6) that

$$\begin{aligned} \mathbf{y}_{k-s:k} = \boldsymbol{\Gamma}^s\mathbf{x}_{k-s} + \mathbf{H}^s\mathbf{u}_{k-s:k} + \mathbf{G}^s\boldsymbol{\omega}_{k-s:k} + \\ \mathbf{M}_u^s\Delta\mathbf{u}_{k-s:k} + \mathbf{M}_y^s\Delta\mathbf{y}_{k-s:k} + \boldsymbol{\varepsilon}_{k-s:k} \end{aligned} \tag{7}$$

where

$$\left(\boldsymbol{\Gamma}^s\right)^T = \begin{bmatrix} \mathbf{C}^T & \mathbf{A}^T\mathbf{C}^T & \cdots & \left(\mathbf{A}^s\right)^T\mathbf{C}^T \end{bmatrix} \in \Re^{m(s+1)\times n},$$

$$\mathbf{H}^s = \begin{bmatrix} \mathbf{D} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{CB} & \mathbf{D} & & \vdots \\ \vdots & & \ddots & \\ \mathbf{CA}^{s-1}\mathbf{B} & \mathbf{CA}^{s-2}\mathbf{B} & \cdots & \mathbf{D} \end{bmatrix} \in \Re^{m(s+1)\times l(s+1)},$$

$$\mathbf{G}^s = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{C} & \mathbf{0} & & \vdots \\ \vdots & & \ddots & \\ \mathbf{CA}^{s-1} & \mathbf{CA}^{s-2} & \cdots & \mathbf{0} \end{bmatrix} \in \Re^{m(s+1)\times n(s+1)},$$

$$\mathbf{M}_y^s = \begin{bmatrix} \mathbf{M}_y & \mathbf{0} & \cdots & \mathbf{0} \\ 0 & \mathbf{M}_y & & \vdots \\ \vdots & & \ddots & \\ 0 & \cdots & & \mathbf{M}_y \end{bmatrix} \in \Re^{m(s+1)\times n_y(s+1)},$$

$$\mathbf{M}_u^s = \begin{bmatrix} \mathbf{DM}_u & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{CBM}_u & \mathbf{DM}_u & & \vdots \\ \vdots & & \ddots & \\ \mathbf{CA}^{s-1}\mathbf{BM}_u & \mathbf{CA}^{s-1}\mathbf{BM}_u & \cdots & \mathbf{DM}_u \end{bmatrix} \in \Re^{m(s+1)\times n_u(s+1)}.$$

Herein, the choice of s is discussed. It must be chosen such that $\text{rank}\left\{\boldsymbol{\Gamma}^s\right\} = n$. There can be many candidates for s to meet such a requirement. We just select a proper value that allows us to perform both fault detection and isolation. This will be elaborated later.

With the establishment of (7), a primary residual vector (PRV) can be calculated for fault detection. Select a matrix $\mathbf{W}_{\text{prv}}$ from the null space of $\mathbf{\Gamma}^s$. Multiplying both sides of (7) by $\mathbf{W}_{\text{prv}}$ leads to

$$\mathbf{W}_{\text{prv}}\mathbf{y}_{k-s:k} = \mathbf{W}_{\text{prv}}\mathbf{H}^s\mathbf{u}_{k-s:k} + \mathbf{W}_{\text{prv}}\left(\mathbf{M}_u^s\Delta\mathbf{u}_{k-s:k} + \mathbf{M}_y^s\Delta\mathbf{y}_{k-s:k}\right) + \mathbf{W}_{\text{prv}}\left(\mathbf{G}^s\boldsymbol{\omega}_{k-s:k} + \boldsymbol{\varepsilon}_{k-s:k}\right) \quad (8)$$

where $\mathbf{W}_{\text{prv}}\mathbf{\Gamma}^s = \mathbf{0}$ has been considered.

Define

$$\begin{aligned}\mathbf{e}_{s,k} &\equiv \mathbf{W}_{\text{prv}}\left(\mathbf{y}_{k-s:k} - \mathbf{H}^s\mathbf{u}_{k-s:k}\right)\\ &= \mathbf{W}_{\text{prv}}\left(\mathbf{M}_u^s\Delta\mathbf{u}_{k-s:k} + \mathbf{M}_y^s\Delta\mathbf{y}_{k-s:k}\right) + \\ &\quad \mathbf{W}_{\text{prv}}\left(\mathbf{G}^s\boldsymbol{\omega}_{k-s:k} + \boldsymbol{\varepsilon}_{k-s:k}\right)\end{aligned} \quad (9)$$

as the PRV for fault detection due to the following facts:

a. In the ideal case, i.e., no fault, no noise, and no disturbance, $\mathbf{e}_{s,k} = \mathbf{0}$.

b. Without any faults, $\mathbf{e}^*_{s,k} = \mathbf{W}_{\text{prv}}\left(\mathbf{G}^s\boldsymbol{\omega}_{k-s:k} + \boldsymbol{\varepsilon}_{k-s:k}\right)$ is a moving average (MA) process of noise vectors $\boldsymbol{\omega}_k$ and $\boldsymbol{\varepsilon}_k$, having zero mean value and covariance matrix

$$\mathbf{R}_e^s \equiv \mathbf{W}_{\text{prv}}\left(\mathbf{G}^s\mathbf{R}_\omega^s\left(\mathbf{G}^s\right)^T + \mathbf{R}_\varepsilon^s\right)\mathbf{W}_{\text{prv}}^T \in \Re^{(m(s+1)-n)\times(m(s+1)-n)} \quad (10)$$

where $\mathbf{R}_\omega^s \equiv \mathbf{I}_{s+1}\otimes\mathbf{R}_\omega$, $\mathbf{R}_\varepsilon^s \equiv \mathbf{I}_{s+1}\otimes\mathbf{R}_\varepsilon$, with $\mathbf{I}_{s+1}$ being an identity matrix and $\otimes$ the Kronecker tensor product [30].

c. In the presence of sensor and/or actuator faults, the PRV is

$$\mathbf{e}_{s,k} = \mathbf{W}_{\text{prv}}\left(\mathbf{M}_u^s\Delta\mathbf{u}_{k-s:k} + \mathbf{M}_y^s\Delta\mathbf{y}_{k-s:k}\right) + \mathbf{e}^*_{s,k}$$

which is a random vector with non-zero mean,

$$\mathbf{W}_{\text{prv}}\left(\mathbf{M}_u^s\Delta\mathbf{u}_{k-s:k} + \mathbf{M}_y^s\Delta\mathbf{y}_{k-s:k}\right)$$

and covariance $\mathbf{R}_e^s$ [22]. Therefore, fault detection is to check if the mean value of the PRV is zero.

So far, no discussion on the calculation of $\mathbf{W}_{\text{prv}}$ has been conducted yet. Considering the dimension and rank of $\mathbf{\Gamma}^s$, it can be figured out that $\mathbf{W}_{\text{prv}}$ has $(s+1)m-n$ rows and $(s+1)m$ columns. Detailed steps to calculate $\mathbf{W}_{\text{prv}}$, which are lengthy but straightforward, are documented in [19].

Assume that $\boldsymbol{\omega}_k$ and $\boldsymbol{\varepsilon}_k$ are Gaussian distributed. It can be easily shown that $\overline{\mathbf{e}}_{s,k} = \left(\mathbf{R}_e^s\right)^{-1/2}\mathbf{e}_{s,k}$ follows a zero mean Gaussian distribution with an identity covariance matrix [22]. Therefore, define

$$d_{s,k} = \overline{\mathbf{e}}_{s,k}^T\overline{\mathbf{e}}_{s,k} \quad (11)$$

as the fault detection index, which follows a χ^2 distribution with degree of freedom $m(s+1)-n$ [22]. With a pre-determined confidence limit, if $d_{s,k}$ exceeds the limit, it indicates that some sensors and or actuators are faulty. Otherwise, everything is normal.

At the end of this section, the issue of disturbance decoupling is discussed. It has been assumed that $\boldsymbol{\omega}_k$ is Gaussian distributed. In reality, this assumption sometimes may be not valid. If this is the case, one can try to decouple the effect of disturbance as much as possible.

Note that in the PRV represented by (9), the disturbance contributed term is $\mathbf{W}_{\text{prv}}\mathbf{G}^s\boldsymbol{\omega}_{k-s:k}$. Perform a singular value decomposition (SVD) on $\mathbf{G}^s$, and design $\mathbf{W}_{\text{prv}}$ such that each of its rows is orthogonal to the principal left singular vectors related to few largest singular values of $\mathbf{G}^s$. As a consequence, effect of $\boldsymbol{\omega}_k$ can be decoupled partially. How much percentage of $\boldsymbol{\omega}_k$ can be decoupled? This depends on the dimension of $\mathbf{W}_{\text{prv}}$ and the rank of $\mathbf{G}^s$. Intuitively, the lower the rank of $\mathbf{G}^s$ is, the easier to decouple $\boldsymbol{\omega}_k$ from the PRV. Even if $\boldsymbol{\omega}_k$ cannot be removed from the PRV completely, it can be decoupled to some extent by using the above-mentioned SVD [23].

4. Isolation of Faulty Sensors/Actuators

While many methods have been available, here contribution analysis [18] is proposed for sensor fault isolation due to its practicality and simplicity. Computationally, the PRV is $\mathbf{e}_{s,k} \equiv \mathbf{W}_{\text{prv}}\left(\mathbf{y}_{k-s:k} - \mathbf{H}^s\mathbf{u}_{k-s:k}\right)$, which is contributed by $\mathbf{y}_{k-s:k}$ and $\mathbf{u}_{k-s:k}$. In the presence of sensor faults, the PRV will deviate from its nominal values. Consequently, by analyzing the contributions in $\mathbf{e}_{s,k}$ from each element of $\mathbf{y}_{k-s:k}$, the faulty sensors can be identified.

However, since $\mathbf{u}_{k-s:k}$ does not include any actuator failures, contribution analysis cannot be used for isolation of actuator failures. We will propose the structured residual-based approach towards actuator fault isolation. Therefore, combining contribution analysis [17-18] and the structured residual-based approach [19] will develop a comprehensive fault isolation methodology for both sensor and actuator failures.

4.1. Contribution Analysis for Isolation of Faulty Sensors

We illustrate how to identify faulty sensors by contribution analysis first. It turns out from (5) and the first line of (9) that

Table 1. Incidence matrix showing actuator fault isolation logic.

SRVs	1st actuator	2nd actuator	…	i^{th} actuator
SRV_1	0	1	1	1
SRV_2	1	0	1	1
⋮	1	⋱	⋮	⋮
SRV_l	1	1	1	0

$$\mathbf{e}_{s,k} \equiv \sum_{i_1=1}^{m}\sum_{i_2=1}^{s} \mathbf{W}_{\text{prv}}\left(:, i_1 + i_2 m\right) \mathbf{y}_{k-s:k}\left(i_1 + i_2 m\right) - \mathbf{W}_{\text{prv}} \mathbf{H}^s \mathbf{u}_{k-s:k} \tag{12}$$

where $\mathbf{W}_{\text{prv}}(:, i)$ is the ith column of matrix $\mathbf{W}_{\text{prv}}$, and $\mathbf{y}_{k-s:k}(i)$ is the ith element of stacked vector $\mathbf{y}_{k-s:k} \in \Re^{m(s+1)}$.

If the i_1^{th} sensor has a failure for any $1 \le i_1 \le m$, the failure-associated term in $\mathbf{e}_{s,k}$ is a vector:

$$\mathbf{f}_{k,i_1} = \sum_{i_2=1}^{s} \mathbf{W}_{prv}(:, i_1 + mi_2)\mathbf{y}_{k-s:k}(i_1 + mi_2) \cdot$$

Moreover, define

$$\overline{\mathbf{f}}_{k,i_1} = \sum_{i_2=1}^{s} \mathbf{W}_{prv}(:, i_1 + mi_2)\overline{\mathbf{y}}_{k-s:k}(i_1 + mi_2)$$

where $\overline{\mathbf{y}}_{k-s:k}(i_1 + mi_2)$ is the scaled counterpart of $\mathbf{y}_{k-s:k}(i_1 + mi_2)$, i.e., mean-centered and divided by its standard deviation.

If a sensor failure begins to occur at time instance k_{f_1} and ends at k_{f_2}, then one calculates the absolute average value, i.e.,

$$\left|\overline{\mathbf{f}}_{k_{f1}:k_{f2},i_1}\right| = \left|\frac{1}{t_{f_2} - t_{f_1} + 1} \sum_{k=k_{f1}}^{k_{f2}} \overline{\mathbf{f}}_{k,i_1}\right| \tag{13}$$

for all $i_1 \in [1, m]$, obtaining m magnitude values.

Any sensor that has the largest magnitude defined by (13) can be identified as faulty. For example, if Sensor 5 has the largest magnitude, then it can be inferred that it is faulty. As a result, the failure direction matrices for the sensor is $\mathbf{M}_y = \mathbf{I}_m(:, 5)$.

4.2. Isolation of Faulty Actuators

In order to isolate faulty actuators, one can also generate a set of structured residual vectors (SRVs) in line with a predetermined isolation logic [18]. For simplicity, assume that at each time, only a single actuator can be faulty. Under this assumption, for the system represented by (2), we design a set of l SRVs such that the ith SRV is insensitive to the ith actuator fault, while of maximized sensitivity to faults in other $l-1$ actuators for $i \in [1, l]$. Sensitivity of all the designed l SRVs to actuator faults can be described by the following incidence matrix shown in Table 1.

In the above incidence matrix shown in Table 1, '0'/'1' means insensitivity/maximized sensitivity of a SRV to an actuator fault. For instance, in presence of a fault, if SRV_i is not affected by the fault, while all the other SRVs are affected, then the ith actuator is faulty.

Mathematically, the ith SRV can be calculated by

$$\begin{aligned} \mathbf{r}_{s,k}^i &= \mathbf{W}_{\text{srv}}^i(\mathbf{y}_{k-s:k} - \mathbf{H}^s \mathbf{u}_{k-s:k}) \\ &= \mathbf{W}_{\text{srv}}^i \boldsymbol{\Gamma}^s \mathbf{x}_{k-s} + \mathbf{W}_{\text{srv}}^i \left(\mathbf{M}_u^s \Delta\mathbf{u}_{k-s:k} + \mathbf{M}_y^s \Delta\mathbf{y}_{k-s:k}\right) + \\ &\quad \mathbf{W}_{\text{srv}}^i \left(\mathbf{G}^s \boldsymbol{\omega}_{k-s:k} + \boldsymbol{\varepsilon}_{k-s:k}\right) \\ &= \mathbf{W}_{\text{srv}}^i \left(\mathbf{M}_u^s \Delta\mathbf{u}_{k-s:k} + \mathbf{M}_y^s \Delta\mathbf{y}_{k-s:k}\right) + \mathbf{r}_{\text{srv}}^{i,*} \end{aligned} \tag{14}$$

where $\mathbf{r}_{s,k}^{i,*} = \mathbf{W}_{\text{srv}}^i \left(\mathbf{G}^s \boldsymbol{\omega}_{k-s:k} + \boldsymbol{\varepsilon}_{k-s:k}\right)$ is fault free, and the matrix $\mathbf{W}_{srv}^i$ is designed such that $\mathbf{x}_{\text{k-s}}$ and the following associated $\{\Delta\mathbf{u}_{k-s:k}(i+l), \Delta\mathbf{u}_{k-s:k}(i+2l), \cdots \Delta\mathbf{u}_{k-s:k}(i+sl)\}$ are excluded from $\mathbf{r}_{s,k}^i$. This indicates that

$$\mathbf{W}_{\text{srv}}^i [\boldsymbol{\Gamma}^s \ \mathbf{M}_u^s(:, i) \ \mathbf{M}_u^s(:, i+l) \ \mathbf{M}_u^s(:, i+2l) \ \cdots \mathbf{M}_u^s(:, i+sl)] = \mathbf{0} \tag{15}$$

for all $i \in [1, l]$. Design of $\mathbf{W}_{\text{srv}}^i$ is documented in [19]. Similar to (11),

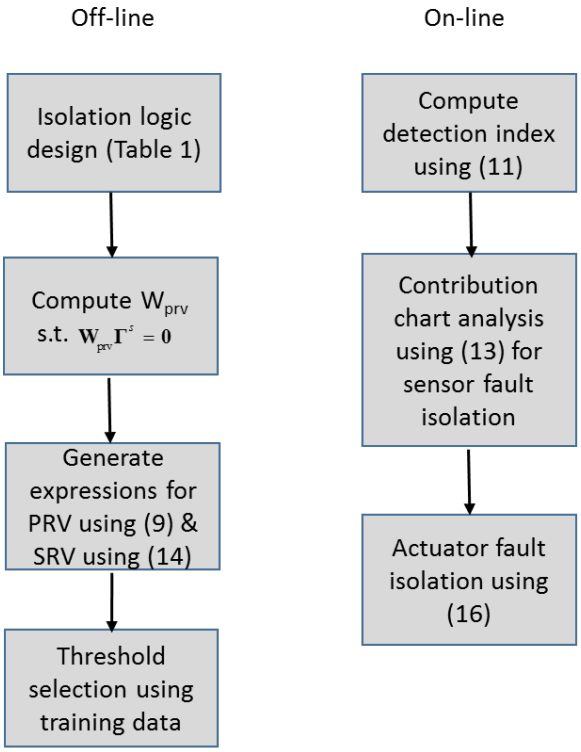
$$I_{s,k}^i = (\mathbf{r}_{s,k}^i)^T (\mathbf{R}_s^{s,i})^{-1} \mathbf{r}_{s,k}^i \tag{16}$$

is defined as the actuator fault isolation index, where $\mathbf{R}_e^{s,i}$ is the covariance of $\mathbf{r}_{s,k}^{i,*}$ for $i \in [1, l]$.

At the end of this section, hereafter we summarize the steps to perform detection and isolation of sensor and actuator faults:

1. Develop an isolation logic under an assumed fault pattern.
2. Design the PRV and a set of SRVs in line with the pre-determined isolation logic.
3. From a set of training data, calculate a sequence of PRV and sequences of SRVs. Calculate their respective thresholds.
4. Check if the detection index $d_{s,k} = \overline{\mathbf{e}}_{s,k}^T \overline{\mathbf{e}}_{s,k}$ given by (11) is triggering an alarm.
5. Use contribution analysis to isolation sensor faults.
6. Use the SRVs to isolate actuator faults.

The above procedures are detailed also in the following flow chart. We divide the procedures into off-line, which is the preparation phase, and the online parts.



5. Reconstruction of Faulty Sensors and Actuators

After identification of faulty sensors/actuators, one has to estimate the fault magnitudes first. Then, from the estimated fault value, faulty sensors can be reconstructed.

5.1. Estimation of Fault Magnitudes

It can be derived from

$$\mathbf{e}_{s,k} = \mathbf{W}_{\text{prv}}\left(\mathbf{M}_u^s \Delta\mathbf{u}_{k-s:k} + \mathbf{M}_y^s \Delta\mathbf{y}_{k-s:k}\right) + \mathbf{e}_{s,k}^*$$

that

$$\mathbf{e}_{s,k}^* = \mathbf{e}_{s,k} - \mathbf{W}_{\text{prv}}\begin{bmatrix}\mathbf{M}_u^s & \mathbf{M}_y^s\end{bmatrix}\begin{bmatrix}\Delta\mathbf{u}_{k-s:k}\\ \Delta\mathbf{y}_{k-s:k}\end{bmatrix} \quad (17)$$

With known $\mathbf{M}_u^s$ and $\mathbf{M}_y^s$, in line with [12], one can estimate $\begin{bmatrix}\Delta\mathbf{u}_{k-s:k}\\ \Delta\mathbf{y}_{k-s:k}\end{bmatrix} \in \Re^{(s+1)(n_u+n_y)}$ by minimizing

$$\left\{\mathbf{e}_{s,k} - \mathbf{W}_{\text{prv}}\begin{bmatrix}\mathbf{M}_u^s & \mathbf{M}_y^s\end{bmatrix}\begin{bmatrix}\Delta\mathbf{u}_{k-s:k}\\ \Delta\mathbf{y}_{k-s:k}\end{bmatrix}\right\}^T \mathbf{R}_{s,e}^{-1}\left\{\mathbf{e}_{s,k} - \mathbf{W}_{\text{prv}}\begin{bmatrix}\mathbf{M}_u^s & \mathbf{M}_y^s\end{bmatrix}\begin{bmatrix}\Delta\mathbf{u}_{k-s:k}\\ \Delta\mathbf{y}_{k-s:k}\end{bmatrix}\right\}$$

giving

$$\left(\mathbf{W}_{\text{prv}}[\mathbf{M}_u^s \;\; \mathbf{M}_y^s]\right)^T \mathbf{R}_{s,e}^{-1}\left(\mathbf{W}_{\text{prv}}[\mathbf{M}_u^s \;\; \mathbf{M}_y^s]\right)\begin{bmatrix}\Delta\mathbf{u}_{k-s:k}\\ \Delta\mathbf{y}_{k-s:k}\end{bmatrix} = \left(\mathbf{W}_{\text{prv}}[\mathbf{M}_u^s \;\; \mathbf{M}_y^s]\right)^T \mathbf{R}_{s,e}^{-1}\mathbf{e}_{s,k} \quad (18)$$

If the matrix $\left(\mathbf{W}_{\text{prv}}[\mathbf{M}_u^s \;\; \mathbf{M}_y^s]\right)^T \mathbf{R}_{s,e}^{-1}\left(\mathbf{W}_{\text{prv}}[\mathbf{M}_u^s \;\; \mathbf{M}_y^s]\right)$ is of full rank, one can get the least squares (LS) solution to $\begin{bmatrix}\Delta\mathbf{u}_{k-s:k}^T & \Delta\mathbf{y}_{k-s:k}^T\end{bmatrix}^T$ easily. If the matrix is of rank deficiency, one can apply a latent variable-based approach, such as the partial least squares (PLS), to estimate the fault magnitude [24].

5.2. Reconstruction of Faulty Sensors

With estimated fault magnitudes, one can correct measurements of the faulty sensors. The measurements in faulty sensors are affected both by $\Delta\mathbf{u}_k$ and $\Delta\mathbf{y}_k$. In order to get the corrected measurements in faulty sensors, one can design a Kalman filter,

$$\begin{aligned}\hat{\mathbf{x}}_{k+1} &= \mathbf{A}\hat{\mathbf{x}}_k + \mathbf{B}\mathbf{u}_k + \mathbf{K}(\mathbf{y}_k - \mathbf{M}_y\Delta\hat{\mathbf{y}}_k - \mathbf{D}\mathbf{M}_u\Delta\hat{\mathbf{u}}_k - \mathbf{C}\hat{\mathbf{x}}_k - \mathbf{D}\mathbf{u}_k)\\ \hat{\mathbf{y}}_k &= \mathbf{C}\hat{\mathbf{x}}_k + \mathbf{D}\mathbf{u}_k\end{aligned} \quad (19)$$

where $\hat{\mathbf{y}}_k$ is the corrected measurement of faulty sensors.

6. Numerical Examples

The proposed methodology for sensor/actuator fault detection, isolation, and reconstruction is applied to a gas turbine jet engine control system [25].

6.1. The Simulated System

The gas turbine engine can be described essentially as a heat engine that uses atmospheric air as a working medium to generate propulsive thrust and mechanical power. The central unit of the mechanical arrangement comprises two main rotating parts, the compressor and the turbine. The control system has the function of coordinating the main burner fuel flow and the propelling exhaust nozzle.

The thermodynamic model of the engine is in continuous-time domain. It has 17 state variables, including pressures, air and gas mass flow rates, shaft speeds, absolute temperatures, and static pressure. This is a highly nonlinear dynamic structure that has grossly different steady-state operation over the entire range of spool speeds, flow rates, and nozzle areas. The linearized 17th model in the continuous-time domain is used here for study. The nominal operating point is set at 70% of the demand high spool speed (N_H). For practical reasons and convenience of design, a fifth-order model is employed [25] to approximate the 17th-order model as represented by (20), where a sampling period of $\Delta T = 0.026$ s was taken.

Using a reduced-order model to approximate the full-order dynamic system leads to modeling errors. Another problem is that the operating point of the system varies according to realistic running conditions. In general, different operating points correspond to different plant models. In the practical situation, the design of the fault detection and isolation scheme is based on a fixed mode. When the operating point changes, a model-plant mismatch occurs. Therefore, the system is represented by a fifth order state space model with two inputs, five outputs, and disturbance, i.e.

$$\mathbf{x}_{k+1} = \begin{bmatrix} -0.9813 & 7.5320 & -0.5983 & 0.4857 & -0.6979 \\ 0.2838 & -0.0826 & 0.0779 & -0.0617 & 0.0928 \\ -6.8588 & 28.9161 & -2.0561 & 1.6083 & -2.2612 \\ 1.2235 & -5.6607 & 0.4020 & -0.3192 & 0.4141 \\ 13.2662 & -53.4047 & 4.7390 & -3.7710 & 5.3669 \end{bmatrix} \mathbf{x}_k + \begin{bmatrix} 0.0001 & 0.0002 \\ 0.0001 & -0.0000 \\ 0.0032 & 0.0006 \\ 0.0078 & -0.0003 \\ 0.0031 & -0.0015 \end{bmatrix} \mathbf{u}_k + \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \boldsymbol{\omega}_k, \quad (20)$$

$$\mathbf{y}_k = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \mathbf{x}_k + \boldsymbol{\varepsilon}_k$$

where the two inputs are main engine fuel flow rate and the exhaust nozzle area. In addition, the five outputs are compressor shaft speed, two pressure measurements, two temperature measurements, and the demand high spool speed. In addition, $\boldsymbol{\omega}_k$ represents the afore-mentioned modeling errors or model plant mismatch and $\boldsymbol{\varepsilon}_k$ is measurement noise. It is assumed that $\boldsymbol{\omega}_k$ and $\boldsymbol{\varepsilon}_k$ are independently distributed Gaussian white noise sequences, with respective covariance matrices $\mathbf{R}_\omega = 0.001\mathbf{I}_5$ and $\mathbf{R}_\varepsilon = 0.001\mathbf{I}_5$.

Given the model parameters, selecting $s = 5$, we calculate the following matrices $\boldsymbol{\Gamma}^s \in \Re^{30\times 5}$, $\mathbf{H}^s \in \Re^{30\times 12}$, $\mathbf{M}_u^s \in \Re^{30\times 12}$, $\mathbf{M}_y^s \in \Re^{30\times 30}$, and $\mathbf{G}^s \in \Re^{30\times 30}$ in line with the formulas in (7).

We then calculate the matrix, $\mathbf{W}_{\text{prv}}$, to generate the PRV for fault detection. In accordance with [19], $\mathbf{W}_{\text{prv}} \in \Re^{25\times 30}$ is calculated such that each of its rows is orthogonal to $\boldsymbol{\Gamma}^s$ while having maximized covariance with the non-zero singular value related-left singular vectors of $[\mathbf{M}_u^s \quad \mathbf{M}_y^s] \in \Re^{30\times 42}$.

With the calculated $\mathbf{W}_{\text{prv}}$, the PRV is generated by

$$\mathbf{e}_{s,k} \equiv \mathbf{W}_{\text{prv}} \left(\mathbf{y}_{s,k} - \mathbf{H}^s \mathbf{u}_{s,k} \right) \in \Re^{25}.$$

Furthermore, we calculate two more matrices $\mathbf{W}_{\text{srv}}^1 \in R^{20\times 30}$ and $\mathbf{W}_{\text{srv}}^2 \in R^{20\times 30}$ to generate two SRVs

$$\mathbf{r}_{s.k}^i \equiv \mathbf{W}_{\text{srv}}^i \left(\mathbf{y}_{s,k} - \mathbf{H}^s \mathbf{u}_{s,k} \right) \in \Re^{20}$$

for $i = 1, 2$ according to the isolation logic listed in Table 1.

Without any faults, we used (20) to generate 10,000 samples of training data, where $\mathbf{u}_k = [100 \quad 100]^T$, and accordingly we calculated three sequences of $\{\mathbf{e}_{s,k}^*\}$, $\{\mathbf{r}_{s.k}^{1,*}\}$, and $\{\mathbf{r}_{s.k}^{2,*}\}$, respectively. Using such sequences, we estimated the covariance, $\mathbf{R}_e^s \in \Re^{25\times 25}$ of $\{\mathbf{e}_{s,k}^*\}$, $\mathbf{R}_e^{s,1} \in \Re^{20\times 20}$, and $\mathbf{R}_e^{s,2} \in \Re^{20\times 20}$.

We calculated a sequence of fault detection index $\{d_{s,k}\}$ from $\{\mathbf{e}_{s,k}^*\}$ and $\mathbf{R}_e^s$ by using (11). It is depicted below with a confidence limit 48.1444 (the red line in the graph). Moreover, we calculated two sequences of $\{I_{s,k}^i\}$, which are depicted in Figure 2, respectively, with their own confidence limits.

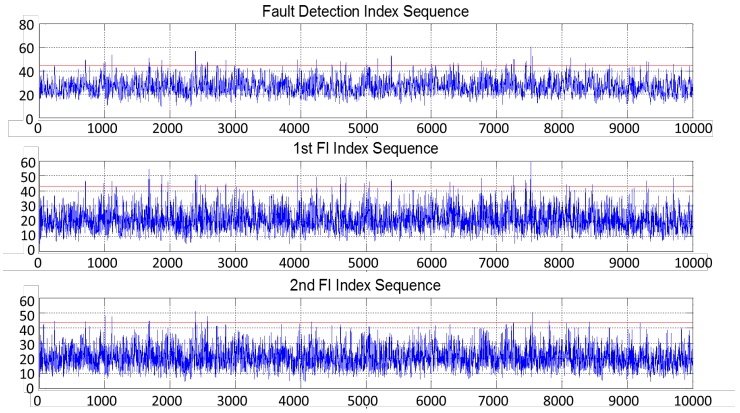


Figure 2. Fault detection and Isolation indices from training data.

$\{d_{s,k}\}$ given a level of confidence 99%. And such a limit will be employed for fault detection in test data.

6.2. Generation of Test Data and FDI Results

Many case studies have been conducted, and few results are presented hereafter.

6.2.1 Case 1: Detection and Isolation of a Faulty Actuator

A time varying fault simulated by

$$\Delta\mathbf{u}_k = \frac{\left(k - k_f\right)}{10000} * \begin{bmatrix} 1\left(k - k_f\right) \\ 0 \end{bmatrix}$$

is introduced to one actuator, where

$$1\left(k - k_f\right) = \begin{cases} 1, & \forall\ k \geq k_f \\ 0, & \forall\ k < k_f \end{cases}$$

and $k_f = 510$. Note that k_f denotes the instant at which a fault begins to occur. In this case, the test data are generated from the following equation:

$$\mathbf{x}_{k+1} = \mathbf{A}\mathbf{x}_k + \mathbf{B}\mathbf{u}_k + \mathbf{B}\mathbf{M}_u\Delta\mathbf{u}_k + \boldsymbol{\omega}_k$$

$$\mathbf{y}_k = \mathbf{C}\mathbf{x}_k + \mathbf{D}\mathbf{u}_k + \mathbf{D}\mathbf{M}_u\Delta\mathbf{u}_k + \boldsymbol{\varepsilon}_k$$

where $\mathbf{u}_k = [100 \quad 100]^T$ as selected before.

From available data sequences $\{\mathbf{u}_k, \mathbf{y}_k\}$, we generated the PRV and two SRVs. At the meanwhile, we calculated the fault detection index sequence $\{d_{s,k}\}$ and two fault isolation index sequences $\{I^1_{s,k}\}$ and $\{I^2_{s,k}\}$. As clearly shown in Figure 3, as time goes by, $\{d_{s,k}\}$ exceeded its limit, indicating that a fault/faults has/have occurred.

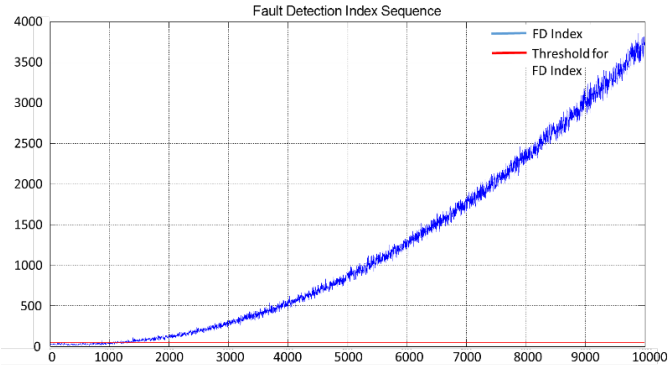


Figure 3. Fault Detection Result from Test Data.

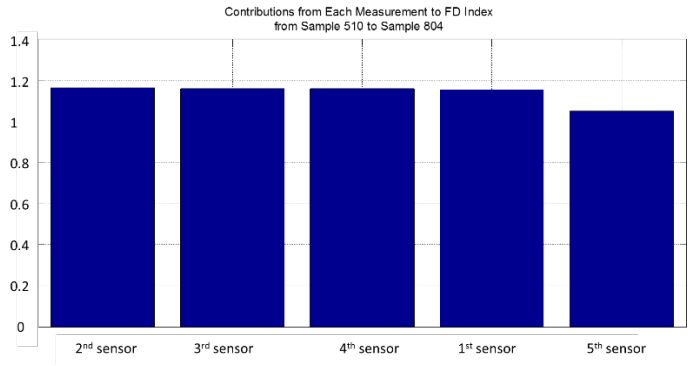


Figure 4. Contribution Analysis Results.

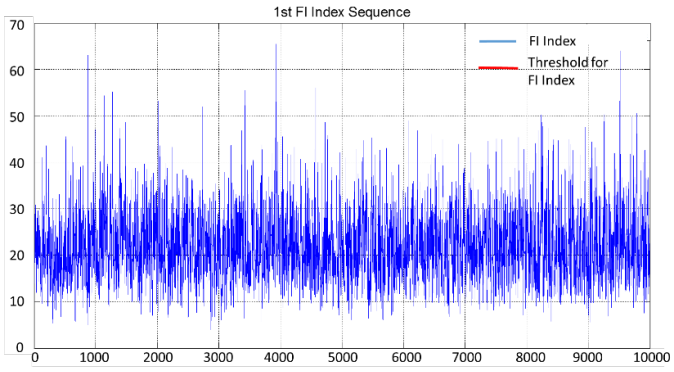


Figure 5. The 1st Fault Isolation Index Sequence Calculated from Test Data.

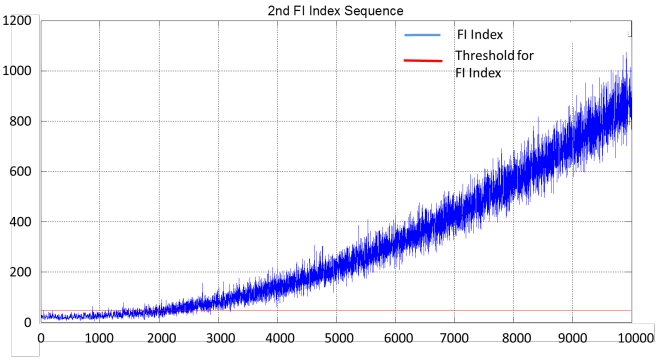


Figure 6. The 2nd Fault Isolation Index Sequence Calculated from Test Data.

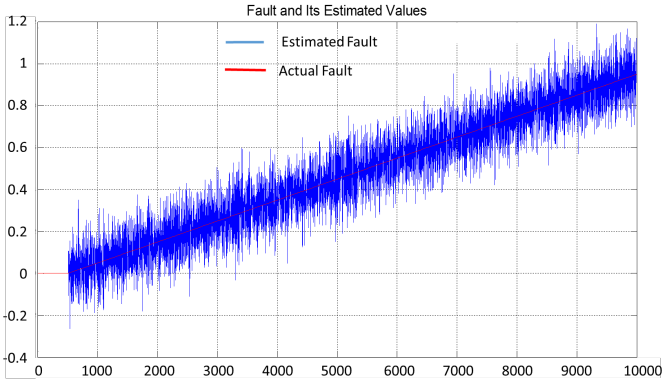


Figure 7. Actual and Estimated Fault Values from Test Data.

A bias with size 1.5 was introduced to a sensor. The fault detection index is displayed in Figure 8, where the $\{d_{s,k}\}$ exceeds its limit at sample 3000, giving $k_{f_1} = 3000$.

Having detected the sensor fault, the next step is to isolate the faulty sensor by conducting the contribution analysis. A bar chart showing the contributions from all five sensors is depicted in Figure 9, where $k_{f_1} = 3000$ and $k_{f_2} = 9995.$.

It is clearly shown that the first output has made the most significant contribution to the sensor failure. Therefore, it can be inferred that the first sensor is faulty.

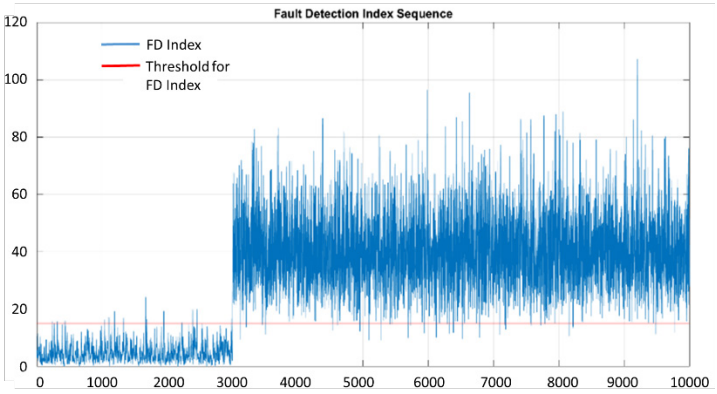


Figure 8. Detection result of a sensor fault

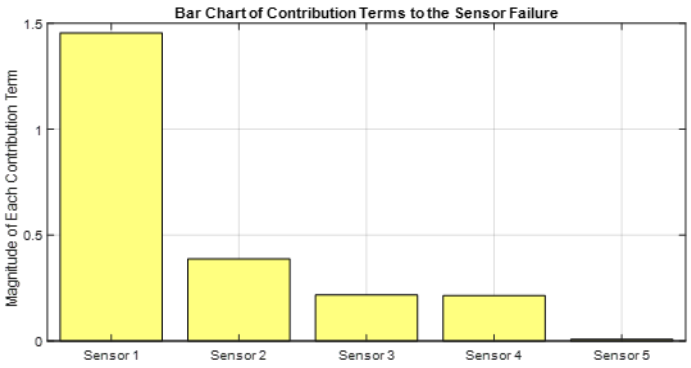


Figure 9. Contribution analysis result of a sensor failure

$p = 0.95$, then one can calculate the probabilities of simultaneous sensor failures. The results are listed in Table 2. For example, the probability that i sensors are faulty simultaneously is

$$\text{Prob}_{i,0.95} = C_5^i p^{5-i}(1-p)^i$$

where C_5^i is the combination of i from 5 for $i \in [1,5]$.

Table 2. Probabilities of simultaneous sensor faults

Number of simultaneously faulty sensors	probability
1	0.2036
2	0.0214
3	0.0011
4	0.0000297
5	3.125×10^{-7}

As shown in Table 2, since the chance for three or more sensors to fail simultaneously is low, one only needs to consider the detection and isolation of two sensor failures for the simulated system. Two faults are introduced to two sensors, respectively. One is a bias with magnitude 0.5, and the other is precision degradation simulated by a random noise signal with zero mean and variance of 0.01. The fault detection result is illustrated in Figure 10.

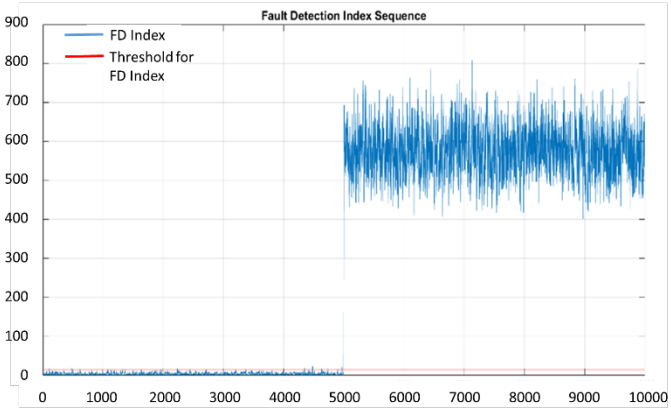


Figure 10. Detection of two simultaneous faults

Contribution analysis is similarly conducted in this case, where $k_{f_1} = 5000$ and $k_{f_2} = 9995$. The associated results are illustrated in Figure 11. In this figure, since Sensor 1 and Sensor 2 have contributed the most to the violation of the fault detection index, it may be concluded that these two sensors are faulty.

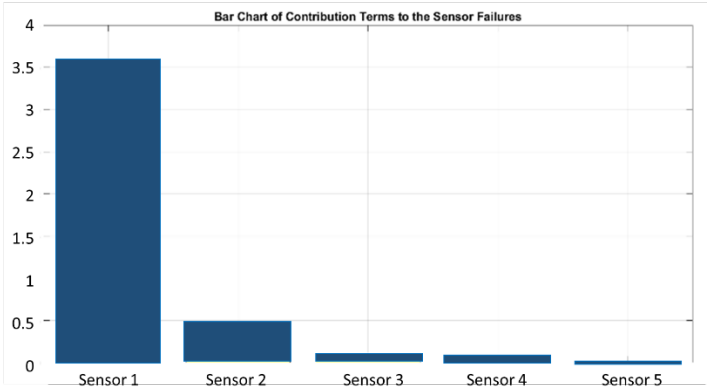


Figure 11. Contribution analysis results of two simultaneous faults

7. Conclusions

Combining analytic redundancy and contribution analysis, a novel methodology for fault detection, isolation, and reconstruction in MIMO dynamical systems has been proposed. It has been applied to a simulated jet engine control system. Numerical results have fully supported the correctness and effectiveness of the developed theory. Even for a slowly evolving incipient fault, the scheme can detect, isolate, and estimate it very effectively. Moreover, different sensor failures are also successfully detected and isolated, further justifying the practicability of the developed methodology.

One future research direction is the integration of our methodology to fault tolerant control systems and apply to power systems [10], aircraft [26], robots [27-28, 31], motors [11,32], planetary rovers [29], and missiles [33,34].

Conflict of Interest

The authors declare no conflict of interest.

References

[1] R. Xu, C. Kwan, "Robust isolation of sensor failures" Asian Journal of Control, **5** (1), 12-23, March 2003. https://doi.org/10.1111/j.1934-6093.2003.tb00093.x

[2] C. Kwan, R. Xu, "A note on simultaneous isolation of sensor and actuator faults" IEEE Trans. Control System Technology, **12** (1), 183-192, 2004. https://doi.org/ 10.1109/TCST.2003.821960

[3] C. Kwan, R. Xu, W. Li, "A New Approach to Sensor Validation" in 21st Power Engineering Symposium, Taipei, Taiwan, 2000.

[4] R. Xu, G. Zhang, X. Zhang, C. Kwan, K. Semega, "Sensor Validation Using Nonlinear Minor Component Analysis" in Third International Symposium on Neural Networks, Lecture Notes in Computer Science, Vol. 3973, 352-357, 2006. https://doi.org/10.1007/11760191_52

[5] J. Cao, C. Kwan, R. Xu, F. Figueroa, "Simultaneous Sensor and Process Fault Diagnostics for Propellant Feed System" in 6th IFAC Symposium on Fault Detection, Supervision and Safety of Technical Processes, 2006.

[6] X. Zhang, R. Xu, C. Kwan, M. Pritchard, L. Haynes, M. Polycarpou, Y. Yang, "Fault Tolerant Formation Flight Control of UAVs" *Int. J of Vehicle Autonomous Systems*, **2** (3/4), 217-235, 2004. https://doi.org/10.1504/IJVAS.2004.006108

[7] M. Polycarpou, X. Zhang, R. Xu, Y. Yang, C. Kwan, "A Neural Network Based Approach to Adaptive Fault Tolerant Flight Control" in *Proc. IEEE Int. Symposium on Intelligent Control*, pp. 61-66, 2004. https://doi.org/10.1109/ISIC.2004.1387659

[8] X. Zhang, M. M. Polycarpou, R. Xu, C. Kwan, "Actuator Fault Diagnosis and Accommodation for Improved Flight Safety" in *Joint IEEE International Symposium on Intelligent Control and Mediterranean Conference on Control and Automation conference*, pp. 640-645, 2005. https://doi.org/10.1109/.2005.1467089

[9] X. Zhang, Y. Liu, R. Rysdyk, C. Kwan, R. Xu, "An Intelligent Hierarchical Approach to Actuator Fault Diagnosis and Accommodation" in *IEEE Aerospace Conference*, 2006. https://doi.org/10.1109/AERO.2006.1656088

[10] J. Zhou, B. Ayhan, C. Kwan, S. Liang, W. Lee, "High performance arcing fault location in distribution networks" *IEEE Trans. Industry Applications*, **48** (3), 1107-1114, 2012. https://doi.org/10.1109/TIA.2012.2190819

[11] C. Kwan, T. Qian, Z. Ren, H. Chen, R. Xu, W.-J. Lee, H. Zhang, J. Sheeley, "A Novel Approach to Corona Monitoring" *Advances in Neural Networks – ISNN 2005: Second International Symposium on Neural Networks, Lecture Notes in Computer Science Volume 3498*, pp 494-500, 2005. https://doi.org/10.1007/11427469_80

[12] J. Qin, W. Li, "Detection, Identification and Reconstruction of Faulty Sensors with Maximized Sensitivity" AIChE J., **45** (9), 1963 – 1976, 1999. https://doi.org/10.1002/aic.690450913

[13] A. Deluca, R. Mattone, "Actuator Failure Detection and Isolation Using Generalized Momenta" in IEEE Conference on Robotics and Automation, pp. 634– 639, September 2003. https://doi.org/10.1109/ROBOT.2003.1241665

[14] X. Zhang, D. Miller, R. Xu, C. Kwan, H. Chen, "A Maximum Entropy based Approach to Fault Diagnosis Using Discrete and Continuous Variables" in 6*th IFAC Symposium on Fault Detection, Supervision and Safety of Technical Processes*, Beijing, 2006. https://doi.org/10.3182/20060829-4-CN-2909.00072

[15] R. Xu, T. Qian, C. Kwan, "An Improved Optimal Pairwise Coupling Classifier" in *International Symposium Neural Network, Advances in Neural Networks – ISNN 2005, Lecture Notes in Computer Science Volume 3497*, pp 32-38 2005. https://doi.org/10.1007/11427445_6

[16] P. Frank, "Fault diagnosis in dynamic systems using analytical and knowledge-based redundancy: A survey and some new results" Automatica, **26** (3), 459-474, 1990. https://doi.org/10.1016/0005-1098(90)90018-D

[17] P. Nomikos, J. MacGregor, "Multivariate SPC charts for monitoring batch processes" Technometrics, **37** (1), 41-59, 1995. https://doi.org/10.2307/1269152

[18] P. Miller, R. E. Swanson, C. E. Heckler, "Contribution plots: a missing link in multivariate quality control" Applied Mathematics and Computer Science, **8** (4), 775-792, 1998.

[19] W. Li, S. Shah, "Structured residual vector based approach to sensor fault detection and isolation" J. of Process Control, **12** (3), 429-443, 2002. https://doi.org/10.1016/S0959-1524(01)00046-4

[20] W. Li, C. Kwan, "A novel approach to sensor and actuator integrity monitoring" in IEEE Conference on Decision and Control, pp. 2140 – 2145, December 2016. https://doi.org/10.1109/CDC.2016.7798580

[21] P. Van Overschee, B. de Moor, "N4SID: Subspace algorithms for the identification of combined deterministic-stochastic systems" **30** (1), 175-93, 1994. https://doi.org/10.1016/0005-1098(94)90230-5

[22] R. Johnson, D. Wichern, Applied Multivariate Statistical Analysis, Prentice Hall, 5th Edition, 2001.

[23] Z. Han, W. Li, S. Shah, "Fault detection and isolation in the presence of process uncertainties" Control Engineering Practice, 13 (5), 587-599, 2005. https://doi.org/10.1016/j.conengprac.2004.04.020

[24] P. Geladi, B. Kowalski, "Partial least-squares regression: A tutorial" Analytica Chimica Acta, 185 (1), 1-17, 1986. https://doi.org/10.1016/0003-2670(86)80028-9

[25] R. J. Patton, J. Chen, "Robust fault detection of jet engine sensor systems using eigenstructure assignment" J. of Guidance, Control, and Dynamics, **15** (6), 1491-1497, 1992. https://doi.org/10.2514/3.11413

[26] M. Ciuryla, Y. Liu, J. Farnsworth, C. Kwan, M. Amitay, "Flight Control Using Synthetic Jets on a Cessna 182 Model," Journal of Aircraft, **44**(2), 642-653, 2007. https://doi.org/10.2514/1.24961

[27] C. Kwan, K. S. Yeung, "Robust Adaptive Control of Revolute Flexible-Joint Manipulators Using Sliding Techniques," Systems and Control Letters, **20**(4), 279-288, 1993. https://doi.org/10.1016/0167-6911(93)90004-P

[28] C. Kwan, "Robust Adaptive Force/Motion Control of Constrained Robots," IEE Proceedings Pt. D, **143**(1), 103-109, 1996. https://doi.org/10.1049/ip-cta:19960090

[29] B. Ayhan, M. Dao, C. Kwan, H. Chen, J. F. Bell III, R. Kidd, "A Novel Utilization of Image Registration Techniques to Process Mastcam Images in Mars Rover with Applications to Image Fusion, Pixel Clustering, and Anomaly Detection," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2017. https://doi.org/10.1109/JSTARS.2017.2716923

[30] Kronecker product, http://www.math.uwaterloo.ca/~hwolkowi/henry/reports/kronthesisschaecke04.pdf

[31] C. Kwan, F. L. Lewis, Y. H. Kim, "Robust Neural Network Control of Flexible-Joint Robots," Asian Journal of Control, pp. 188-197, vol. 1, no.3, September, 1999. https://doi.org/10.1111/j.1934-6093.1999.tb00019.x

[32] P. Ballal, A. Ramani, M. Middleton, C. McMurrough, A. Athamneh, W. J. Lee, C. Kwan, F. L. Lewis, "Mechanical Fault Diagnosis using Wireless Sensor Networks and a Two-Stage Neural Network Classifier," Proc. IEEE Aerospace Conference, 2009. https://doi.org/10.1109/AERO.2009.4839671

[33] C. Kwan, R. Xu, W. Liu, R. Tan, L. Haynes, "Nonlinear Control of Missile Dynamics," in Proc. IEEE Conference on Decision and Control, pp.4685-4691, 1998. https://doi.org/10.1109/CDC.1998.762073

[34] D. Li, J. B. Cruz, Jr., G. Chen, C. Kwan, H. Chang, "A Hierarchical Approach to Multi-Player Pursuit-Evasion Differential Games," in 44th IEEE Conference: Decision and Control, 2005 and 2005 European Control Conference, pp. 5674-5679, 2005. https://doi.org/10.1109/CDC.2005.1583067

Appendix: Nomenclature and Variables

SVD – Singular Value Decomposition

PRV – Primary Residual Vector

MI - Multiple inputs

MO - Multiple outputs

SRV – Structured Residual Vector

MIMO - Multiple inputs Multiple outputs

k : time index

$\mathbf{x}_k \in \Re^n$: State variable vector

$\mathbf{u}_k \in \Re^l$: System input vector at time k

$\mathbf{y}_k \in \Re^m$: System output vector, at time k

$\{\mathbf{A},\ \mathbf{B},\ \mathbf{C}, \mathbf{D}\}$: System matrices with compatible dimensions

$\boldsymbol{\omega}_k$ and $\boldsymbol{\varepsilon}_k$: Process disturbance and measurement noise

$\mathbf{R}_\omega$, and $\mathbf{R}_\varepsilon$: Covariance matrices of $\boldsymbol{\omega}_k$ and $\boldsymbol{\varepsilon}_k$

$\Delta\mathbf{y}_k \in \Re^{n_y}$: Sensor failures

$\Delta\mathbf{u}_k \in \Re^{n_u}$: Actuator failures

$\mathbf{M}_u$ or $\mathbf{M}_y$: Failure direction matrix for actuators or sensors

$\boldsymbol{\Gamma}^s$: Observability gramian

$\mathbf{W}_{prv}$: A matrix from the null space of $\mathbf{\Gamma}^{s}$

$\mathbf{e}_{k}^{s}$: Primary residual vector

$d_{s,k}$: Fault detection index

$\mathbf{r}_{s,k}^{i}$: i[th] structured residual vector

$\mathbf{f}_{k,i_1}$: Term associated with faults. It is used in contribution analysis

ASTES Advances in Science, Technology and Engineering Systems Journal
Vol. 2, No. 3, 1758-1770 (2017)
www.astesj.com
Special Issue on Recent Advances in Engineering Systems 2017

ASTES Journal
ISSN: 2415-6698

A Comparison of Mean Models and Clustering Techniques for Vertebra Detection and Region Separation from C-Spine X-Rays

Anum Mehmood*, M. Usman Akram, Anum Tariq, Ayesha Fatima

Department of Computer Engineering, College of Electrical and Mechanical Engineering, NUST, Rawalpindi, ZIP Code # 45200, Pakistan

ARTICLEINFO

Article history:
Received: 31 May, 2017
Accepted: 20 August, 2015
Online: 20 September, 2017

Keywords :
Vertebra Localization
Detection
Region Separation
Generalized Hough Transform
Clustering
Comparison

ABSTRACT

In Computer Aided Diagnosis (CAD) tools, vertebra localization and detection are the essential steps for the diagnosis of cervical spine injuries. The accurate localization leads to accurate treatment, which is more challenging in case of poor contrast and noisy radiographs. This paper targets c-spine radiographs for the localization of vertebra using different vertebra templates, vertebra detection at each level using two different clustering techniques and gives a comparison between them. Moreover it separates the regions for each individual c-spine. It takes the poor contrast x-ray as input, enhance the contrast and detect the edges of enhanced image. After the edge detection, manually selected Region of Interest (ROI) helps in getting the edges of area covering $C_3 - C_7$ only. These edges along with 4 different template models of vertebra are used for the localization by Generalized Hough Transform (GHT). The results obtained are analyzed visually for the best localization template. Then, on voted points obtained after pruning, two clustering techniques Fuzzy C Means and K-Means are applied separately, to form clusters and centroids for each vertebra. Another part of this paper is to separate vertebra regions. For this, intervertebral points are calculated and then along these points, centroids are rotated using Affine Transformation. It gives parallel lines to vertebrae and joining them gives region for each vertebra. The comparison and testing of proposed technique has been performed using dataset 'NHANES II' publicly accessible at 'The National Library of Medicine', total 150 cervical spine scans are used securing accuracies 93.76%, 84.21% and 83.1% for FCM, K-Means and region separation, respectively.

1 Introduction

Spinal cord proves to be the most important element of human anatomy. It works as a backing for other segments of human body so that one can bend, twist and move easily. If this most essential segment of human structure is damaged in any way, a person cannot even stand up straight. Figure 1 illustrates the division of human spinal cord which consists of 33 individual bones known as vertebra. First 7 cervical vertebra ($C_1 - C_7$) collectively form the neck, next 12 thoracic vertebra ($T_1 - T_{12}$) form the chest, next 5 lumbar vertebra ($L_1 - L_5$) form the lower back, next 4 sacral vertebra ($S_1 - S_4$) form the pelvis and the last 4 fused vertebra form the tail bone of human structure are known as coccyx bones.and fall down [1][2].

In the past few decades, medical imaging has been developed in dramatic way and helping radiologists for the diagnosis and treatment of cervical spine disorders including osteoporosis, cervical spine injuries and trauma [3]. The most common sources for c-spine injuries (CSI) are accident of vehicles and outdoor games. In youngsters age ranging 15-25, hyperextension and vehicles accident where as in senior citizens and osteoarthritic victims, sudden fall is reported as the most common reason of CSI [4]. Latterly, a signifi-

*Anum Mehmood, College of Electrical and Mechanical Engineering, NUST, Peshawar Road, Rawalpindi, +923345804039 & anummehmood81@ce.ceme.edu.pk

https://dx.doi.org/10.25046/aj0203215

cant amount of Traumatic Spinal Cord Injuries (TSCI) has been reported which may become more problematic and critical if the patient is not treated on time by accurate diagnosis [5].

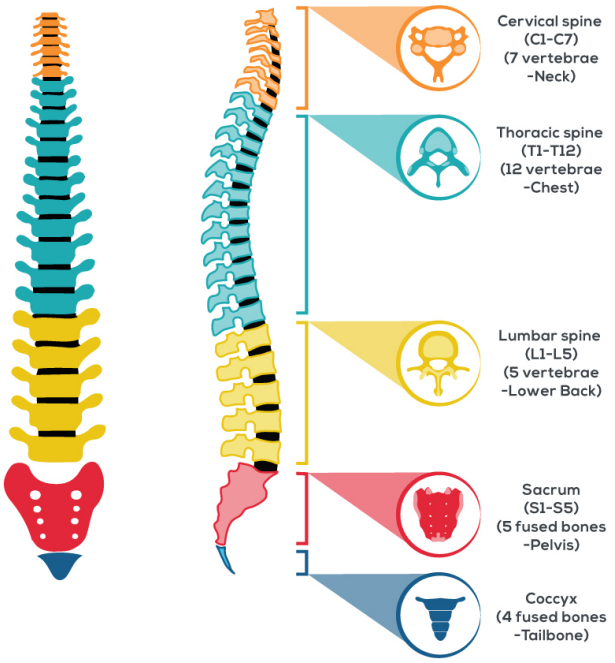


Figure 1: Spinal Column: Anatomical Features

In Computer Aided diagnosis (CAD) tools, vertebra localization and detection at each cervical level are initial and essential steps to be performed and it becomes more difficult when images are low contrast and noisy like X-rays. These steps work significantly in many orthopedics and neurological applications for the diagnosis and treatment of spinal column disorders. The accurate localization and detection leads to accurate treatment of abnormalities. Thus, great success rate in vertebra localization and detection with separated regions for each vertebra would be very helpful for radiologists' community, not only for the diagnosis but also in surgeries. Figure 2 shows the (a) Normal cervical spine x-ray with perfect relationship between posterior, anterior and spinolaminar lines. They are aligned representing a healthy radiograph where as (b) shows loss of alignment at C_6 and C_7 representing unhealthy or fractured X-ray.

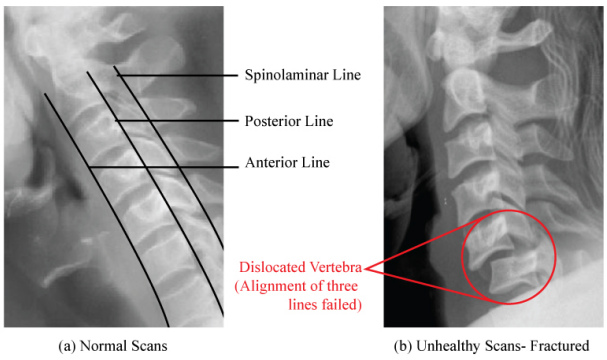


Figure 2: C-Spine Radiographs (a) Healthy (b) Fractured

This figure shows that the fractures can be located using parameters like posterior, anterior and spinolaminar lines by getting the alignment information between them. In literature, many techniques like Generalized Hough Transform (GHT) [6], Active Shape Model (ASM) [7], discrete dynamic contour model (DDCM) [8] and Template Matching [9] have been described for the localization, detection and segmentation of spinal cord using different set of images including magnetic resonance imaging (MRI) and computed tomography (CT) scans. Klinder et al. [10] described a methodology for the extraction of vertebra shape using CT scans. They designed a methodology for the detection, identification and segmentation of vertebra from CT scans. They applied Generalized Hough Transform (GHT) with the combination of adapted triangulation shape for the localization and segmentation of vertebra. They tested the proposed technology on 64 CT Scans with success rate of 70%. Alomari et al. [11] presented a methodology of localization using MRI images. They used two datasets of 50 and 55 scans attaining success rate of 87% and 89.1%. Their proposed model was based on two steps targeting intervertebral discs. Korez et al. [12] presented an automated technique for the detection and segmentation of vertebra and spine using 3D CT Scans. They used interpolation theory for the localization of spine which is further used to locate the individual vertebra. The localization results obtained helped in the segmentation of each vertebra by enhanced shape-constrained deformable model approach. They tested the technique using two CT Scans databases of 50 lumbar and 170 thoracolumbar vertebrae and achieved high success rate. Lecron et al. [13] presented a methodology taking benefit of edge polygonal approximation to locate the vertebra and perform the segmentation using Active Shape Model (ASM). They enhanced the performance of the methodology by parallel computing and heterogeneous architectures for the vertebra extraction. They used X-ray images for the testing of proposed methodology. Larhmam et al. [14] described a technique for the vertebra localization targeting x-rays image. They used a novel combination of GHT and K-means clustering for the localization of vertebra using template matching theory. Benjelloun et al.[15] presented an analysis of vertebra based on segmentation from x-ray images. They described a relative study of two algorithms to segment out vertebra. Lecron et al. [16] presented a model targeting the vertebra localization from X-rays using SVM model and SIFT descriptor. They evaluated their methodology using 50 scans and attained satisfactory success rate of 81.6%. Dong et al. [17] presented a graphical model based vertebra identification which need no training steps for processing. They designed it to automatically detect the total number of visible vertebra in scans and localize them. Larhman et al. [18] described a localization technique by offline training of template model as a preprocessing step and applied GHT to localize the vertebra. Then they performed the post processing of model using adaptive filter and attained success rate

of 89% when tested by 200 vertebrae. Benjelloun et al. [19] proposed a scheme based on polygon regions for the extraction of vertebra boundries and perform segmentation. Later in [20], they presented a technique for the localization and segmentation based on ASM.

This paper is basically an extension of model originally presented in IEEE International Conference on Communication, Computing and Digital Systems (C-CODE 17)[21]. This paper is extended in terms of comparison of template models for better localization and then a comparison and analysis of clustering techniques used for the detection of centroids. A section which is extension to original paper, is dedicated to the region separation for each individual vertebra which can be used further for the segmentation process and will be very helpful for the diagnosis of several c-spine disorders. Section II describes the methodology of proposed technique, Results are analyzed and discussed in Section III and Section IV presents the conclusion of this paper.

2 Methodology

The presented technique comprises five stages to give a comparison of techniques for better localization and detection results along with separated regions for each vertebra. The flow chart of presented model is shown in Figure 4, representing the main steps of model including Preprocessing, Creation of Mean models, Vertebra Localization, Centroid detection and Region separation.

2.1 Preprocessing

Preprocessing is the first and most significant stage of the presented model for attaining high success rate. The steps involved in this stage are contrast enhancement, detection of edges and manual selection of Region of Interest (ROI). These steps are performed to get enough data for our vertebra localization stage.

2.1.1 Contrast enhancement

The input X-ray images are poor contrast and noisy set of images which cannot be processed directly to achieve high accuracy of presented technique. The superimposition of other bony structures and level of brightness at C_6 and C_7 leads to wrong localization results. So 'Contrast Limit Adaptive Histogram Equalization'(CLAHE) [22] is applied to input radiographs so that the contrast can be enhanced for better results. CLAHE works like ordinary adaptive histogram equalization, the only difference is it uses a specific threshold to clip the histogram. For X-ray images, CLAHE is one of the most efficient contrast enhancement algorithm when compared with Gamma Correction (GC) and Histogram Equalizer (HE)[23].

2.1.2 Detection of edges

After the contrast enhancement, the edges of input image are detected using 'Canny Edge Detector' [24]. The reason behind the edge detection is that the GHT algorithm works with edges only. The process of localization takes place using edges. So we detected the edges of input image for any further processing.

2.1.3 Manual selection of Region of Interest (ROI)

The edge detection step gives us the edges of whole input image, but the proposed technique requires edges for the $C_3 - C_7$ only. so in this step we constructed the Region of Interest (ROI) covering the area of lower 5 cervical vertebrae only. Then this ROI is multiplied with the edges to get the edges of this area only. This is the last step of our Preprocessing stage.

2.2 Creation of Mean Models

The creation of mean models is next stage of proposed technology. In this stage, 4 different template models M_1, M_2, M_3 and M_4 are constructed and shown in Figure 3. For the construction of these template models, 25 vertebra images for each template model are created using some drawing tools and then mean image is created using these 25 images. Eq. 1 is used to create mean model of each shape.

$$M_i = \frac{1}{V_t} \sum_{x=0}^{V_t} V_x \tag{1}$$

where 'V_t' presents the total vertebra images (25) created manually and 'V_x' is one vertebra image with 'x' varying from 1 to 25.

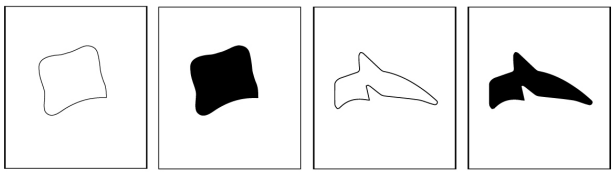


Figure 3: Vertebra Models created: (a) M1 (b) M2 (c) M3 (d) M4

The reason behind using these 4 mean models is just to analyze the results and give a comparison between them. So that one can select best mean model for localization of vertebra securing high accuracy. M1 and M2 represents the body of vertebra whereas M3 and M4 represents complete vertebra shape including body. These created models will be used as a template image for the detection process of GHT. The results are generated using all four template images one by one and then analyzed visually to select the best template image.

2.3 Vertebra Localization(GHT)

The third and most important stage of proposed methodology is Vertebra Localization which helps in the detection of centroids. In this stage, the selected

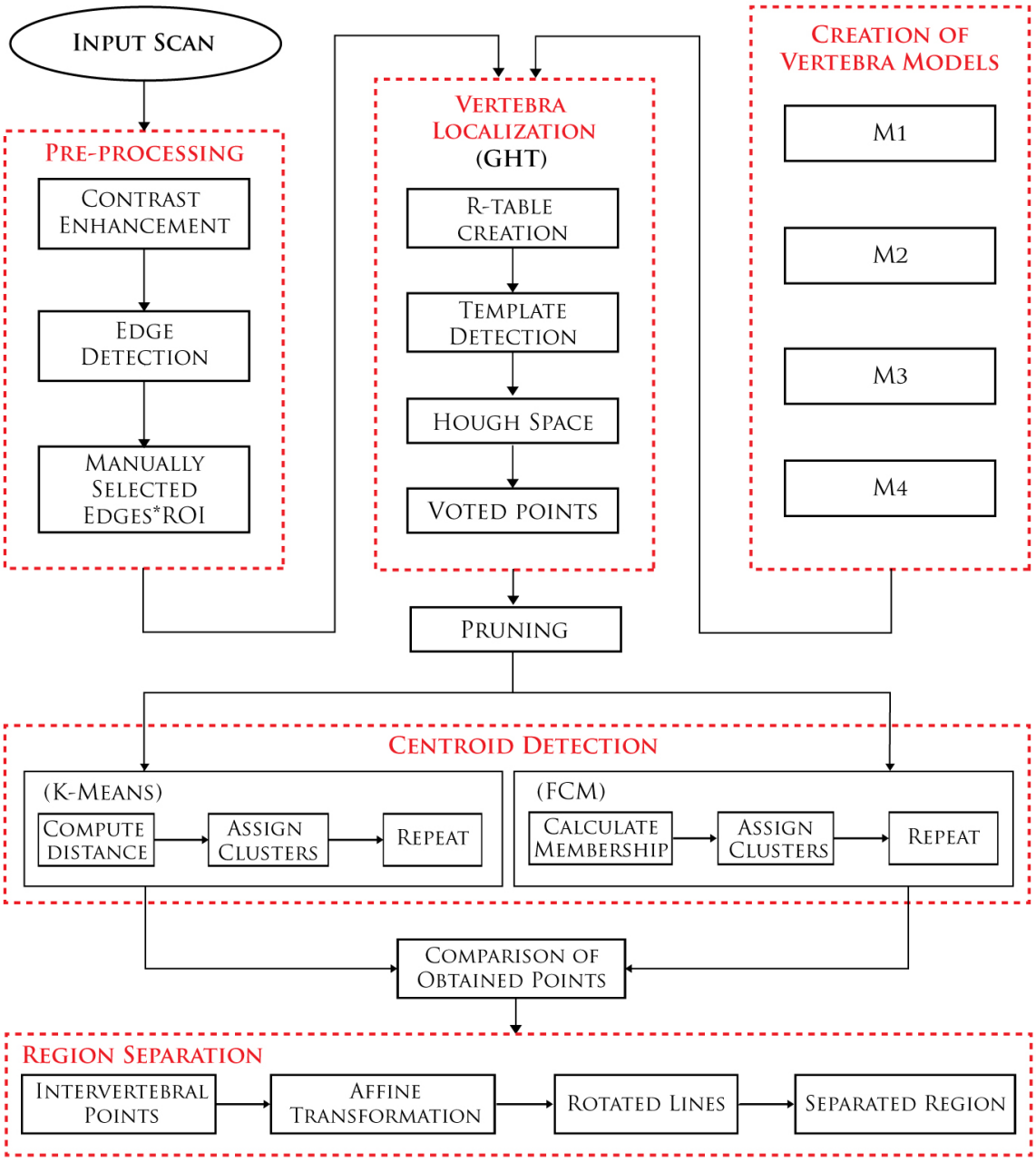


Figure 4: Flow Chart: Main stages of the proposed methodology

edges of Preprocessing stage are used and GHT is applied. Ab initio, Hough Transform [25] was introduced as a technique for the detection of lines, circles, elipse etc following a modified concept by Ballard, known as generalized Hough Transform [6] used in various applications of image processing and computer vision. GHT is a concept used widely for the detection of arbitrary shapes and pattern recognition in all fields of imaging including biomedical imaging. The most positive point of this technique is it is invariant to any kind of transformations. It works with two factors *R-Table* which basically represents the template model of vertebra which is to be detected in the image and second one is *Accumulator*, which is Hough Space and stores the voted points localizing the area of $C_3 - C_7$. GHT is based on the template matching theory and uses the information stored in R-Table for the detection process.

2.3.1 R-Table creation

For the construction of R-table, a reference point (r_x, r_y) is selected. The coordinates of extracted edges helped in this regard. The reference point is selected by taking the mean of all the edges and next, gradient is calculated. Then for each point, gradient direction is computed and saved as a function of gradient. There are more than one entries of gradient direction

for each value of gradient. So, in R-table each entry is the difference between the coordinates of reference point and boundary point w.r.t the direction of gradient. In this way, the template model of vertebra is stored as a R-Table. Table 1 [6] represents the generic form of R-table, where Φ, d_r and β represents gradient, distance and orientation, respectively. The distance and orientation are calculated using Eq. 2 and Eq. 3, respectively.

$$d_r = \sqrt{(r_x - b_x)^2 + (r_y - b_y)^2} \tag{2}$$

$$\beta = tan^{-1} \frac{b_y - r_y}{b_x - r_x} \tag{3}$$

where in both equations (r_x, r_y) is the reference point and (b_x, b_y) is the boundary point.

Orientation Φ	**Position** (r, β)
0	$(r_x, \beta_x)/\Phi_x = 0$
Φ	$(r_x, \beta_x)/\Phi_x = \Delta\Phi$
$2\Delta\Phi$	$(r_x, \beta_x)/\Phi_x = 2\Delta\Phi$
$3\Delta\Phi$	$(r_x, \beta_x)/\Phi_x = 3\Delta\Phi$
...	...

Table 1: R-Table: Generic Form

2.3.2 Template detection and Hough Space

Now, for the detection process an Accumulator is created which represents the Hough Space and stores the points known as Voted points. These points mark the positions containing the template model. So, for each edge point obtained in preprocessing stage, gradient is computed and the corresponding value of R-Table stored as r(Φ) is added into the coordinates of edge point. Then the model check the resultant value if it is within the scope of Accumulator, if yes it increments the location of accumulator defined by this new point. All the voted points obtained by this process give the positioning of vertebra in X-rays. However, few outliers are also calculated by this process which are removed in next step.

2.3.3 Pruning

Pruning [26] is a technique used to remove unwanted data points known as outliers. These outliers may effect the success rate of the proposed model. The window scheme is applied and voted points are removed if they are less than the specified number. Then the voted points obtained after pruning are used for the centroid detection. For each candidate point, a window placed such that candidate point lies at center of window and all points lying inside window are computed. Now, if the count of points is less than a specific number then that candidate point is removed.

2.4 Centroid Detection of vertebra body

After the localization of vertebra from C_3–C_7, the next stage of proposed technique is to detect the centroids of each vertebra using the voted points obtained by all four template models. The centroid detection is performed using two different clustering techniques K-Means and Fuzzy C Means. Each clustering technique forms clusters using the voted points and then gives the centroid for each cluster representing each vertebra ($C_3 - C_7$).

2.4.1 K-Means clustering

K-Means [27] is a clustering algorithm works simply to segment out the data into clusters by calculating the euclidian distance between them. Initially, the cluster's center are selected randomly from obtained voted points, it randomly selected the 5 cluster centers representing 5 cervical vertebra ($C_3 - C_7$). Then for each voted point, euclidean distance is calculated between the voted point and all the cluster centers. And the cluster at minimum distance is assigned to that voted point. This process is performed for all the voted points, completing the first iteration of k-means algorithm. For next iteration, cluster centers are selected by calculating the mean of all the points belonging to the same cluster. Then again distance computation is carried out and clusters are assigned and this process continues until two or more consecutive iterations give same results. The cluster centers obtained in last iteration represents the centroid of each vertebra.

2.4.2 Fuzzy C Means clustering

Fuzzy C-Means (FCM) is another clustering technique developed by Dunn [28] in 1973, then in 1981, Bezdek [29] enhanced the working of this algorithm. It works with the concept of fuzziness and assign clusters in such a way that one data point may belong to more than one cluster at a time. It assign membership to each data point for all the clusters. The range of membership varies from 0 to 1. This factor of fuzziness makes FCM less sensitive to outliers. Initially, FCM selected the 5 cluster centers randomly representing 5 cervical vertebra. Then for each voted point, total 5 memberships are computed, between this point and all the cluster centers. The membership varies from 0 to 1 and is computed using Eq. 4 which basically gives the information about the point that how much this point relates to which cluster. The cluster with greatest membership is the most closes to that voted point.

$$m_{jx} = \frac{1}{\sum_{i=1}^{t} \frac{\|b_j - c_x\|}{\|b_j - c_i\|}^{\frac{2}{t-1}}} \tag{4}$$

where 'i' varies from 1 to 't' representing the total clusters i-e 5, 'c_x' and 'b_j' represents the cluster and voted point for which membership is to be calculated, respectively. This process of assigning memberships creates a level of fuzziness and one voted point belongs to more than one cluster at a time. First iteration is completed, after calculating the membership

for all the voted points and new cluster centroids are selected using Eq. 5

$$c_x = \frac{\sum_{i=1}^{t_p} m_{ix}^j . p_j}{\sum_{i=1}^{t_p} m_{ix}^j} \quad (5)$$

where 'c_x' and 't_p' represents the new cluster computed and total number voted points, respectively. Then in next iteration for these new computed cluster centroids, membership is calculated and this process continues until we get similar centroids in 2 or more consecutive iterations, representing the 5 cervical vertebra from C_3 to C_7.

2.5 Separation of vertebra regions

The separation of each vertebra region is another stage and extension to original work, which allows us to extract the region for vertebra individually using the center points of vertebra. These extracted regions are very helpful in the segmentation process which further leads to the diagnosis of many spinal cord disorders. The steps involved in this stage are calculation of Intervertebral Points and Affine Transformation.

2.5.1 Intervertebral Points

To extract the regions covering the area for each vertebra,intervertebral points are required. The intervertebral points are the location between the two vertebra. In our case we required total 6 intervertebral points, from which 4 are easy to find out by just calculating the mean of centroids from $C_3 - C_7$ which we already have. Eq.6 represents the calculation of these 4 intervertebral points.

$$IV_{C_i} = (C_i + C_{i+1})/2 \quad where\ i = 3, 4, 5, 6 \quad (6)$$

The other two intervertebral points, one before C_3 and second after C_7 couldn't be computed using this equation as it require centroids before and after these points. So, a simple technique using equations 7 and 8 is applied to obtained these two intervertebral points. The difference between C_3 and IV_{C_3} is subtracted from the coordinates of C_3 to get an estimated intervertebral point before C_3 i-e IV_{C_2}

$$IV_{C_2} = C_3 - (IV_{C_3} - C_3) \quad (7)$$

Similar technique using C_7 and IV_{C_6} is applied for IV_{C_7} but instead of subtracting the difference, it is added into the coordinates of C_7 to get points greater than it's location.

$$IV_{C_7} = C_7 + (C_7 - IV_{C_6}) \quad (8)$$

Now, for next step there are total 11 points, 6 intervertebral and 5 centroids in the same sequence as IV_{C_2}, C_3, IV_{C_3},...,C_7, IV_{C_7}.

2.5.2 Affine Transformation

A transformation following the rule of collinearity is known as Affine Transformation [30]. The term collinearity defines that points stay on a line at same distance and the parallel lines remains parallel, before and after the transformation. But in case of parallel lines, collinearity is not applicable on the angles between them. The term Affine Transformation is also called Affinity.

In proposed technique, affine transformation is used to form lines parallel to the vertebra. For each line, 3 points of vertebra are required, and location of two points is transformed in such a way that they stay on same line before and after the transformation, Affine Map helps in this regard. The calculation of Affine Map is shown in Eq. 9 where 'a', 'R' and 'c' are rotational point, rotational matrix and directional matrix, respectively. These three matrices are the general Affine transformation matrix.

$$M = a * R * c \quad (9)$$

The rotational matrix requires angle Θ to rotate the lines accurate enough and make them parallel to each vertebra. To obtain this angle, initially two angles are calculated: one between the first intervertebral point and centroid and second between the centroid and next intervertebral point, then taking average of these two angles and adding 90 into it, gives an exact angle for making lines parallel. Using this angle $\Theta + 90$, Affine Map is constructed using Eq. 9 and multiplied with each point to rotate them along intervertebral points. As a result, total 6 lines are formed covering C_3 to C_7. Then these lines are joined to make separate regions for each vertebra.

3 Results and Evaluation of proposed technique

The proposed technique has been implemented in MATLAB R2013a. The dataset NHANES II [31] used for the evaluation of our proposed technique is publicly accessible at U.S. National Library of Medicine, incorporating 17,000 scans (10,000 cervical spine and 7,000 lumbar) of patients age ranging 25-74, captured in 1976-1980. These X-rays images had been taken under varying scale, orientation and environment. For evaluation of our proposed technique we used 150 cervical spine scans of this dataset, selected on random basis. In literature, several methods has been tested using the same dataset but varying number of images. In these papers, no one ever mentioned the criteria following which they has selected the specific number of images. Neither they defined exactly which images has been selected. So it is clear that everyone is using different images of the same dataset. For fair testing, we selected subset with larger number of images than literature. For the comparison of results, a ground truth is maintained by visually examining each image. For each image, 5 center points

representing each vertebra body are marked manually, targeting C_3 to C_7. These annotated points are stored and used for the testing purposes. A comparison of FCM and K-Means is given in this section using Statistical and Visual results.

3.1 Statistical Results

The parameters used to obtain the statistical results include pixel distance, mean, standard deviation and accuracy.

3.1.1 Pixel distance

Figure 8 shows the pixel distance of 50 images for both K-means and FCM. For better visualization, the result of 50 images is shown in figure but is calculated for all 150 images. Each graph represents each vertebra with red dotted line representing the mean distance of K-means and black dotted line is for FCM, for all 150 images. The difference is calculated between the annotated points and detected centroids, representing how much detected centroid is away from the center point. The mean error lines show that at each vertebra level FCM works better than K-Means, the mean error for FCM is less than that of K-Means, but at level C_7 there is a minor difference between the two. The presence of high brightness and superimposition of other structures make the false detection at this cervical level.

3.1.2 Mean

Figure 5 shows the calculated mean of the distance. The green bar shows the mean of k-means which is larger than FCM at every level, represented by blue bar. This parameter also marked FCM as a better

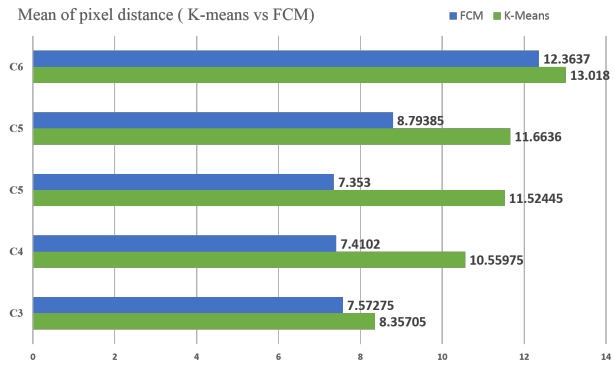


Figure 5: Mean of calculated distance: K-Means vs FCM

clustering technique for centroid detection of vertebra from X-rays. The eq. 10 is used for the calculation of mean.

$$Mean_{error} = \frac{\sum_{i=1}^{T} d_i}{T} \tag{10}$$

Where 'd_i' represents the distance at each cervical and 'T' represents total images used.

3.1.3 Standard Deviation

The standard deviation is another statistical parameter of our presented methodology. Figure 6 shows the calculated standard deviation in both cases.

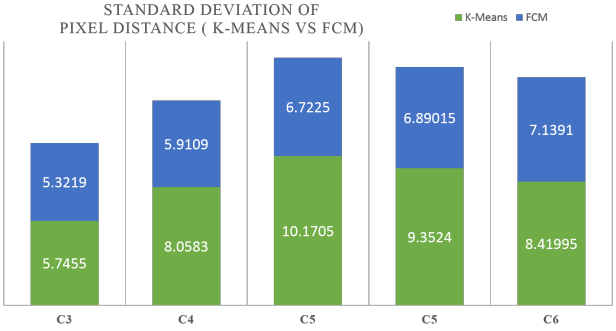


Figure 6: Standard Deviation: K-Means vs FCM

Each level shows less variation in case of FCM than K-means, making FCM better than the K-Means. The standard deviation is calculated using Eq. 11

$$SD = \sqrt{\frac{\sum_{i=1}^{T}(d_i - Mean_{error})^2}{T-1}} \tag{11}$$

where 'd_i', '$Mean_{error}$' and 'T' represents distance, calculated mean and total number of images, respectively.

3.1.4 Accuracy

The accuracy of presented technique is measured at two levels *Visual* and *Experimental*. In case of visual accuracy, labels are assigned to all 150 images at each cervical label.

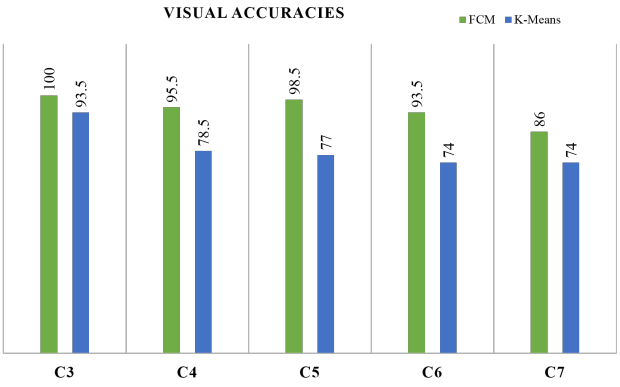


Figure 7: Visual accuracies reported at each cervical level

If the detected vertebra is correct label '1' is assigned and in case of wrong detection label '0' is assigned. The presented model consider the vertebra detection as correct if it is within the body of vertebra.So in visual examination, this point is followed for labeling process and each vertebra has been given label '1' if it is inside or on the boundary of vertebra body and '0' is it is outside the vertebra body. Using this criteria all the images has been assigned total 5 labels representing 5 vertebra and then using these labels overall accuracy of 94.7% and 79.4% are measured for FCM and K-Means, respectively.

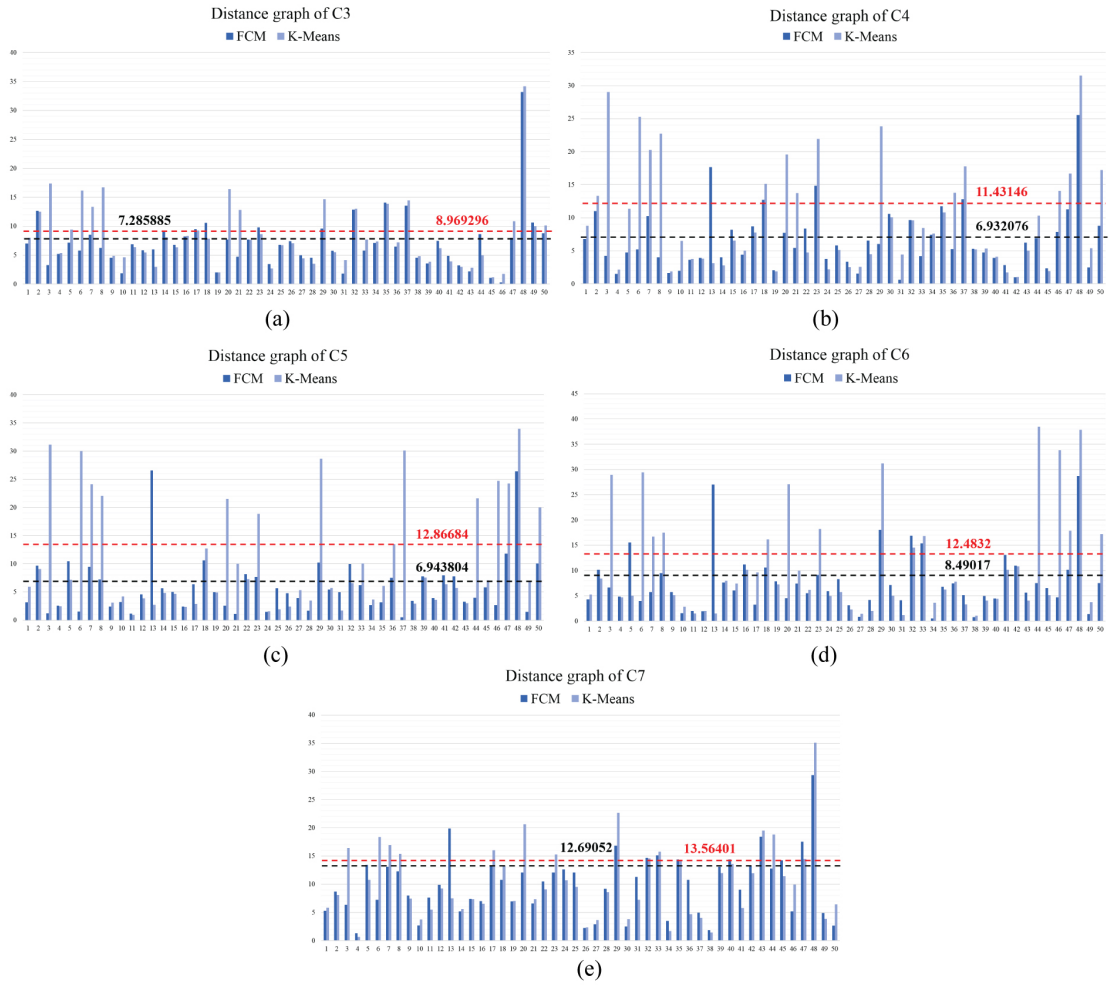


Figure 8: Pixel wise Distance Graph: (a)C_3 (b)C_4 (c)C_5 (d)C_6 (e)C_7

The accuracies at each cervical level are shown in figure 7 where green bars are representing greater accuracies for FCM than the K-means represented by blue bars.

Similar method of labeling is used for measuring the accuracies of separated regions. An overall accuracy of 83.1% is reported for separated regions with 90.5%, 93%, 88.5%, 73%, 70.5% at each cervical level from C_3 to C_7, respectively. In case of **Experimental** accuracies, ROC curves are formed for both FCM and K-Means shown in figure 9. Each line represents the accuracy measure of each cervical level from C_3 to C_7 with varying threshold from 0 to 20. The threshold here represents the width of vertebra. We selected threshold 14 for the accuracies, the reason behind choosing threshold 14 is, it is the radius of vertebra which means all the point inside the vertebra body are correctly detected. At threshold 14 the proposed technique attained accuracy of [93.12, 85.79, 81.30, 81.52, 79.30]and [96.74, 96.65, 95.51, 95.33, 84.55] representing C_3, C_4, C_5, C_6 C_7 for K-Means and FCM, respectively.In both cases, visual and experimental accuracies FCM secured more accuracy than k-means.

The accuracy when compared with other techniques in literature, it has been observed that only Larhman [14] secured greater accuracy of 97.5% than proposed technique i-e 93.6%. But it is not a fair comparison as both are using different subset of same dataset. They used 66 images of NHANES II dataset whereas presented methodology has been tested using 150 images of NHANES II dataset.

3.2 Visual Results

The visual analysis of each stage are shown in this section including steps of each stage. All the results shown are of 5 different cases of NHANES II dataset. Each column represents different case, whereas each row represents each step.

3.2.1 Preprocessing

Figure 10 shows the visual results of first stage i-e Preprocessing of proposed technology. The first row

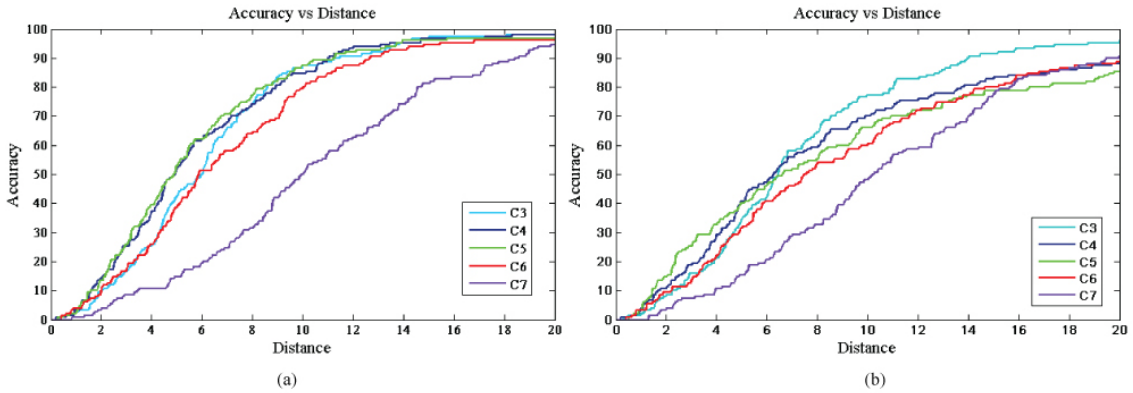


Figure 9: Receiver Operating Characteristic (ROC): (a) FCM Clustering (b) K-Means Clustering

shows the low contrast input x-rays which are enhanced by using CLAHE and shown in second row, it is clearly observed that the contrast of second row is improved and now these images can be used for better results. Next row shows the result of edge detection step, the edges are extracted for complete input image which are reduced to the ROI $C_3 - C_7$ and shown in 4th row.

3.2.2 Creation and selection of vertebra Models

The next stage of our model is creation of different vertebra models and figure 11 shows the results obtained by these models. Each row shows the detected centroids obtained using different models but same cases and it is clearly observed that $M1$ has given the most accurate results than other models.

3.2.3 Vertebra Localization (GHT)

The visual results obtained in vertebra localization stage are shown in figure 12, first row represents the hough space created by each case and voted points are represented in row 2, row 3 shows the voted points obtained after pruning where many outliers are removed, which can be observe in row 2.

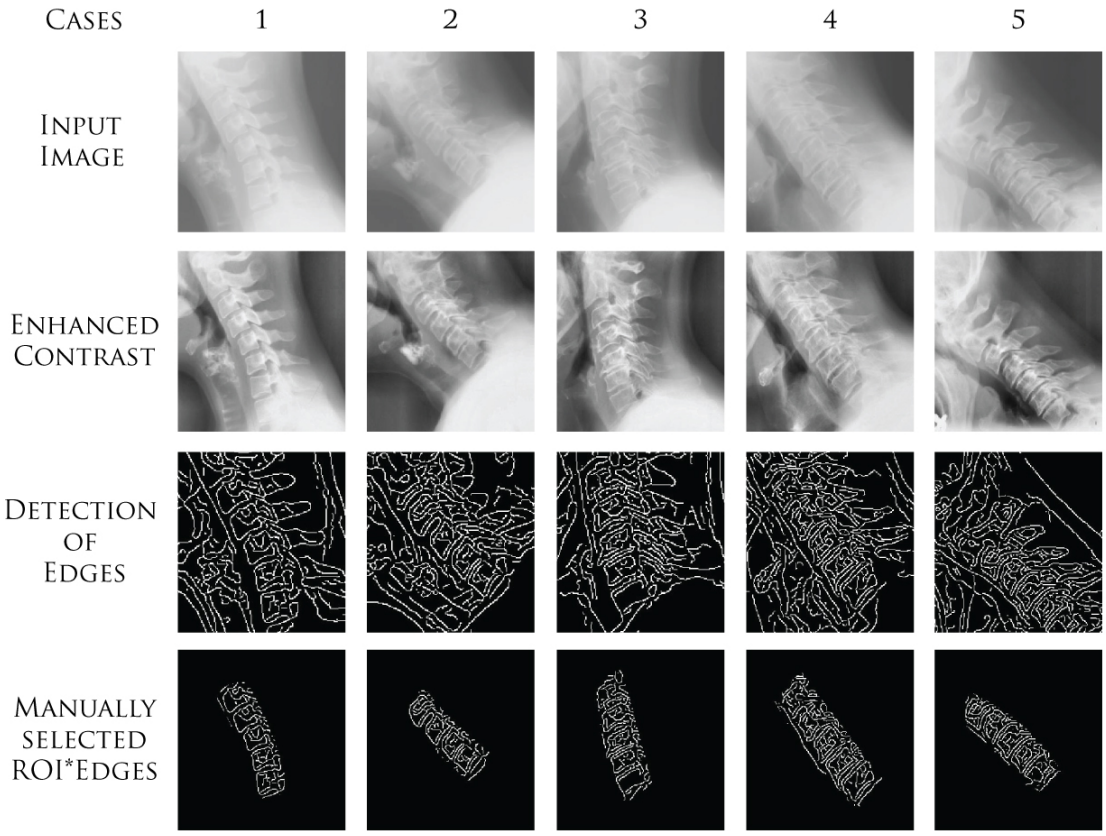


Figure 10: Phase 1: Preprocessing

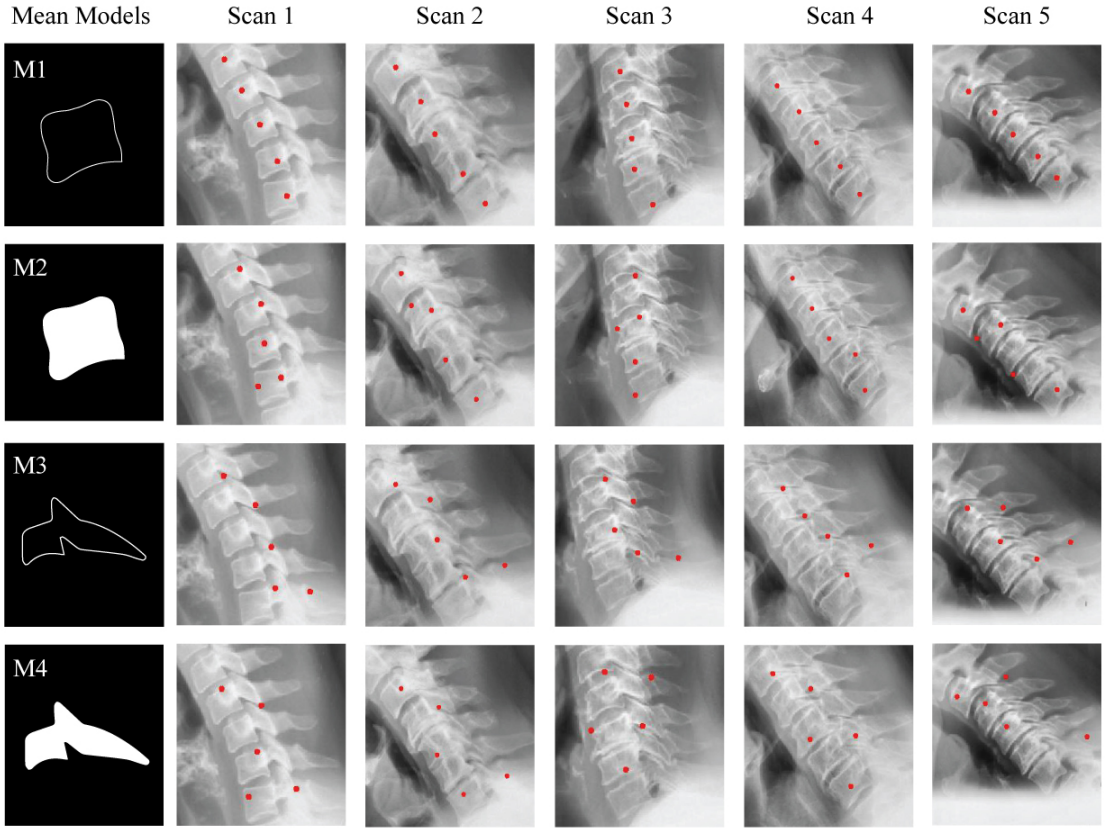


Figure 11: Phase 2: Comparison of different vertebra models

3.2.4 Centroid Detection

The centroid detection is carried out using two clustering techniques and their results are shown in figure 13 where first two rows represent the clusters and next two rows represents the detected centroids for each vertebra using K-Means and FCM clustering, respectively. Last row represents a comparison between annotated centroids (blue points), FCM centroids (red points) and K-Means (magenta point). It is clearly observed that FCM works way much better than the K-Means. In same cases FCM marked the correct centroids where as K-Means failed to do so.

3.2.5 Separation of vertebra regions

The steps involved in the separation of vertebra regions are shown in figure 14 where first row shows the blue centroids and 4 intervertebral points marked by red star. In next row, two additional intervertebral points are shown by yellow stars. The rotated lines by Affine transformation can be observed in 3rd row. Then after joining lines the separated regions are shown in 4th column.

4 Conclusion

This paper presents a comparison between 4 different vertebra template models for localization and 2 clustering techniques for centroid detection of cervical vertebra from x-ray images. The proposed model gives the separated regions for 5 targeted vertebra $C_3 - C_7$. The comparison of clustering techniques, vertebra template models and region separation technique is the contribution to this paper. For the localization of vertebra, GHT is applied along with each template model and results are analyzed. For GHT, $M1$ proves to be the best vertebra template. The voted points obtained after pruning are used for centroid detection. For the detection procedure two different clustering techniques are used FCM and K-Means. The results obtained by these two clustering techniques are analyzed at statistical and visual level and FCM proves to be much better than K-Means clustering for the detection of c-spine from x-rays. The results at level C_7 are less than the other levels in both clustering techniques, reason behind this is the presence of high brightness and other structures which are superimposed. The testing and comparison of proposed technique has been performed using 150 scans of NHANES II dataset accessible publically at U.S National Library of Medicine. The obtained accuracies of centroid detection are 84.21 and 93.76 for K-means and FCM, respectively. From all the experiments it is concluded that the FCM is more accurate than the K-means for the localization and detection of cervical verterba from X-rays. It is due to the fact that the fuzziness of FCM makes it less sensitive to outliers

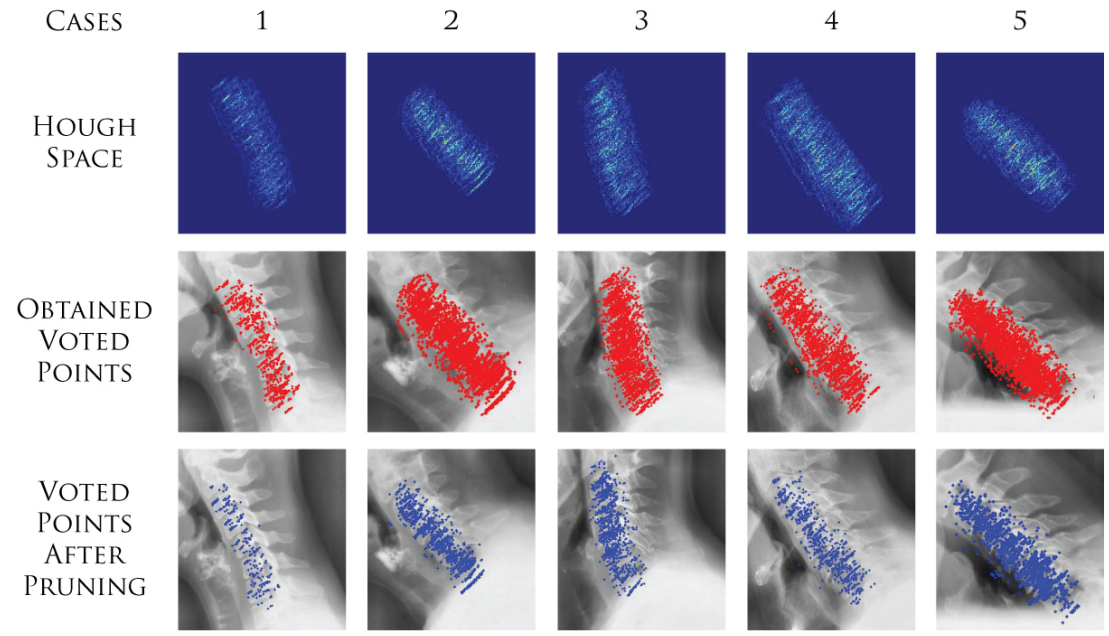


Figure 12: Phase 3: Visual results of vertebra localization steps

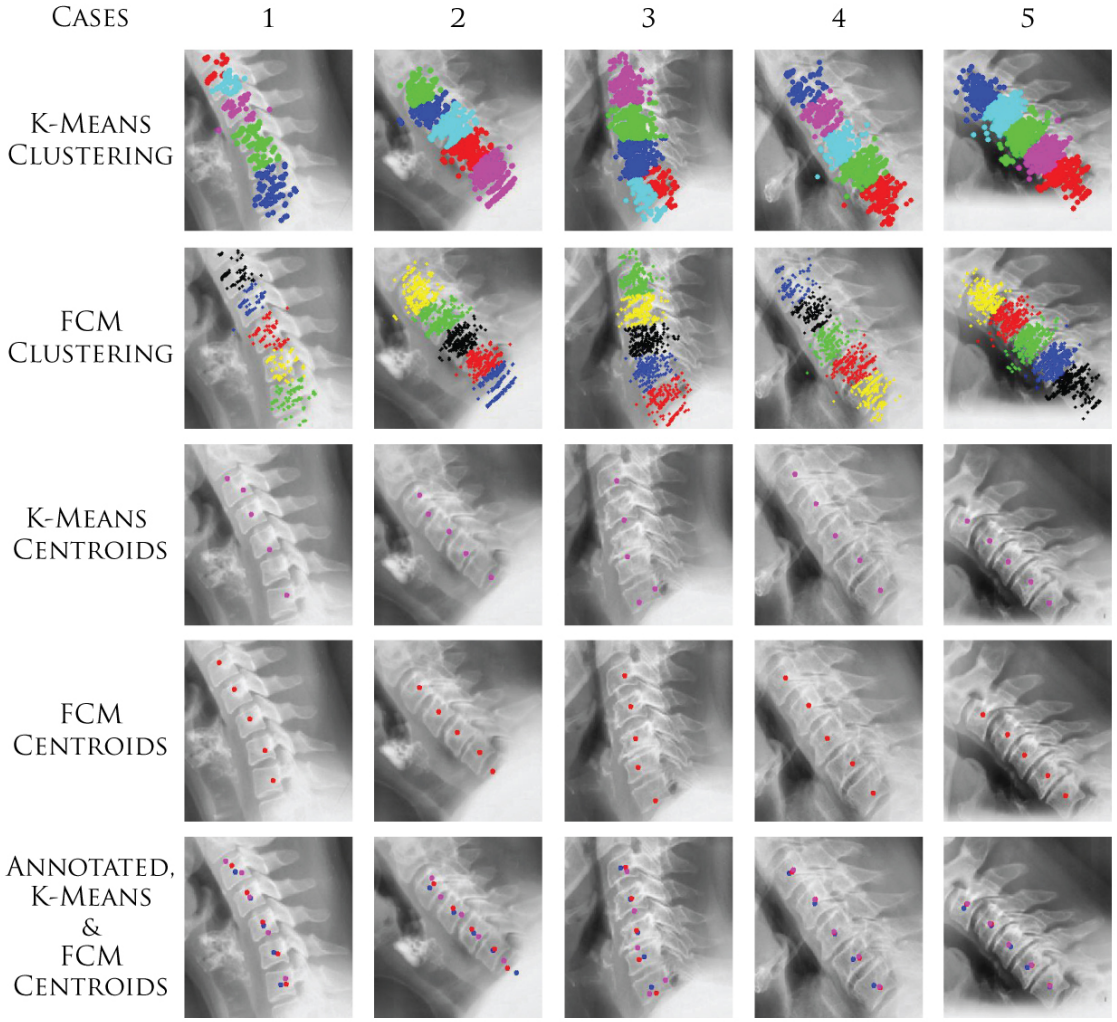


Figure 13: Phase 4: Centroid detection and comparison of clustering techniques

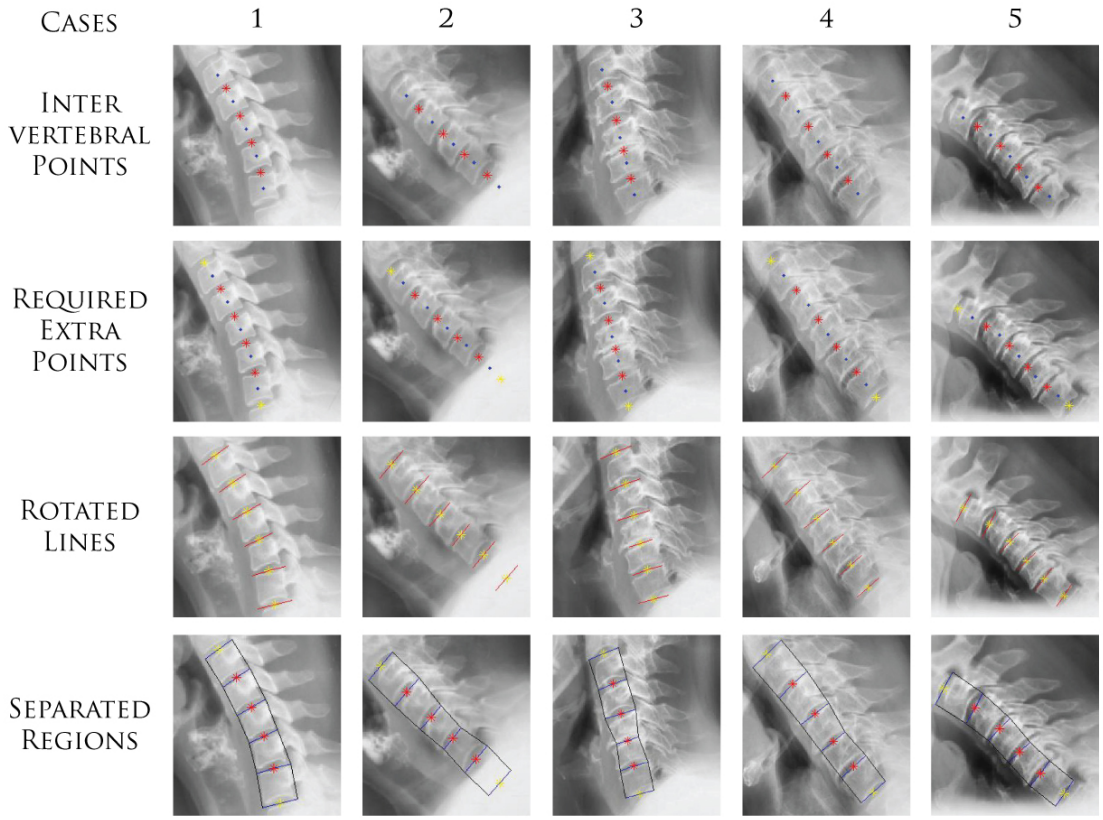


Figure 14: Phase 5: separation of vertebra regions

than K-means. Moreover, the proposed method secured satisfactory success rate of 83.1% for the separated regions which can be increased by increasing the width of the parallel lines. In future, our work is focused on segmentation using the separated regions and detected centroids.

Conflict of Interest The authors declare no conflict of interest.

References

[1] R. M. Hasler, A. K. Exadaktylos, O. Bouamra, L. M. Benneker, M. Clancy, R. Sieber, H. Zimmermann, and F. Lecky, "Epidemiology and predictors of cervical spine injury in adult major trauma patients: a multicenter cohort study," *Journal of trauma and acute care surgery*, vol. 72, no. 4, pp. 975–981, 2012.

[2] J. W. Tee, P. C. Chan, R. L. Gruen, M. C. Fitzgerald, S. M. Liew, P. A. Cameron, and J. V. Rosenfeld, "Early predictors of mortality after spine trauma: a level 1 australian trauma center study," *Spine*, vol. 38, no. 2, pp. 169–177, 2013.

[3] L. A. Nadalo, "Thoracic Spinal Trauma Imaging," Accessed on: April 06,2016.

[4] P. J. Richards, "Cervical spine clearance: a review," *Injury*, vol. 36, no. 2, pp. 248–269, 2005.

[5] V. Rahimi-Movaghar, M. K. Sayyah, H. Akbari, R. Khorramirouz, M. R. Rasouli, M. Moradi-Lakeh, F. Shokraneh, and A. R. Vaccaro, "Epidemiology of traumatic spinal cord injury in developing countries: a systematic review," *Neuroepidemiology*, vol. 41, no. 2, pp. 65–85, 2013.

[6] D. H. Ballard, "Generalizing the hough transform to detect arbitrary shapes," *Pattern Recognition*, vol. 13, no. 2, pp. 183–194, 1991.

[7] T. F. Cootes, C. J. Taylor, *et al.*, "Statistical models of appearance for computer vision," 2004.

[8] S. Lobregt and M. A. Viergever, "A discrete dynamic contour model," *IEEE transactions on medical imaging*, vol. 14, no. 1, pp. 12–24, 1995.

[9] R. Brunelli, *Front Matter*. Wiley Online Library, 2009.

[10] T. Klinder, J. Ostermann, M. Ehm, A. Franz, R. Kneser, and C. Lorenz, "Automated model-based vertebra detection, identification, and segmentation in {CT} images," *Medical Image Analysis*, vol. 13, no. 3, pp. 471 – 482, 2009.

[11] A. Raja'S, J. J. Corso, and V. Chaudhary, "Labeling of lumbar discs using both pixel-and object-level features with a two-level probabilistic model," *IEEE transactions on medical imaging*, vol. 30, no. 1, pp. 1–10, 2011.

[12] R. Korez, B. Ibragimov, B. Likar, F. Pernuš, and T. Vrtovec, "A framework for automated spine and vertebrae interpolation-based detection and model-based segmentation," *IEEE transactions on medical imaging*, vol. 34, no. 8, pp. 1649–1662, 2015.

[13] F. Lecron, S. A. Mahmoudi, M. Benjelloun, S. Mahmoudi, and P. Manneback, "Heterogeneous computing for vertebra detection and segmentation in x-ray images," *Journal of Biomedical Imaging*, vol. 2011, p. 5, 2011.

[14] M. A. Larhmam, M. Benjelloun, and S. Mahmoudi, "Vertebra identification using template matching modelmp and k-means clustering," *International journal of computer assisted radiology and surgery*, vol. 9, no. 2, pp. 177–187, 2014.

[15] M. Benjelloun and S. Mahmoudi, "X-ray image segmentation for vertebral mobility analysis," *International journal of computer assisted radiology and surgery*, vol. 2, no. 6, pp. 371–383, 2008.

[16] F. Lecron, M. Benjelloun, and S. Mahmoudi, "Fully automatic vertebra detection in x-ray images based on multi-class svm," in *SPIE Medical Imaging*, pp. 83142D–83142D, International Society for Optics and Photonics, 2012.

[17] X. Dong and G. Zheng, "Automated vertebra identification from x-ray images," *Image Analysis and Recognition*, pp. 1–9, 2010.

[18] M. A. Larhmam, S. Mahmoudi, and M. Benjelloun, "Semi-automatic detection of cervical vertebrae in x-ray images using generalized hough transform," in *Image Processing Theory, Tools and Applications (IPTA), 2012 3rd International Conference on*, pp. 396–401, IEEE, 2012.

[19] M. Benjelloun and S. Mahmoudi, "Spine localization in x-ray images using interest point detection," *Journal of digital imaging*, vol. 22, no. 3, pp. 309–318, 2009.

[20] M. Benjelloun, S. Mahmoudi, and F. Lecron, "A framework of vertebra segmentation using the active shape model-based approach," *Journal of Biomedical Imaging*, vol. 2011, p. 9, 2011.

[21] A. Mehmood, M. U. Akram, and A. Tariq, "Vertebra localization and centroid detection from cervical radiographs," in *Communication, Computing and Digital Systems (C-CODE), International Conference on*, pp. 287–292, IEEE, 2017.

[22] K. Zuiderveld, "Contrast limited adaptive histogram equalization," in *Graphics gems IV*, pp. 474–485, Academic Press Professional, Inc., 1994.

[23] I. A. M. Ikhsan, A. Hussain, M. A. Zulkifley, N. M. Tahir, and A. Mustapha, "An analysis of x-ray image enhancement methods for vertebral bone segmentation," in *Signal Processing & its Applications (CSPA), 2014 IEEE 10th International Colloquium on*, pp. 208–211, IEEE, 2014.

[24] P. Zhou, W. Ye, Y. Xia, and Q. Wang, "An improved canny algorithm for edge detection," *Journal of Computational Information Systems*, vol. 7, no. 5, pp. 1516–1523, 2011.

[25] R. O. Duda and P. E. Hart, "Use of the hough transformation to detect lines and curves in pictures," *Commun. ACM*, vol. 15, pp. 11–15, Jan. 1972.

[26] E. R. Dougherty and R. A. Lotufo, *Hands-on morphological image processing*, vol. 59. SPIE press, 2003.

[27] K. Alsabti, S. Ranka, and V. Singh, "An efficient k-means clustering algorithm," 1997.

[28] J. C. Dunn, "A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters," 1973.

[29] P. Wang, "Pattern-recognition with fuzzy objective function algorithms-bezdek, jc," 1983.

[30] C. C. Stearns and K. Kannappan, "Method for 2-d affine transformation of images," Dec. 12 1995. US Patent 5,475,803.

[31] "NHANES II X-ray Images," Jan 27,2016.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1771-1776 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Monitoring power breakers using vibro acoustic techniques

Horia Balan*,[1], Traian Varodi[1], Mircea I. Buzdugan[2]

[1]*Technical University Cluj-Napoca, Department of Energy and Management, F.I.E, 400114, Romania*

[2]*Technical University Cluj-Napoca, Department of Building Services Engineering, F.I., 400114, Romania*

ARTICLE INFO

Article history:
Received: 05 April, 2017
Accepted: 06 June, 2017
Online: 23 September, 2017

Keywords:
power breaker
connecting
disconnecting
reliability
maintenance

ABSTRACT

Speaking about the commutation's equipment, it can be said that the best solution in increasing reliability and lowering the maintenance costs is a continuous monitoring of the equipment. However, if the price/quality ratio is considered, it is obvious that, for the moment, the diagnosis can be also an acceptable solution. Nowadays the predictive maintenance for equipment's diagnosis is currently replacing the preventive diagnosis. An efficient modality of lowering the maintenance costs is to online monitoring the power breakers, during their operation in the power systems. Consequently any connecting/disconnecting operations may be used in diagnosing a power breaker. Thus any supplementary and superfluous tests and/or maintenance maneuvers are avoided. The paper presents the operational maintenance in a power station with three high voltage active breakers, Areva type. The method of establishing the state of a breaker consists in the comparison between the signature of the acoustic signal provided by the manufacturer and the signal issued from the testing operation of the breaker's state. The software processing procedure and the methodology of determining the faults of the monitored equipment are also developed. All the tests on the circuit breaker are made according the prescriptions of normative.

1. Introduction

Nowadays there is a major concern in using acoustic vibrations for monitoring medium and high voltage breakers. The results of this method are comparable to the traditional test results of power breakers. Moreover, the vibration method gives online supplementary information, unavailable when using traditional methods.

The use of the vibro acoustic method is possible because even though the same number of connecting/disconnecting maneuvers is made, in the signature of the vibro acoustic signal there are much more distinctive events. These issues can be solved using supplementary processing and filtration of the vibration signals.

Measurements performed in a transformer substation compared to those performed in the laboratory highlight some low duration interference of acoustic signals in the first set of measurements. These interferences have led to errors in the results, but after finding out the noise sources, i.e. of the perturbations, and processing the registered signals using software analysis and filtering, consistent conclusive results were obtained.

Mainly, the conclusion was that the vibration monitoring methods operate similarly both in the field and in the laboratory.

2. Breakers' Maintenance

Nowadays the maintenance policy of power breakers is oriented to corrective maintenance, acting in the moment that a fault occurs and on planning maintenance, which uses pre-established maintenance criteria periodically programmed.

Power breakers represent one of the key components in every electric power transmission and distribution system. It is the very last component in the control system that has to protect the system against damages determined by faults. In the moment of a fault in the power system operation, the breaker should operate safely, even though there was a long period of previous lack of maneuvers.

*Corresponding Author: Horia Balan, Technical University Cluj-Napoca, Department of Energy and Management, F.I.E, 400114, Romania.
Email: horia.balan@eps.utcluj.ro

For a proper operation, medium and high voltage power breakers are periodically redrawn from the power systems in order to be submitted to measurements and functional tests [1].

The main components of the breakers represent their so called "active part" and have the role to maintain and break the current in the electrical installations in which they are commissioned. These are:

- the extinguishing chamber (the breaking chamber),
- the ancillary equipment,
- the electric insulation equipment (the insulating column, air, oil, vacuum, SF6, etc.).

The control and the auxiliary element have the role to transmit the commutation command (connecting/ disconnecting) to the breaker. This category is composed by:

- connecting/disconnecting circuits,
- auxiliary commuters,
- contactors,
- relays,
- thermostats,
- the density control device SF6,
- fuses,
- blocking devices, etc.

The driving mechanism represents the component that provides the mechanical energy necessary to move the contacts. It consists in:

- the transmission mechanism,
- the energy accumulator,
- the control elements (the non electric ones),
- the control devices,
- the dumping elements,
- the compressors,
- the pumps, etc.

The most used types of driving mechanisms are:

- devices accumulating energy in springs
- devices with hydraulic drive
- devices with oleo pneumatic drive [2]

3. Acoustic Vibrations of Breakers

The aim of measuring vibrations at a power breaker is to make possible the diagnosis of its operational status without performing any traditional tests. For a proper processing and a real result, the vibration signal is registered during a connecting/disconnecting process and is compared with a reference signal of the same type. The reference signal must be a known previous registration corresponding to a proper operation of the breaker. The point is that when the vibration fingerprints are performed, one can detect the contact imperfections, the mechanical faults or any other faults which are obvious in the vibration's signature modification of the breaker.

Some practical aspects of the experiments [3] must be considered, like as an overvoltage in the data recording system. In order to prevent the failure of the recording system, an overvoltage protection system should be installed, and the accelerometer should be insulated with respect to the ground.

For a breaker that had been previously tested in the sense of vibration signal during its long time operation, the time errors may be compared in the terms of commutation times. The advantage of the dynamic time warping (DTW) analysis is that the time error is related to a certain moment of the maneuver [4]. The traditional measurements of the commutation time, according to the existing standards, provide details about the total commutation time, but provide no information about the moment and the evolution area of the maneuver due to the time error.

When the vibration signals determined by the movement of the operation mechanism is measured, it is possible to find some hidden and serious faults of the breakers, even though they were tested using traditional methods and have been considered in a proper functional status.

The measurement of the current drawn by the operating inductor cannot be replaced by the DTW analysis. However if the accelerometer is located in the vicinity of the inductor, it is possible to register vibrations, when the inductor is energized.

When the impulse is recorded with a higher sampling frequency, in this case more than 500 kHz, it is possible to estimate the impulse duration. See Fig. 1 in which the duration of the pulse is 2 μs (where x represents the measured value, the red line represents the evolution of the signal and the dotted line represents the width of the impulse).

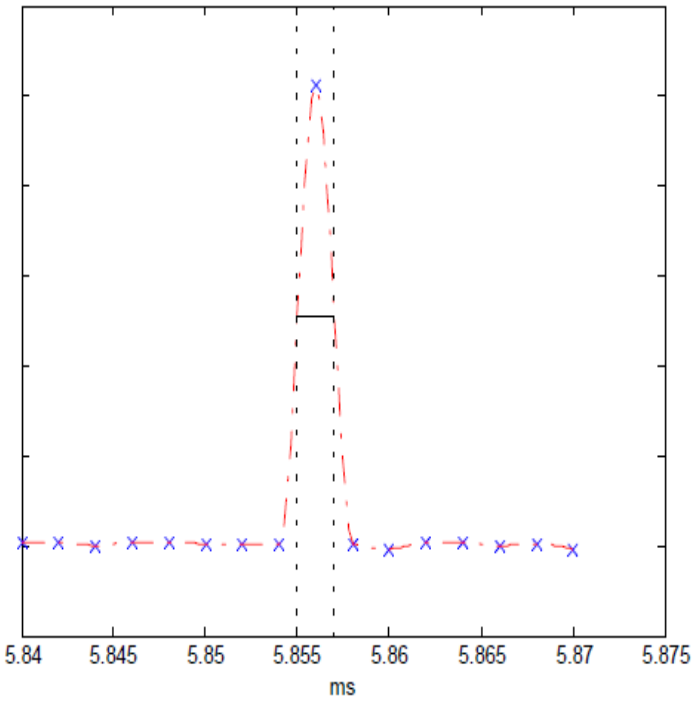


Figure 1. Impulse recorded at a sampling frequency of 500 kHz [5]

In the case of such pulses it is very difficult to mention a fault or a non functionality, because when measuring with accelerometers, the measurements may be affected by high voltage pulses captured by the cables linking the piezoelectric sensors and

the measurement device, even though they are shielded and/or separately supplied.

If sound recordings are performed and processed, one can obtain information on the operating anomalies and of their causes. Applying an appropriate electro acoustic measurement and analysis method, conclusions on several faults can be drawn.

Due to the interdiction of mounting sensors on commissioned breakers, sound records may be suitable for the diagnosis of several types of faults.

4. Methods of Processing Acoustic Signals

Current or vibration signals measurements are in general performed in time domain. In order to identify the frequency components of signals it is necessary to use the Fourier transform. The Fourier analysis is useful in the study of the signals in general and in the study of harmonic functions in particular (Fig. 2).

Any periodic function $f(t) = f(t+T)$, may be written as an infinite sum of sinusoidal harmonic oscillations, each of them having frequencies equal to integer multiples of a minimum frequency, namely the fundamental frequency, according to eq. 1 [7]:

$$f(t) = a_0 + \sum_{n=1}^{\infty} (a_n \cos n \cdot \omega t + b_n \sin n \cdot \omega t) \quad (1)$$

Components appearing in the sums of the eq. 1 are called harmonic components and deserve to determine the signal spectrum, a_0 being the continuous component. The coefficients a_n and b_n are determined using the following equations [7]:

Figure 2.

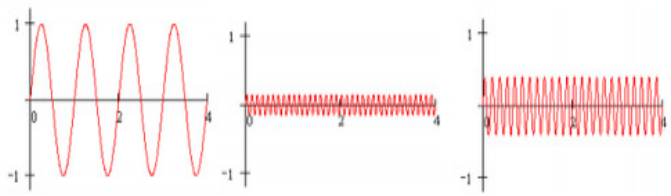


Figure 2. Sinusoidal signals [6]

$$a_0 = \frac{1}{T}\int_0^T f(t) \cdot dt \quad (2)$$

$$a_n = \frac{2}{T}\int_0^T f(t) \cdot \cos n \cdot \omega t \cdot dt \quad (3)$$

$$b_n = \frac{2}{T}\int_0^T f(t) \cdot \sin n \cdot \omega t \cdot dt \quad (4)$$

The Wavelet technique is more and more used in the analysis and the synthesis of the sounds, due to the advantages with respect to the fast Fourier transform [8].

The continuous wavelet transform of a finite energy signal $x(t)$, $\Psi(t)$ represents the "mother Wavelet function" and it is the convolution product of $x(t)$ with the wavelet scaled and conjugated function Ψ^*, according to the eq. (5) [9]:

$$\Omega_{\xi}(a,b) = \frac{1}{\sqrt{a}}\int_{-\infty}^{\infty} x(t) \cdot \Psi^*\left(\frac{t-b}{a}\right) dt \quad (5)$$

In order to be a Wavelet function, every function must satisfy the following admissibility condition [9]:

$$C_{\Psi} = \int_{-\infty}^{\infty} \frac{|\Psi(\omega)|^2}{|\omega|} \cdot d\omega < \infty \quad (6)$$

where $\Psi(\omega)$ is the Fourier transform of $\Psi(t)$ and C_{Ψ} depends on the chosen Wavelet transform [10]. The Wavelet functions are asymmetric, having a finite duration and a zero mean value.

$$|\Psi(\omega)| = \int_{-\infty}^{\infty} \Psi(t) \cdot dt = 0 \quad (7)$$

Using the calculated coefficients of the Wavelet functions, the original signal $x(t)$ can be restored.

The Wavelet coefficients appreciate the resemblance between the signal $x(t)$ and the function $\Psi_{a,b}(t)$. The use of the Wavelet transform consists in the choice of "mother Wavelet" functions. A certain number of "mother Wavelet" functions exist, but for the choice and the use of these functions, several particularities application depending must be considered.

Dynamic Time Warping is the method comporting two signals, the first one unknown and the second one a reference signal. The result of the comparison of the two signals is a function describing the temporal variation of the misalignment between the two signals. For an optimal result, the cumulated distance of the differences between the two signals must reach a minimum. In other words, considering the function from its initial point to its final point and cumulating at each step the distance between the characteristics of the frames, in the last point, the distance has to be the minimum one [11].

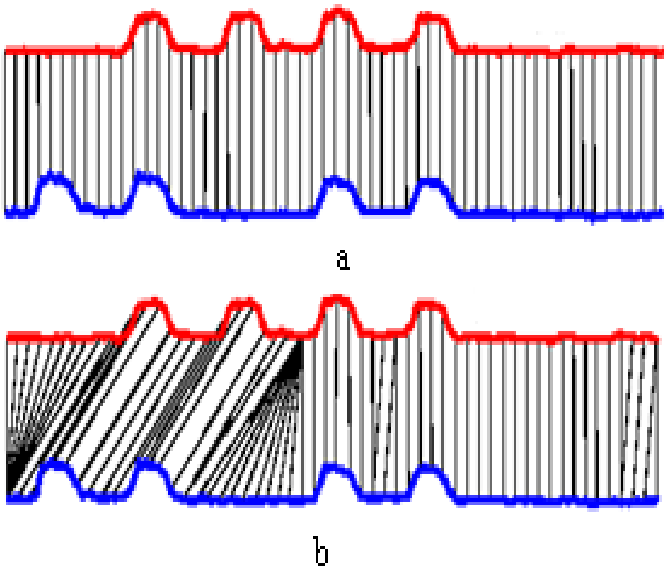


Figure 3. Alignment of two signals by the Euclidian method (a) and by the DTW method (b) [11]

Dynamic Time Warping (DTW) is an algorithm of aligning two time series, which was at its origins developed for the speech recognition. In the time series analysis, the dynamic deformation in time is an algorithm of measurement of the similitude of two temporal sequences that can be variable in time and in speed [12].

How resembling could be for instance the signals A and B from Fig. 3 (signal A is denoted with a red line and signal B with a blue line).

A method that modifies the time axis by contraction or dilatation is needed in order to superimpose the two signals without modifying the phase of the signals or to displace the time axis. The approximation used in the Euclidian method is used quite often, but the method is not able to perform an accurate comparison between signals which are not in phase.

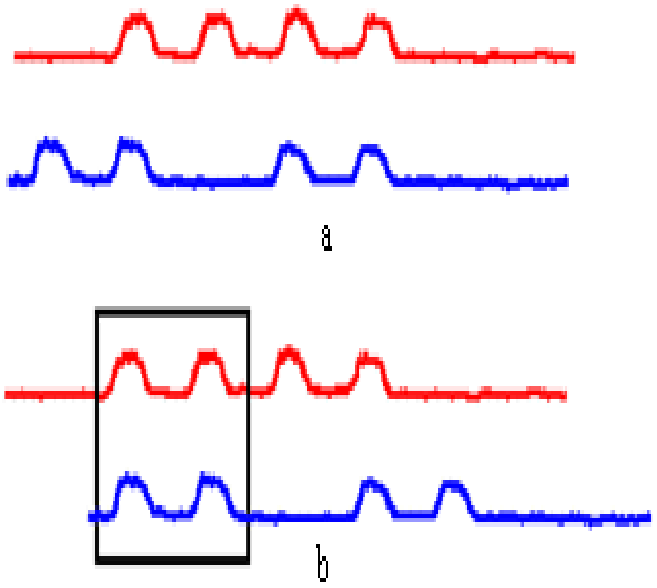


Figure 4. Alignment of two signals by displacement on the time axis [11]

The dynamic time warping (DTW) method is able to perform an alignment of events between the two signals, even though they are not in phase (Fig. 4).

5. Experimental Results

The experimental determinations are performed on a high voltage power breaker, Areva, GL 310 F3/4031 O/VR, with SF6, consisting in two columns of poles, the base frame and a driving mechanism. The columns with poles consist in an insulator which is isolating the operating voltage with respect to the ground and an insulator of the chamber in which the breaking unity is located.

The experiment consisted in recording the sound depicted in Fig. 5, in the stages of connecting arming and locking of the power breaker.

The objective of the determinations is to automatically identify the existence of the three stages (connection, arming and locking), of the operating mode, (the normal operating, or the abnormal operating mode) of the whole ensemble consisting in the three 110 kV power breakers, on the three phases L_1, L_2 and L_3.

The dynamic warping method is used and the spectrograms are depicted in Figs. 6 and 7.

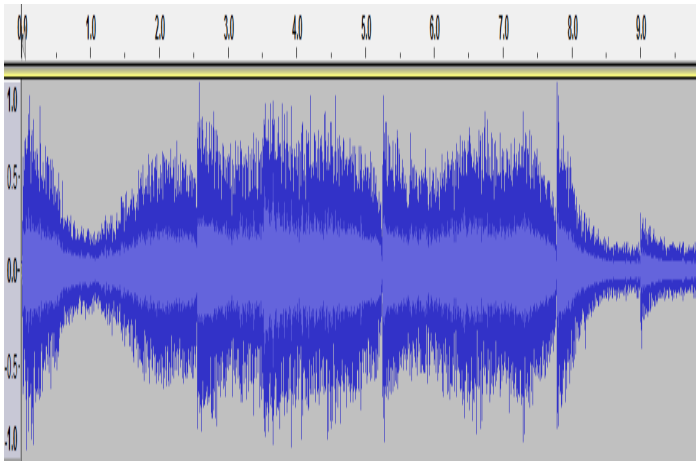


Figure 5. Audio signal

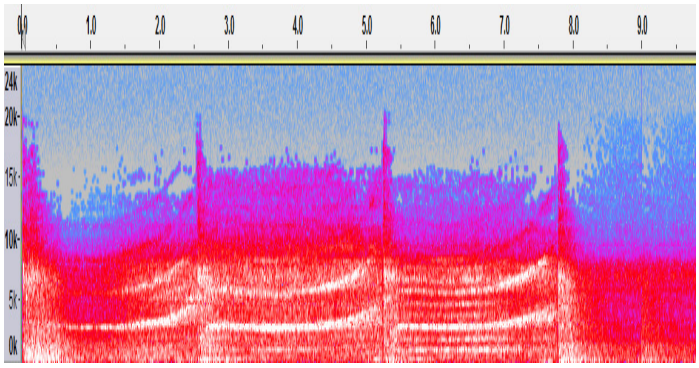


Figure 6. Audio signal spectrogram

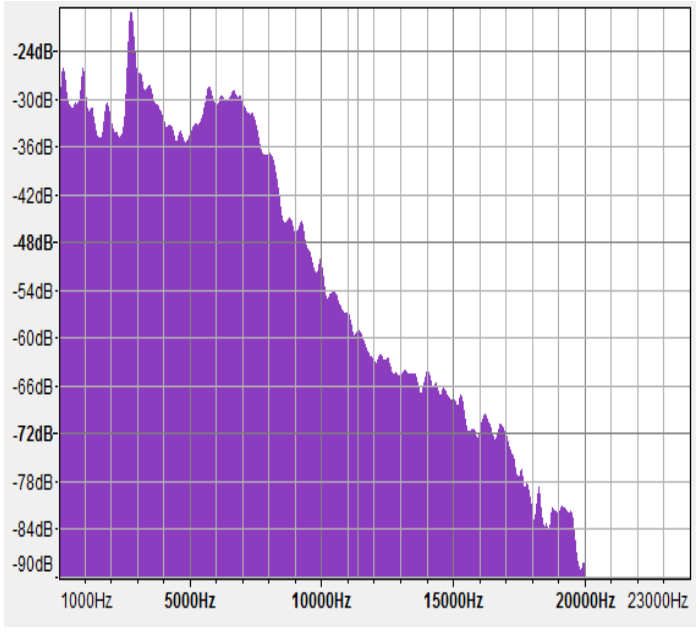


Figure 7. Magnitude versus frequency spectrum

The method that is checking out the breaker status consists in using the software Test_DTW. The software allows the comparison of each line signal with respect to the reference signal. The values obtained after the time alignment and the comparison, highlight the status of the power breaker.

The simulations have been performed using the Matlab platform.

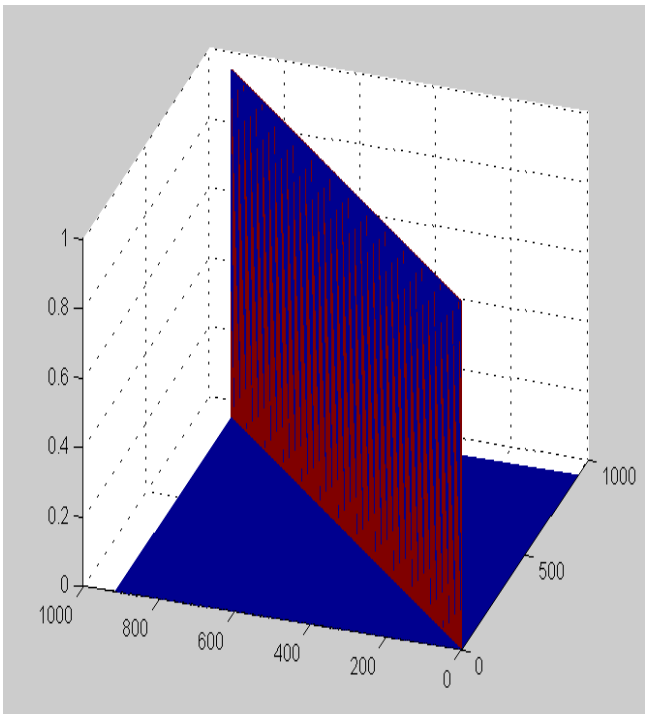


Figure 8. The match matrix M. Comparison of the reference signal **O_R1** with itself.

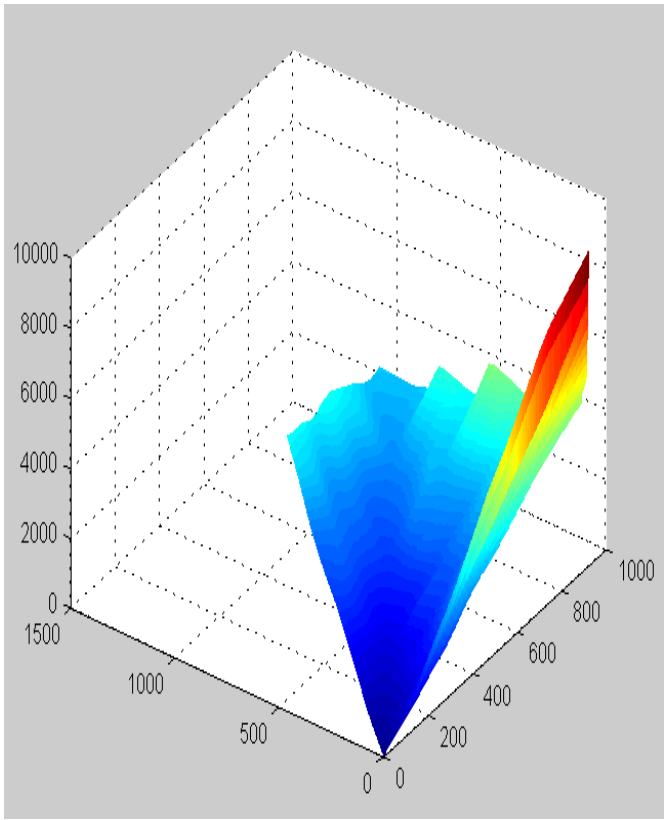


Figure 9. Cumulated distance D matrix

In order to highlight the accuracy of the method, Fig. 8 depicts a sort of a self test, consisting in a comparison between the reference signal with itself. Due to the fact that the values of the algebraic sum of the cumulated magnitudes are perfectly superimposed on the diagonal of the base square, the validity of the solutions is confirmed. The lack of superposition on the diagonal of the matrix highlights the existence of failures.

On the other hand, Fig. 9 indicates a breaker fault, highlighted by the comparison between the reference signal OR-1 and the signal OR 2.

6. Conclusions

Due to the permanent development of the electrical energy generating sources, the problem of the modernization of the transmission and distribution networks became a matter of high necessity. Consequently the need of higher reliability of equipment the cost reductions are compulsory. At the same time the maintenance problem is an acute one.

The costs with the maintenance process can be influenced by the reduction of the regular maintenance operations which in several cases are even not necessary. These costs can be optimized by passing from the classic corrective maintenance to the preventive maintenance using efficient methods.

The development and the implementation of the monitoring systems for power breakers became a necessity and a constant concern for the equipment manufacturers who provide equipment with embedded sensors. Due to the nowadays trend of monitoring and of the diagnosis of the medium and high voltage power breakers, the research topics should be focused on the monitoring, measurement and analysis of the sounds made be the vibration signals of the power breakers. This method represents a good choice because on one hand the vast majority of the power breakers commissioned until now has no embedded sensors and on the other hand the sensors cannot be mounted on the power breakers.

The sound analysis method proposed is a non invasive one which comply the safety norms, the measuring equipment being kept at a safety distance with respect to the operating equipment.

The monitoring systems presented in the paper have the following advantages:

- obtain information according the operating faults or anomalies;
- significantly mitigate the maintenance costs, by increasing the time duration between two faults;
- simplify the diagnosis process;
- increase significantly the life time of the power breakers;
- optimize the maintenance costs;
- maximize the investments efficiency;
- reduce the reaction time in case of functional faults.

Highlighting the specific type of fault that occurs at the commutation equipment will represent the topic of a further work. The present paper highlight only the benefits of the monitoring method that is able to offer the software processing procedure and the methodology of determining the faults of the measurements outcomes.

References

[1] Huang Lingjie, Wang Wei, Wu Zhensheng, Xu Lijie, Research on the model of HV SF6 circuit breaker fault diagnosis based on fuzzy theory,

International Conference on Condition Monitoring and Diagnosis, Beijing, China, 2008;

[2] M. Adam, A. Barboi „Electric Equipment II" Ed. Gh. Asachi, Iasi 2002 (in Romanian);

[3] H.K. Hoidalen; M. Runde, "Continuous Monitoring of Circuit Breakers Using Vibration Analysis", IEEE Transaction on Power Delivery, Vol. 20, No. 4, 2005, pp 2458 – 2465;

[4] M. Runde et al., "Vibration analysis for diagnostic testing of circuit-breakers", IEEE Trans. On Power Delivery, vol. 11, 1996;

[5] Tat Wai (Alan) Leung. Predictive Maintenance of Circuit Breakers. Waterloo, Ontario, Canada, 2007;

[6] http://ro.math.wikia.com/wiki/Analiz%C4%83_Fourier;

[7] http://newton.phys.uaic.ro/data/pdf/Introducere%20in%20analiza %20Fourier.pdf;

[8] P. S. Addison, "Low-Oscillation Complex Wavelets", Journal of Sound and Vibration (2002) 254(4), pp. 733-762;

[9] M. Răducanu, L. Stanciu, A. Mateescu, Time- frequency transformers used in telecommunications: Fast Fourier transform and Wavelet transform, Telecomunicaţii, no.1/2007, (in Romanian)

[10] Shie Qian, "Introduction to Time-Frequency and Wavelet Transforms", Upper Saddle River, New Jersey: Prentice Hall PTR, 2001;

[11] E. Keogh Exact indexing of Dynamic Time Warping http: /www.cs.ucr. edu;

[12] M. Runde; G.E. Ottesen; B. Skyberg; M. Ohlen, "Vibration Analysis for Diagnostic Testing of Circuit Breakers" *IEEE Transactions on Power Delivery* Vol. 11, No. 4, 1996, pp.1816-1823;

Advances in Science, Technology and Engineering SystemsJournal Vol. 2, No. 3, 1777-1783 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

A Cost Effective Security Technology Integrated with RFID Based Automated Toll Collection System

Rafiya Hossain*, Moonmoon Ahmed, Hasan Uz Zaman

Department of Electrical and Computer Engineering, North South University, Dhaka, Bangladesh

ARTICLE INFO

Article history:
Received: 31May, 2017
Accepted: 29August, 2017
Online: 23 September, 2017

Keywords:
RFID technology
Text message
Toll charge

ABSTRACT

Crime statistics and research on criminology show that under similar circumstances, crimes are more likely to occur in developing countries than in developed countries due to their lack of security measures. Transport crimes on highways and bridges are one of the most common crimes in the developing nations. Automation of various systems like the toll collection system is being introduced in the developing countries to avoid corruption in the collection of toll, decrease cost and increase operational efficiency. The goal of this research is to find an integrated solution that enhances security along with the advantage of automated toll collection. Inspired by the availability of many security systems, this research presents a system that can block a specific vehicle or a particular type of vehicles at the toll booths based on directives from the law enforcement agencies. The heart of the system is based on RFID (Radio Frequency Identification) technology. In this system, by sending a text message the law enforcement agency or the authority that controls the toll booths can prevent the barrier from being lifted even after deduction of the toll charge if the passing vehicle has a security issue. The designed system should help the effort of reducing transport crimes on highways and bridges of developing countries.

1. Introduction

Developing countries are still lagging behind in technological development compared to developed countries. Security systems and safety measures to stop crime in developing countries are still dependent on trivial manual systems rather than automated systems. A major portion of crime happens at the highways and bridges. Most of the hijackers, kidnappers, murderers, anddrug dealers usually try to change their location by travelling from one state to another. As most of the roads and bridges have toll booth in developing countries, they have to pass through those areas. But police sometimes fail to recognize the criminals and their transport vehicles. However, in our work we tried to develop a system at the toll booth area to mitigate this problem. For this, we have integrated an advanced security system along with automated toll collection system based on the RFID technology.This paper presents the completed work of a research whose initial results were published in a conference paper [1].

For implementing this system, we need RFID tag in every vehicle. Moreover, every vehicle owner's basic information will be recorded in the database along with their prepaid based account number. So, it will be easier for the authority to find the vehicle owner's details in a faster way in case of emergency. For instance, suppose a car has been stolen from a state. Usually stolen cars are removed secretly out of the state and for this it has to pass any of the bridge. However, the victim can write a General Diary (GD) at the police station and the police can inform the authority about the specified RFID number plate and thus the particular vehicle can be prevented from crossing the exit point at the toll bridge. Thus, the missing car can be found easily.

The overall system is a hassle free and user friendly secured system. The benefits of the system also include congestion reduction, time saving, operating cost saving, payment flexibility, and incident reduction.

2. Literature Review

A considerable research work has been going on regarding automation of toll collection system to replace manual toll collection system. Most of them are based on RFID technology as it is cost effective and also safer at the same time. Prepaid based account is needed for deducting the toll charge from the vehicle

*Corresponding Author: Rafiya Hossain, Department of Electrical and Computer Engineering, North South University, Dhaka, Bangladesh.
Email: rafiya_2012@yahoo.com

https://dx.doi.org/10.25046/aj0203217

owner's account. This system is much more convenient, safer and it also reduces traffic congestion at the highways and bridges.

In 2005, Jerry.L and Barbara.C proposed a technique which was based on smart card [2]. Later, in 2011 Hitachi.S had proposed a system using ARM-7 processor and that reduced the overall process time to a great extent [2]. The vital part of the system is the RFID. A study related to sustainability of passive RFID technology for fast moving vehicle identification was found in a research paper [3]. This paper talks about the ALOHA based anti-collision algorithms to identify moving vehicles. The simulation work was developed using C++ and three configurations which includes one, two and multi traffic lane [3].

However, some research is also going on in the area of intelligent safety and security of automobiles by implementing different types of sensor and wireless locking and unlocking systems to reduce road accidents [4][5]. Cyber security is another big concern nowadays. Research on data security is also going on in many ways [6].

We have analyzed some research on car parking systems and we found RFID based access control system that is used in some confidential areas. A tracker based automatic detection of suspicious behavior of pickpockets in a shopping mall related work is also been proposed in a research paper [7].

RFID identity card system is already available in hostels, universities, medical centers, and shopping malls to restrict the entry of unauthorized people [8]. Combining all the ideas, we have aimed at developing an advanced security system which will be integrated with the RFID based toll collection system and will provide higher security on highways or bridges.

3. Description of the Proposed System

The overall system model is designed based on our previous project ideas and some new concepts regarding RFID technology. For implementing this system in a country, all vehicles must hold RFID number plates or tags. After the vehicle enters the toll booth area, the RFID reader at the toll booth instantly reads the vehicle's identification number and detects the vehicle type from memory. Then, according to the preset toll charge for each type of vehicle, it deducts toll charge from the vehicle owner's prepaid based account.

In case of truck or other weight carrying vehicles, like-cargo, it uses a formula to calculate the toll charge amount as the weight varies. However, a confirmation text is sent immediately to the vehicle owner's phone. So, to execute the system, the database has to record all the information including National ID card number, phone number, and account number etc. of the vehicle owner.

If any sort of crime occurs at the highways or bridges and the criminal wants to escape by moving into another state, he has to cross some toll booths. For example, if a murderer or a drug dealer is crossing a certain toll booth and the police somehow get informed about it, they can simply resist them. For this, they will only send the RFID tag code or number of the vehicle to the authority if they want to stop that vehicle. Then the system controller will send the code text to the system and block that vehicle even after paying the toll amount. In this way, if a security concern issue arises, the authority can also block all the vehicles from passing through that booth.

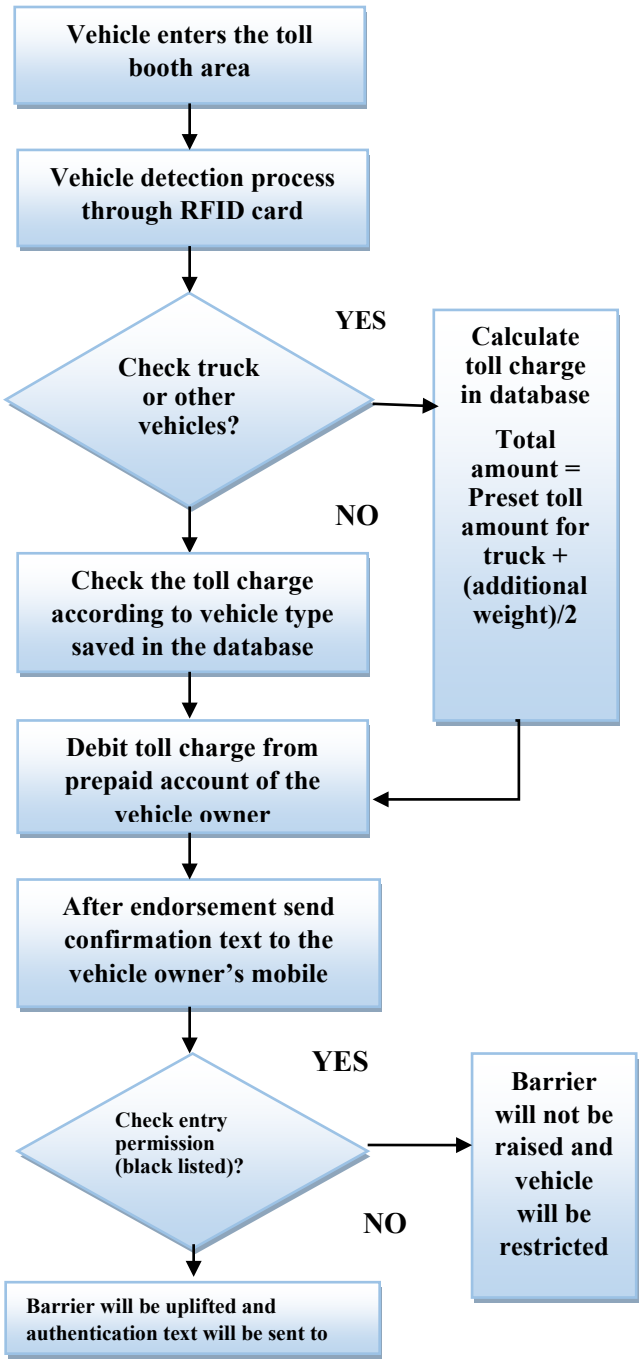


Figure 1: The Proposed System

4. Design Materials

The prototype of the system has been developed by using RFID system which includes the RFID Reader and RFID Tag, Microcontroller-Arduino Mega 2560, Force Sensitive Resistor (FSR), LCD Display, GSM Module and Servo Motor. The specifications of the components are briefly given below:

4.1 RFID Reader

A RFID reader is a network connected device (fixed or mobile) with an antenna that sends power as well as data and commands to the tags. RFID systems allow the unique identification of items [8] [9] or any other products which allows

tracking of individual items throughout the supply chain. In our system, RFID reader is used to detect the RFID tag attached to the vehicle. The reader used in our prototype can read RFID tag within 3.5 cm. However, for the real time model we will use more powerful reader which can detect the RFID tag from a longer distance.

4.2 RFID Tag

Radio-Frequency Identification (RFID) is the use of radio waves to read and capture information stored on a tag attached to an object. In RFID system, every individual item is equipped with an inexpensive and small RFID tag. The tag contains a transponder with digital memory chip. An antenna packaged with a transceiver and decoder emits a signal activating the tag so that it can read and write data into it. When RFID tag passes through the reader zone, it detects the reader activation signal [8] [9]. For our prototype five different active RFID tags are used for identifying five types of vehicle

4.3 Arduino Mega

Arduino Mega is a microcontroller board based on the ATmega2560. It has 54 digital input/output pins (of which 14 can be used as PWM outputs), 16 analog inputs, 4 UARTs (hardware serial ports), a 16 MHz crystal oscillator, a USB connection, a power jack, an ICSP header, and a reset button [10]. In our prototype, two Arduino mega have been used.

4.4 Force Sensitive Resistor (FSR)

FSR is basically a resistor that changes its resistive value (in ohms Ω) based on the pressure it is given [11].In our system we have used a square shape FSR in place of weight sensor for the practical demonstration purpose.

4.5 LCD Display

A 16x2 LCD is used for displaying the deducted toll amount and vehicle types at the toll booth.

4.6 GSM Module

The technical design of the system includes a GSM module which will help to send a confirmation text to the vehicle owner about the toll deduction. It is also used to communicate with the security personnel hence control the access of vehicles.

4.7 Servo motor

Servo motor has been used in our prototype which rotates at 90 degree angle to control the barrier.

5. Hardware Design

The system contains two Arduino Mega 2560 as the microcontrollers. The first Arduino is connected with a Force Sensitive Resistor (FSR), GSM module and the RFID Reader. On the contrary, the other Arduino is connected with a GSM module, servo motor and a Liquid Crystal Display (LCD). Then, the Arduino-1 and Arduino-2 are connected through serial communication. However, we need to power up the two Arduino Mega 2560 with 9V~12V and 2A. To activate the GSM shield, we need to insert a SIM card with balance recharged and hold the button until the green light stop blinking and settle down. Finally, we need to pass the RFID tag or card in front of the RFID reader.

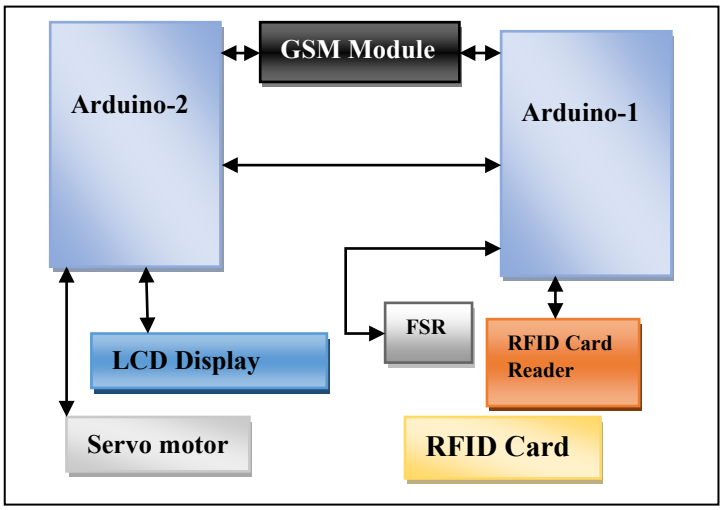


Figure 2: Hardware architecture

6. Software Design

The software design is constructed with two microcontrollers. The first processor is used to communicate with the RFID system, GSM module and FSR. However, the second processor combines the functionalities of GSM module, LCD display and servo motor. In the first processor, the system embarks by sending a tag request. If the vehicle carries a valid RFID tag, it sets or resets corresponding bits to the second processor [7]. After executing all these steps, it reveals the information in Liquid Crystal Display.

If valid tag is missing, then the former process backtracks again and thus the infinite loop continues. In the second processor, system activates by reading the analog voltage values A0, A1, A2 and then examines to determine the type of vehicle. For example, in case of truck it records the weight of the truck and the calculation is conducted by using the method below:

Total amount = Preset toll amount for truck + (additional weight)/2

For instance, if the vehicle is 50kg overweight, then the user has to pay half of the weight value along with the initial fixed amount.

If valid tag is missing, then the former process backtracks again and thus the infinite loop continues. In the second processor, system activates by reading the analog voltage values A0, A1, A2 and then examines to determine the type of vehicle. For example, in case of truck it records the weight of the truck and the calculation is conducted by using the method below:

Total amount = Preset toll amount for truck + (additional weight)/2

For instance, if the vehicle is 50kg overweight, then the user has to pay half of the weight value along with the initial fixed amount.

After performing the calculation, it displays the corresponding data and sends confirmation text using AT Commands. Then, it sends command to rotate servo motor at 90 degree angle. After 30 sec of delay, it will rotate back into 0 degree angle again.

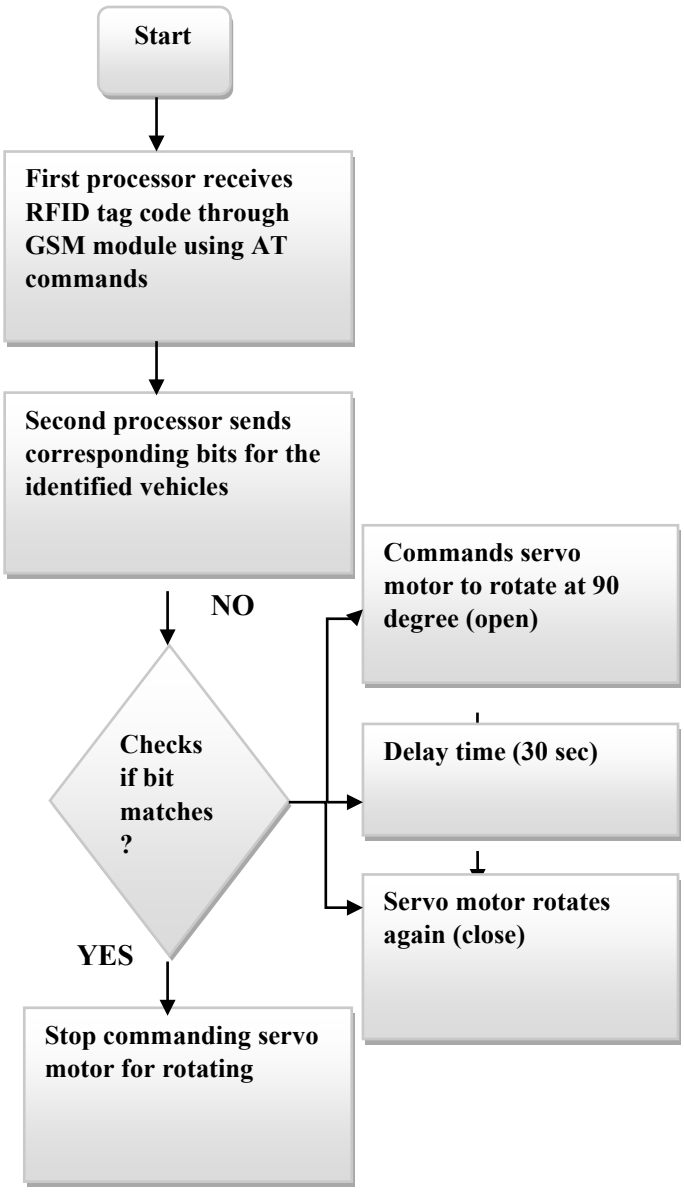


Figure 3: Steps for integrating between two processors

The security system starts when the first processor receives the RFID tag code by using GSM module. By using some AT commands, this process is done. After that, the second processor sends the currently available tag code. Then the tag code is matched with the currently available vehicle's RFID tag code in the database. If it matches then the system stops commanding for the rotation of the servo motor. So, the barrier remains closed. However, if it doesn't match, then the servo motor commands to rotate at 90 degree angle and eventually open the barrier. After 30 sec of delaying, the servo motor again rotates back at 90 degree angle to close the barrier.

To control the barrier for security purpose, we have used a code text message with the help of GSM module. The code text format for blocking the system is given below:

SMS Format: **Block<space>Car/Bus/Truck/CNG/Bike**

For example, to block a car the authority should write a text message to the system like Block Car and send it to the system.

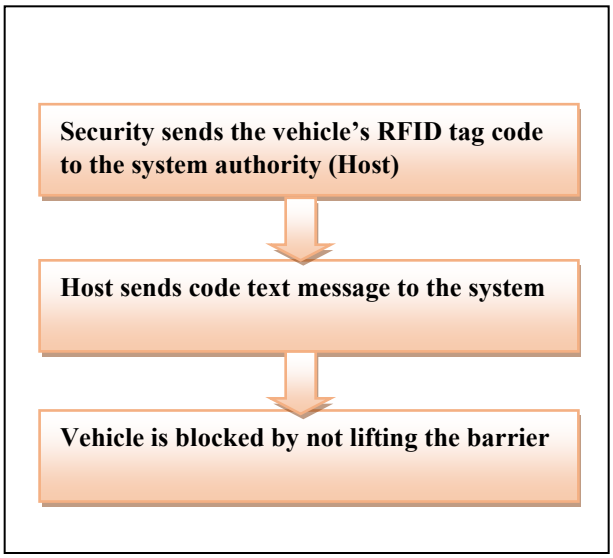


Figure 4: Blocking procedure

7. Results and Discussion

The designed prototype system has arranged a RFID booth over the road where the RFID reader was positioned for showing the demonstration. We had utilized a 16x2 Liquid Crystal Display for exhibiting purpose and also applied a 10k potentiometer to regulate the contrast of LCD display. We have used five different RFID tags or cards which represent five different types of vehicle, i.e., Bike, CNG, Car, Bus and Truck. A SIM card was inserted in the GSM module using which text messages were sent to the phone numbers that were already stored in the code for each RFID card.

The satisfactory level of our demonstration was adequate enough. The process started when a vehicle crossed RFID booth area, it was scanned instantaneously and informed by means of some LED's in the circuit. At the beginning the LCD exhibited in the display '*no vehicle……charge is 0'*.In contrast, in case of vehicle without proper RFID tag, the LCD display showed '*no tag present, use alternate path'*. Rather than truck for the other four types of vehicle (Bike, CNG, Car, Bus), it showed in the LCD display the type of vehicle and the corresponding charge.

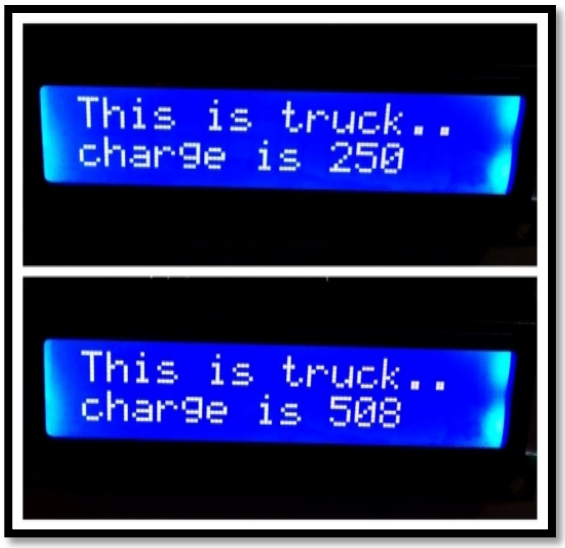


Figure 4: Toll charge for truck

We charged different amount of toll for the 5 types of vehicle. A confirmation text was received immediately in the phone number we had given in our system code. For truck the weight was calculated by varying the weight using FSR. For additional weight the extra toll amount that has to be paid was calculated using the following formula:

Previous Toll Charge for truck + (1/2) * weight = Total Toll charge for truck.

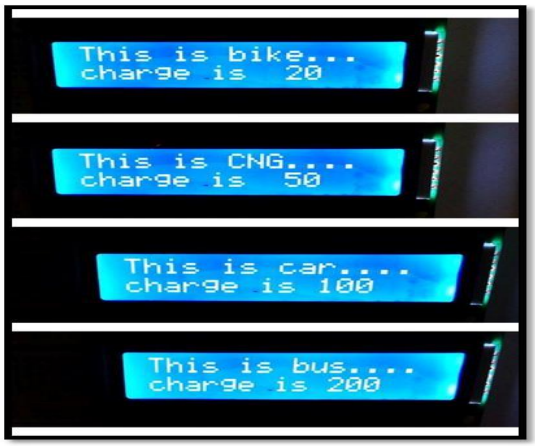


Figure 5: Toll charge based on different types of vehicle

Figure 6: The example of authentication messages

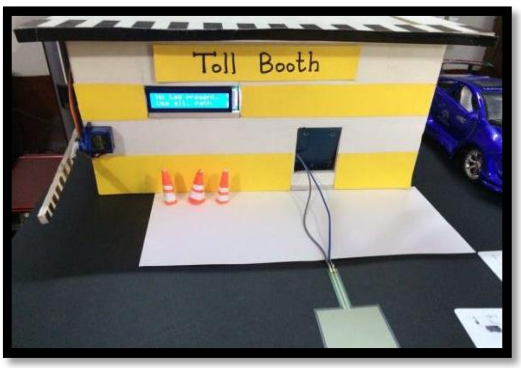


Figure 7:The prototype of proposed system

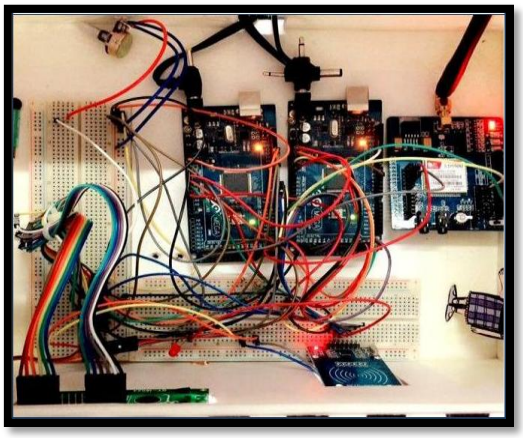


Figure 8: Internal circuit of the proposed system

7.1 The result of the security feature

In our demonstration, first of all we recorded five RFID cards in the database for five different types of vehicle. Now, we wanted to block the RFID card which was registered in our database system as 'car'. For this, we sent a text message writing 'Block Car' to the system and then the LCD showed '*Car Blocked'* and barrier was not lifted.

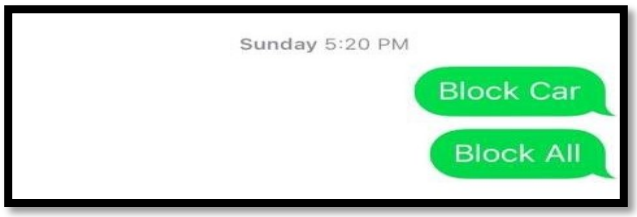


Figure 9: The example of message format for blocking vehicle

Figure 10: The example of screen for blocking vehicle

Then againwe wanted to block the whole system so that no vehicle can pass the toll booth area. For testing, we used the format of the SMS "All<space>Block" and sent a message to the system. After receiving the SMS, when we passed the RFID cards, the system blocked all the five types of vehicle and showed in the display "*All Blocked"*.

Figure 11: The example of screen for blocking all types of vehicle

```
int8_tsendATcommand(char* ATcommand, char* expected_answer, unsigned int
timeout){
uint8_t x=0, answer=0;
char response[100];
unsigned long previous;

memset(response, '\0', 100);
delay(100);
// Initialice the string

while( Serial2.available() > 0) Serial2.read();
// Clean the input buffer
```

Figure 12: Part of coding used in the proposed prototype

The overall programming is done by using Arduino code. A part of our code is shown in Figure 12.

In order to determine the performance, the key matrices are the distance and the range of RFID reader and Card. Time is also the parameter depending on which the overall performance of the system was determined.

In our prototype, we have used an active RFID reader which can measure up to 3.5 cm distance. We did the distance vs efficiency calculation by checking how fast the RFID card is detected by the reader while passing the booth area and also kept the record when it was failed to do the detection process.

For the improvement of efficiency we need to use higher frequency ranged RFID reader. Taking some sample data we tried to generate the graph using MATLAB to see the performance approach over the distance.

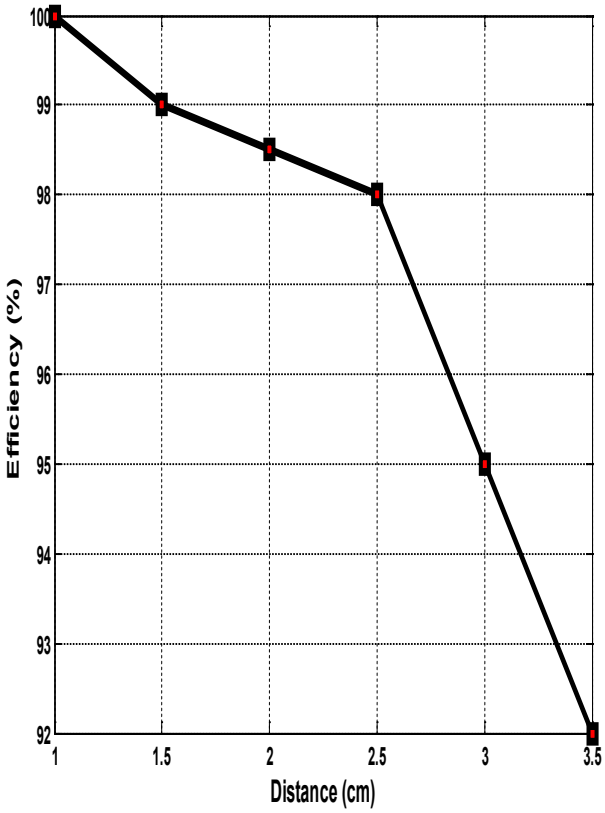


Figure 13: The comparison of efficiency and distance

From the graph, it is clearly visible that the efficiency of the system slowly decreases from 1cm to 2.5cm. After that the graph drastically decreases between 2.5 cm to 3.5 cm. Therefore, the finding is higher the distance lower the efficiency.

8. Cost Analysis based on the proposed prototype

One of the major steps for developing a compact system is taking into consideration about its building cost. As, our aim was to model the system for developing countries or underdeveloped countries, we tried to keep in mind about the cost issue. The system is very cost efficient and feasible for developing nations like Bangladesh, India, and Myanmar. Apart from that the components are easily available too. The whole circuit can be implemented on a PCB design board to have a compact design. The list and calculation of our total cost of the whole system are givenin Table 1.

9. Constraints

Junk messages are a problem which can hang the system. To erase the junk messages, the Reset button of the Arduino Mega 2560 is pressed. The latteris connected with the Liquid Crystal Display.

The implemented system has a short range for RFID tag detection, which are only a few centimeters at this time. This works fine in the small prototype system. Improvement of the system is possible by using more powerful RFID, which can be used for practical deployment with actual objects.

Table 1.Price of components and cost of the prototype system

Components	Quality	Amount(BDT)
RFID Reader	1	700
RFID Card	5	5x10=50
Arduino Mega	2	(2x1200)=2400
16x2 LCD	1	180
GSM Shield	1	3500
Servo Motor	1	320
Potentiometer	1	20
Force Sensitive Resistor(FSR)	1	912
Breadboard	2	(2x120)=240
Other materials		678
Total 9000 BDT = 112.5 USD		

10. Conclusion

The paper has presented the successful design of an embedded security system along with the automated toll collection system for reducing transport related crimes on highways and bridges. The overall system is based on RFID technology. Hence, there is no chance of using duplicate number plates as RFID tag is unique for each vehicle. The designed system should help the law enforcement agencies in many ways in reducing and preventing highway crimes. The system is cost effective and sustainable. It is expected that the system will sustain for a long time. It is also resistive to all external functions and parameters. For developing and underdeveloped countries, this automated security system along with automatic toll collection system should help greatly in reducing crime rates.

11. Future Work

A new feature can be added to the system which may include a database that stores vehicle data that includes vehicle owner's photo, name, national identification number and residential address. After the vehicle with security problem is detected, the information in the database related to the owner will be immediately pulled and sent to the police close to the area via text messages.

Another feature can be added to the system in where CCTV can be placed at the toll booths. They will capture the image of the driver. The system will match the image with the images of known criminals with outstanding warrants. If there is a match, the toll boothwill block the passage of the vehicle and notify the law enforcement agency immediately.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] R. Hossain, M. Ahmed, M. Alfasani, H. Zaman and R. Islam,"An Advanced Security System Integrated With RFID Based Automated Toll Collection System," Proc. 3rd Asian Conference on Defence Technology(ACDT 2017), pp. 59-64, Phuket, Thailand, 18-20 January, 2017, DOI: 10.1109/ACDT.2017.7886158

[2] R. Kumar, A. Bennet, L.P and R.S, "Automatic Toll Gate System using Advanced RFID and GSM Technology", Int. Journal of Recent Scientific Research,vol 7,issue 4, April, 2016.

[3] H.Khali, A. Araar and Z. Abdulla, "Sustainability PassiveRFID Technology for Fast Moving Vehicle Identification", Journal of Imerging trends in Computing and Information Sciences, Vol 5, No. 1. Jan 2014.

[4] S. Nandhini and P. Kumar," Automatic Toll Gate System Using, Advanced RFID and GSM Technology" International Journal Of Advanced Research in Electrical, Electronics and Instrumentation Engineering, Vol 3, 2014.

[5] T. Aphale, R. Chaudhari, and J.Bansod ,"Automated Toll Plaza Using RFID and GSM," International Journal On Recent and Innovation Trends in Computing and Communication, Vol 2, No. 9.

[6] Y. Tseng and H.Pan ," Data hiding in 2 color images", "IEEE Transaction on Computers, Vol 51, Issue 7, Jul 2002

[7] H. Bauma, J. Bann, G. Burghouts and J. Rest," Automatic detection of suspicious behavior of pickpockets with track-based features in a shopping mall", Proc. SPIE, Amsterdam, September,2014,DOI:10.1117/12.2066851

[8] A.M.Win, C.M. New ,andK. Z.Latt ,"RFID Based Automated Toll Plaza System," International Journal of Scientific and Research Publications, Vol 4, No. 6, June 2014.

[9] R. Hossain, M. Ahmed, M. Alfasani, H. Zaman and R. Islam,"Automated Toll Collection System," Proc. 8th Intl. Conference on Electrical, Electronics and Civil Engineering (ICEECE'16), pp. 41-44, Dubai (UAE), Planetary Scientific ResearchCenter(PSRC)

[10] The Mega 2560 is a microcontroller board based on the ATmega2560. Retrieve from: http://www.arduino.cc/en/Main/ArduinoBoardMega2560

[11] FORCE-SENSITIVE RESISTOR (FSR). Retrieved from https://www.adafruit.com/products/166

[12] "Working Paper on future RFID Research Needs", Cluster of European RFID Projects, September, 2007.

[13] P. Salunke, P. Malle, K. Datir and J. Dukale, "Automated Toll Collection System Using RFID", IOSR J. of Comput. Eng., vol 9, issue 2, pp. 61-66 (January-February. 2013).

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1784-1798 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Investigating the Expertise Indicators of Vulnerability Discovery Professionals

Mortada Al-Banna*, Boualem Benatallah, Moshe Chai Barukh

UNSW Australia, Computer Science and Engineering, Sydney, Australia

ARTICLE INFO

Article history:
Received: 15 June, 2017
Accepted: 29 August, 2017
Online: 23 September, 2017

Keywords:
Crowdsourcing
Expertise
Vulnerability Discovery
Professionals

ABSTRACT

In crowdsourcing, selecting the person with suitable expertise is very important; especially since the task requester is not always in direct contact with the worker. Recently, this has become increasingly important particularly when the crowdsourced tasks are complex and require skillful workers (e.g. software development, software testing, vulnerability discovery, and open innovation). In this paper, we investigate the expertise indicators of vulnerability discovery professionals in a crowdsourcing vulnerability discovery platform. We conduct a systematic literature review, we review online contents, conduct interviews with domain experts, and survey vulnerability discovery professionals involved in the task of vulnerability discovery. We discuss the indicators we have found, and we provide some recommendations to help improve the process of selecting vulnerability discovery professionals to perform crowd tasks related to vulnerability discovery.

1. Introduction

The complexity of software-based systems is increasing dramatically as development becomes even more distributed across multiple heterogeneous, autonomous, and evolving cloud services. More specifically, the increased reliance on third-party software-based systems (e.g., cloud services, open APIs, external programming libraries and black-box software packages) makes it very difficult for in-house IT experts to deal with the inherent risks of using external software. In order to overcome potential vulnerability issues, several organizations outsource tasks such as vulnerability discovery to third-party providers. More recently, the approach of crowdsourcing vulnerability discovery has emerged. Companies like Google, Facebook, and Microsoft have their own crowdsourced vulnerability discovery programs (aka vulnerability reward programs). At present, even the government sector is adopting this approach (e.g., United States Department of Defense). In Addition, several crowdsourcing platforms for vulnerability discovery have emerged (e.g., Bugcrowd 'Bugcrowd.com', HackerOne 'hackerone.com', Synack 'synack.com', and Cobalt 'cobalt.io'). Similarly, competitions have been conducted for discovering vulnerabilities (e.g., Pwn2Own). This paper is an extension of work originally presented in IEEE 2nd International Conference on Collaboration and Internet Computing (CIC) [1].

Crowdsourcing vulnerability discovery allows organization to benefit from the merits of crowdsourcing (e.g., diversity, scale and speed). Alongside the benefits of crowdsourcing, some quality control concerns have emerged (e.g., clarity of the outcome, trustworthiness of the crowd) [2]. Facebook, in their vulnerability discovery program, highlights the following : "We received 13,233 submissions in 2015. Of these, only 526 were valid and eligible to receive rewards". Similarly, Bugcrowd mentions in the state of bug bounty report for 2016 [3], the platform received from January 1, 2013 to March 31, 2016, a total of 54,114 submissions, of which only 9,963 contained valid vulnerabilities. Although the variation in expertise is considered desirable for discovering vulnerabilities [4], organizations may be overwhelmed by the number of contributions and hence need to adopt techniques to ensure the quality of submitted vulnerability reports. One technique used in crowdsourcing to mitigate quality concerns is worker selection. Research shows that recruiting suitable workers enhances the quality of contributions [2,5].

In this respect, we are not advocating that limiting vulnerability discovery to selected vulnerability discovery professionals from the crowd is superior over open programs. The scope of this paper is to examine how Vulnerability Discovery Professionals (henceforth referred to as VDPros) are selected in crowdsourcing vulnerability discovery platforms. Organizations that crowdsource the task of vulnerability discovery are not always in direct contact with VDPros whom perform the task. Additionally, traditional methods of recruitments are not always feasible (e.g., interviews, probation periods). Hence, understanding expertise indicators of VDPros participating in the crowdsourced task of vulnerability discovery becomes even more important. We aim to understand how to measure the expertise of VDPros in order to make sure that the selected VDPro would yield a high-quality outcome for the vulnerability discovery task. In practice this is critical, since poorly

*Corresponding Author: Mortada Al-Banna, UNSW Australia, Computer Science and Engineering, Sydney, Australia | Email: m.al-banna@unsw.edu.au

managed vulnerabilities could cause huge losses to organizations whether financially or to the reputation . Our study aims to answer the following research questions:

– RQ1: What are the perceived indicators of expertise for selecting VDPros?

– RQ2: Which sources are accepted as reliable to extract these indicators of expertise for selecting VDPros?

While there is a large body of research for worker selection in crowdsourcing [6], to the best of our knowledge none yet addresses the selection of VDPros in crowdsourced vulnerability discovery. Giboney et al. developed a conceptual-expertise-based tool that can be used to discriminate between novice and expert VDPros [7]. They rely on self-reported skills and testing the knowledge of the VDPros to measure their expertise. This could be suitable in traditional recruitment but for crowdsourcing systems it might have some shortcomings (e.g., recurrent questions, automating simple question-answering) [8]. On the other hand, Hafiz and Fang performed an empirical study on VDPros who have disclosed vulnerabilities [9]. They investigated methods and tools used by VDPros to discover the vulnerabilities and how the community of VDPros tend to focus on certain vulnerability types. While they have touched some aspects in regard to the expertise of VDPros involved in discovering vulnerabilities, they have not explored the criterion for indicators of expertise with respect to VDPros.

Accordingly, this paper makes the following main contributions: we have conducted an extensive review of literature using a "systematic literature review" methodology [10], along with investigating a diverse range of online contents (e.g., blogs, technical reports and articles). We have then adopted "exploratory research methods" [11,12], where we relied on the insights we gained from the literature and the online contents, to conduct semi-structured interviews with domain experts. We also conducted an online survey for the VDPros involved in the task of vulnerability discovery to validate our findings. We employed "qualitative data analysis" guidelines [12,13] to analyse our data collected and scientifically present our findings. And finally, based on our findings we propose recommendations that we believe will pave the way for further advancement in the effective selection of security experts in vulnerability discovery crowdsourcing. While arguably personality traits are important characteristic in the selection of VDPros, in this paper we only focus on the indicators to expertise. An indicator to expertise is a signal about the knowledge of the VDPros in a certain area and their competency to solve problems within that area.

The rest of this paper is organized as follows: In Section 2, we provide some background knowledge. In Section 3 we discuss our review of literature and online contents. The methodology we adopt in this work is thereby illustrated in Section 4. In Section 5, we outline our findings; and in Section 6 we provide recommendations to help improve selecting VDPros for vulnerability discovery tasks. We conclude with a discussion and directions for future work in Section 7.

2. Background

In this section, we examine the topic of crowdsourcing and its application to vulnerability discovery. We provide an illustrative example to further clarify the research problem.

2.1. Crowdsourcing Vulnerability Discovery

A vulnerability is a security flaw, which arises from system design, implementation, or maintenance (e.g. SQL injection vulnerability). By exploiting these vulnerabilities, malicious parties could gain unauthorized access to protected resources. The field of vulnerability discovery and repair has been very active in recent years, especially with the increase in number of security threats and incidents [14].

Crowdsourcing (also known as human computation) harnesses the wisdom of large groups and communities working independently to solve problems, much as open source does for software development. From the service customers' perspective, crowdsourcing is a form of digital service (notwithstanding its essentially human infrastructure), yet its unique value draws from its effective way to perform tasks that remain difficult for, or even beyond the reach of machine computation [15] (e.g., image tagging, natural language translation, and transcription). Several organizations including DARPA, NASA, Honda and various other organizations use platforms such as MTurk and Ushahidi to crowdsource information gathering. Other commercial crowdsourcing platforms have recently emerged alongside MTurk; platforms like Innocentive, TopCoder, Kaggle, and uTest use crowdsourcing for tasks that require skilled workers (Web design, testing, Web development tasks, R&D challenges).

Crowdsourced vulnerability discovery programs allow vulnerability discovery professionals to play the role of an attacker to discover vulnerabilities. In a crowdsourced vulnerability discovery program, software providers submit vulnerability discovery tasks to a community of VDPros. This approach is gaining increasing popularity recently [1] . Crowdsourcing vulnerability discovery could be in the form of an open call and managed directly by an organization (e.g., Facebook VRP, and Google VRP), or directed toward members of specialized platform for crowdsourcing vulnerability discovery (e.g., Bugcrowd, Cobalt, HackerOne, or Synack). Crowdsourced vulnerability discovery programs could be publicly available for everyone and VDPros self-select and decide to participate according to their preference (e.g., Uber program in HackerOne[2]), or could be private where the organization invite only selected VDPros to participate according to a specific criterion (e.g., LinkedIn Private program[3]). There are also unregulated methods for the crowd to submit discovered vulnerability information (e.g., black and grey vulnerability markets[16], and VDPros disclosing vulnerabilities publicly[4]).

Several researchers investigated crowdsourced vulnerability discovery from different angles. Laszka et al investigated the problem with invalid submitted vulnerability reports and their relation to incentives misalignment [17]. Zhao et al. investigated the web vulnerability discovery echo system and analyse vulnerability trends, response behaviour and reward structures [18]. Finifter et al. investigated two crowdsourced vulnerability discovery programs [19]. The authors illustrated that crowdsourced vulnerability discovery programs are economically efficient and explored the difference in the reward strategies and the effect on the engagement with VDPros.

1 www.vulnerability-lab.com/list-of-bug-bounty-programs.php

2 hackerone.com/uber

3 security.linkedin.com/vulnerability-disclosure

4 schneier.com/essays/archives/2007/01/schneier_full_disclo.html

2.2. Illustrative Example

Let's assume that a company called AwsomeAPI has decided to delegate the vulnerability testing for its new API to a crowdsourcing platform for vulnerability discovery. AwsomeAPI is concerned about the quality of the output of the task. As the quality of the output from the crowdsourced task is related to the expertise of the crowd worker [2], AwsomeAPI has thereby decided to limit participation of the program to only expert VDPros. AwsomeAPIs has decided that the scope of the program is to protect its API from the injection attacks (e.g. SQL injection, and script injection) [20]. The internal team in AwsomeAPI has identified that in order for the VDPro to be competent in the vulnerability discovery task, they need to have: knowledge about the technology (e.g., the API to be tested, and protocols such as http, and SSL/TLS); programming skills (e.g., low level programming, and scripting); knowledge about security practices (e.g. security controls such as authentication and authorization, and software vulnerabilities).

AwsomeAPI is facing some difficulties in regard to selecting VDPros with the required level of expertise, and how can they assess the expertise of the VDPro.

3. Expertise Indicators from Literature and Online Contents

We reviewed literature using a systematic literature review (SLR) approach where we followed the guidelines illustrated in [10,21]. As an additional source of information, we have also examined a range of online content, such as companies' blogs, and technological articles, and rely on the online research methods [22]. We provide a brief description and summarize our findings in this section.

3.1. Steps of the Systematic Literature Review

Various approaches amongst literature have been adopted to determine expertise. Hence the use of literature review helps in identifying the indicators that have been relied on in these approaches and how relevant they are in measuring the expertise of VDPros involved in the task of vulnerability discovery. We adopt the methods of literature review on the recommendations of Kitchenham [10] and Okoli et al. [21].

I. Selecting the Articles

We searched for relevant articles addressing people selection according to expertise. We included both journal and conference articles as the conference articles have been found to be of important role in the field of information systems [23]. The search was conducted for the period between January 2009 and January 2016 for articles in IEEE Digital Library, ACM Digital Library, and Scopus databases. We focused on articles available in full text and in English language. We used the following keywords for the search: "Expert" OR "Expertise" OR "Reputation", "Selection" OR "Finding", "Crowd" OR "Crowdsourcing". We checked that searches using similar keywords like 'crowd-source', 'crowdsourced', and 'crowdsource', identify the same corpus and therefore can be discarded.

The result of the search was 172 articles. After removing the duplicates, editorial introductions, posters, tutorials, workshop summaries, and articles that just mention the search keywords and that are not relevant to discovering expertise, a total of 96 articles remain in the initial pool of resources.

II. Filtering Articles and Additional Search

We applied the following inclusion criteria: (i) articles concerned with selecting experts to perform a task; (ii) articles measuring expertise of workers in a platform or members in a community; and (iii) articles measuring reputation for workers in a platform or members in a community. We also applied the exclusion criteria for articles with the main focus of relying on self-assessments (e.g., self-reported skills or qualities).

By applying the inclusion and exclusion criterion, the number of articles becomes 23. We performed additional literature search in a less systematic approach, focusing on articles that we were already aware of ourselves, articles found by a full text search using Google scholar, and relying on back snowballing where we looked at references cited in the articles we have. We believe that this addition has provided a more comprehensive review of scientific contributions in this area. The final number of articles that were reviewed was 34.

III. Data Extraction and Synthesis

We reviewed the articles in order to identify the indicators of expertise relied on in literature. The first author codified the review articles and the other authors checked the extracted data and modifications were made according to the comments. In any cases of disagreement on data extraction, the three researchers discussed until reaching consensus.

We synthesized the extracted data in order to answer the research question. We reviewed the extracted data and built a list of the indicators of expertise that were relied on in each article. We then merged the indicators that were related together and narrowed down our list of indicators to five indicators.

3.2. Emerging Indicators from the Literature Review and Online Contents Investigation

Five indicators to expertise have emerged from the literature (illustrated in Fig. 1). In the following, we briefly discuss each of these indicators, and we explain the resources that motivated us to derive them as appropriate indicators. Table 1 illustrate a summary of the indicators and the supporting research work.

Figure 1. Selecting Vulnerability Discovery Professionals (Expertise Indicators)

I. Certification

This is one of the most known indicators to knowledge in both educational bodies (academic record certification) and industrial bodies (industrial professional certifications). Relying on a proof issued by a trusted party (whether being tested or not) the crowdsourced task owner will know that the VDPro have the required knowledge to perform the task of discovering

vulnerabilities. Crowdsourcing platforms like MTurk allow the task requester to issue a qualification test for each of the crowd workers to assert their eligibility to perform the task. Online freelancers market UpWork (previously known as oDesk) rely on certifications to assess the competency of the freelancer along with online testing for the skills. Similarly, Synack which is a crowdsourcing vulnerability discovery platform, administer to VDPros a written and practical evaluation to certify that they are eligible to join the platform [24].

Ejiaku et al. in their work discussed the importance of Information technology certification and its impact on the recruitment process [25]. In a similar context, Bishop and Frincke in their work discussed general professional certifications and compared them with academic certifications. They also looked at how each of these certifications reflect the knowledge required for certain jobs and their effect in increasing the possibility of success on the job [26]. On the other hand, McGill et al. addressed the importance of recertification in order to ensure the validity of the certification and express the true knowledge and expertise of the certification holder [27].

Kanij et al. in their work surveyed several software testers for factors they considered most important to the task of software testing [28]. In their study, 56.7% of the respondents to the questionnaire indicated that they had done training or certification in software testing in the last five years and the majority of the participants found certification and training useful. In a similar context but in different field, Botella et al. identified certification as a key requirement for persons who aim to become an expert in usability evaluation [29].

II. Referrals

Referrals or recommendations are also well known in the domain of recruitment as a way to gather and confirm expertise. Many organizations rely on referrals and encourage the employees to recommend potential candidates for employment. Salesforce has indicated in its official blog that its main strategy for recruitment for 2015 is based on referrals [30].

Many online professional communities and professional social platforms have recognized the importance of referrals in determining expertise. For instance, professional social media platform LinkedIn facilitates endorsements in order to tag a specific skill to the professional [31]. CoderWall allow users to endorse and assert any skill that the profile owner has mentioned in her profile [32]. Repcoin a novel reputation market platform rely solely on recommendations to build up the users' reputation of having expertise in a certain field, and as it is built on top of the bitcoin technology the platform facilitates having the reputation as a portable accreditation that could be presented anywhere [33].

Some research has addressed the impact of referral as an indicator to expertise. Schall et al. in their research introduced a context-sensitive trust-based algorithm called ExpertHITS inspired by the concept of hubs and authorities in Web-based environments [34]. They rely mostly on referrals from experts to delegate a task to another expert in her network. In the same context, Zhang et al. proposed a mechanism for finding experts by a referral chain from the initiator to the expert [35]. Similarly, the work of Pushpa et al. has addressed searching for expertise in a network of co-authorship where a query to find an expert is seen by an agent that manages the expertise model of the users. It thereby makes a decision to forward the request to relevant experts in the neighbourhood of co-authors if the query is not matched [36].

Within the enterprise work environment, finding an expert to perform certain task has been a problem addressed by several researchers. Braub et al. have proposed people tagging where they rely on employees to tag certain skill or knowledge to their colleagues [37]. Ghosh et al. have proposed relying on user curated Twitter lists to determine topic experts, since users include experts on topics that interest them in their lists [38].

III. Performance Review

Reviewing the performance of crowd workers has been relied on as an indicator for knowledge and expertise in many crowdsourcing platforms [39,40]. Similarly, the performance of VDPros within the crowdsourcing vulnerability discovery platform could be used to assert their level of expertise (reflected by a high reputation score). BugCrowd calculates a reputation score for each VDPro according to the number of discovered vulnerabilities, the speed of discovery, and abiding by the rules of the vulnerability discovery program [41]. Similarly, Cobalt (former known as Crowdcurity) relies on the same aspects but they add to them the feedback from the task requester and the quality of the vulnerability discovery report [42]. Hackerone in addition to relying on the number of the vulnerabilities discovered, speed and abiding by the rules, they add extra points for VDPros who acquire a bonus credit from the task requester (e.g., for clever vulnerabilities, or critical ones) [43].

Zhao et al in their work highlighted the performance of VDPros in a web vulnerability disclosure program as a measure of their expertise [44]. Similarly, Algarni and Malaiya in their work relied on the performance of VDPros to define the top discoverers of vulnerabilities [45]. Anvik et al. proposed an approach that apply machine learning algorithm to an open bug repository to recommend the suitable developer to fix a software bug according to their past performance [46]. In the same context, Hassan and Curry in their work proposed extracting the crowd worker expertise from the previously accomplished tasks of certain type (e.g. image tagging) in order to predict the performance of the crowd worker in tasks of different types (e.g. image recognition) [47]. Sarasua and Matthias however, proposed the use of a cross-platform Curriculum Vitae in their work to record the crowd worker expertise and skills according to the performance of the crowd worker in several crowdsourcing platforms [48].

Ebay, as an online market, set reputation for each of the sellers according to the quality of the goods they provide and their performance (e.g. speed of delivery, how they handle returns) and accordingly they categorize the seller as being expert [49]. Resnick and Zeckhauser in their work looked at the reputation system in eBay and how predictive it is for future performance of the sellers [50]. Similarly, UpWork[5] rely on the freelancer's job success rate and client feedback to assess her expertise.

IV. Artefacts

The artefact could be in the form of a software code in GitHub, an answer to a question in a Q&A platform like Stackoverflow, or a published article. Vasilescu et al. have looked at the relation

[5] www.upwork.com/i/how-it-works/client/

between the activity of users in Stackoverflow and GitHub, they have discovered that active GitHub committers ask fewer questions and provide more answers than others in Stackoverflow [51]. In the same context, Hauff et al. in their work proposed an approach for matching job advertisements with expert developers according to their artefacts contributed in GitHub [52].

Table 1. Indicators and Supporting Articles

Indicator	Supporting Articles
Certification	[9] [11] [21] [34] [46] [62]
Referral	[12] [26] [44] [52] [56] [71] [75] [76]
Performance Review	[2] [5] [28] [32] [33] [53] [54] [57] [64] [69] [72]
Artefacts	[8] [15] [16] [23] [30] [37] [50] [63] [65]
Association	[25] [43] [54] [58] [61] [68] [70] [74]

Kolari et al. in their work looked at blogs related to certain enterprise as a source of evidence of that expertise for potential employees. The authors investigated whether depending on blogs has similar effect as depending on emails but with less privacy concerns and they even have the added value of allowing implicit voting via comments by the community [53]. Balog et al. have presented strategies for finding experts relying on document repositories in the enterprise [54]. In their work, they have proposed two approaches, the first determines the employee expertise based on the documents (e.g., reports, manuals) that they are associated with, whilst the second locates documents based on topic, and then finds associated experts.

Pelechrinis et al. [55] proposed an approach to infer expertise and reliability of the answers provided in a Q&A platform depending on the pattern of the authors answering to other questions. The authors investigated users' consistent responses to questions related to particular topic and their expertise in that topic, as well as determining whether the user is answering too many unrelated topics, it would therefore be safe to assume they are an amateur in each of those areas.

Demartini investigated whether in order to find experts with certain knowledge, it is better to look at where knowledge is created. The author proposed an algorithm to find experts according to their contribution in Wikipedia [56]. Chang et al. in their reputation based access control for crowdsourcing platform relies on the quality of the artefacts submitted by the crowd worker in order to assert the quality of the worker herself [57].

In the academic field, academics have been evaluated by their publications such as the number and quality of publications, number of citations, and the h-index. Tang et al. have presented an academic search system, called ArnetMiner which relies on the publication of academics as an indicator to their knowledge and expertise in a particular field amongst other sources [58]. Similarly, Freund et al proposed an expert-finding system in the academic domain where they also rely on the publications of academics as one of the measures of expertise [59].

V. Association

Relying on associations to determine expertise is an approach that is recently emerging. Fu et al. in their work proposed strategies to discover the associations amongst people from emails and Web pages. Relying on these associations, a candidate can acquire extra expertise probability from a reliable expert who has strong relationship with the candidate [60]. Li et al investigated documents' co-authorship as a relation between experts, and they assessed if the probability of being an expert increases if they have co-authored a topic relevant document together with a well-known expert [61]. Similarly Zhang et al. proposed a propagation-based approach to expert finding in a social network where they relied on local information (e.g. published articles in certain topic) and then propagated possible expertise of a candidate according to their relationships (e.g. co-authored or supervised) [62].

Daniel Schall introduced a ranking approach called Dynamic Socially-Aware PageRank where the expert ranking depends on context aware interactions. As an example, an expert of certain topic may attract a large number of help requests in regard to that topic [63]. In the same context, in [64] the authors investigated the usage of social relationships in a microblogging platform as a potential evidence that a person is a real expert. In the sense that if user A who is an expert follow user B, then user B is most likely to be an expert too. Similarly, Zhao et al. in their work have addressed the problem of lacking information about the crowd worker's past performance (cold-start worker). They proposed an approach to determine the expertise of the worker from social networks, such as, if the cold-start worker was associated with another worker who is considered to be an expert then the cold start worker is also to be considered and expert [65].

Pujol et al. in their work proposed an algorithm called NodeRanking, which rely on modelling the community as a social graph relying on relationships. Such as, being mentioned in the user's website or having email exchange between the two users. Their approach is to infer expertise according to the degree of connectedness of each of the nodes in the graph [66]. Suryanto et al. looked at Q&A platforms and proposed that users who ask high quality questions that attract experts are considered experts and users who provide answers to questions asked by experts are considers experts as well [67].

4. Methodology

In our study we rely on exploratory research methods [11,12]. Our study consists of three phases of data collection: (i) Literature review and online content investigation; (ii) Interviews; and (iii) Online surveys. Figure 2 illustrates the phases of the study. Our methodology consisted of data collection and iterative phases of data analysis.

In the first phase all the authors contributed to the discussion and analysis of the literature, and online content following the SLR guidelines [10]. For the second phase, the interview questions were discussed by the authors and refined through multiple iterations. The interviews were conducted by one of the authors. Summaries of the interviews were discussed by the authors to draw insights and identify the key emerging themes. For the third phase, the survey questions were designed by the same author who conducted the interviews. The survey questions were revised through multiple iterations of refinement by the authors and three external VDPros before being published. This helped us refine the terminology used in the survey questions and to ensure the study was clear and

motivating to VDPros. The results of the survey were summarized and discussed by authors. In this section, we discuss the research questions, study design, data collection methods and data analysis approach.

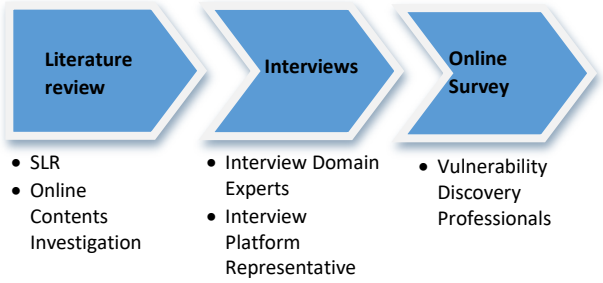


Figure 2. Stages of the Study

4.1. Study Design

In the first phase of our study, we reviewed scientific literature using systematic literature review methodology [10]. As mentioned above, we also investigated a diverse range of online contents (e.g., companies' blogs, experts' blogs, and technological articles) [22]. We utilized the findings in this initial phase to refine our interview and survey questions for the second and third phases. The results of the initial phase were already presented in the previous section.

In the second phase of our study, we interviewed 32 participants. We also approached one of the platforms for crowdsourcing vulnerability discovery for additional insights. Through iterative analysis of the data collected from this phase we were able to gain better understanding of what domain experts perceived as indicators to measure expertise in VDPros and what are the reliable sources to extract them.

In the third phase, by relying on the results of previous phases, we approached VDPros involved in the task of vulnerability discovery with an online survey. The purpose of this survey was to acquire the opinion of VDPros who are involved in the task of vulnerability discovery – and to thereby validate our findings from previous phases.

4.2. Study Design

In the first phase of our study, we reviewed scientific literature using systematic literature review methodology [10]. As mentioned above, we also investigated a diverse range of online contents (e.g., companies' blogs, experts' blogs, and technological articles) [22]. We utilized the findings in this initial phase to refine our interview and survey questions for the second and third phases. The results of the initial phase were already presented in the previous section.

In the second phase of our study, we interviewed 32 participants. We also approached one of the platforms for crowdsourcing vulnerability discovery for additional insights. Through iterative analysis of the data collected from this phase we were able to gain better understanding of what domain experts perceived as indicators to measure expertise in VDPros and what are the reliable sources to extract them.

In the third phase, by relying on the results of previous phases, we approached VDPros involved in the task of vulnerability discovery with an online survey. The purpose of this survey was to acquire the opinion of VDPros who are involved in the task of vulnerability discovery – and to thereby validate our findings from previous phases.

4.3. Data Collection

For the interviews, we invited participants via e-mail invitation, LinkedIn messages, and advertising our study in information security groups on LinkedIn (e.g., Information Security Community [6]). We also leveraged snowballing to get recommendation from people about their colleagues that would also be interested in participating [68]. Snowballing was the most effective approach since direct approaches may be considered spamming – especially as spamming attacks are well known by security experts who are much more cautious to accept invitations from an unknown party. The interviews lasted between 30-60 minutes and were conducted via Skype, Google Hangouts, telepresence systems, phone calls, and e-mail exchange. Audio was recorded when the interviewee consented and notes were taken. The interviews were semi-structured based on the findings of Phase 1 of the study, and the interviewer asked additional questions when appropriate. Summary of interview questions are outlined in Table 2 and the full interview guide is published online[7].

We also interviewed a senior representative from one of the platforms that supports crowdsourcing vulnerability discovery. This interview provided us with indispensable insight from a typical platform that is responsible for managing vulnerability discovery programs. Hence, they provided very valuable input in accordance to their experience from interacting with over 250 clients and thousands of VDPros.

Finally, we targeted VDPros involved in the task of vulnerability discovery with an online survey. We designed the questions of the survey based on the results from Phase 1 and 2. We shared the preliminary survey with three of the VDPros to acquire feedback before releasing it. The feedback helped us to modify some of the terminology used and modify some of the questions that were considered vague. The survey included both closed and open questions [69]. While closed-form questions are easier to summarize, open questions allowed respondents to include their remarks and comments.

We analysed the responses by applying structural codes and then aggregating the results for common trends [70]. Most closed-form questions used a Likert scale [71] with five possible responses from "Very Important" to "Not Important". The actual form of the survey is published online[8]. We sought participants via direct e-mail or tagging on twitter, LinkedIn messages, posting in Bugcrowd forum[9], Offensive Community site[10], and LinkedIn information security groups.

[6] www.linkedin.com/groups/38412

[7] Interview Guiding Questions: https://goo.gl/jVyOCe

[8] Survey Questions: https://goo.gl/forms/kXjIHjAEkZdwUvIt2

[9] forum.bugcrowd.com

[10] offensivecommunity.net/

Table 2. Summarized Interview Question for the Domain Experts

Characterization

1. Tell us about your organization (size, business area)?
2. What is your role in the organization?
3. Are you familiar with the approaches for vulnerability discovery (management, practices, decision making)?

Indicators to Expertise

1. What are the Indicators you rely on to measure expertise in vulnerability discovery professionals involved in the task of vulnerability discovery?
2. How much do you value certification as an indicator that could be relied on to measure the expertise in Vulnerability discovery professionals?
3. How much do you value referrals to measure the expertise of vulnerability discovery professionals?
4. How much do you value performance as an indicator to the expertise of vulnerability discovery professionals?
5. How much do you value artefacts generated by the vulnerability discovery professionals as an indicator to their expertise?
6. How much do you value association networks as an indicator to expertise for vulnerability discovery professionals?

Reliable Sources

1. What sources of certification you consider reliable to indicate expertise of vulnerability discovery professionals?
2. What forms of referrals do you rely on to measure expertise in vulnerability discovery professionals?
3. How much do you value past performance as an indicator to measure expertise in vulnerability discovery professionals?
4. What are the artefacts you think would be a reliable source to express expertise?
5. What forms of association do you think are reliable to express expertise?

4.4. Participants

Due to the time constrains and the exploratory nature of the study, we followed the recommendations of Adler and Adler and aimed for a sample size between 12 and 60 [72]. We reached saturation in the opinions of the interviewees and this gave us confidence that we captured sufficient depth to enable us to assert the solidity of our findings. We designed a comprehensive set of questions for the interviews and surveys relying on the literature review and online contents.

I. Interview Participants

Table 3 shows the details about the domain experts who have participated in the interviews. We refer by domain experts to personnel who oversee the security posture of organizations. Although they do not perform vulnerability discovery tasks as their daily job, they nonetheless influence (directly or indirectly) setting the criterion for selecting VDPros to perform the task.

For the purpose of analysis, participants are referred to by an anonymous identifier (P#). We invited participants from different disciplines (e.g., finance, entertainment, telecoms, etc.), so that we capture a diverse set of views sourced from different approaches for measuring expertise. Among the interviewees P3, P11, P13, P14, P17, P24, and P29 have experience in crowdsourcing vulnerability discovery. Furthermore, we interviewed a representative of one of the crowdsourcing vulnerability discovery platforms.

Table 3. Information about Interview Participants

P#	Industry	Size of Company (# Employees)	Position
1	Entertainment	450-600	CISO
2	Finance	350-500	CIO
3	Security Consultancy	30-50	Senior Security Consultant
4	Backup solutions	30-50	Quality Assurance manager
5	Networking solution	Global company	Senior Security Analyst
6	Multi-disciplinary	Global company	Security Testing Lead
7	Telecommunication	350-500	Information Security Expert
8	Telecommunication	350-500	Security Analyst
9	Payment systems	100-150	CISO
10	SaaS provider	30-50	Security Team Leader
11	Multi-media	100-150	Platform Security Lead
12	Security Consultation	50-100	Senior Security Engineer
13	Charity Foundation	15-30	Internal Security Assessor
14	Mobile Gaming	10-15	Security Lead
15	Video production	50-100	Principle Security Advisor
16	SaaS provider	100-150	Senior Security Engineer
17	Social Media	50-100	Security Lead
18	Financial Institute	150-250	CIO
19	Security Consultation	25-50	Penetration Testing Lead
20	Security Software	15-30	CTO
21	Medical	250-300	CISO
22	Education	300-500	IT Manager
23	Security Consultation	30-50	Senior Consultant
24	Software Provider	450-500	Security Architect
25	Financial Institute	250-300	Security & System Engineer
26	SaaS provider	250-300	Senior Security Analyst
27	Telecom Equipment	Global Company	Security Evaluation lead
28	Transportation	Global Company	Security Engineering Mngr
29	Bitcoin Exchange	150-200	Product Manager
30	SaaS Provider	25-50	Cyber Security Lead
31	Software Provider	400-450	CSO
32	SaaS Provider	600-700	CTO

II. Online Survey Respondents

We targeted VDPros who have been involved in the task of vulnerability discovery. We also targeted members of crowdsourcing vulnerability discovery platforms (Hackerone, Bugcrowd, and Cobalt), members of online cyber-security community along with relying on snowballing as well. The number of participants was 78 but we excluded 12 responses. In particular, 8 responses were submitted by participants that were not directly involved in vulnerability discovery; we used test questions in the

survey to confirm that. Another 4 responses were apparently artificial, since we noticed that these participants selected the first choice of all the questions, which would otherwise create incorrect inconsistencies in the data.

For the demographics of the participants in the online survey, 45% of the participants were aged 18-25 years old, 30% were aged 25-32, 13% were aged more than 33 years old, and the rest were aged less than 18 years old. With respect to location, the majority of participants were from the United States 37%, 21% were from India, 12% were from Pakistan, 9% were from Europe, 7% from Egypt, and the rest did not share their country of origin. Furthermore, 45% of the participants work in the security industry, 30% were students, 10% work in system administration, and the rest work in software development.

4.5. Data Analysis

We analysed the data following the qualitative data analysis guidelines [12,13]. The following stages were included in the analysis: (i) Transcription of the data recorded; (ii) Organization of the transcripts data and notes taken into easily retrievable sections; (iii) Getting more familiar with the data through reading and re-reading and writing down notes and summaries; (iv) Labelling data segments (we relied on structural coding to annotate the responses[70]); and (v) contrasting different views of the interviewees and re-coding to develop a more refined understanding of how the indicators are valued by the participants. As part of the analysis, similarities and differences of the compiled codes were clustered together in order to create categories. Conceptual saturation was reached when no new categories were generated.

5. Findings

In this section, we address the research questions and present the themes that emerged from our study. We also discuss the findings from the interviews and online survey, and annotate with selected quotes to help appreciate how we derived our conclusions.

5.1. The Interviews of Domain Experts

I. Indicators to the expertise of Vulnerability Discovery Professionals as Perceived by Domain Experts

There are four indicators for expertise that our study participants agreed on: (i) Certification; (ii) Referral; (iii) Performance review; and (iv) Artefacts. Relatively, there is weak support in regard to relying on association (see Fig 3).

Certification as an Indicator to Expertise: 75% of the interviewees value certification as an important indicator of expertise. One of the interviewees described it as a clear indicator to commitment "To get a formal education in the field from an academic body or being certified to have knowledge in the field via acquiring an industrial certification, it shows that the VDPro is passionate and committed enough to invest this much time and effort" [P1]. The remaining 25% of the interviewees did not consider certification as an indicator to expertise for VDPros. One of the interviewees stated that "Some skills required for the VDPro involved in the task of vulnerability discovery are not taught in universities or professional courses" [P19].

Referral as an Indicator to Expertise: Referral is highly valued by 90% of the interviewees since it provides attestation for the expertise of the VDPro from a person who they have known or worked with. One of the interviewees mentioned "Knowing that the person doing the referral has interacted with and observed the VDPro is a great way to measure expertise" [P1]. One of the interviewees indicated that when they trust a person doing the referral it would give very valuable credit to the VDPro especially since roles in the security field are sensitive ones. "Knowing and trusting the person doing the referral would mean minimized risks for us" [P8]. 57% of the interviewees mentioned that even if they do not know the person making the referral, if s/he is well-known in the industry, this would add substantial credit to the VDPro.

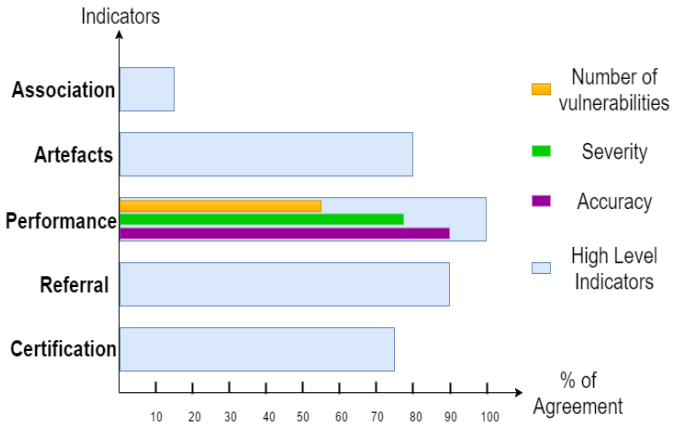


Figure 3. Indicators of Expertise (Domain Experts' Perspective)

Performance Review as an Indicator to Expertise: We found 84% of the interviewees assigned performance review as the top indicator while the remaining 16% considered it of high value. An interviewee stated, "If the VDPro has discovered vulnerabilities successfully many times before then why would s/he stop doing that now?" [P7]. Some considered it of high importance since it means the VDPro have used their knowledge in performing the task. One interviewee said, "Showing knowledge is good, but showing how the knowledge is implemented is more impressive" [P9].

We investigated further and asked about how the performance of VDPros is considered within the task of vulnerability discovery. Several themes have emerged. The first theme is considering the severity of the vulnerabilities discovered. 78% of the interviewees agreed that the severity of the discovered vulnerabilities is the best measure for performance. One of the interviewees mentioned "Discovering a critical vulnerability, that could cause a great loss or embarrassment to the company, is highly appreciated by any company and it means that the VDPro is an expert in what s/he is doing" [P15]. While the remaining 24% of interviewees considered it a good measure. Another theme is considering the number of vulnerabilities discovered by the VDPro. 56% of the interviewees considered it an important measure to performance. While 24% considered it somewhat an important measure. 19% considered it useful only when a good number of vulnerabilities discovered were of high severity. An interviewee has mentioned "If the VDPro has discovered thousands of vulnerabilities that of very low severity for me it does not tell they are good in what they are doing. May be they just have a good automated tool" [P12]. The last theme that has emerged is considering the accuracy in discovering vulnerability (i.e., the ratio of the valid vulnerabilities reported over all the vulnerability reports submitted by the VDPro) as a measure for performance. 90% of the interviewees agreed that accuracy is an important measure for the performance of the

VDPros. One of the interviewees mentioned "That is why we target VDPros with good expertise, we need to minimize the false positives since they consume a lot of resources" [P14].

Artefacts as an Indicator to Expertise: 81% of the interviewees agreed that contributed artefacts are important indicator for expertise. There was a contrast in opinions about the type of artefacts. 40% of the interviewees considered code artefacts and contribution to open-source projects to be the most valuable artefacts. An interviewee stated, "Developing your own testing tools is a clear indicator that you deeply understand the concepts" [P4]. Another 24% of the interviewees considered published articles, and blogging about vulnerabilities as a more important type of artefacts. One of the interviewees said "VDPros are supposed to be good in breaking systems and understanding how they work. I would consider a VDPro with good publications about security as an expert in the field" [P6]. While the opinion of the rest of the interviewees is that it is more effective to look at all artefacts and assess the quality as whole. An interviewee mentioned "Whether it is a publication, source code, or a blog, if I see many experts citing the work or using the open-source tool then it means the VDPro is not generating useless artefacts and that s/he is acknowledged for the contribution" [P4].

Association as an Indicator to Expertise: 84% of the interviewees considered association as a less important indicator of expertise of a VDPro. Sources like membership to a community, affiliation with a group, professional relationships with an expert were considered as indicators of passion and interest rather than expertise. One of the interviewees mentioned "Being affiliated with a group focused on security, indicates passion but not necessarily expertise" [P3]. Another interviewee shared with us "if the VDPro is well connected it could be useful since it indicates that the VDPro could reach out to experts if needed, but I would not say it is a must have" [P6]. The remaining 16% of the interviewees considered it a good indicator to expertise. An interviewee stated, "It is definitely not enough to see the number of connections or followers to determine expertise, but when I see interactions such as a post in LinkedIn being discussed or a tweet being retweeted by experts in the field then it must be that they acknowledge what the VDPro has produced" [P10].

II. Perceived reliable Sources to Extract the Expertise Indicators of Vulnerability Discovery Professionals?

As sources of reliable certifications, 87% of the interviewees valued certification from industrial bodies (e.g., $(ISC)^2$) the highest, followed by certifications from academic bodies (e.g., Universities) which was selected by 72% of the interviewees. On the other hand, accredited MOOCs and online learning were considered good to show eagerness to learn but not as much credible by 75% of the participants in the study.

As sources of referral, all interviewees valued direct referral the most (i.e. a direct form of communication). We asked the interviewees if they would value indirect forms like LinkedIn endorsements or LinkedIn recommendations and 85% of the interviewees did not consider LinkedIn endorsements as reliable. While 62% of the interviewees thought that recommendations in LinkedIn could be useful. One of the interviewees shared with us "People in LinkedIn could just be nice and click 'endorse' for certain skills that appeared under the name of the VDPro, but making some effort and writing some detailed recommendation could be worthwhile to look at" [P4].

For performance review, 75% of the interviewees considered vulnerability databases to be the most reliable source. A vulnerability database is a platform to collect and disseminate information about discovered vulnerabilities (e.g., the National Vulnerability Database). Vulnerability Databases are considered an important source especially since they provide details of the vulnerabilities discovered (e.g., description of the vulnerability discovered, vulnerability severity score, references for security advisories related to the vulnerability). Furthermore, 83% of the participants in the study considered being mentioned in the hall of fame or 'thank-you' pages (e.g., Facebook white hat thanks page) is of high value. An interviewee said, "If the VDPro is mentioned in the hall of fame of Facebook or Google then they pack enough skills to be an expert" [P5]. Also 83% of the interviewees considered the measured performance within the crowdsourcing vulnerability discovery platform as a good source to extract expertise. While the rest agreed that it is a good source only if the crowdsourcing vulnerability discovery platform discloses the details of the discovered vulnerabilities by the VDPros. One of the interviewees stated, "With vulnerability databases, I can have a complete idea about the vulnerability discovered and its impact, but sometimes with the vulnerabilities discovered within a crowdsourcing vulnerability discovery platform we can only see limited information" [P16].

As reliable sources for artefacts, 39% considered GitHub to be the most reliable source. 60% considered security conferences, security forum, and blogs to be the most reliable sources. On another hand, 75% of the interviewees considered Q&A platforms (e.g., Security StackExchange) not very important sources. Additionally, 72% of the interviewees mentioned that they consider publishing exploits for vulnerability as a high-level indicator of expertise. An interviewee mentioned "By showing that you know how to exploit vulnerabilities, you show that you deeply understand how the vulnerability works and what its impact is" [P13].

5.2. The Perspective of Crowdsourcing Vulnerability Discovery Platforms

To gain additional insights into the perspective of the crowdsourcing vulnerability discovery platforms, we interviewed a representative of one of the platforms. The interview allowed us to gain some valuable insights about the programs the platform is managing and the importance of selecting VDPros. The platform's representative mentioned "Almost all of our customers now start with a private program with vetted VDPros before moving to a public program. It's only a small fraction of companies are ready to work with thousands of VDPros". About the effect of selecting VDPros in minimizing the false positives in reported vulnerabilities the platform's representative mentioned, "Public bug bounties is 95% noise and 5% signal. Unless you have a very mature security organization it is not recommended to use public bounties. Private bounties with vetted VDPros are the way to go"

Regarding the indicators of expertise, the platform relies on the past performance, artefacts and will also look at professional social network. It is quoted that: "We rely on past performance in private or public bounties, on our platform or other public bounty programs (Facebook/Google). Also, we check LinkedIn, GitHub profiles as well as conference talks"

When we asked for elaboration about the sources the platform rely on, the platform's representative mentioned: "We look at LinkedIn, GitHub, and Conferences. This is a manual process and it is on a case-by-case basis. E.g. a researcher might have given a

presentation on Blackhat but have no connections on LinkedIn etc. And the other way around".

About certification and whether clients ask for specific requirements, the platforms' representatives mentioned: "We typically don't see requirements like "university degree," but occasionally we will get requests based on geographical region, or professional accreditation (CISSP[11], CISA[12], etc.)".

5.3. The Online Survey of Vulnerability Discovery Professionals

In this section, we briefly discuss the results of the online survey. Figure 4 illustrates the high-level results of the online survey. 60% of the participants agreed that certification is an important indicator, while the rest considered it between somewhat important and useful. We asked about the sources of certification. 62% of the participants thought industrial certification were important, while only 30% considered certification from academic body as important. When we asked the participants about the importance of referral as a practice in the security domain, 72% considered it important while 11% considered it useful and the rest considered it not important. We also asked the participants their opinion about relying on artefacts as indicators to expertise: 83% considered it an important indicator. 74% considered publishing vulnerability exploits as an important indicator to expertise. Followed by, publishing security articles and presentations or videos both at 63%. They considered the least important source of artefacts is answering questions in Q&A platforms (e.g., Security StackExchange) where only 30% considered it "somewhat important". Regarding performance review, all participants considered it very important. 84% of the participants considered the number of vulnerabilities discovered by a VDPro as a very important measure for performance. While 75% considered the severity of vulnerability discovered as an important measure to performance. On the other hand, 56% considered the accuracy in discovering vulnerability as an important measure to performance, 36% considered it useful but not important, and 8% considered it not important. Finally, we asked the participants about how much they value associations and being part of communities or groups as an indicator to expertise. 81% of the participants considered it an important indicator, and the rest considered it somewhat important.

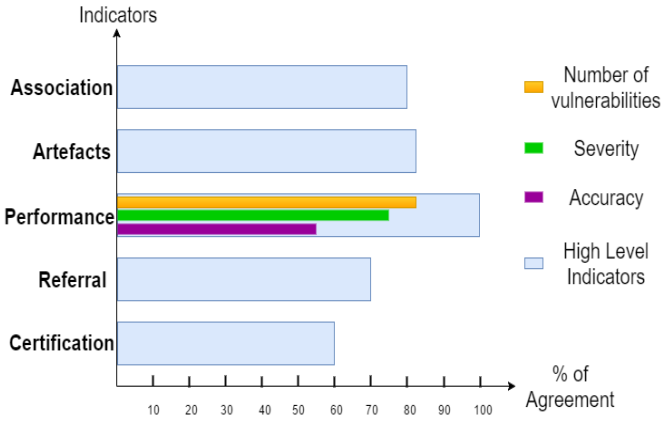


Figure 4. Indicators of Expertise (Vulnerability Discovery Professionals' Perspective)

5.4. Analyzing Public Data about Vulnerability Discovery Professionals

As an additional experiment, we have looked at public information of 100 VDPros from the leader boards (among the top 100 ranked in each platform) of three different crowdsourcing vulnerability discovery platforms: Bugcrowd, Cobalt, and Hackerone. For each VDPro we aggregated additional publicly available data from several external sources (LinkedIn, Twitter, GitHub, StackExchange (SE), blogs, and personal/professional websites). In this section, we discuss the result of the experiment in the light of the indicators we have discovered and illustrate some of the insights we have acquired. As we have relied only on the public data, there are possible threats to validity. The vulnerability discovery professionals could be providing falsified data to the public or hiding some information. We believe that these cases do not have much impact on the analysis; VDPros in the mentioned platforms seek to maintain good reputation to the potential business partners. Additionally, we aggregate the information from multiple sources to minimize this concern.

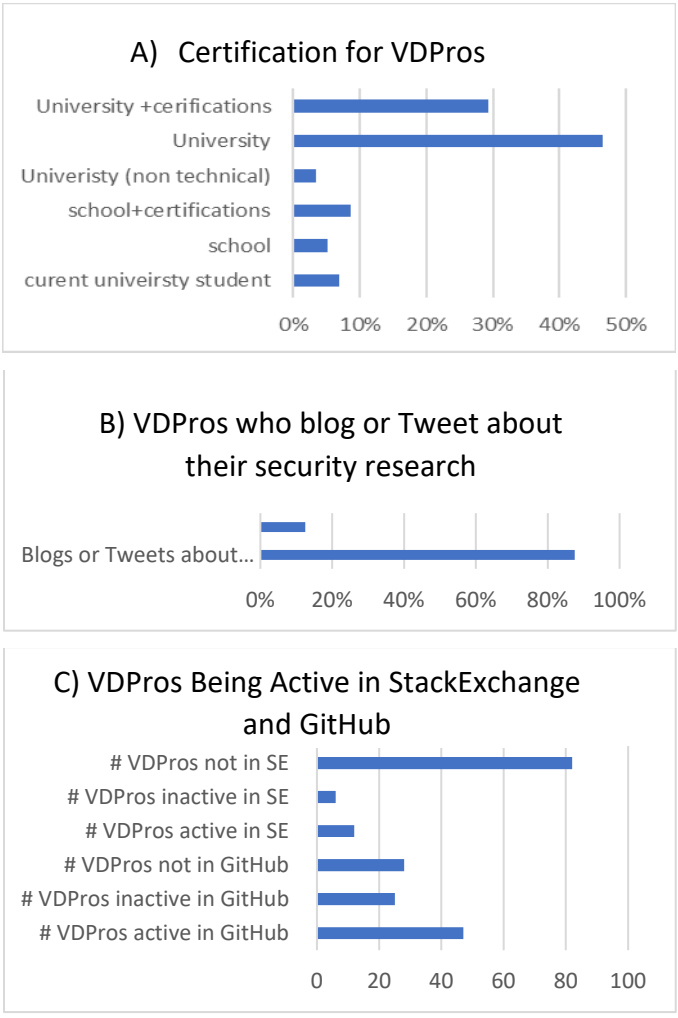


Figure 5. Insights from the public information about Vulnerability Discovery Professionals

[11] www.isc2.org/cissp/default.aspx

[12] http://www.isaca.org/certification/cisa-certified-information-systems-auditor/pages/default.aspx

We have noticed in this analysis of public data, 76% of the VDPros are university graduates of computer and technology fields. In addition (42%) of the VDPros have certifications in the security field (e.g., Certified Ethical hacker). We also noticed that 88% of the VDPros tend to publicly publish information about the vulnerabilities they discover, methods of discovery, and exploitation of vulnerabilities. They rely on their personal blogs or Websites in addition to twitter. For artefacts created in platforms like GitHub or StackExchange, we discovered that 82% of the VDPros have no profiles in this platform or have inactive accounts. While for GitHub, 73% of the VDPros have profiles in Github and 47% are active. Figure 5 illustrates the insights from the public data about the VDPros.

6. Discussion

In this section, we: (i) summarize our findings; (ii) provide some recommendations based on these findings; and (vi) discuss potential limitations of our study.

6.1. Analysis of our Findings

It is clear from our findings that domain experts and VDPros disagree in regard to association being an indicator to expertise. A possible reason is that the cyber security community is a strong community where teaming with other VDPros and pair practice is considered one of the factors to success[13]. So, in the opinion of VDPros, being associated with expert VDPros reflects expertise.

We observed from the results of the interviews that participants in our study valued certification from industrial parties more than from academic bodies. According to the input from the representative of the crowdsourcing vulnerability discovery platform we have interviewed, the platform has received requests to involve VDPros with industrial certifications before, but never a request to specify academic certification. A possible reason for this is that some academic bodies are not progressing as fast as the industrial bodies in the field of information security. One of the interviewees mentioned "some universities are not really up to date with the requirements of the security industry" [P5].

For measuring the performance of VDPros within the task of vulnerability discovery there are contrasting views in the way the number of vulnerabilities and accuracy thereof are valued. A possible reason for this contrast could be that domain experts look at how to utilize the resources to verify and mitigate the vulnerabilities. If VDPros are submitting a high number of vulnerability reports with only few of them being valid vulnerabilities, this defeats one of the purposes in selecting VDPros with the required expertise. On the other hand, VDPros may have the impression that all the vulnerability reports they submit (even if the task requester considered them invalid for being duplicates, or out of scope) illustrate their expertise. The reason for this is because they have already demonstrated what is required to discover the vulnerabilities.

6.2. Recommendations

I. Vulnerability Discovery Professionals Analytics

Information about VDPros are scattered within different sources mentioned earlier (e.g., crowdsourcing vulnerability discovery platforms, professional communities, research articles, social coding platforms, Q&A sites). Therefore, in order to facilitate effective selection of VDPros to perform the task of vulnerability discovery, we recommend leveraging VDPros analytics. Analytics rely on characterizing features, which are variables that grasp and encode information from raw or curated data, thereby enabling to derive meaningful inferences from data. Security expertise features could be extracted from within vulnerability discovery platforms (e.g., number of vulnerabilities discovered in the platform) and external sources (e.g., number of contribution to security related projects in GitHub, number of contributed records in ExploitDB[14]). It is possible to extract security expertise features using crowdsourcing (e.g., crowdsourcing tasks to judge the quality of comments, questions, and answers by VDPro in a security forum like Bugcrowd Forum[15]), as well as automated extraction of security expertise features (e.g., relying on GitHub API calls to extract the number of security related code repositories). Properly leveraged efforts in vulnerability discovery professional and content analytics [73,74] could potentially allow improving the selection of VDPro, ensure compliance with governance rules, and provide insights into the task of crowdsourcing vulnerability discovery (e.g., work effort estimation and time tracking, VDPro interests, or malicious VDPros). In addition, we recommend providing task requesters with a customizable language to specify rules (i.e., selection expressions) based on indicators they deem relevant to include in the selection process, as well as indicator aggregation strategies. As an example, let us assume Organization-A has a business rule stating that when contracting external parties, they prioritize certification from academic bodies. The security officer responsible for the task considers the exploits published in ExploitDB along with the number of repositories in GitHub as indicators to expertise. In this case, the task requester may specify the following rule to acquire a ranked list of VDPros: (Have University degree) AND (Number of records in ExploitDB) AND (Number of repositories in GitHub)). Moreover, we expect new indicators to be incrementally introduced and customized. Security is a complex area and there could be several different types of indicators, from generic to more specific knowledge (e.g., operating systems, databases, devices, browsers, and frameworks). This rule-based language will allow for the specification of fine-grained selection indicators and hence will improve achieving a high probability in matching the task with the VDPro.

II. Harvesting Expertise Indicators Systematically

It is clear from our findings that it is difficult to homogenize the knowledge acquired from different sources (domain experts, VDPros, crowdsourcing vulnerability discovery platforms) in order to determine the indictors to measure expertise in VDPros. We propose to consider techniques that systematically harvest the knowledge from these different perspectives. We believe this could be done by leveraging a knowledge-driven approach to create a security expertise feature graph. We adopt the notion of a feature from the field of machine learning and data science [75]. The security expertise feature graph will serve to store available features for reuse as part of future curation processes. Organizing features into high-level indictors (e.g., certifications, performance review) and low-level features (e.g., security course, number of vulnerabilities discovered) will help simplify their representation,

[13] hackerone.com/blog/what-great-hackers-share

[14] A database of vulnerability exploits (www.exploit-db.com).

[15] forum.bugcrowd.com

and aid searching and using features and indicators. Experts in the security community would be invited to curate this graph by incrementally and collectively: adding new features (e.g., a sub-feature of "artefacts" could be the number of published security articles); vote up/down; or add resources and comments as annotations to these features. Moderators may be assigned to verify amendments to the graph, reject, or accept new features. A similar approach has been proposed by Mark Klien which is a large-scale argumentation system for decision making in large communities [76].

III. Cyber Security Competitions

We observed that there are contrasting views between the opinions of domain experts in industry and VDPros involved in the task of vulnerability discovery. In order to minimize this contrast, we recommend for industrial organizations and crowdsourcing vulnerability discovery platforms to strengthen their ties with VDPros. A possible step to achieve this is to rely on cyber security competitions (e.g., Google capture the flag[16], Cyber Security Challenge Australia[17]) while emphasizing on educating VDPros about important practices that are relevant (e.g., providing high rewards or bonus to the VDPros with higher accuracy, providing higher rewards for high quality vulnerability reports). Similarly, involving universities in these competitions will be more beneficial. This would not only attract perspective students to pursue careers in the security industry and educate them about the rights skills and practices, it would also increase the focus of academic bodies to progress further in the field. Bridging the gap between what students is being offered at universities in the field of cyber security and what the security industry needs.

6.3. Limitations and Threats to Validity

When investigating online content, it is possible to come across material that is biased or inaccurate (since blogs and online articles are not peer reviewed). In order to minimize any bias or inaccuracies, we checked the background of the author to ensure are based on expert opinions. We also excluded articles that are marketing motivated (i.e. promoting or demoting certain technique or approach for personal gain).

For the interviews, the number of participants is not high so it is possible that we missed some insights. We tried to minimize this threat by targeting various industries in order to obtain a broader view of possible opinions. We also considered only one participant from each organization. The strong background we acquired from performing the literature review and online content investigation helped us to formulate clear questions for the interviews and online survey. We observed saturation in the opinions of domain experts, and this indicated we have reached a proper depth to draw solid conclusions.

For the online survey, although we obtained responses from the VDPros, we cannot assume that this number covers all the possible opinions from the security community. One of the authors attended security conferences (e.g., Ruxcon[18], BSides[19]) to discuss with VDPros directly. While some feedback was acquired about the study from VDPros, it was mostly an off-the-record approach. We also reached out to some of the VDPros for clarification of their answers in the survey, and this provided us with some additional insights.

As the questions from the interview and the online survey relied on the results from surveying the literature and online contents, another threat of validity is having the study suffer from confirmatory bias. In order to minimize this threat, we started each interview with the open question of what the interviewees rely on to measure the expertise of VDPros and we also closed the interview with asking if there are any additional indictors we missed. Similarly, for the online survey we deliberately added some open-ended questions to allow the VDPros to add any additional indicators that were not mentioned in the survey. Some VDPros added comments but they fall within the same indicators that we have already included (e.g., publishing exploits, which we already considered part of the artefacts generated by the VDPro).

Another important aspect to consider in the experiment we conducted for analysing public profiles of VDPro is time constrains. As a result, we were only able to look at a limited number of vulnerability discovery professionals; there could thus be insights we did not acquire due to this limitation. We intend to do a large-scale analysis and survey as future work.

7. Conclusion and Future Work

The contribution of this research is threefold: (i) Exploring the indicators to measure expertise of VDPros involved in the crowdsourced task of vulnerability discovery; (ii) Discovering the reliable sources that can be used to extract information about these indicators; and (iii) Recommendations to improve the selection process for VDPros involved in tasks related to discovering vulnerabilities. Our findings will help in the process of selecting VDPros that are more likely to contribute higher quality outcomes in this type of task.

We observed that domain experts and vulnerability discovery professionals, who have participated in our study, agree unanimously on certification, referral, performance review, and artefacts as good indicators to measure expertise in VDPros. There is some conflict about the importance of association as an indicator of expertise. We also observed that the opinions of domain experts and VDPros involved in the task of vulnerability discovery diverge in some areas. Being aware of this divergence is an important step to build a better ecosystem.

This is the first step for investigating the problem of selecting VDPros in Vulnerability Discovery Platforms. As future work, an additional aspect we will pursue is identifying malicious VDPros, and methods to successfully block them from participating in the task. We will develop techniques to model and capture VDPros behaviour patterns and abstract them into meaningful concepts (e.g., honest or malicious). We will investigate indicators and powerful aggregation functions and operators to characterize and identify malicious VDPros.

Acknowledgment

We greatly acknowledge all participants that provided valuable feedback and insight in interviews and online survey. A special appreciation to Jacob Hansen, Jeffrey Stapleton, Troy Hunt,

[16] https://capturetheflag.withgoogle.com/

[17] https://www.cyberchallenge.com.au/

[18] ruxcon.org.au/

[19] www.securitybsides.com/w/page/12194156/FrontPage

Dimitrios Stergiou, and Roee Hay for their valuable input and help in reaching out to members of the security community.

References

[1] Mortada Al-Banna, Boualem Benatallah, and Moshe Chai Barukh. 2016. Software Security Professionals: Expertise Indicators. In *2016 IEEE 2nd International Conference on Collaboration and Internet Computing (CIC)*, 139–148. DOI:https://doi.org/10.1109/CIC.2016.030

[2] Mohammad Allahbakhsh, Boualem Benatallah, Aleksandar Ignjatovic, Hamid Reza Motahari-Nezhad, Elisa Bertino, and Schahram Dustdar. 2013. Quality control in crowdsourcing systems: Issues and directions. *IEEE Internet Comput.* 17, 2 (2013), 76–81. DOI:https://doi.org/10.1109/MIC.2013.20

[3] BugCrowd. 2017. *2017 State of Bug Bounty Report*. Retrieved July 26, 2017 from https://pages.bugcrowd.com/hubfs/Bugcrowd-2017-State-of-Bug-Bounty-Report.pdf

[4] Anne Edmundson, Brian Holtkamp, Emanuel Rivera, Matthew Finifter, Adrian Mettler, and David Wagner. 2013. An empirical study on the effectiveness of security code review. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 197–212. DOI:https://doi.org/10.1007/978-3-642-36563-8_14

[5] Edith Law, Conner Dalton, Nick Merrill, Albert Young, and Krzysztof Z. Gajos. 2013. Curio: A Platform for Supporting Mixed-Expertise Crowdsourcing. *First AAAI Conference on Human Computation and Crowdsourcing*. Retrieved from http://www.aaai.org/ocs/index.php/HCOMP/HCOMP13/paper/view/7534

[6] LE Celis, SP Reddy, IP Singh, and S Vaya. 2016. Assignment Techniques for Crowdsourcing Sensitive Tasks. *19th ACM Conf. ...* (2016). Retrieved September 3, 2016 from http://dl.acm.org/citation.cfm?id=2835202

[7] Justin Scott Giboney, Jeffrey Gainer Proudfoot, Sanjay Goel, and Joseph S. Valacich. 2016. The Security Expertise Assessment Measure (SEAM): Developing a scale for hacker expertise. *Comput. Secur.* 60, (2016), 37–51. DOI:https://doi.org/10.1016/j.cose.2016.04.001

[8] DE Difallah, G Demartini, and P Cudré-Mauroux. 2012. Mechanical Cheat: Spamming Schemes and Adversarial Techniques on Crowdsourcing Platforms. *CrowdSearch* (2012). Retrieved September 5, 2016 from http://ftp.exascale.info/sites/default/files/Difallah_CrowdSearch_2012.pdf

[9] Munawar Hafiz and Ming Fang. 2016. Game of detections: how are security vulnerabilities discovered in the wild? *Empir. Softw. Eng.* 21, 5 (October 2016), 1920–1959. DOI:https://doi.org/10.1007/s10664-015-9403-7

[10] Barbara Kitchenham. 2004. Procedures for performing systematic reviews. *Keele, UK, Keele Univ.* 33, TR/SE-0401 (2004), 28. DOI:https://doi.org/10.1.1.122.3308

[11] Steve Easterbrook, Janice Singer, Margaret-Anne Storey, and Daniela Damian. 2008. Selecting Empirical Methods for Software Engineering Research. *Guid. to Adv. Empir. Softw. Eng.* (2008), 285–311. DOI:https://doi.org/10.1007/978-1-84800-044-5_11

[12] C.B. Seaman. 1999. Qualitative methods in empirical studies of software engineering. *IEEE Trans. Softw. Eng.* 25, 4 (1999), 557–572. DOI:https://doi.org/10.1109/32.799955

[13] Anne Lacey and Donna Luff. 2007. Qualitative data analysis. *Trowbridge, Wiltsh. Cromwell ...* (2007), 13–14.

[14] An ISACA and RSA Conference Survey. 2015. *State of Cybersecurity: Implications for 2015*. Retrieved August 22, 2016 from http://www.isaca.org/cyber/Documents/State-of-Cybersecurity_Res_Eng_0415.pdf

[15] Thomas W. Malone, Robert Laubacher, and Chrysanthos Dellarocas. 2010. The collective intelligence genome. *IEEE Engineering Management Review 38*, 38. DOI:https://doi.org/10.1109/EMR.2010.5559142

[16] Andy Ozment. 2004. Bug auctions: Vulnerability markets reconsidered. In *Workshop on Economics of Information Security (WEIS)*, 1–23. Retrieved from http://www.andyozment.com/papers/weis04-ozment-bugauc-slides.pdf

[17] Aron Laszka, Mingyi Zhao, and Jens Grossklags. 2016. Banishing Misaligned Incentives for Validating Reports in Bug-Bounty Platforms. . Springer, Cham, 161–178. DOI:https://doi.org/10.1007/978-3-319-45741-3_9

[18] Mingyi Zhao, Jens Grossklags, and Peng Liu. 2015. An Empirical Study of Web Vulnerability Discovery Ecosystems. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security - CCS '15*, 1105–1117. DOI:https://doi.org/10.1145/2810103.2813704

[19] Matthew Finifter, Devdatta Akhawe, and David Wagner. An Empirical Study of Vulnerability Rewards Programs.

[20] Mitre. CAPEC - CAPEC-152: Inject Unexpected Items (Version 2.9). Retrieved February 22, 2017 from https://capec.mitre.org/data/definitions/152.html

[21] C Okoli and K Schabram. 2010. A guide to conducting a systematic literature review of information systems research. *Sprouts Work. Pap. Inf. Syst* (2010). Retrieved February 20, 2017 from http://www.academia.edu/download/3250666/OkoliSchabram2010SproutsLitReviewGuide.pdf

[22] Nina Wakeford and Kris Cohen. 2008. Fieldnotes in Public: Using Blogs for Research. In *The SAGE Handbook of Online Research Methods*. 307–327. DOI:https://doi.org/10.4135/9780857020055

[23] J Freyne, L Coyle, and B Smyth. 2010. Relative status of journal and conference publications in computer science. *Commun. ACM* 53.11, (2010), 124–132. Retrieved February 21, 2017 from http://dl.acm.org/citation.cfm?id=1839701

[24] Synack. Join a Global Hacker Team | Synack Red Team. Retrieved February 22, 2017 from https://www.synack.com/red-team/

[25] SA Ejiaku and MA Badamas. 2010. An examination of information technology certification: A measure of professional qualification. *Bus. Res. Yearb.* 17, (2010), 119–125. Retrieved February 22, 2017 from http://lakishasimmons.com/images/publications/bothsimmonsandaiken_2010v1bry.pdf#page=132

[26] Matt Bishop and Deborah Frincke. 2004. Academic degrees and professional certification. *IEEE Security and Privacy 2*, 56–58. DOI:https://doi.org/10.1109/MSP.2004.91

[27] Tanya McGill and Michael Dixon. 2013. An investigation of the impact of recertification requirements on recertification decisions. In *Proceedings of the 2013 annual conference on Computers and people research - SIGMIS-CPR '13*, 79. DOI:https://doi.org/10.1145/2487294.2487310

[28] Tanjila Kanij, Robert Merkel, and John Grundy. 2014. A preliminary survey of factors affecting software testers. In *Proceedings of the Australian Software Engineering Conference, ASWEC*, 180–189. DOI:https://doi.org/10.1109/ASWEC.2014.32

[29] Federico Botella, Eloy Alarcon, and Antonio Peñalver. 2014. How to classify to experts in usability evaluation. In *Proceedings of the XV International Conference on Human Computer Interaction - Interacción '14*, 1–4. DOI:https://doi.org/10.1145/2662253.2662278

[30] Sarah Boutin. 2015. Behind the Scenes at Salesforce: Our #1 Recruiting Secret - Salesforce Blog. Jan. Retrieved February 22, 2017 from https://www.salesforce.com/blog/2015/01/behind-scenes-salesforce-our-1-recruiting-secret.html

[31] LinkedIn. Skill Endorsements - Overview | LinkedIn Help. Retrieved February 22, 2017 from https://www.linkedin.com/help/linkedin/answer/31888?lang=en

[32] A community of great programmers and their programming tips. Retrieved February 22, 2017 from https://coderwall.com/

[33] Repcoin: Find your expert. Retrieved June 12, 2015 from http://www.repcoin.com

[34] Daniel Schall, Florian Skopik, and Schahram Dustdar. 2012. Expert Discovery and Interactions in Mixed Service-Oriented Systems. *IEEE Trans. Serv. Comput.* 5, 2 (April 2012), 233–245. DOI:https://doi.org/10.1109/TSC.2011.2

[35] Lan Zhang, Xiang Yang Li, Yunhao Liu, QiuYuan Huang, and Shaojie Tang. 2012. Mechanism design for finding experts using locally constructed social referral web. In *Proceedings - IEEE INFOCOM*, 2896–2900. DOI:https://doi.org/10.1109/INFCOM.2012.6195723

[36] S. Pushpa, K. S. Easwarakumar, Susan Elias, and Zakaria Maamar. 2010. Referral based expertise search system in a time evolving social network. *Proc. Third Annu. ACM Bangalore Conf. - Comput. '10* (2010), 1–8. DOI:https://doi.org/10.1145/1754288.1754294

[37] Simone Braun, Christine Kunzmann, and Andreas Schmidt. 2010. People tagging and ontology maturing: Toward collaborative competence

management. In *From CSCW to Web 2.0: European Developments in Collaborative Design - Selected Papers from COOP 2008*, 133–154. DOI:https://doi.org/10.1007/978-1-84882-965-7_7

[38] Saptarshi Ghosh, Naveen Sharma, Fabricio Benevenuto, Niloy Ganguly, and Krishna Gummadi. 2012. Cognos: crowdsourcing search for topic experts in microblogs. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval - SIGIR '12*, 575. DOI:https://doi.org/10.1145/2348283.2348361

[39] Man-Ching Yuen, Irwin King, and Kwong-Sak Leung. 2012. Task recommendation in crowdsourcing systems. In *Proceedings of the First International Workshop on Crowdsourcing and Data Mining - CrowdKDD '12*, 22–26. DOI:https://doi.org/10.1145/2442657.2442661

[40] Hyun Joon Jung and Hyun Joon. 2014. Quality assurance in crowdsourcing via matrix factorization based task routing. In *Proceedings of the 23rd International Conference on World Wide Web - WWW '14 Companion*, 3–8. DOI:https://doi.org/10.1145/2567948.2567951

[41] Kymberlee Price. 2015. How We Measure Crowd Performance. Retrieved February 22, 2017 from https://blog.bugcrowd.com/measure-crowd-performance/

[42] Jacob Hansen. 2015. Introducing Rep. Retrieved February 22, 2017 from https://blog.cobalt.io/introducing-rep-ecbacb04fa96#.sm5eltcs7

[43] Hackerone. 2014. Introducing Reputation. Retrieved February 22, 2017 from https://www.hackerone.com/blog/introducing-reputation

[44] Mingyi Zhao, Jens Grossklags, and Kai Chen. 2014. An Exploratory Study of White Hat Behaviors in a Web Vulnerability Disclosure Program. In *Proceedings of the 2014 ACM Workshop on Security Information Workers - SIW '14*, 51–58. DOI:https://doi.org/10.1145/2663887.2663906

[45] A Algarni and Y Malaiya. Software vulnerability markets: Discoverers and buyers. *waset.org*. Retrieved May 27, 2016 from http://www.waset.org/publications/9997791

[46] John Anvik, Lyndon Hiew, and Gail C. Murphy. 2006. Who should fix this bug? *Proceeding 28th Int. Conf. Softw. Eng. - ICSE '06* 2006, (2006), 361. DOI:https://doi.org/10.1145/1134285.1134336

[47] Umair ul Hassan, Edward Curry, U. Hassan, and Edward Curry. 2013. A Capability Requirements Approach for Predicting Worker Performance in Crowdsourcing. *Proc. 9th IEEE Int. Conf. Collab. Comput. Networking, Appl. Work.* (2013), 429–437. DOI:https://doi.org/10.4108/icst.collaboratecom.2013.254181

[48] Cristina Sarasua and Matthias Thimm. 2013. Microtask available, send us your CV! In *Proceedings - 2013 IEEE 3rd International Conference on Cloud and Green Computing, CGC 2013 and 2013 IEEE 3rd International Conference on Social Computing and Its Applications, SCA 2013*, 521–524. DOI:https://doi.org/10.1109/CGC.2013.87

[49] Paul Resnick, Richard Zeckhauser, John Swanson, and Kate Lockwood. 2006. The value of reputation on eBay: A controlled experiment. *Exp. Econ.* 9, 2 (June 2006), 79–101. DOI:https://doi.org/10.1007/s10683-006-4309-2

[50] Paul Resnick and Richard Zeckhauser. 2002. Trust among strangers in Internet transactions: Empirical analysis of eBay's reputation system. *Adv. Appl. ...* 11 (2002), 127–157. DOI:https://doi.org/10.1016/S0278-0984(02)11030-3

[51] Bogdan Vasilescu, Vladimir Filkov, and Alexander Serebrenik. 2013. StackOverflow and GitHub: Associations between Software Development and Crowdsourced Knowledge. In *2013 International Conference on Social Computing*, 188–195. DOI:https://doi.org/10.1109/SocialCom.2013.35

[52] Claudia Hauff and Georgios Gousios. 2015. Matching GitHub developer profiles to job advertisements. *Proceedings of the 12th Working Conference on Mining Software Repositories*, 362–366.

[53] Pranam Kolari and Kelly Lyons. 2008. Expert Search using Internal Corporate Blogs. *Sigir* (2008), 2–5.

[54] Krisztian Balog, Leif Azzopardi, and Maarten de Rijke. 2006. Formal models for expert finding in enterprise corpora. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '06*, 43. DOI:https://doi.org/10.1145/1148170.1148181

[55] Konstantinos Pelechrinis, Vladimir Zadorozhny, and Vladimir Oleshchuk. 2011. A Cognitive-based scheme for user reliability and expertise assessment in Q&A social networks. In *2011 IEEE International Conference on Information Reuse & Integration*, 545–550. DOI:https://doi.org/10.1109/IRI.2011.6009614

[56] Demartini and Gianluca. 2007. Finding experts using Wikipedia. *Proceedings of the 2nd International Conference on Finding Experts on the Web with Semantics - Volume 290*, 33–41.

[57] Jian Chang, Peter Gebhard, Andreas Haeberlen, Zack Ives, Insup Lee, Oleg Sokolsky, and Krishna K. Venkatasubramanian. 2013. TrustForge: Flexible access control for collaborative crowd-sourced environment. In *2013 11th Annual Conference on Privacy, Security and Trust, PST 2013*, 291–300. DOI:https://doi.org/10.1109/PST.2013.6596065

[58] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. 2008. ArnetMiner. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 08*, 990. DOI:https://doi.org/10.1145/1401890.1402008

[59] Luanne Freund, Kristof Kessler, Michael Huggett, and Edie Rasmussen. Exposing and Exploring Academic Expertise with Virtu.

[60] Yupeng Fu, Rongjing Xiang, Yiqun Liu, Min Zhang, and Shaoping Ma. 2007. Finding Experts Using Social Network Analysis. In *IEEE/WIC/ACM International Conference on Web Intelligence (WI'07)*, 77–80. DOI:https://doi.org/10.1109/WI.2007.14

[61] Juanzi Li, Jie Tang, Jing Zhang, and Q Luo. 2007. EoS: expertise oriented search using social networks. *Proc. 16th Int. Conf. World Wide Web* (2007), 1271–1272. DOI:https://doi.org/10.1145/1242572.1242803

[62] Jing Zhang, Jie Tang, and Juanzi Li. 2007. Expert Finding in a Social Network. In *Advances in Databases: Concepts, Systems and Applications*. Springer Berlin Heidelberg, Berlin, Heidelberg, 1066–1069. DOI:https://doi.org/10.1007/978-3-540-71703-4_106

[63] Daniel Schall and Daniel. 2012. Expertise ranking using activity and contextual link measures. *Data Knowl. Eng.* 71, 1 (January 2012), 92–113. DOI:https://doi.org/10.1016/j.datak.2011.08.001

[64] Xiu Xiu Li, Jianguo Jianguo Ma, Yujiu Yujiu Yang, and Dongzhi Dongzhi Wang. 2013. A Service Mode of Expert Finding in Social Network. In *2013 International Conference on Service Sciences (ICSS)*, 220–223. DOI:https://doi.org/10.1109/ICSS.2013.48

[65] Zhou Zhao, James Cheng, Furu Wei, Ming Zhou, Wilfred Ng, and Yingjun Wu. 2014. SocialTransfer. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management - CIKM '14*, 779–788. DOI:https://doi.org/10.1145/2661829.2661871

[66] Josep M. Pujol, Ramon Sangüesa, and Jordi Delgado. 2002. Extracting reputation in multi agent systems by means of social network topology. In *Proceedings of the first international joint conference on Autonomous agents and multiagent systems part 1 - AAMAS '02*, 467. DOI:https://doi.org/10.1145/544741.544853

[67] Maggy Anastasia Suryanto, Ee Peng Lim, Aixin Sun, and Roger H. L. Chiang. 2009. Quality-aware collaborative question answering. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining - WSDM '09*, 142. DOI:https://doi.org/10.1145/1498759.1498820

[68] R Atkinson and J Flint. 2001. Accessing Hidden and Hard-to-reach Populations: Snowball Research Strategies. *Soc. Res. Updat.* (2001). Retrieved September 9, 2017 from http://eprints.gla.ac.uk/37493/

[69] Martyn Denscombe. 2010. *The good research guide for small-scale social research projects*.

[70] J Saldaña. 2015. *The coding manual for qualitative researchers*. Retrieved September 3, 2016 from https://books.google.com.au/books?hl=en&lr=&id=ZhxiCgAAQBAJ&oi=fnd&pg=PP1&dq=The+Coding+Manual+for+Qualitative+Researchers&ots=yHYb8GPZjU&sig=7ZFrhcKUzg7vhCugj7NHPMKCmHc

[71] Clare M. Lewandowski, New Co-investigator, and Clare M. Lewandowski. 2015. *Statistics In a Nutshell: A Desktop Quick Reference 2th*. DOI:https://doi.org/10.1017/CBO9781107415324.004

[72] Sarah Elsie Baker and Rosalind Edwards. 2012. How many qualitative interviews is enough: Expert voices and early career reflections on sampling and cases in qualitative research. *Natl. Cent. Res. Methods Rev. Pap.* (2012), 1–43. DOI:https://doi.org/10.1177/1525822X05279903

[73] Tamsyn P Waterhouse, Spear Street, and San Francisco. 2013. Pay by the Bit : An Information-Theoretic Metric for Collective Human Judgment. *Comput. Support. Coop. Work* (2013), 623–637. DOI:https://doi.org/10.1145/2441776.2441846

[74] Jane Hunter, Abdulmonem Alabri, and Catharine van Ingen. 2013. Assessing the quality and trustworthiness of citizen science data. *Concurr. Comput.*

Pract. Exp. 25, FEBRUARY 2013 (2013), 454–466. DOI:https://doi.org/10.1002/cpe.2923

[75] Avrim L. Blum and Pat Langley. 1997. Selection of relevant features and examples in machine learning. *Artif. Intell.* 97, 1 (1997), 245–271. DOI:https://doi.org/10.1016/S0004-3702(97)00063-5

[76] Mark Klein. 2012. How to Harvest Collective Wisdom on Complex Problems: An Introduction to the MIT Deliberatorium. *Cent. Collect. Intell. Work. Pap.* (2012), 15. Retrieved from http://cci.mit.edu/docs/working_papers_2012_2013/kleinwp2013.pdf

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1799-1803 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Optimization of Depth-Based Routing for Underwater Wireless Sensor Networks through Intelligent Assignment of Initial Energy

Tanveer Ahmed[1], Muhammad Kaleem[*1], Khurram Saleem Alimgeer[1], Mustafa Shakir1 and Sajid Nazir[2]

[1]*Department of Electrical Engineering, COMSATS Institute of Information Technology Islamabad, Pakistan*

[2]*Firstco Ltd., London, W2 6EU, UK*

ARTICLE INFO

Article history:
Received: 29 May, 2017
Accepted: 20 July, 2017
Online: 05 October, 2017

Keywords:
Routing
Energy Efficiency
Wireless Sensor Networks
Network lifetime
Coverage Hole

ABSTRACT

Underwater Wireless Sensor Networks (UWSNs) are extensively used to explore the diverse marine environment. Energy efficiency is one of the main concerns regarding performance of UWSNs. In a cooperative wireless sensor network, nodes with no energy are known as coverage holes. These coverage holes are created due to non-uniform energy utilization by the sensor nodes in the network. These coverage holes degrade the performance and reduce the lifetime of UWSNs. In this paper, we present an Intelligent Depth Based Routing (IDBR) scheme which addresses this issue and contributes towards maximization of network lifetime. In our proposed scheme, we allocate initial energy to the sensor nodes according to their usage requirements. This idea is helpful to balance energy consumption amongst the nodes and keep the network functional for a longer time as evidenced by the results provided.

1. Introduction

Underwater wireless sensor networks (UWSNs) have attained much attention nowadays due to their use in monitoring underwater environment. Underwater communication is characterized by larger propagation delay, high error rate and lower bandwidth. Acoustic signals are thus used for communications in underwater environment due to channel characteristics. Furthermore, nodes in UWSNs are powered with batteries that have limited lifetime. It is impractical to change the batteries of the nodes once deployed. The energy of the sensor nodes should be optimally used to avoid coverage holes i.e., nodes with no energy. Due to these reasons, routing in UWSNs is quite challenging and we need to address these issues while developing routing strategies for UWSNs [1].

A number of routing schemes are reported in literature to address the issue of efficient routing in UWSNs. One such protocol is depth based routing (DBR) protocol [2] which uses deployment depth of sensor nodes to determine the route of a data packet from its source node to the sink. Sensor nodes are usually deployed randomly in underwater environment and the sink is placed at the surface of water. Each sensor node is responsible to gather data and then send it to the sink in a cooperative manner. Some sensor nodes are deployed at a higher depth and some at a lower depth hence consituting a sensor network. Nodes with higher depth sense the required information and send it to a node present at lower depth in the form of data packets. Receiver nodes then send these packets to the next node located above. By repeating this process, data packets are passed from the sensor nodes to the sink which is placed at the water surface. It is important to note that nodes which are at a greater distance from the sink, only send their own data packets to low-depth nodes hence they consume lesser amount of energy and remain alive for a longer period of time. Whereas, sensor nodes that are near to the sink have to relay data packets received from deeper nodes in addition to sending their own data packets. So the nodes near the sink consume more energy than other nodes located deeper and thus die earlier. In this way an area with dead nodes i.e., a coverage hole is created in the network and no more data packets can be received at sink.

We propose an intelligent depth based routing (IDBR) protocol that overcomes coverage holes by assigning initial energy to the nodes in proportion to their distance from the sink. This work is an extended version of our previous work [3] in which we used idea of optimal energy assignment for improving lifetime of UWSNs and now we have enhanced this idea to make energy assignment more intelligent.

Rest of the paper is organized as follows. Section 2 gives an overview of the related work. Section 3 provides motivation for

[*]Corresponding Author: Muhammad Kaleem, Department of Electrical Engineering, COMSATS Institute of Information Technology Islamabad, Pakistan
Email: mkaleem@comsats.edu.pk

this research problem and we present our proposed protocol IDBR in Section 4. In section 5 performance of the proposed protocol is evaluated using simulations and Section 6 concludes the paper.

2. Literature Review

In this section, we present an overview of some of the protocols that have been reported in the literature to address the issues regarding efficient routing in UWSNs.

A hybrid protocol is presented in [4] to contribute towards balanced energy consumption in UWSNs. This technique is limited to the sparsely deployed networks only. Proposed scheme uses direct as well as hop-by-hop communication modes among sensor nodes during its operation. All nodes in the network are deployed linearly. Authors divide energy of each node into small chunks known as energy grades. When the protocol operation starts, all nodes are communicating in hop-by-hop mode. As nodes closer to the sink bear more data traffic, their energy falls earlier. Remaining energy grades determine the residual energy of a node. When the energy level of a low-depth node falls, it informs high depth sender nodes. Sender nodes then do not send their data packets to that node; instead they send it directly to the sink. This reduces data traffic load on the low-depth node and increases its lifetime. This process continues for all nodes and energy of the whole network is utilized in a balanced way.

Tayyaba et al. present another technique [5] for balanced energy consumption for UWSNs. The protocol is termed as depth based energy balanced hybrid (DB-EBH) protocol. In this protocol, all nodes are provided with equal amount of initial energy and then both direct and hop-by-hop communication modes are used to utilize the node energy in a balanced manner to maximize nodal lifetime. Sensor nodes in [4] are deployed linearly while in [5] these are deployed randomly in underwater environment.

Authors in [6] propose two energy balancing techniques for UWSNs. First technique is named as Efficient and Balanced Energy consumption Technique (EBET). EBET is based on the residual energy level of the nodes but instead of using direct communication mode authors propose route changes to reduce traffic load on receiving nodes. When residual energy level of some receiving node falls, it informs the sending node. Sending node does not send its data packet to that receiving node; instead it looks for some other receiving node with higher residual energy level and sends its data packet to it. In this way data traffic load on the nodes with lower residual energy is decreased and they remain alive for a longer period of time. Other technique called Enhanced Efficient and Balanced Energy consumption Technique (EEBET) uses the concept of depth threshold to enhance significance of the proposed technique.

EEDBR [7] is energy efficient version of DBR in which authors address limitations of DBR by suppressing transmission of extra data packets. Data packets are sent to the low-depth nodes and held for a short period of time called holding time. If the same packet is forwarded to the sink by some other node, its transmission is suppressed otherwise it is forwarded to the sink upon expiration of holding time. EEDBR performs better than DBR in terms of energy efficiency.

Another balanced energy consumption scheme is proposed by the authors in [8]. This protocol is based on optimal distance acquisition to achieve higher energy efficiency in Wireless Sensor Networks (WSNs). Authors in [9] and [10] also contribute towards balanced energy utilization in UWSNs using hybrid technologies with slightly different approaches from each other. H2-DAB [11] meets the challenges of UWSNs by implementing the dynamic addressing scheme among the sensor nodes without requiring the localization information. Another scheme known as RERP2R [12] uses the routing metric based on the physical distances between the sensor nodes and to achieve higher throughput in UWSNs. This protocol also gives the energy efficient solution for data forwarding along with better link quality. QELAR [13] is another routing scheme for UWSNs which strives to achieve longer network lifetime by selecting adequate receiving nodes for sender nodes. PULRP [14] introduces layered architecture and detailed algorithm to improve throughput and achieve energy efficiency in dense underwater conditions. HH-VBF [15] uses the vectors assumptions between the source and the destination nodes and proposes a vector-based algorithm to achieve low end-to-end delay in UWSNs.

3. Motivation

DBR uses deployment depth of the sensor nodes to perform routing operation. This fact makes it a very simple and easy to implement routing protocol. But DBR has a limitation that nodes deployed closer to the sink are out of energy after a short period of operation due to greater traffic load on them. This limitation causes creation of coverage holes in the network and data packets fail to reach the sink despite many alive nodes in the network. Coverge holes disrupt the network operation and available energy of the alive nodes cannot be utilised. Ideally, all nodes of the network should have lifetime closer to each other to avoid this energy-mismatch problem. EEDBR adresses energy wastage problem to some extent by suppressing extra transmissions. In EEDBR, nodes nearer to the sink are allocated with the same amount of energy as all other nodes. Ultimately these nodes die out because of greater load of data traffic. We address these limitations by proposing a more efficient scheme called IDBR which uses intelligent initial energy assignment to avoid coverage holes, improve throughput and prolong lifetime of the network.

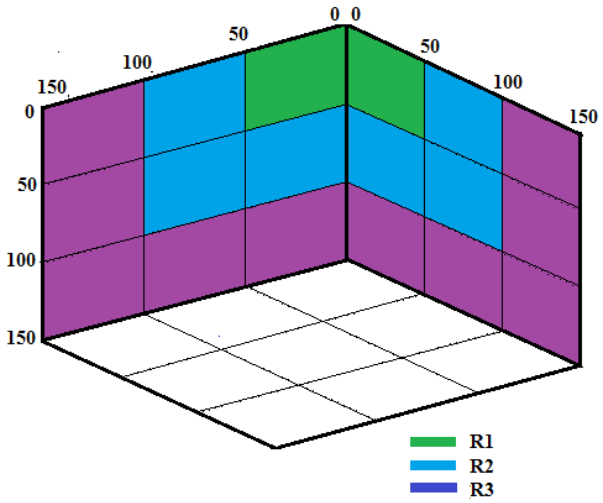


Figure 1. Network model

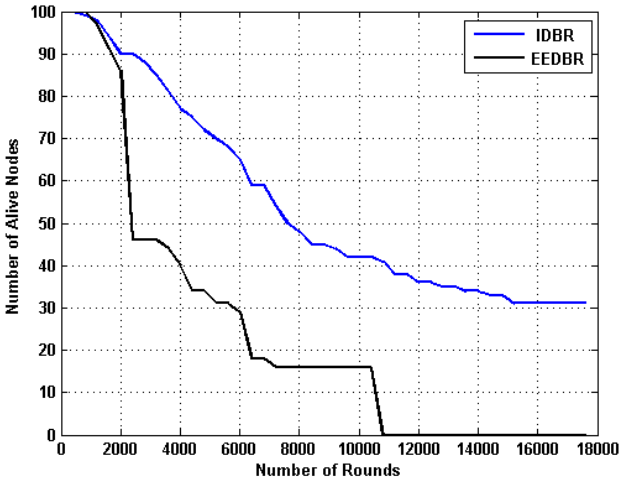


Figure 2. Comparison of alive nodes after each round for IDBR and EEDBR

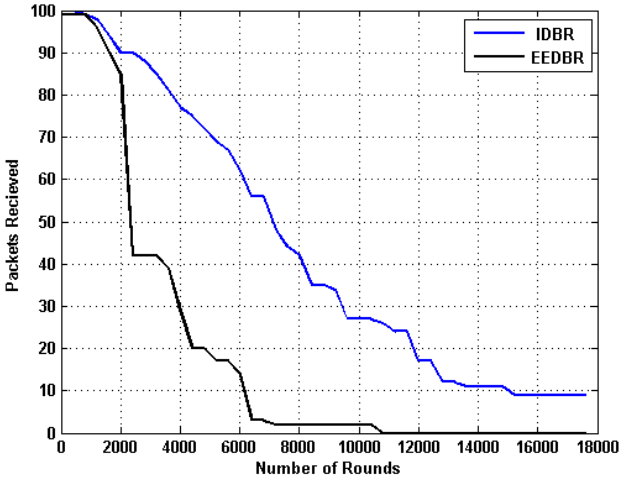


Figure 3. Comparison of packets received after each round in IDBR and EEDBR

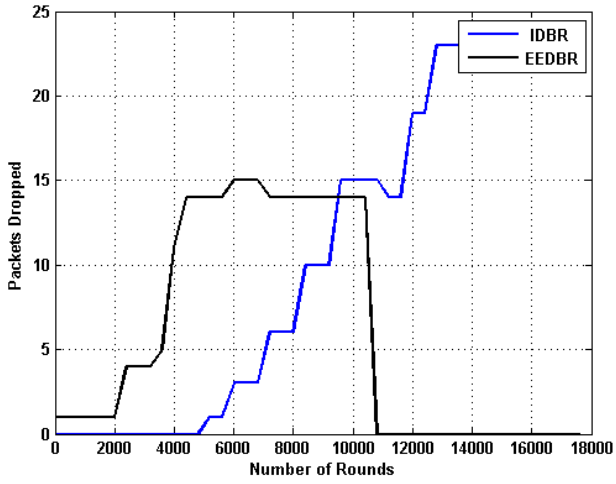


Figure 4. Comparison of packets dropped after each round for IDBR and EEDBR

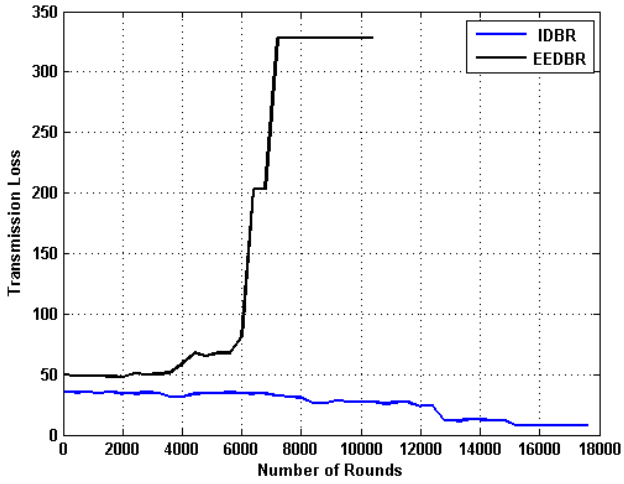


Figure 5. Comparison of transmission loss after each round for IDBR and EEDBR

4. Proposed Scheme

Energy efficiency is one of the most important issues in UWSNs. In this paper, we present an energy efficient routing protocol for UWSNs.

Figure 1 shows 3-Dimensional view of the network under consideration. It is assumed that nodes are deployed randomly in an area with equal length, width and depth.

In IDBR, we divide network area in three equal regions. Region nearest to the sink is named as R1. Second region is named as R2 at relatively greater distance from the sink compared to R1. Region at the maximum distance from the sink is named as R3. These regions are distinguished with the help of different colours in Figure 1.

Nodes belonging to R1 consume maximum amount of energy as they are responsible to receive and transmit data packets to the sink from all other nodes along with their own data packets. In IDBR, we assign maximum amount of initial energy to the sensor nodes belonging to R1. This assignments reflects the core idea of our scheme i.e allocation of initial energy keeping usage requirements of nodes in view.

Nodes belonging to R2 are assigned with relatively lesser amount of energy because they share lesser traffic load as compared to those nodes present in R1.

R3 nodes are provided with the least amount of initial energy as they have to send their own data packets only and hence utilize least amount of energy.

Total initial energies of IDBR and EEDBR are set be approximately equal to make realistic comparison between protocols. For IDBR, we assume that nodes in each region are one third of the total nodes. Equation (1) is used to realize this supposition:

$$n/3(E_{R1}+ E_{R2}+ E_{R3}) \approx nE_0 \quad (1)$$

Where:

n = Total number of nodes

E_{R1} = Energy of each node in R1 for IDBR

E_{R2} = Energy of each node in R2 for IDBR

E_{R3} = Energy of each node in R3 for IDBR

E_0 = Energy of each node for EEDBR

E_{R1}, E_{R2} and E_{R3} are set in a way that equation (1) is satisfied. This relation ensures that total initial energies assigned to the entire network are approximately equal for both the protocols. The only difference is that; in EEDBR all nodes are assigned equal amount of energy but in IDBR nodes belonging to the different zones carry different initial energies. IDBR operation is summarized as follows:

i. After deployment, all nodes in the network need to know about depth information of each other. This is accomplished through exchange of a hello packet. Hello packet contains depth information of each node. When nodes exchange this packet, they obtain the depth information of each other. This step is repeated before each round of data transmission.

ii. Nodes that lie within the transmission range of a particular node, are known as neighbour nodes of that node. Each node in the network identifies its neighbor nodes when a hello packet is exchanged.

iii. A node which has a sensed data packet to send is known as a sender node. All sender nodes compare their own depth with the depth of their neighbor nodes. Sender node passes its data packet to a neighbor node for which two conditions demonstrated by inequalities (a) and (b) are satisfied. Neighbour node to whom the data packet is sent, is called receiver node.

$$d_r < d_s \quad (a)$$

$$d_s - d_r > d_{th} \quad (b)$$

where:

d_s = Depth of sender node

d_r = Depth of reciever node

d_{th} = Depth threshold

Inequalities (a) ensures that reciever node is a low-depth node so the data packet is being fowarded upwards i.e., towards the sink. While inequality (b) ensures that the distance between sender and reciever nodes is less than the depth threshold. Depth threshold is the minimum distance to which a node is allowed to send a data packet. Depth threshold makes sure that sender and reciever nodes are not too close. This idea avoids too many hops during the routing action of a data packet.

iv. Receiver nodes now become sender nodes for their low-depth neighbors. Steps (ii) and (iii) are repeated until data packets are passed to the sink.

5. Results and Discussion

In this section, performance of proposed scheme IBDR is evaluated using simulations. Performance of IDBR is compared with that of EEDBR. Simulations are carried out using MATLAB. Length, width and depth of the network under consideration are set to be 150 meters each. Transmission range and depth thresholds are taken as 50 and 25meters respectively. 100 nodes are deployed in the network and simulations are run for 18000 rounds of data transmission.

Network Lifetime

Network lifetime is defined as the time period from initialization of the network operation till death of the last node. In simulations, number of data transmission rounds determine network lifetime. More rounds of data transmission mean longer lifetime and vice versa. Number of alive nodes after each round for IDBR and EEDBR is shown in Figure 2.

It is clear from the graph that number of alive nodes in EEDBR drop to zero when around 11000 data transmission rounds are run. Whereas, in IDBR almost 30% nodes are alive after 18000 data transmission rounds. This makes a huge difference between lifetimes of our proposed protocol IDBR and EEDBR. It is thus confirmed that intelligent assignment of initial energy to the nodes in IDBR increases operational lifetime of the network significantly as compared to that of EEDBR scheme.

Throughput

Data packets received at sink in unit time define throughput of a routing scheme. Comparison of data packets received after each round in IDBR and EEDBR is shown in Figure 3.

Figure 3 shows that throughput of IDBR is much better than that of EEDBR. In EEDBR number of received data packets falls drastically and no more packets are being received after 11000 data transmission rounds. While in IDBR, number of ecieved packets decreases rather smoothly and data packets are being received till the last round.

Data packets dropped after each round for both the protocols are compared in Figure 4. Number of packets dropped in EEDBR increases rapidly and finally drops to zero near 11000 rounds because at this stage no alive nodes are present in the network and no more data transmission is occurring. In IDBR, lesser data packets are being dropped than those of EEDBR. But more data packets are being dropped in IDBR during last rounds. The reason lies in the fact that more number of nodes in IDBR die with the passage of time.

Energy Consumption

One of the most important parameters to be considered in UWSNs is consumption of energy. Figure 5 shows comparison of transmission losses of IDBR and EEDBR.

Figure 5 shows that transmission loss in EEDBR is much greater than that of IDBR. Transmission loss of IDBR is almost constant during initial rounds but goes decreasing in last rounds because sensor nodes go on dying as the time passes. Transmission loss of EEDBR is much higher in the later rounds.

This comparison justifies the efficiency of IDBR in term of energy usage. Lesser and uniform energy consumption makes IDBR a much better routing protocol for UWSNs than EEDBR.

6. Conclusion

The paper compares the proposed protocol IDBR and EEDBR and shows significant improvement in lifetime, throughput and energy consumption of UWSNs. The intelligent assignment of initial energy to the sensor nodes enables us to distribute energy to the nodes according to their communications utilization. This idea improves the performance of depth based routing protocols for UWSNs and requires energy estimation of sensor nodes for realistic initial energy assignments.

References

[1] F. Akyildiz, D. Pompili, and T. Melodia, "Underwater acoustic sensor networks: research challenges," Ad Hoc Networks 3, pp. 257279, 2005.

[2] Y. Hai, Z. J. Shi, and J. H. Cui, "DBR: depth-based routing for underwater sensor networks," International Conference on Research in Networking Springer Berlin Heidelberg, pp. 72-86, 2008.

[3] T. Ahmed, M. Chaudhary, M. Kaleem, and S. Nazir, "Optimized depth-based routing protocol for underwater wireless sensor networks." IEEE International Conference on Open Source Systems & Technologies (ICOSST), pp. 147-150, 2016.

[4] L. Tayyaba, N. Javaid, S. M. Ali, M. Imran, and M. Alnuem, "Depth-Based Energy-Balanced Hybrid Routing Protocol for Underwater WSNs," International Conference on Intelligent Networking and Collaborative Systems (INCOS) Taiwan, pp. 262-267, 2015.

[5] L. Hanjiang, Z. Guo, K. Wu, F. Hong, and Y. Feng, "Energy balanced strategies for maximizing the lifetime of sparsely deployed underwater acoustic sensor networks," Sensors 9, vol. 9, pp. 6626-6651, 2009.

[6] J. Nadeem, M. Shah, A. Ahmad, M. Imran, M. I. Khan, and A. V. Vasilakos, "An Enhanced Energy Balanced Data Transmission Protocol for Underwater Acoustic Sensor Networks," Sensors 16, vol. 4 , pp. 487, 2016.

[7] A. Wahid, S. Lee, H. Jeong, and Dongkyun Kim. "Eedbr: Energy-efficient depth-based routing protocol for underwater wireless sensor networks," Advanced Computer Science and Information Technology, Springer Berlin Heidelberg, pp. 223-234, 2011.

[8] L. Xuxun, "An optimal-distance-based transmission strategy for lifetime maximization of wireless sensor networks," IEEE Sensors Journal 15, vol. 6, pp. 3484-3491, 2015.

[9] D. Jinfeng, G. Zhang, Z. Guo, and J. Cao, "PAS: probability and sub-optimal distance-based lifetime prolonging strategy for underwater acoustic sensor networks," Wireless Communications and Mobile Computing 8, vol. 8, pp. 1061-1073, 2008.

[10] C. Jiabao, J. Dou, and S. Dong, "Balance transmission mechanism in underwater acoustic sensor networks," International Journal of Distributed Sensor Networks, pp. 2, 2015.

[11] M. Ayaz, A. Abdullah, I. Faye, and Y. Batira, "An efcient dynamic addressing based routing protocol for underwater wireless sensor networks", Computer Communications, vol. 35, no. 4, pp. 475486, 2012.

[12] A. Wahid, S. Lee, and D. Kim, "A reliable and energy-efficient routing protocol for underwater wireless sensor networks", International Journal of Communication Systems, pp.2048-2062, 2012.

[13] T. Hu and Y. Fei, "Qelar: a machine-learning-based adaptive routing protocol for energy-efcient and lifetime-extended underwater sensor networks", Mobile Computing, IEEE Transactions on, vol. 9, no. 6, pp. 796809, 2010.

[14] S. Gopi, G. Kannan, D. Chander, U. Desai, and S. Merchant, "Pulrp: Path unaware layered routing protocol for underwater sensor networks", IEEE International Conference on Communications, pp. 3141-3145, 2008.

[15] X. Peng, Z. Zhong, N. Nicolas, C. Jun-Hong, and S. Zhijie, "Efcient vector-based forwarding for underwater sensor networks", EURASIP Journal on Wireless Communications and Networking, pp.195910, 2010.

ASTES

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1804-1814 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Improved Hybrid Opponent System for Professional Military Training

Michael Pelosi*[1], Michael Brown[2], Kinza Ahmad[3]

[1]*Professor, East Central University, 1100 E 14th St, Ada, OK 74820, USA*

[2]*Program Director, University of Maryland University College, 1616 McCormick Dr, Upper Marlboro, MD 20774, USA*

[3]*Department of Computer Science, University Institute of Information Technology, PMAS Arid Agriculture University, 46300, Pakistan*

ARTICLE INFO

Article history:
Received: 28 June, 2017
Accepted: 13 September, 2017
Online: 05 October, 2017

Keywords:
Artificial intelligence
Expert system
Hybrid AI
Professional training
Software engineering

ABSTRACT

Described herein is a general-purpose software engineering architecture for autonomous, computer controlled opponent implementation in modern maneuver warfare simulation and training. The implementation has been developed, refined, and tested in the user crucible for several years. The approach represents a hybrid application of various well-known AI techniques, including domain modeling, agent modeling, and object-oriented programming. Inspired by computer chess approaches, the methodology combines this theoretical foundation with a hybrid and scalable portfolio of additional techniques. The result remains simple enough to be maintainable, comprehensible for the code writers as well as the end-users, and robust enough to handle a wide spectrum of possible mission scenarios and circumstances without modification.

1. Introduction

"There is no substitute for a human opponent." — Vincent "T.J." Taijeron, USMA Warfighting Simulation Center, West Point, NY. When one is lacking, however, we attempt to offer a usable substitute. In this paper, we describe an architecture and a methodology for software engineering a Computer Opponent Artificial Intelligence (COAI) for professional military training. Such an architecture ideally should meet the design goals of being frugal and efficient in code, easily maintainable, and produce an acceptable level of realism and flexibility for military training personnel and administrators. A truly low-overhead and low-impact solution to the vexing "AI" problem for professional military training at echelons below division and corps level. At the current time, a paucity of software exists, either commercial off-the-shelf computer games or DoD produced and acquired software, for low-cost training in this regard [1].

1.1. Background

Army simulation training has typically used extremely complex, sophisticated, and costly software that necessitates set-up time and planning, large staffs, large budgets, training, and Herculean scenario design efforts. Recently, there has been a shifting of emphasis to what is called low-overhead/low-impact computerized training that lower-level echelons, which traditionally did not have access to large-scale simulation support, can utilize effectively and efficiently. The offerings in this area are slim, and typically commercial computer wargames are wholly inadequate for many reasons. In particular, realistic and useful computer opponent "AI's" are virtually completely lacking. Tasking organization staff to "play the part" of opposing forces is a plausible solution, but necessarily involves a huge commitment of resources when theoretically the CPU can be doing the same thing at little to no cost. Architecting and implementing a targeted solution to the "AI" problem at the appropriate level of simulation and modeling fidelity has been a persistent issue for more than a decade [2,3].

1.2. Current Practice

Large-scale simulation exercises are frequently conducted at the higher levels of the Army command structure. These include division, corps, and army echelon levels. Lower-level training has largely been restricted to manual map exercises, or expensive field training and wargames [4]. Software tools at the company, battalion, and brigade training levels have been sparse. One frequently utilized piece of software is the VBS simulation, which

*Corresponding Author: Michael Pelosi, Professor, East Central University, 1100 E 14th St, Ada, OK 74820, USA | Email: michael@mauisolarsoftware.com

is commercially marketed as "Arma". This is a first-person shooter type game that has been utilized for squad and platoon level infantry type training. However, at the next higher echelons software training tools are virtually nonexistent. Even the widely utilized VBS platform has a woefully inadequate AI — routinely cooks and mechanics are tasked to drive trucks, fly planes and helicopters, and play civilians, as part of the simulation exercise.

1.3. Envisioning the Solution

How can the computer opponent AI problem be solved? Several AI approaches are possible. Hierarchical Task Network planning could be utilized. Unfortunately, HTN planning would typically result in one monolithic game-play solution largely predictable between sessions. Expert systems are possible, but structuring the implementation and gathering expert responses to all possible inputs is difficult and time consuming. Computer chess type solutions offer great promise, but it took over 50 years to get a computer to play a good enough chess to beat the best human players.

1.4. The Computer Chess Analogy

Despite 50 years of advancements in computer technology, computer chess AI still relies on a largely brute force approach. Using the Min-Max algorithm, transposition tables, and other optimizations, the chess AI scans through the game tree, analyzing millions of potential moves, before producing the highest scored next move [5,6]. Thankfully for chess, there are only 64 squares and a maximum of 32 pieces. In a professional military simulation, a map may consist of millions of individual 100-meter sized grid squares, thousands of units, and a completely unpredictable terrain and mission. In other words, searching through the game tree of all possible actions is tens of magnitudes more difficult. The chess modeling approach is not scalable, not nearly so. The chess analogy AI cannot be made to fit, yet we can adopt some of the lessons learned from computer chess. This includes dividing the session into phases: opening, middle-game, and endgame. Scoring the value of different pieces (maneuver unit groups), and evaluating final desired states, and how to get there, have proven to be extremely useful concepts.

1.5. A Scalable Hybrid Approach

The approach we have implemented and described in this paper could be considered as a hybrid AI approach. Relying on the chess analogy and metaphor, particularly the game phase concept, we add on an expert system that models and abstracts the actual units in decision-making processes wherever possible. For example, the AI groups subunits into their actual mission structure, including companies and battalions. These are controlled as in the military force structure chain of command. For the creation of plans, it is possible to adopt the actual military staff procedure and adapt this into a software planning sequence. Further, lower-level planning details such as route determination can be supplemented by the ubiquitous A* Algorithm. Through trial and error over a development lifecycle of several years, and through feedback from professional users, we have cobbled together a hybrid approach that is efficient, robust, and easily implemented. Flexibility for future features and maintainability has been realized.

1.6. Presentation Goals

Our goals in this presentation are to describe to the simulation community, and perhaps computer gaming community, the derived computer opponent AI architecture. More than 10 years of effort have been devoted to its creation. Much of this time was spent ruminating on the final ideal implementation — a big part of the solution was multiple layers of modeling abstraction. Isolation between the layers was and is also extremely important for comprehension and maintainability. Further, and to a large degree, the AI layers actually model in parallel the military command echelons.

1.7. Engineering the New Approach

There are individuals who spend vast portions of their waking hours operating and contending with AI code in simulations and games. We have been careful to consider their recommendations and concerns in engineering an appropriate and efficient AI system for land warfare simulation. One of the quotes that comes to mind is the following: "The AI must perform the mission, the AI must be a good soldier not a general." In other words, unexpected or completely unpredictable behavior is out of the question. This is both for realism and utility. If the trainee utilizing the AI cannot understand what it is doing it may be time wasted. In addition, the scenario designer may want to script certain actions to instruct certain doctrinal principles. Thus, scripting of the AI should also make it possible to funnel behavior into a certain group of possible actions. At the same time, the AI cannot be overly complex: "…and one turn requires 15 minutes of preparation and then crashes! Good job!" Clearly there is a trade-off between complexity and robustness.

2. Methodology

A realistic AI useful for professional training purposes should both model and mimic the military decision-making process at various echelons. In that light, as a robust foundation the AI goes straight to the U.S. Army field manuals for guidance. Fortuitously, more than 100 years of modern warfighting experience has distilled down Army planning doctrine to a few formulaic processes in the Military Decision-Making Procedures (MDMP). These include TLP, METT-TC, and OCOKA. These processes can and have been modeled almost directly in code.

2.1. Military Decision-Making Procedures

U.S. Army Field Manual 101–5, Staff Organization and Operations, explains in detail the Army MDMP (U.S. Army 1997). "The MDMP is an adaptation of the Army's analytical approach to problem solving. The MDMP is a tool that assists commanders and staff in developing estimates and plans. The full MDMP is a detailed, deliberate, sequential, and staff-intensive process used when adequate planning time and sufficient staff support are available to thoroughly examine numerous friendly and enemy courses of action (COAs). This staff effort has one objective—to collectively integrate information with sound doctrine and technical competence to assist the commander (in our case the COAI "commander") in decisions, leading ultimately to effective plans. The analytical aspects of the MDMP continues at all levels during operations."

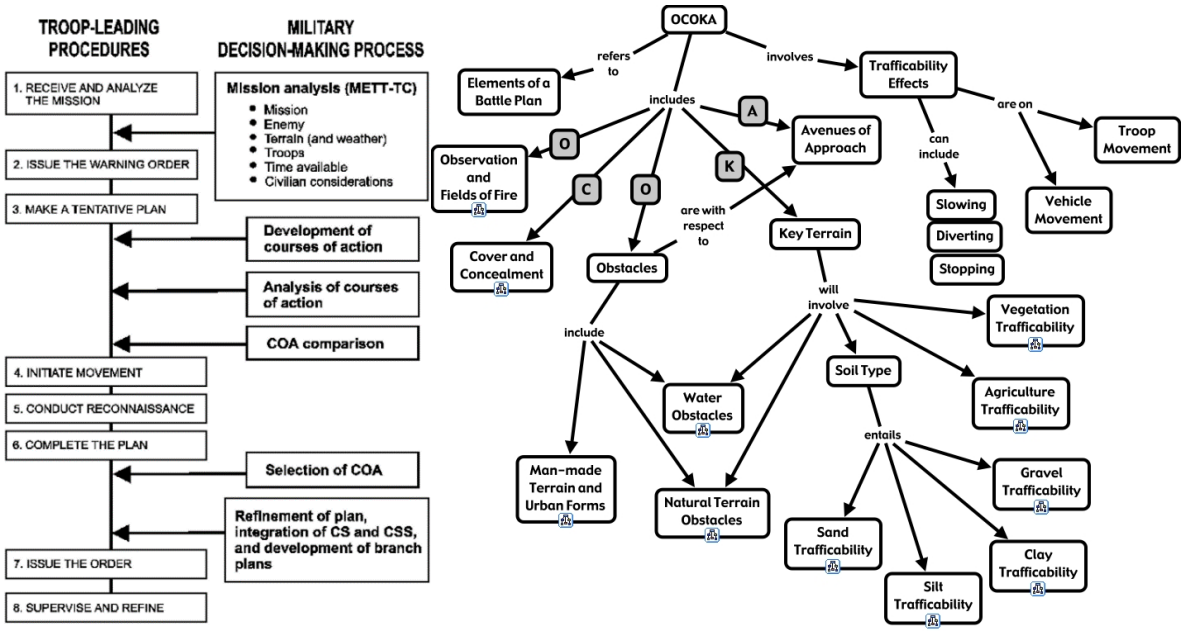


Figure 1: TLP, MDMP, METT-TC, and OCOKA Depictions [7]

2.1.1. *Military Missions and Objectives*

The COAI is presented a military mission that is contained within scenario and AI option specification files. The files include information on task organization, friendly forces, and a timeline. Objectives are also specified with points values for various objective type such as occupying a location, clearing an area of enemy forces, moving friendly forces past a certain demarcation zone, or searching for a hidden target.

2.1.2. *Military Decision Making Procedures*

Where possible the military decision-making process (MDMP) is then followed both in modeling and implementation. The COAI is designed to follow a similar process to what is recommended doctrine in the Army field manuals. Likewise, parallel modeling and decision-making takes place at each of the important unit echelon levels: platoon, company, battalion or task-force, and support. The following further outlines MDMP aspects:

TLP (*T*roop *L*eading *P*rocedures) consist of the following steps: 1. receive the mission and conduct METT-TC and OCOKA, 2. prepare for the mission and issue preliminary orders, 3. make a tentative plan: identify goals, gather information, generate/analyze/compare possible solutions, and implement the best tentative plan, 4. start movement, 5. conduct reconnaissance, and 6. follow through with execution of the final plan.

METT-TC is: *M*ission analysis, *E*nemy analysis, *T*errain analysis, *T*roops analysis, *T*ime limit analysis, *C*ivilian impact analysis. *OCOKA* is conducted as part of terrain analysis.

OCOKA stands for *O*bservation and fields of fire, *C*over and concealment, *O*bstacles, *K*ey terrain, and *A*venues of approach. This constitutes a more detailed terrain analysis. Obstacles can include man-made and urban terrain obstacles, natural terrain obstacles, and water obstacles. Key terrain may involve, for example, high elevation or easily traversed terrain near objectives. Avenues of approach include roads and otherwise clear areas. Trafficability can be evaluated both for vehicle and troop movement regarding slowing, diverting, or stopping movement.

2.1.3. *Course of Action Adoption*

The Course of Action (COA) for the COAI that is produced is a result of the analysis of the above factors and constraints. Reducing all considerations to a quantitative scoring allows a brute force solution, that randomly generates various plans and each plan can be scored for its suitability and feasibility. The highest scoring feasible plan is selected as the best COA. Since plans are randomly generated, differing and unique plans are generated during each new instance of the same scenario mission design. This is important for replay value.

Execution requires close supervision and monitoring, as well as continuous analysis, the updating of intelligence, and refinement of the COA plan. In certain cases, the plan must be discarded and regenerated completely.

The COAI has certain unit groups assigned to it in the scenario and mission design. This allows the possibility of multiple instances of the AI, each controlling its own respective force grouping. Likewise, human participants would each be controlling various force groupings.

2.2. *Mission Analysis*

Scenario design inputs to the COAI for information analysis include: time limit constraints, enemy OOB, friendly OOB, quantified scenario objectives, along with AI options and settings. Five major phases are accomplished during a preliminary mission analysis:

1. Analysis and calculation of the goal state to satisfy objectives. Often this will involve the ideal placement of forces by the scenario end time, such as the occupation of objective locations.
2. Analysis and calculation of known enemy dispositions and force allocations. Here relative points values are calculated

relying on "combat power" summations for known enemy unit types in a catalog database of unit types. Friendly unit group combat power totals are likewise analyzed to create favorable force match ups. In general, a ≥ 3:1 points total advantage will be necessary for successfully taking the occupation of locational objective from a defender [8].

3. Analysis and calculation of a tentative plan to reach the goal end state. This is accomplished by using a brute force approach like the solving of the Traveling Sales Person problem [9]. Several thousand likely plans are randomly generated and scored. Top scoring plans are then further evaluated and selected.
4. Surplus time and resources are evaluated. If the plan can be accomplished before the scenario mission end time, further refinements and optimizations can be preliminarily executed. This can include actions such as further intelligence gathering and reconnaissance, softening up of target locations through preliminary bombardments and airstrikes, and conducting feign attacks or deep pincer movements.
5. Finally, initial tentative movements and actions for the first game phase are calculated, however these may change when the final COA is adopted.

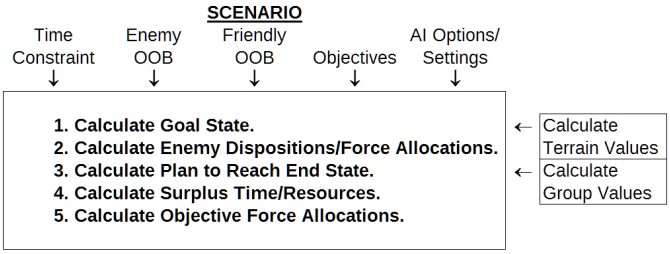


Figure 2: Scenario and Mission Specifications and Analysis.

2.2.1. *Discrete Mapping Zones*

As part of the mission analysis and execution, map zones are segmented and demarcated. Map zones are segmented based on the center mass coordinates of friendly forces, and known or likely center mass coordinates of enemy forces. With these two points fixed in space, a relative "mapping" of center, left and right flanks, and depth can take place. If portions of these areas are off the given terrain area, they are ignored and become notional. Areas which are untrafficable (for vehicular and/or foot units respectively) are also ignored for deployment or movements. The point halfway between friendly center mass and enemy center mass becomes the Forward Edge of the Battle Area (FEBA) anchor. Ten discrete zones are then demarcated from this mapping: forward screen left, left, center, right, screen right, and corresponding rear areas respectively. Reconnaissance missions will be routed into the far side of the FEBA, and missions will be allocated to specific zones. Combat group deployments take place generally in the left, center, and right, with screening units assigned to screen left and screen right areas. Supporting units, including headquarters, artillery, and logistics, are routed toward the rear areas. Spacing between units is calculated by the frontage span of the respective maps zone. Reserves are held in the rear area as well. Thus, any map size from 5 km × 5 km up to 500 km × 500 km can be automatically mapped into a convenient "scenario" mission and planning space, based on the map size, force size, and initial deployments.

2.2.2. *Map and Terrain Analysis and Updating*

As part of the mission analysis, a detailed terrain analysis is conducted and the results are stored in a map database of grid squares. Each map grid square is assigned a weighted and then normalized score value for specific characteristics. These include the following:

Map static terrain analyses:

- *Objectives* – value and proximity to objective locations.
- *Water* – water obstructions to ground movement — note this may have an important lack of effect on the many amphibious vehicles in operation.
- *Elevations* – in many circumstances higher elevation locations are more valuable to occupy.
- *Grades* – steep uphill and downhill grade serve as detriments to mobility.
- *LOS* – line-of-sight to nearby grid squares; some locations can observe much more of the surrounding terrain.
- *Blocks* – blocks can include highly dense vegetation, urban locations, as well as man-made obstacles.
- *Cover* – cover provides shelter from blast effects and observation.
- *Avenues* – key avenues for movement, central locations networks to objectives are preferred.
- *Concealment* – concealment has low line of sight visibility as well as good cover.
- *Defense* – combination of effects from above for defensibility.
- *Survivability* – cover, concealment, and defensibility modified for survivability aspects.
- *Ambush* – areas with good visibility/survivability, nearby to key avenues of movement.
- *Counter mobility* – places where the enemy's movement can be stopped efficiently.

Map dynamically updated analyses:

- *Valuable Areas* – avenues, high elevations, nearby to objectives, etc.
- *Friendly Proximity* – weighted/normalized value for staying close to friendly concentrations.
- *Enemy Proximity* – weighted/normalized value for known/updated enemy concentrations.

The COAI user interface produces shaded map graphics depicting weighted and normalized values for each of the terrain analyses based on grid square location. The last three terrain analyses are dynamically updated as the simulation progresses, and reflect new and current information.

2.2.3. *Order of Battle (OOB) Analysis*

Because of the COAI not considering individual unit entities at the lowest level (platoon and section sized entities), the COAI is only aware of the unit groupings, relative combat power, and the group types. Using these characteristics, Order of Battle (OOB) analysis can assign various groups to specific objective missions. For example, an armor company may only constitute 14 tanks but have triple the combat power of a 90-soldier infantry company. The combat power is based on points totals from the unit catalogs. As part of the COA production, the COAI analyzes the most

optimal assignment of groups to objectives using its limited knowledge. With regards to enemy forces — given adequate knowledge, OOB analysis will endeavor to produce adequate match ups, specifically the greater than 3 to 1 advantage of an attacker over a defender in terms of combat power.

2.2.4. Spot Table and Updates

The COAI must keep track of known and likely enemy force locations. Initially, the training scenario designer can choose to reveal as much or as little about enemy force locations as desired. This can be loaded into an initial enemy spot table at scenario start, and be used for COAI planning. After that the COAI is on its own gleaning, updating, and aging information as it comes in. It keeps track of this in a dynamic spot table that contains coordinates, unit type and points value, and most recent spot time. As spots are aged they are reduced in weighting importance for relevance and accuracy. The simulation produces a listing of current force enemy spots and hands it off to the COAI, which collates and posts the information in the spot table.

2.2.5. Force Allocations

As previously mentioned, planning the force allocations of groups to objectives involves a brute force planning approach. For each of several thousand plan iterations, groups are randomly assigned to objectives. Then, based on the group data, objective data, and enemy locations, a special function calculates the feasibility of each allocation. Time to reach the objective, attack or defense ratio, and other factors, can cull out infeasible assignments. Of the remaining feasible plans, these are scored based on minimum cost in terms of movement time, and attractiveness of force match ups (desired minimum 3:1 on offense) and other factors. A final "minimum time to complete the plan" duration is calculated, and from this any surplus time available for further measures, such as reconnaissance or softening up of targets, is then known. Force allocation intrinsically calculates the end game phase plan. Since each scenario objective is assigned a relative points value, scoring of plans takes into consideration the satisfaction of more valuable objectives, as well as the distance from each respective group to its assigned objective. For further information on brute force planning using this approach, traveling salesperson solutions are a good starting point [9].

2.2.6. Course of Action Decisions, Types, and "Playbooks"

Once final endgame force allocations have been favorably calculated, if sufficient surplus time and resources exist, a middle game "playbook" COA can be adopted by the COAI to further shift the favorable odds preliminary to the endgame phase. In the case of a defensive posture this can include securing objectives, static defense, or in-depth defense [10]. For attack postures, broad front attacks, counterattacks, or deep attack "playbooks" can be adopted. A reserve force can possibly be selected. The playbook selected is not optimal, but suitable for the given situation and circumstances. This is analogous to the football play: a running play may not be any better than a deep pass, but it keeps the other team guessing. Seemingly random intelligent plans and actions in terms of time and execution are an important part of a realistic and engaging COAI with suitable replay value.

2.3. Major "Game" Execution Phases

Once the COA has been finalized execution will be transitioned through a series of major game phases, relying on the chess motif. These include the opening, middle-game, and endgame. Assuming little slack time exists for achieving the mission objectives — the execution phase will be shifted immediately to "endgame". Endgame can be considered the all-out effort to achieve the objectives immediately. Otherwise, if slack time exists, perhaps an opening and middle-game phase will be adopted as part of the execution.

2.3.1. The OODA Loop: Observe, Orient, Decide, and Act

The execution phases are analogous to the very important military precept of the OODA loop [11]. The opening is comparable to the *O*bserving phase. Here, advantages are to be acquired in terms of additional intelligence and other measures, such as occupation of key terrain. Middle game is analogous to *O*rienting — the major reorienting of friendly forces to further tip the balance for further movements and attacks. Endgame is the *D*ecide and *A*ct portion, where commitment to a decisive outcome is adopted. Final actions are wagered here. Replanning is necessitated between the opening phase and the middle-game phase, as well as between the middle-game and endgame phase.

2.3.2. Execution Phase Transitions

Transitions between the major game phases are based on the characteristics of what should be taking place generally in each phase — once again relying on the chess metaphor. The opening takes place between the scenario start and first enemy contact, first weapons fire, and/or first friendly casualties. The endgame is transitioned to after the middle-game when the previously calculated time deadline for accomplishing the mission objectives is reached. This also includes a time safety factor built-in. In other words, final execution is committed to when there is still enough time to safely accomplish the objectives. That said, to preserve verisimilitude, there exists the possibility of a "lightning battle" COA adoption where the COAI will skip the opening and/or middle-game phases and progress directly to an endgame phase. This is analogous to a surprise execution, which forces the training audience to consider all possibilities. Skipping phases is easily incorporated into the COAI options as probability factors for skipping middle game, and skipping opening and middle-game. At the juncture of each major game phase, replanning takes place based on the phase goals described further below.

2.3.3. The Opening Phase

The opening phase is largely characterized by observing the enemy's respective force deployments and dispositions. Goals here include grouping and further deploying friendly forces, exploiting terrain based on mission analysis terrain calculations, reconnaissance missions are executed, counter reconnaissance missions are executed, and the seizing of easy objectives closer than the enemy which may not be occupied.

2.3.4. The Middle-Game Phase

The middle-game is characterized by major force movements to orient deployment for final attacks and/or defense. Seizing of key intermediate terrain is conducted. Harassment missions are

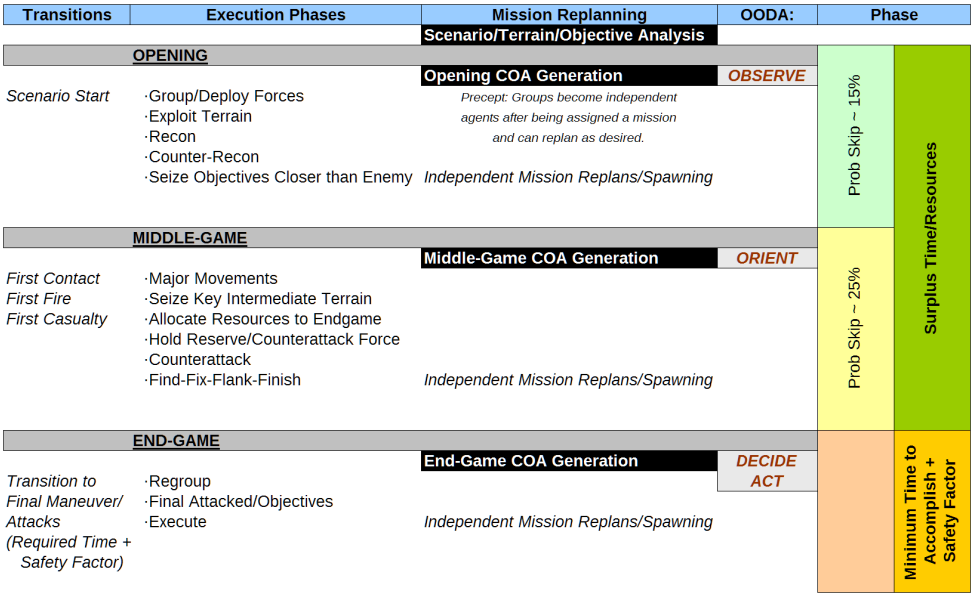


Figure 3. COAI Execution Phases, and major aspects of each phase.

perhaps selected, these would include randomized probes or artillery fire missions. Allocation of resources to the endgame is recalculated. Further, most COAs will hold a major reserve and/or counterattack force for unforeseen events. Counterattacks can also take place. Attacks use as a basis the "4F's" for planning and execution: *find – fix – flank – finish* [6].

2.3.5. The End Game Phase

The endgame embarks on achieving the final scenario objectives. Thus, a regrouping of scattered forces may be necessary for the execution of the final plan. Final attacks are enacted, if necessary, and objectives are occupied. The plan is irretrievably executed at this point — for either final success or failure. Ideally, the opening and middle-game phases have set up the COAI for uncontested victory at this point through incremental and methodical gaining of advantage. As mentioned, the endgame is analogous to the Decide and Act portions of the OODA loop, and opening and middle-game phases only take place if surplus time and resources exist for the satisfaction of the mission goals. Otherwise, the COAI would need to embark on an endgame plan immediately at the scenario start.

2.4. Modeling Layer Architecture

Analogous to the OSI model [12] for computer networking (and its inherent division of responsibilities and functionality), there are at least six layers and levels of modeling that are used in the COAI architecture. Most of these parallel a corresponding layer in the military decision-making process and unit echelon structure, in real world military forces. This leverages the concept of object modeling for real-world abstractions, as well as organizes and simplifies the architecture software code. At the lowest level are the simulation entities, nominally platoon down to section sized organic units. The simulation in usage models each of these uniquely as a C++ class entity. Typically, these are grouped into company to battalion sized units, each consisting of 4 to 10 subunits. It is these groups that constitute the "unit groups" that are under direct control of a scripting engine layer. The scripting engine layer is responsible for issuing the entire group order directives discussed in more detail below. Above the scripting layer is mission control, more directly controlled by the COAI. Once the group has been assigned to a mission, the mission instance is responsible for autonomous control over the group through the scripting engine. Mission sequences, are in turn controlled by the overall COA class plan that is being implemented. And finally, the *COA class* is planned in response to the overall scenario objectives, available resources, game phase, terrain map, and options.

Table 1: Modeling Layers of Abstraction, Planning, and Control.

1.	Execution Phase: Open, Middle, Endgame → Determines strategic approach/goals.
2.	Course of Action (COA) → Self-contained master plan for phase, controls Layer 3.
3.	AI Mission Sequence Collection → Insertion, deletion, reordering possible.
4.	AI Independent Mission Control Agent → An autonomous OOP class, with reports.
5.	Group Scripting Engine → Programmed Sequences of Events/Actions/Responses.
6.	Unit Entity Grouping → Company/Battalion/Task Force, abstracts Layer 7.
7.	Simulation Entity → Platoon/Section/Section/Battery/Vehicle. Hi-fidelity modeling.

2.4.1. Section-to-Platoon-Sized Entities

Section to platoon entities are the lowest level of fidelity in the simulation in usage. The COAI does not control this echelon directly. The group and scripting engine issues direct orders and

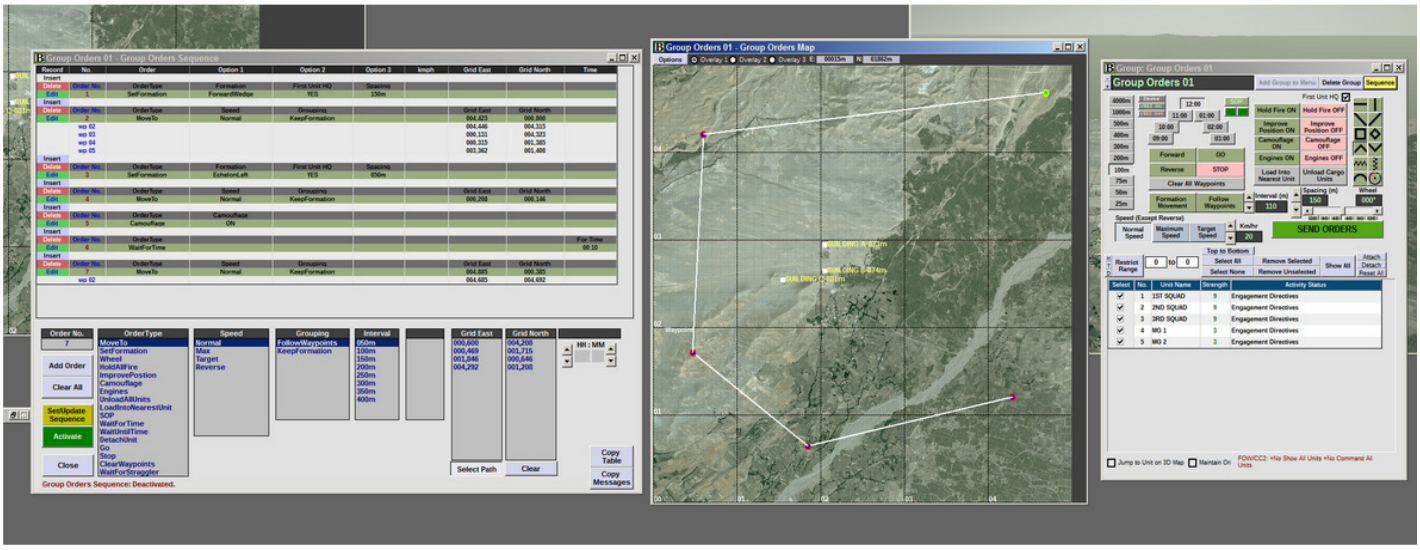


Figure 4: Group Orders Scripting User Interface, which allows the inputting of sequences of orders.

commands to entities at this level. In summary, sections and platoons are characterized by locations, ammunition and fuel levels, strengths and casualty levels, current orders and status, among other data in a cornucopia of minutiae. Accurately modeling this spectrum of characteristics is an extremely labor-intensive task, that requires copious research and data entry. Accounting for hundreds of data items into COAI considerations is architecturally untenable, hence the abstraction to larger unit groupings is necessary to accomplish the goal of a robust and usable AI with a frugal amount of code. Codewise, these entities are modeled as classes.

2.4.2. Unit Groups

Scenario design creates company and task force level unit groupings that normally model individual combat companies or battalions, artillery batteries, helicopter flights, and other unit groupings that would normally be controlled by a battalion or brigade level task force organization. Unit groups are modeled as a class and are the owner of combinations of the platoon and lower level entity grouping.

2.4.3. Group Orders, SOPs, and Formations

Group orders and tasks are relatively straightforward and implemented easily by the controlling mission class. The controlling mission class merely instructs the scripting engine to calculate and implement the command order. Group orders contain such simplistic directives as: move *n* meters, change facing, set speed, dismount infantry, improve position, camouflage, discharge smoke, or set formation. Set formation automatically orients the group in, among others, line, column, box, diamond, forward wage, reverse wedge, and echelon formations. Company and battalion sized formations will typically orient themselves in one of the aforementioned formations to advantageously engage likely targets. The scripting engine handles the details, while the COAI concentrates on decision making at the next echelon above. Lower level units are responsible for handling their own engagement of targets of opportunity. Standard operating procedure (SOP) allows independent decision-making for units in regards to firing smoke or vehicle engine exhaust smoke systems for defense, reversing on enemy sightings, or aggressively attacking new contacts.

2.4.4. Group Scripting

Group order scripting consists of a collection of sequential orders. As mentioned earlier, movements, formation changes, camouflage orders, orders to "dig-in", etc., can be added to a scripting sequence. The scripting engine automatically executes the orders serially until completion. The COAI process communicates into the simulation the desired scripting sequence and parameters through active mission classes. The user interface of the group orders scripting window is shown below in Figure 4. Scripting can be manually controlled as desired, and saved to file.

2.4.5. COAI Control

The COAI enters a control main loop after completing the mission analysis and creating a preliminary COA. Each time the loop executes, more information is extracted from the simulation, this is processed, and the COAI may modify directives or give additional orders to the group scripting engine. Groups with scripting orders pending can have those sequences cleared if necessary. The main loop continues until the scenario end time is reached and objective condition scores calculated.

2.4.6. Course of Action (COA) Modeling

COAs are modeled as a class and store their own data and update themselves inherently. Additionally, given certain circumstances they are capable of canceling themselves which will necessitate and involve an automatic regeneration of a COA. This can optionally be done at random, or in the case of catastrophic goal failure. COAs contain enough information to be considered a high-level plan, with very little implementation details.

2.5. Mission Types

In satisfaction of the current COA, unit groups are assigned sequences of missions that fall into various categories. Each of the

missions is defined in a C++ class, and the sequences of missions are collections of missions. The unit group conducts the next mission listed in its respectively assigned mission collection, until each one is completed. If necessary, a new mission can be spawned and inserted at the top of the collection, at which time the unit grouping will embark on the new mission, and resume the second mission once the newly spawned mission has been completed. For example, a grouping on a movement mission toward an objective can be assigned a newly spawned mission to attack a target of opportunity. Once this attack mission has been completed, the movement toward the objective mission will be resumed. Further, since missions are modeled as autonomous agents, they can spawn their own new missions as necessary which may supersede the current mission. The hierarchical breakdown of various COAI mission classes developed totals over 40 at the present time.

2.5.1. *Mission Agents (as Self-Planning, Autonomous, Self-Updating, and Reporting OOP Classes)*

As mentioned, missions are modeled as classes and have a decoupled implementation. Each mission has a pair of classes closely related: a planner class and an implementer class. The planner class plans the mission and hands over the implementation details to the implementer class. If the implementer class runs into problems, it will call the planner class to once again reinitialize and replan the mission. Once a mission is spawned, it is initialized with several goal variables and the mission class code itself calculates how to correctly carry out the mission. During each iteration update (periodically calculated based on an AI update time step), the mission updates itself and commands to the mission unit grouping as necessary. Upon mission completion, the mission class instance is removed from the mission sequence collection and ceases to exist. Some of the major mission taxonomy types, which rely on C++ class inheritance from the mission base class, include movement missions (which activate A* pathfinding), attack, defense, recon, and support missions. Missions are queryable for public properties such as mission start time, estimated completion time, status codes, and other information. As a result, it is straightforward to keep the user interface updated with graphical status for users.

2.5.2. *Movement*

Movement missions are generally tactical movement or road movement missions. Tactical movement will move in a tactical formation using advantageous routes to the goal endpoint. Road movement will simply travel by the most trafficable route to the goal location. Generally, movement missions will respond to RTC (React To Contact) events using a SOP, which may include attack, evade, stop movement, or retreat doctrines. Recon missions are similar but will move to advantageous locations for observation based on the terrain analysis, among other doctrinal differences.

2.5.3. *Pathfinding*

Pathfinding to mission waypoints along a movement route is calculated using a modified version of the A* [13]. The implementation considers multiple goals, for example the cost function can include factors for avoiding or approaching the enemy, attractiveness for traveling on roads, moving through high line-of-sight grid squares, or maintaining terrain cover. For example, if one of the mission goals is concealment during movement, lower movement cost can be assigned for terrain covered by forested areas or buildings. Pathfinder estimated time of arrival results are based on the speed over terrain of the slowest unit in the group.

An important consideration to note is that pathfinding algorithm code must take place in its own CPU thread. As a result, when a mission needs a pathfinding route, it sends desired goal coordinates as well as group data and route preference to a collection of pathfinder processes executing on the machine. The pathfinding request is queued and the mission class waits until a result is returned. Pathfinding is by far the most computationally intensive element of the entire COAI architecture, other than initial scenario and terrain analysis. Route movement of mission groups is 80% of what the COAI does. The importance of this cannot be under-stressed: a brick-house architecture for planning and implementing movement has been essential.

2.5.4. *Attack Missions*

Various type of attack missions are implemented based on group composition; vehicular, foot, aircraft, etc. Generally for attack missions an endpoint location is assigned as well as a casualty threshold. The group will generally conduct tactical movement toward the objective, execute the attack several times as necessary, and regroup between attacks. During movement react-to-contact standard operating procedure is enacted.

2.5.5. *Attack Precept: Find, Fix, Flank, and Finish*

Generally, attacks are coded to take place using the well-established military metaphor for success which elaborates on the "4F's": find 'em, fix 'em, flank 'em, and finish 'em [14]. Therefore, planned attacks would normally involve coordinated movements and attacks by two or more unit groupings, comprising at least a $\geq$ 3:1 combat power points advantage. Surprisingly, attacks are tractable to plan with acceptable realism once an enemy defender has been located. The fixing force approaches directly within weapons distance and begins firing. Meanwhile, the flanking force(s) conduct flanking movements around the left, right, or rear, and engage the enemy from that direction. The finishing force can be either the flanking force or another available group, which will move in for a final clobbering. A casualty loss threshold is established preliminarily, and if it is reached the attack will be broken off as unsuccessful and the forces reallocated.

2.5.6. *Defend Missions*

Isolated group defend missions are more simple and will merely move to the endpoint location and prepare a defense, normally by "digging in". If a casualty threshold is met the group will withdraw to a safer location.

2.5.7. *Support Missions*

Support missions include being held in reserve, artillery fire support, logistics and supply, and headquarters missions. Forces allocated to support missions, in addition to conducting their primary mission, will relocate periodically to maintain a relative position behind the FEBA. Most support missions will also

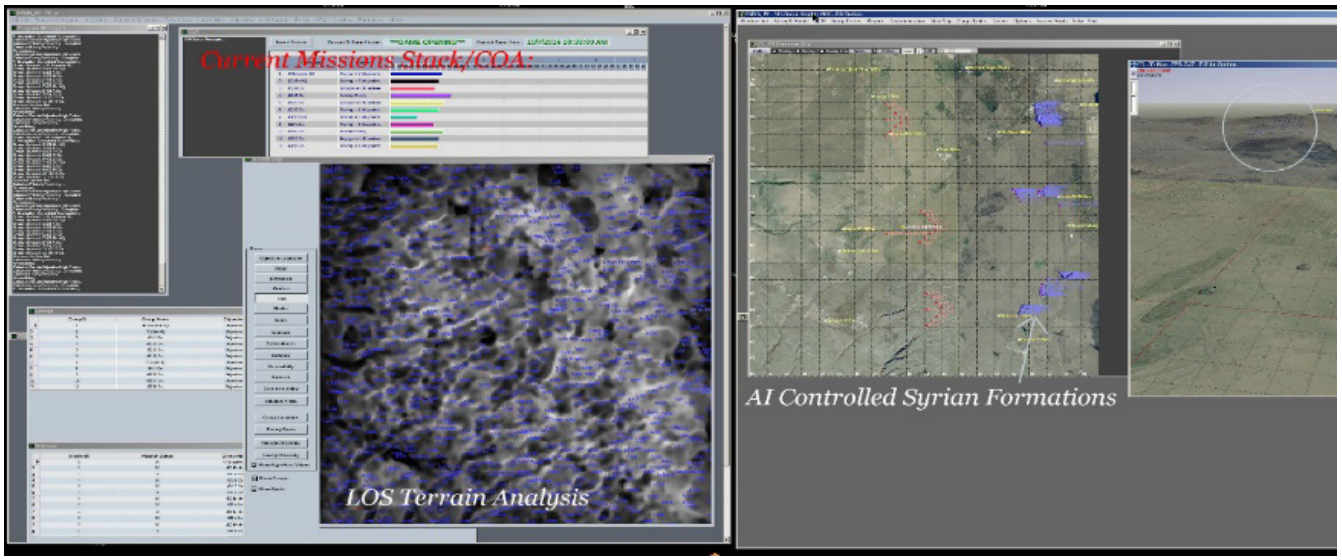


Figure 5: COAI User Interface with mission Gantt chart, terrain maps, and mission status logs.

periodically relocate based on their proximity to the enemy, for example headquarters units will maintain a safe distance between themselves and the nearest enemy, or advance to maintain a general distance to the FEBA. Artillery monitors the current spot table and fires on lucrative targets of opportunity, or fires in support of attacking or defending units when advantageous.

In the special cases of artillery and attack aviation support missions, available direct fire artillery as well as available attack aviation assets, are placed in pools. Group missions can request fire or aviation support based on their circumstances, in which case it is added to a request listing. Artillery and aviation assets periodically evaluate the listing and prioritize their response based on likely effectiveness and proximity. Once satisfied or determined infeasible, requests are removed from the queue listing.

2.6. Group Force Types

Missions are assigned based on unit group type, which is known from the scenario design specification. In general, most unit groupings can be categorized into one of the following major groups: armor, mechanized, mechanized with dismounts, infantry, artillery, recon, screening, aviation (attack/transport/recon), refueling support, ammunition support, or headquarters. The sum of the combat power points for various units is used for calculation of overall combat power and force match up ratios. Periodically, groups may find themselves unassigned to any particular mission. In this case the COA class will score and produce their next best mission assignment.

2.7. Software Implementation

The adopted architecture by necessity uses an object-oriented language. Prolific usage of C++ classes and inheritance features are realized. In particular, the ability to create object collections is critical, as well as the ability for data hiding, decoupling, and multi-threading.

2.7.1. Code Architecture

It is a very important design principle that the COAI code takes place in a separate process in isolation from simulation code. This allows a very important decoupling and higher-level abstraction from the lower-level details generally inherent in the simulation. Additionally, the general time step in the simulation is a tiny fraction of the time step necessary for the COAI calculation and update. For example, if the simulation time step is 30 seconds per iteration, AI is easily updated at 10 minute intervals.

As mentioned, COAs are implemented as a class object, and own mission sequence collections of mission classes. Each unit group has its own mission sequence collection. The current mission is the mission on the first position in the mission sequence collection. This mission class is responsible for controlling its assigned mission group. It passes high-level scripting orders to the simulation scripting interface. The scripting interface controls sequences of orders which include items such as move to waypoint, change formation, change facing, move at speed, and weapons tight or weapons free, among others.

2.7.2. Sequence Execution

When the COAI process is initialized with data, it begins scenario analysis, terrain analysis, and OOB analysis. Typically, this process can take several minutes. When preliminary analysis is completed, an endgame COA is produced as well as the group maneuver plan. Assuming surplus time and resources exist, an opening and/or middle-game COA may be implemented instead. At this point the COA is formally implemented, missions are spawned (which are responsible for planning their own independent execution, and updating themselves, or spawning new missions as replacements). Missions are then implemented. Current missions are posted to a Gantt chart in the user interface which includes each unit grouping. The Gantt chart depicts the mission stack for each group, as well as an estimated completion time for respective missions. The COA is then controlled until the next game phase is reached. The scenario ends at the scenario end time and the victor is calculated based on objective points totals.

2.7.3. COAI User Interface and Processes

It is important to give the training audience and training administrators a detailed window into the internal workings of the

COAI; this is both so that they can understand it as well as appreciate the realism inherent in the modeling. Further, they can more intelligently tweak the COAI settings and options and locate defects and probable improvements. The COAI has its own process independent UI running in parallel with simulation code. Some of the available graphical interfaces include an event log, static terrain map, updated maps with group locations, objective table listing, COAI options, and scenario options windows.

User interface values are updated at each calculation iteration of the COAI, typically at one minute or greater intervals (clock time). The current game time and phase, and information on the COA in execution, is also displayed. It is possible to bring up child windows displaying detailed information on missions, along with corresponding mission log windows. For debugging, it has been necessary to display pathfinding windows with the status of individual route planning efforts.

The user has the option to manually toggle the next game phase, and go from opening to middle-game, or middle-game to endgame. Further, the user can manually toggle a complete COA replan for the current game phase. Figure 5 below shows the COAI user interface.

2.7.4. Allowable Matchups: H2H, H2AI, AI2AI, and MH2MAI

Flexibility for opponents is maintained. Human versus AI contests are possible, as well as AI versus AI, multiple human teams versus AI, and multiple human teams versus multiple AI instances. All that is necessary is instantiation of a unique simulation user interface, and the AI interface that sits on top of it. AI implementation is basically built on top of a simulation user interface. The COAI takes advantage of and issues UI controls and commands not unlike a human user.

2.7.5. Incorporating Plan and Action Randomness

Is important to note that the AI solution to the scenario problem does not have to be optimal. In order to be realistic and acceptable as a computer opponent, it is well known that human opponents are far from optimal. However, they can be counted on for a unique solution to most specific problems that will vary significantly between occurrences. If optimal mission solutions were calculated, this would result in a decreased training benefit and replay value for the AI implementation. Using randomness, crucially, also greatly simplifies the coding and modeling requirement necessary. Missions can be randomly specified within certain acceptable limits; in type, time, and space. For example, a recon mission can set a goal 0.9 km distant from a similar mission, and the overall results will not be that much different or unrealistic. A battalion can attack at 21:00 hours instead of 19:30 hours, and the effect is perhaps more realistic than if all attacks took place at 19:30 hours in similar circumstances.

2.8. Example Execution

As an example of a typical execution, assume a mission with three objectives: A, B, and C. The AI has 5 unit groupings available, and conducts a preliminary mission and terrain analysis resulting in an "endgame" COA. In this COA the 5 unit groupings would proceed directly to their respective objective, and would be assigned using the brute force plan generation methodology previously described. Since several hours of surplus time is found to exist before the scenario deadline, the AI embarks on an opening and middle game phase. During this phase reconnaissance is conducted as well as the seizing of key terrain, as calculated by the terrain analysis. Upon first enemy contact or casualties, the AI transitions into a middle-game phase and conducts preliminary attacks and deeper movements, preparing for final endgame execution. Once the time threshold for proceeding into the endgame phase is reached, the AI transitions into the final execution and replans the endgame movements from the current situation. Forces are allocated and final movements and attacks are conducted.

3. Results and Conclusions

The model described in this paper constitutes a computer opponent AI system conceived for the conducting of professional military training. Although many approaches exist, it is important that the envisioned solution be capable of a low-overhead implementation in usage. In other words, a large stimulation staff and large budget must not be required for end-users to conduct their own training. No special hardware, facilities, or preparation must be required. Bringing low-cost computer-assisted training to the echelon targeted (company, battalion, and brigade level training) requires this characteristic.

3.1. Advantages

The approach outlined has been able to achieve a robust and realistic computer opponent, while at the same time maintaining a reasonable level of coding overhead and debugging requirement. Much of this is due to prolific abstraction and layer isolation. The implementation as a result remains simple enough to understand at each layer, and exists in segregation from the decoupled simulation engine. Grafting a user interface over the AI has allowed users to understand what is going on and to better tweak settings and provide improvement recommendations and defect reports. The AI produces significant enough mission solution differences between instances to allow valuable replay value of identical scenarios.

3.2. Disadvantages

Inability to do a major replans during one of the transitional game phases intelligently is perhaps a disadvantage. Once the AI commits to a certain plan in any of the major game phases, only a triggered event can cause a major replan. In other words, tweaking the plan significantly does not happen within a respective game phase. Certain thresholds can be set, such as an overall casualty factor (as in the event of a catastrophic plan failure) in order to reinstitute a new "ground-up" planned COA. However this is not completely congruent with human level decision-making. Ironically, it may come close — and close enough is what we have been after. Another disadvantage is inflexible unit groupings. Generally, the past grouping that the scenario designer has created in the design files is what the AI will have to implement in force allocations.

3.3. Future Work

Future work includes the addition of more mission types and implementing them in respective classes. Further refinement of

the COA production process and the addition of wider spectrum of playbook templates is envisioned. Better coordination between attacking unit groups is also necessary. This may require an additional layer of abstraction constituting a class (below the COA class level) that controls multiple mission classes — this layer of modeling does not yet exist. Incorporating genetic algorithms and neural networks into mission planning is under investigation. Additionally, implementing graphical output depicting mission maps and plans according to guidelines put forth in specifications outlined in FM101-5-1 Staff Organization and Operations [15] is desired. Black Box AI is a thing of the past. The audience deserves and should demand Clear Boxes.

References

[1] M. J. Pelosi and M. S. Brown, "Software engineering a multi-layer and scalable autonomous forces “A.I.” for professional military training," 2016 Winter Simulation Conference (WSC), Washington, DC, 2016, pp. 3122-3133. doi: 10.1109/WSC.2016.7822345

[2] Johnston, J. H., Goodwin, G., Moss, J., Sottilare, R., Ososky, S., Cruz, D., & Graesser, A. (2015). Effectiveness Evaluation Tools and Methods for Adaptive Training and Education in Support of the US Army Learning Model: Research Outline (No. ARL-SR-0333). Army Research Lab, Aberdeen Proving Grouond, MD.

[3] Lane, H., Core, M., van Lent, M., Solomon, S., and Gomboc, D. (2005) *"Explainable Artificial Intelligence for Training and Tutoring."* 12th International Conference on Artificial Intelligence in Education, Amsterdam, The Netherlands.

[4] U.S. Army (2003). *"FM 7–1, Battle Focused Training*." Washington, DC: GPO.

[5] Newborn, M. (2012). Kasparov versus Deep Blue: Computer chess comes of age. Springer Science & Business Media.

[6] Shannon, Claude (1950). *"Programming a Computer for Playing Chess"*, Philosophical Magazine 41 (314)

[7] U.S. Army (1997). "*FM 101–5, Staff Organization and Operations".* Washington, DC: GPO.

[8] U.S. Army (2002). "*FM 3–06.11, Combined Arms Operations in Urban Terrain".* Washington, DC: GPO.

[9] Russell, S., Norvig, P. (2009). *"Artificial Intelligence: A Modern Approach."* (3rd ed.). Pearson Education.

[10] U.S. Army (2001). "*FM 3–90, Tactics".* Washington, DC: GPO.

[11] Boyd, John Richard (1976), *Destruction and Creation*, US Army Command and General Staff College.

[12] International Organization for Standardization (1989). *"ISO/IEC 7498-4:1989 -- Information technology-Open Systems Interconnection-Basic Reference Model".* ISO Standards Maintenance Portal. ISO Central Secretariat.

[13] Hart, P. E., Nilsson, N. J., Raphael, B. (1968). *"A Formal Basis for the Heuristic Determination of Minimum Cost Paths".* IEEE Transactions on Systems Science and Cybernetics SSC4 4 (2): 100–107. doi:10.1109/TSSC.1968.300136

[14] U.S. Army (2007). *"FM 3–21.8, The Infantry Rifle Platoon and Squad*." Washington, DC: GPO.

[15] U.S. Army (2004). "*FM 101–5–1, Operational Terms and Graphics."* Washington, DC: GPO.

Advances in Science, Technology and Engineering Systems Journal Vol. 2, No. 3, 1815-1822 (2017)

www.astesj.com

Special Issue on Recent Advances in Engineering Systems

ASTESJ

ISSN: 2415-6698

Smartphone Based Heart Attack Risk Prediction System with Statistical Analysis and Data Mining Approaches

M. Raihan[*,1], Saikat Mondal[2], Arun More[3], Pritam Khan Boni[2], Md. Omar Faruqe Sagor[4]

[1]*Biomedical Engineering, Khulna University of Engineering & Technology, Khulna-9203, Bangladesh*

[2]*Computer Science & Engineering, Khulna University, Khulna-9208, Bangladesh*

[3]*Department of Cardiology, Ter Institute of Rural Health and Research, Murud-413510, India*

[4]*Computer Science & Engineering, Khulna University of Engineering & Technology, Khulna-9203, Bangladesh*

ARTICLE INFO

Article history:
Received: 31 May, 2017
Accepted: 03 October, 2017
Online: 03 November, 2017

Keywords:
Cardiology
Heart Disease
Ischemic Heart Disease
Heart Attack
Mobile Computing
Data Mining
C4.5 Decision Tree
Smartphone
Statistics
Android

ABSTRACT

Nowadays, Ischemic Heart Disease (IHD) (Heart Attack) is ubiquitous and one of the major reasons of death worldwide. Early screening of people at risk of having IHD may lead to minimize morbidity and mortality. A simple approach is proposed in this paper to predict risk of developing heart attack using smartphone and data mining. Clinical data from 835 patients was collected, analyzed and also correlated with their risk existing clinical symptoms which may suggest underlying non detected IHD. A user friendly Android application was developed by incorporating clinical data obtained from patients who admitted with chest pain in a cardiac hospital. Upon user input of risk factors, the application categorizes the level of IHD risks of the user as high, low or medium. It was found by analyzing and correlating the data that there was a significant correlation of having an IHD and the application results in high & low, medium & low and medium & high categories; where the p values were 0.0001, 0.0001 and 0.0001 respectively. The experimental results showed that the sensitivity and accuracy of the proposed technique were 89.25 % and 76.05 % respectively, whereas, using C4.5 decision tree, accuracy was found 86% and sensitivity was obtained 91.6%. Existing tools need mandatory input of lipid values which makes them underutilized by general people; though these risk calculators bear significant academic importance. Our research is motivated to reduce that limitation and promote a risk evaluation on time.

1. Introduction

Cardiovascular diseases (CVDs) including Ischemic Heart Disease (IHD) are one of the major causes of death globally. As per the report of World Health Organization (WHO), in 2014, the death rate is 6.96% in Bangladesh due to Ischemic Heart Diseases, which ranks it first as the cause of death [1], [2]. WHO also forecasts that CVDs would be contributing to 24% as cause of death by 2030 and recommended to take extensive initiatives at various levels to reduce the mortality and morbidity out of IHDs [3].

For many diseases, prevention is better than cure and IHD is one of them and it can be primarily, primordially and secondarily prevented [1]. A good approach to prevent the improvement of risk factors or control of the risk factors will decline the risk of IHD. Having episode of IHD has long lasting impacts on the individual from having repeated episodes of chest pain to having limited quality of life due to heart failure and even death. There will be significant advantages if the people are made aware for health and preventive health checks for IHD [4]. With the known risk factors, calculating and categorizing a person into a risk level may help the person to be motivated for a preventive health check so that mortality can be reduced by detecting possibility of IHD earlier.

[*]Corresponding Author: M. Raihan, Department of Biomedical Engineering, Khulna University of Engineering & Technology, Khulna-9203, Bangladesh.
Email: raihanbme@gmail.com, raihan@biophotomatix.org

In current age, smartphones are one of the most widely utilized technologies worldwide and various smartphone based health apps are benefitting to people. Development of a smartphone based tool to predict heart attack risk would benefit masses of people. Currently, there are several tools like Framingham risk calculators, ASCVD risk calculator [5] available in the world to forecast risk of IHD which bear excellent academic significance but are underutilized by general people due to few limitations like mandatory inputs of Lipid profile and Blood Pressure (BP) values moreover, these risk scores do not individualizes the risk.

We developed a robust tool to forecast the threat of developing Ischemic Heart Disease (Heart Attack) using smartphone based application in android platform that may push much of population to check their own risk. This application generates a score based on the input and classifies the person into low, medium or high risk. The application also guides the person for further work up based on the result.

This paper is an extension of our previous work titled "Smartphone Based Ischemic Heart Disease (Heart Attack) Risk Prediction using Clinical Data and Data Mining Approaches, a Prototype Design" originally published and presented in 19th International Conference on Computer and Information Technology (ICCIT) 2016. We collected and added 130 more data to enrich our sample size and now, our total sample size is 917 which was 787 in the previous one. Moreover, we performed some significant statistical analysis like finding sensitivity, specificity, positive predictive value, and f-score here in this paper which were not in the earlier one. We surveyed more relevant literature and also compared our existing system with c4.5 algorithm.

2. Related Works

In order to identify risk factors for coronary heart disease (CHD), in [6], the authors introduced a coronary diseases prediction strategy using categorical feature, where blood pressure and dyslipidemia were categorized and correlated with the threat of coronary heart disease. They set a community based group to collect data and provide feedback after analyzing that data. Their total sample size was 5345. Throughout the twelve years of follow-up a total of 383 men and 227 women were found who had CHD. Multivariate logistic regression was used to predict the CHD as well as identify the risk factors. Sex-specific prediction equations were formulated to predict CHD. In case of middle aged white population sample, their suggested strategies can predict CHD risk effectively.

In [7], the researchers proposed a model to predict CHD risks perfectly. They developed Framingham functions and compared it with the performance of risk functions, originally developed from the individual clinical data. Cox regression coefficients and chi-square test as statistical measures were used in their methodology. The Framingham functions achieved good result for both black and white men and women. This event was taken within 5 years of follow up.

In [8], the scientists studied how mobile health can play significant role in prevention of cardiovascular disease. After that study, the team represented many statistics and provided some ideas how mobile health (mHealth) can prevent cardio vascular disease. They recommend some future research such as mobile application for treating obesity encouraging regular physical activity, smoking cessation, control of hypertension and dyslipidemia; and treating diabetes mellitus.

In [9], the researchers used different classification algorithm such as Decision Trees, Naive Bayes and K Nearest Neighbor in their research. For their analysis they used UCI machine learning database and WEKA (Waikato Environment for Knowledge Analysis), a data mining tool developed by the University of Waikato, New Zealand. Now, WEKA is really a popular suite of machine learning software. The experimental result showed that their true positive rate and accuracy is very good. In case of Naive Bayes algorithm they got 84.2% accuracy and sensitivity.

A system that can automatically predict heart disease was proposed in [10]. They used Decision Tree (C4.5) based algorithm for heart disease prediction. They used Cleveland heart disease dataset in their system which is available in UCI machine learning repository. They had designed a system that can efficiently predict the risk level of patients based on the given parameter about their health. The performance of their system was evaluated in terms of classification accuracy and the results showed that the system has great potential in predicting the heart disease risk level more accurately. They had used KEEL (Knowledge Extraction based on Evolutionary Learning) tool for prediction.

In [11], the scientists used several data mining techniques to predict heart disease. In their system, they implemented Decision Tree, Neural Network and Naive Bayes to find out interesting patterns from data collected from heart patients. For attribute selection they used Best first search and used WEKA as data mining tool. WEKA workbench contains a collection of visualization tools and various machine learning algorithms for data analysis. From Naive Bayes algorithm they got 82.914% accuracy with all attributes and 82.077% with selected attributes.

In [12], the authors presented that up to 75% of Cardiovascular diseases (CVD) in diabetes can be attributable to hypertension. In that paper, they also showed that atherosclerosis, dyslipidemia, micro-albuminuria, endothelial dysfunction, platelet hyperaggregability, coagulation abnormalities, obesity and diabetic are most important risk factors responsible for heart disease.

3. Methodology

Data of 917 patients was collected, where, there were 636 patients admitted in two cardiac hospitals with symptoms like chest pain, dyspnea, palpitations and syncope and remainder 281 individuals' data was collected from health camps irrespective of their symptoms and presence of IHD. The above mentioned symptoms are suggestive of underlying diagnosis of IHD. In Figure 1 the overall analysis flowchart has been represented and the existing system's architecture [1] of mobile application has been shown in Figure 2 and Figure 3. Here, a total number of 70 attributes have collected.

The variables analyzed for the purpose of this study were age, gender, presence or absence of hypertension, diabetes, dyslipidemia, smoking status, a significant family history [13], habit of exercise, abnormal ECG, positive ETT, presence or absence of symptoms suggestive of IHD were considered which are shown in Figure 4 and Figure 5. Chi square test, fisher's exact test, probability, percentage and ratios were used to calculate risk score. The data and generated risk score tree have been integrated to the android application named **PredictRisk**. In the application, the risk has been categorized as per score generated for variables of risk factors but if the user gives an input of having one or more symptoms, the risk level ascends up by one. This is designed to

avoid bias of not having known about presence of diabetes and/or hypertension. The mobile screen shots with active application **PredictRisk** as shown in Figure 6 to Figure 11.

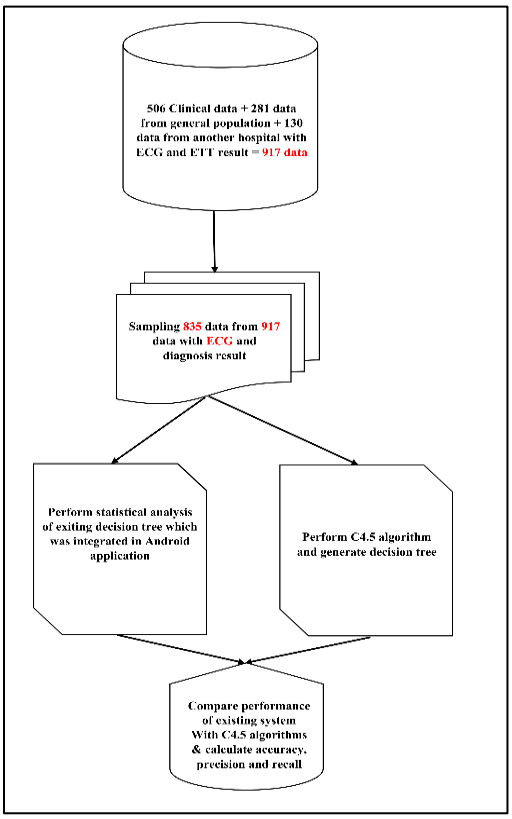


Figure 1: Overall Analysis

3.1. Chi-square test and Fisher's exact test

For nominal data correlation between two attributes can be discovered by chi-square test. If A has c distinct values, namely a_1, a_2, ……. , a_c and B has r distinct values namely b_1, b_2, …….. , b_r, the data tuples c_n can be shown as table, where with the c values of A making up the columns and the r values of B making up the rows [14] [15]. The value (also known as Pearson chi-square statistic) is computed as [14]:

$$\chi^2 = \sum_{i=1}^{c}\sum_{j=1}^{r}\frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

Fisher's exact test is a statistical significance test used in the analysis of contingency tables. For fisher's exact test the p is calculated as [1] [16] [17]:

$$\mathrm{p} = \frac{(a+b)!\times(c+d)!\times(a+c)!\times(b+d)!}{a!\times b!\times c!\times d!\times n!}$$

3.2. Risk Score Calculation

One of the risk factors, Age, was sub-categorized into four and given scores based on level of significance and association with having heart attack. In our dataset, **62.5%** Age is greater or equal to 55 years and out of that heart attack → yes is **75.3%.** This experimental result has been shown in Figure 12. So we have given the highest score 4 for this sub category. Our highest category Age has 4 subcategories and we have given score 1 for the age range 15-30, 2 for the age range 30-40, 3 for the age range 40-55 and 4 for the age greater or equal to 55 based on ratios and percentage with respect to heart attack.

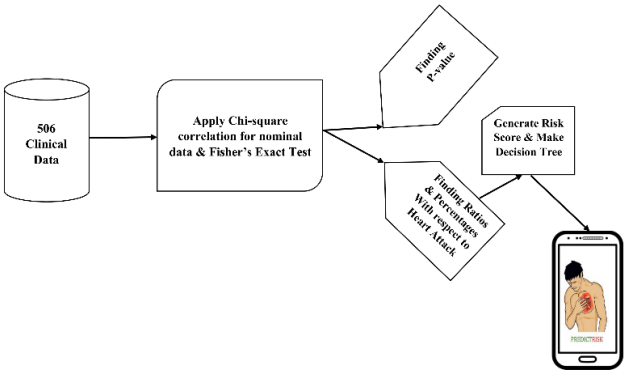


Figure 2: System Architecture of existing system

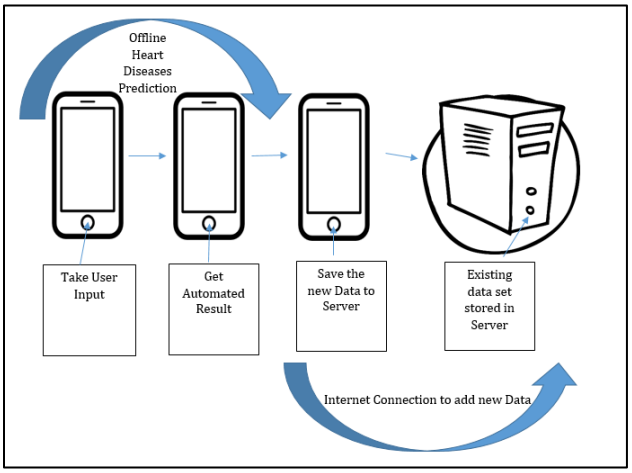


Figure 3: Android application process

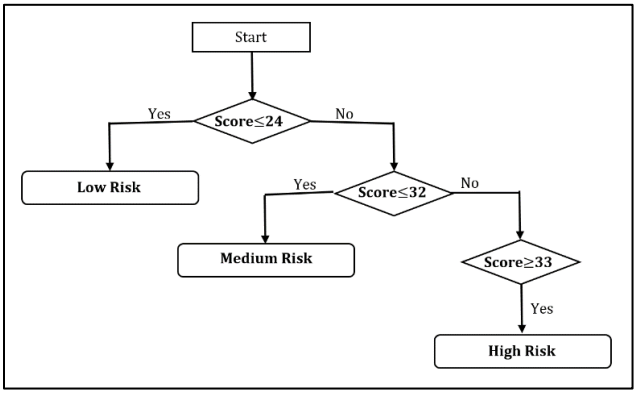


Figure 4: Risk Score tree

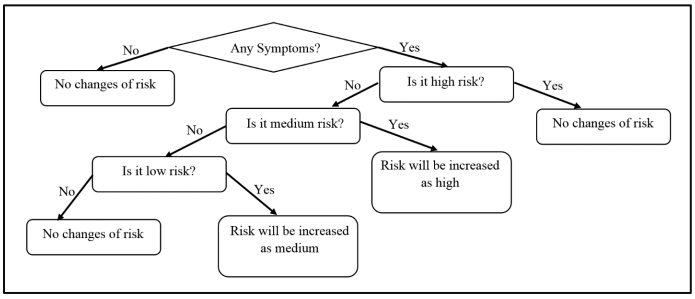


Figure 5: Symptom tree

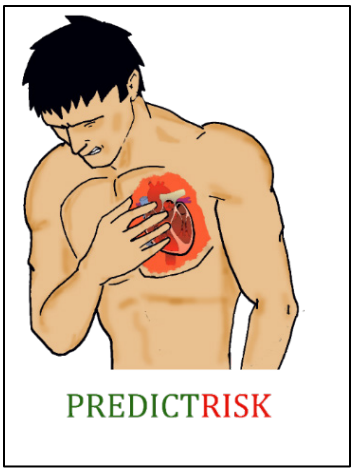


Figure 6: Application screen shot 1

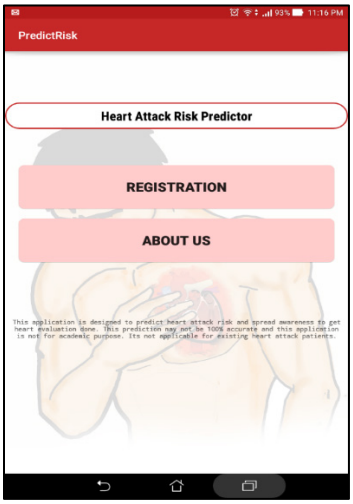


Figure 7: Application screen shot 2

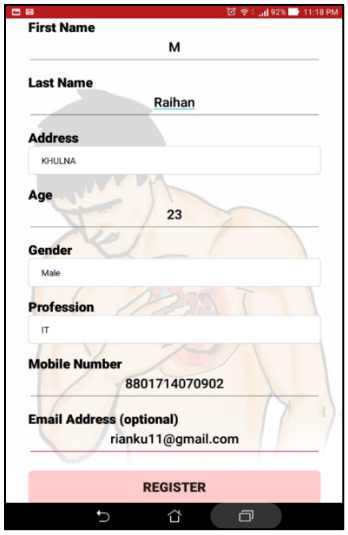


Figure 8: Application screen shot 3

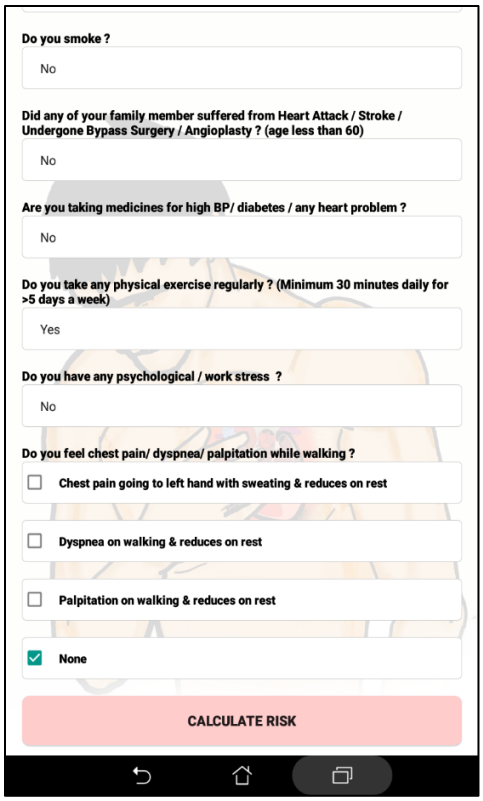


Figure 9: Application screen shot 4

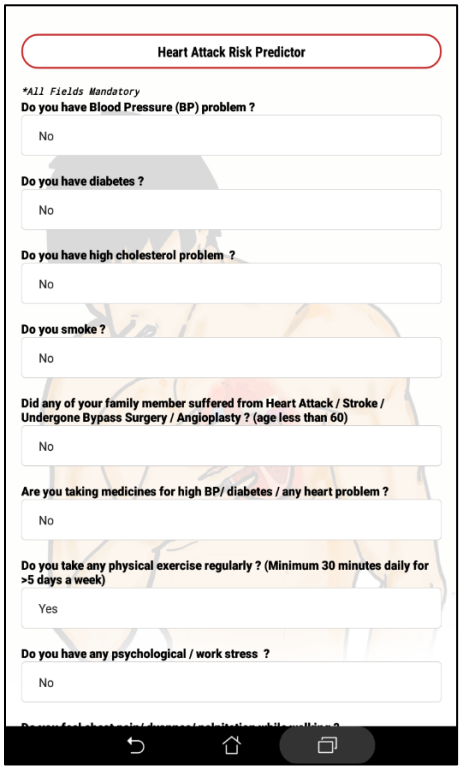


Figure 10: Application screen shot 5

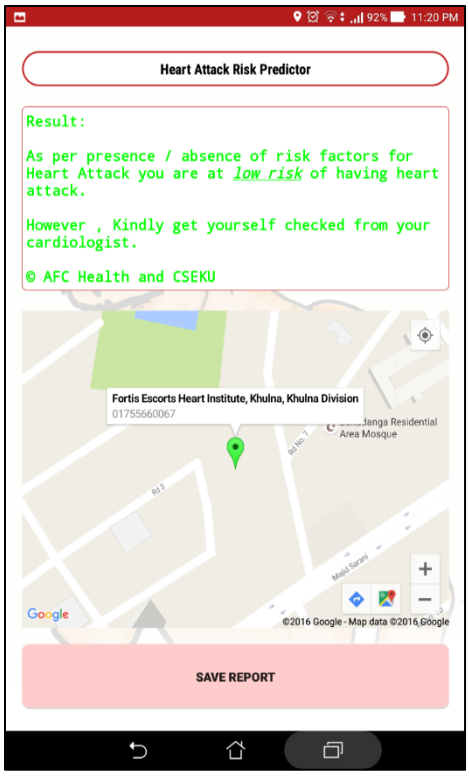


Figure 11: Application screen shot 6

The confusion matrix is an appropriate tool for analyzing how well your classifier can identify tuples of different classes has been shown in Table 2. TP (true positive) and TN (true negative) tell us when the classifier is getting things right, while FP (false positive) and FN (false negative) tell us when the classifier is getting things wrong [14] [15] [18].

3.4. Statistical Metrics for Evaluating Classifier Performance

Statistical metrics such as sensitivity, specificity, accuracy, precision etc. are very significant to evaluate a classifier performance [14]. In this research paper we used sensitivity, specificity, accuracy, positive predictive value (PPV), F-measure, negative predictive value (NPV) as measurement.

a. *Sensitivity: The sensitivity or true positive rate or recall of an experimental test refers to the ability of the test to correctly identify those patients with the disease [19] [20] [21].*

$$Sensitivity = \frac{TP}{TP+FN}$$

b. *Specificity: The specificity or true negative rate of a medical test refers to the ability of the test to correctly identify those patients without the disease [14] [20].*

$$Specificity = \frac{TN}{FP+TN}$$

c. *Positive Predictive Value (PPV): The Positive Predictive Value (PPV) or Precision can be thought of as a measure of exactness [14] [19]. The PPV answers the question: 'How likely is it that this patient has the disease given that the test result is positive?'[20].*

$$PPV = \frac{TP}{TP+FP}$$

d. *Negative Predictive Value (NPV): NPV is the percentage of patients with a negative test who do not have the disease [19].*

$$NPV = \frac{TN}{FN+TN}$$

e. *Accuracy: The accuracy of a classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier [14] [15] [21].*

$$Accuracy = \frac{TP+TN}{P+N}$$

f. *Error rate: the error rate or misclassification rate of a classifier which is simply calculated as [1][19][21]:*

$$Error\ rate = 1 - Accuracy$$

g. *F-measure: F-measure or F-score is the harmonic mean of precision and recall [14].*

$$F\text{-}Score = \frac{2 \times precision \times recall}{precision+recall}$$

Table 1: Risk Score

Attribute	Subcategory	Score
Age	15-30	**1**
	30-40	2
	40-55	3
	Above>=55	**4**
Sex	female	2
	male	4
Smoking	yes	4
	ex	3
	no	1
HTN	no	2
	yes	4
DLP	no	2
	yes	4
DM	no	2
	yes	4
Physical Exercise	no	4
	yes	1
Family History	no	2
	yes	4
Drug History	no	2
	yes	4
Psychological Stress	no	2
	yes	4

Table 2: Confusion Matrix

Predicted Class				
Actual Class		yes	no	Total
	yes	TP	FN	P
	no	FP	TN	N
	Total	P′	N′	P+N

3.5. C4.5 Decision Tree

C4.5 constructs a classifier in the form of a decision tree. In order to do this, C4.5 is given a set of data representing things that are already classified [22] [23] [24]. A classifier is a tool in data mining that takes a bunch of data representing things we want to classify and attempts to predict which class the new data belongs to [14].

Pseudo code of C4.5 decision tree [23] [25] [26]:

i. *Check for base cases*
ii. *For each attribute a : Find the normalized information gain ratio from splitting on a*
iii. *Let a_best be the attribute with the highest normalized information gain*
iv. *Create a decision node that splits on a_best*
v. *Recur on the sublists obtained by splitting on a_best, and add those nodes as children of node*

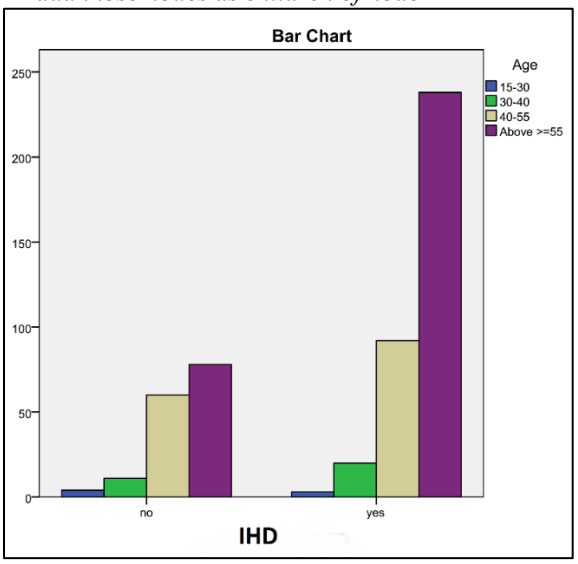


Figure 12: Bar chart between Age and IHD

4. Experimental Results and Statistical Analysis

We calculated the significance level between IHD and risk type (app result).We had 3 categories in risk types: high, low and medium.

4.1. Significance level of IHD & Risk Type:

Here, the association between presence or absence of IHD and risk type (high & low) have been considered to be extremely statistically significant at **5%** significance level [1] [16] [27].We used fisher's exact test to calculate the p value.

Table 3: Confusion Matrix IHD*Risk (High & Low)

Risk Type (App Result)					P-Value
IHD		High	Low	Total	
	Yes	271	53	324	**0.0001**<0.05
	No	75	123	198	significant
	Total	346	176	522	

Table 4: Confusion Matrix IHD*Risk (High & Medium)

Risk Type (App Result)					P-Value
IHD		High	Medium	Total	
	Yes	271	179	450	**0.0001**<0.05
	No	75	134	209	significant
	Total	346	313	659	

Table 5: Confusion Matrix IHD*Risk (Medium & Low)

Risk Type (App Result)					P-Value
IHD		Medium	Low	Total	
	Yes	179	53	232	**0.0001**<0.05
	No	134	123	257	significant
	Total	313	176	489	

The association between presence or absence of IHD and risk category high & medium and medium & low is considered to be extremely statistically significant at 5% significance level as shown in Table 4 to Table 6.Bar chart between IHD, ECG and risk category was shown in Figure 13 and Figure 14.

Table 6: Chi Square Test between IHD and Risk Type

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	115.150	2	**.000**
Likelihood Ratio	117.744	2	**.000**
N of Valid Cases	**835**		

4.2. Statistical Measurements

By analyzing **835** clinical data we calculated number of TP, FP, FN and TN as shown in Table 7. Accuracy, sensitivity, specificity, f-score, error rate and precision of our existing system as shown in Table 8 were calculated.

Table 7: Confusion Matrix of Existing System

Predicted Class				
Actual Class		yes	no	Total
	yes	523	63	586
	no	137	112	249
	Total	660	175	835

Table 8: Statistical Measurement of Existing System

Statistic	Formula	Value
Sensitivity	TP/(TP+FN)	89.25%
Accuracy	(TP+TN)/(P+N)	76.05%
Error Rate	1-accuracy	23.95%
Specificity	TN/(FP+TN)	44.98%
PPV	TP/(TP+FP)	79.24%
NPV	TN/(FN+TN)	64.00%

We found, sensitivity was **89.25%** and specificity was **44.98%**. Under the category high **53.9%** had IHD, under low **10.5%** had IHD while under medium category **35.6%** had IHD.

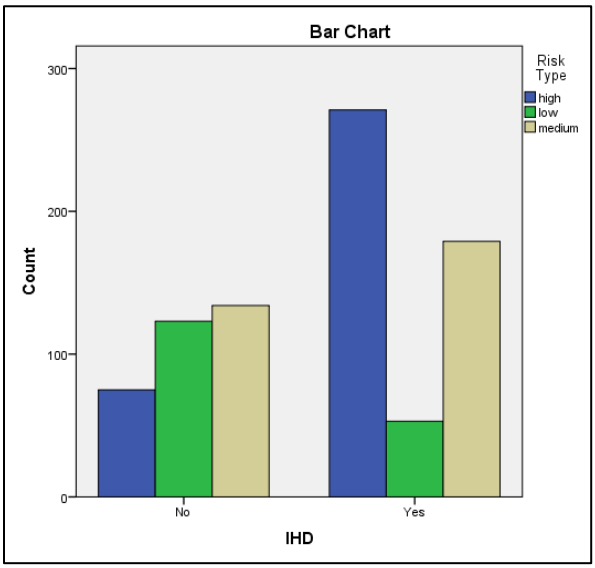


Figure 13: Bar chart between IHD and Risk Type

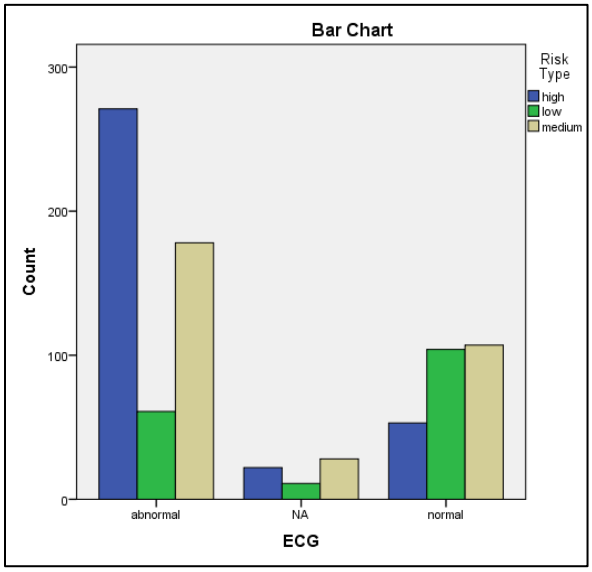


Figure 14: Bar chart between ECG and Risk Type

4.3. C4.5 Decision Tree

We used WEKA data mining tool to apply C4.5 on our clinical data [28]. By applying C4.5 algorithm our decision tree looks like this:

ECG = abnormal: Yes

ECG = normal

| Chest Pain = yes

| | Risk Type = medium: No

| | Risk Type = high

| | | Dyspnea = no: No

| | | Dyspnea = yes: Yes

| | Risk Type = low: No

| Chest Pain = no: No

Here IHD is the output class. We use **70%** data as training set and **30%** for test set. The results of C4.5 are shown in Table 8 and confusion matrix in Table 9.We got **86%** accuracy and **91.6%** sensitivity from C4.5 algorithm.

Table 9: Statistical results of C4.5

	Precision	Recall	F-Measure	ROC Area	Class
	0.865	0.916	0.89	0.852	Yes
	0.851	0.771	0.809	0.852	No
Weighted Average	0.859	0.86	0.859	0.852	

Table 9: Confusion Matrix

	Yes	No
Yes	141	13
No	22	74

From C4.5 algorithm number of correctly classified instances are **215** and number of incorrectly classified instances are **35**. Mean absolute error is **0.2702.**

5. Conclusion

IHD is a major killer worldwide. Using data generated from health care systems and mining them leads to development of mobile applications which are being used for predicting health and health related issues. **PredictRisk** categorizes the risk into clinically and socially acceptable parameters. **PredictRisk** generates and spreads awareness on the prevention and control of risk factors leading to IHD. In experimental analysis significant results have been found among the compared groups and users will be benefited heavily if this app will available in open source platform. Our research is motivated to make simple approach to detect the IHD risk and aware the population to get themselves evaluated by a cardiologist to avoid sudden deaths and morbidities.

Currently available tools have mandatory input of lipid values which makes them underutilized by population; though, those risk calculators bear excellent academic importance. Our research application **PredictRisk** reduces this limitation and promotes a risk evaluation on time.

Our future target is to collect more patients' data who are admitted in hospitals with heart disease and analyzed then using deep learning techniques. Moreover, our approach does not work well when the risks belong to the medium category, so we would like to work more to improve this section also.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgment

Computer Science and Engineering Discipline; AFC Fortis Escorts Heart Institute, Khulna, Bangladesh and Rural Health Progress Trust (RHPT), Murud, Latur, India for providing us technological and clinical support.

References

[1] M. Raihan, S. Mondal, A. More, M. Sagor, G. Sikder, M. Majumder, M. Al Manjur and K. Ghosh, "Smartphone Based Ischemic Heart Disease (Heart Attack) Risk Prediction using Clinical Data and Data Mining Approaches, a Prototype Design", in *19th International Conference on Computer and Information Technology*, North South University, Dhaka, Bangladesh, 2016, pp. 299-303.

[2] "Coronary Heart Disease in Bangladesh", *World Life Expectancy*, 2014. [Online]. Available: http://www.worldlifeexpectancy.com/bangladesh-coronary-heart-disease.

[3] M. Abu Sayeed, H. Mahtab, S. Sayeed, T. Begum, P. Khanam and A. Banu, "Prevalence and risk factors of coronary heart disease in a rural population of Bangladesh", *Ibrahim Med. Coll. J*., vol. 4, no. 2, 2010.

[4] A. Islam and A. Majumder, "Coronary artery disease in Bangladesh: A review", *Indian Heart Journal*, vol. 65, no. 4, pp. 424-435, 2013.

[5] "Risk assessment tool for estimating a patient's 10-year risk of developing cardiovascular disease", *Patient Monitoring and Support CVD Risk Check*. [Online].Available:https://www.cvdriskchecksecure.com/framinghamriskscore.aspx.

[6] P. Wilson, R. D'Agostino, D. Levy, A. Belanger, H. Silbershatz and W. Kannel, "Prediction of Coronary Heart Disease Using Risk Factor Categories", *Circulation*, vol. 97, no. 18, pp. 1837-1847, 1998.

[7] R. D'Agostino, Sr, S. Grundy, L. Sullivan, P. Wilson and for the CHD Risk Prediction Group, "Validation of the Framingham Coronary Heart Disease Prediction Scores", *JAMA*, vol. 286, no. 2, p. 180, 2001.

[8] L. Burke, J. Ma, K. Azar, G. Bennett, E. Peterson, Y. Zheng, W. Riley, J. Stephens, S. Shah, B. Suffoletto, T. Turan, B. Spring, J. Steinberger and C. Quinn, "Current Science on Consumer Use of Mobile Health for Cardiovascular Disease Prevention", Circulation, vol. 132, no. 12, pp. 1157-1213, 2015.

[9] S. Joshi and M. Nair, "Prediction of Heart Disease Using Classification Based Data Mining Techniques", Computational Intelligence in Data Mining - Volume 2, pp. 503-511, 2014.

[10] Purushottam, K. Saxena and R. Sharma, "Efficient heart disease prediction system using decision tree", International Conference on Computing, Communication & Automation, 2015.

[11] U. Shafique, F. Majeed, H. Qaiser and I. Ul Mustafa, "Data Mining in Healthcare for Heart Diseases", International Journal of Innovation and Applied Studies, vol. 10, no. 4, pp. 1312-1322, 2015.

[12] J. Sowers, M. Epstein and E. Frohlich, "Diabetes, Hypertension, and Cardiovascular Disease: An Update",Hypertension, vol. 37, no. 4, pp. 1053-1059, 2001.

[13] M. R. Kolber and C. Scrimshaw, "Family history of cardiovascular disease", Canadian Family Physician, vol. 60, no. 11, p. 1016, 2014.

[14] J. Han, M. Kamber and J. Pei, *Data mining*. Amsterdam: Elsevier/Morgan Kaufmann, 2012.

[15] P. Tan, M. Steinbach and V. Kumar, *Introduction to data mining*, 1st ed. Dorling Kindersley: Pearson, 2015.

[16] F. Dorey, "In Brief: The P Value: What Is It and What Does It Tell You?", Clinical Orthopaedics and Related Research®, vol. 468, no. 8, pp. 2297-2298, 2010.

[17] Fisher's exact test", En.wikipedia.org. [Online]. Available: https://en.wikipedia.org/wiki/Fisher%27s_exact_test.

[18] J. Ab Rahman, *Brief Guidelines for Methods and Statistics in Medical Research*. Singapore: Springer Singapore, 2015.

[19] R. Parikh, A. Mathai, S. Parikh, G. Chandra Sekhar and R. Thomas, "Understanding and using sensitivity, specificity and predictive values", *Indian Journal of Ophthalmology*, vol. 56, no. 1, p. 45, 2008.

[20] A. Lalkhen and A. McCluskey, "Clinical tests: sensitivity and specificity", *Continuing Education in Anaesthesia, Critical Care & Pain*, vol. 8, no. 6, pp. 221-223, 2008.

[21] K. van Stralen, V. Stel, J. Reitsma, F. Dekker, C. Zoccali and K. Jager, "Diagnostic methods I: sensitivity, specificity, and other measures of accuracy", *Kidney International*, vol. 75, no. 12, pp. 1257-1263, 2009.

[22] S. Joshi and M. Nair, "Prediction of Heart Disease Using Classification Based Data Mining Techniques", *Computational Intelligence in Data Mining - Volume 2*, vol. 2, pp. 503-511, 2014.

[23] P. Sharma, K. Saxena and R. Sharma, "Heart Disease Prediction System Evaluation Using C4.5 Rules and Partial Tree", *Advances in Intelligent Systems and Computing*, vol. 411, pp. 285-294, 2015.

[24] M. Gandhi and S. Singh, "Predictions in heart disease using techniques of data mining", *2015 International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE)*, pp. 520-525, 2015.

[25] "C4.5 algorithm", *En.wikipedia.org*. [Online]. Available: https://en.wikipedia.org/wiki/C4.5_algorithm#cite_note-3.

[26] S. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques", *Informatica (Slovenia)*, vol. 31, no. 3, pp. 249-268, 2007.

[27] A. More, M. Raihan, M. Sagor, G. Sikder, S. Mondal and M. Manjur, "Smartphone Based "Heart Attack" Risk Prediction; Initiation of A Clinically Simple Approach", *Journal of Hypertension*, vol. 34, p. e274, 2016.

[28] I. Witten, E. Frank and M. Hall, *Data mining*. Burlington, MA: Morgan Kaufmann, 2011.